Understanding Meaning and Knowledge Representation

Understanding Meaning and Knowledge Representation:

From Theoretical and Cognitive Linguistics to Natural Language Processing

Edited by

Carlos Periñán-Pascual
and Eva M. Mestre-Mestre

Understanding Meaning and Knowledge Representation:
From Theoretical and Cognitive Linguistics to Natural Language Processing

Edited by Carlos Periñán-Pascual and Eva M. Mestre-Mestre

This book first published 2016

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-8461-8
ISBN (13): 978-1-4438-8461-7

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

INTRODUCTION

The purpose of the book is to examine and discuss recent work in meaning and knowledge representation within theoretical linguistics and cognitive linguistics, particularly that research which can be reused to model natural language processing (NLP) applications. Today there is a need to develop NLP systems from a linguistically aware approach. Although there are many NLP applications that can work without taking into account any linguistic theory, this type of systems can only be described as "deceptively intelligent". On the other hand, those computer programs requiring some language comprehension capability should be grounded in a robust linguistic model if we want them to display the expected behaviour. Therefore, this book is concerned with the in-depth study not only of the multiple dimensions of language, e.g. morphology, syntax, semantics, pragmatics, concept formation, lexicon, and many others, but also of the interfaces between the components of the architecture of the language system and the processes underlying language comprehension and production under varying circumstances and situations. The new insights from this type of research can undoubtedly help model more robust NLP systems. This book is divided into three thematic parts.

From functionalist and/or cognitivist approaches, PART 1 deals with various theoretical linguistic issues that have the potential to enhance NLP systems. For example, Role and Reference Grammar (Van Valin 2005) plays an important role in the first three chapters of the book. In chapter one, Brian Nolan explores the causal relata underpinning the concepts of transfer, let/allow and permission and their argument realization in Modern Irish (Nolan 2012, 2013). In this regard, he examines a number of syntactic construction patterns associated with the argument realization, considering factors such as control and volition (Dixon 2010) in causal event chaining. In chapter two, Elke Diedrichsen describes the *lassen* ('let') construction in German, whose semantics can vary within a spectrum of meanings involving direct causation, permission and non-intervention. The syntax and semantics of the construction are discussed extensively, being represented in a Constructional Schema that displays its features. In chapter three, Aoife Finn studies if the meaning of transitive verbs in Māori is actually the same as that of neuter verbs, as proposed in

traditional grammars (Harlow 2007). It seems that, although neuter verbs and transitive verbs have similar meanings, the syntactic realization of these two types of verbs is quite different. Given their different syntactic realizations, this chapter preliminarily considers the logical structure of neuter verbs via Aktionsart tests. Thus, the findings in chapters one, two and three may be very valuable, for example, for machine translation in general and for word sense disambiguation in particular. On the other hand, and from the cognitive realm, the analysis of lexical polysemy has a special treatment in the next two chapters. In chapter four, Svetlana Kiseleva and Nelly Trofimova examine the mechanisms of meaning extension in the polysemous English verbs of part-whole relation. Their theory is based on the idea that the meaning of any word can be explained by means of an exact paraphrase composed of simpler, more intelligible lexical components than the original (Wierzbicka 1972). Thus, their main statement is that every complex word has a substantial core, being the essential basis that provides its semantic integrity. In chapter five, Aysun Balkan analyses the applicability of the Principled Polysemy model (Tyler and Evans 2003) to the Turkish *üst* construction ('on' and 'over') including spatial nouns suffixed with dative case marker. Despite the fact that Turkish and English are typologically distinct languages and express spatial relations using very different linguistic elements, the current study shows a surprising amount of overlap with the primary and extended senses found in the polysemy networks of 'üst + Dative' and English 'over'. Again, a treatment of polysemy such as the ones proposed in these two chapters can help machines solve lexical ambiguity. Likewise, chapter six can be very relevant for anaphora resolution, where Alana Poncarová presents various methods of reconstructing the meaning of co-referential chains in Czech from the perspective of Centering Theory (Brennan, Friedman and Pollard 1987; Grosz, Joshi and Weinstein 1995). Her research shows that the information structure (Topic-Focus articulation) and the constituent structure (Subject-Object function) are key factors in this process. Moreover, in chapter seven, Ciro Antunes de Medeiros describes the human ability to produce superordinate categories in the context of the relationship between conceptual organization and lexical acquisition, a cognitive modelling topic which can be of interest to researchers in ontology development. Finally, in chapter eight, Elena Even-Simkin describes an iconic phono-morphological analysis of the internal-vowel-alternation phenomenon in plural nouns and past tense forms in English based on the theory of Phonology as Human Behavior (Diver 1979; Tobin 2009). The findings presented in this chapter may have

implications in the field of Web search engines or in human-robot interaction.

PART 2 consists of two chapters that intend to demonstrate the reader about the need for NLP research groups to have linguists collaborating with computer engineers. To a layman, this could be a non-issue, since at first sight computational linguistics is deemed as a sub-discipline of applied linguistics. However, theoretical linguistics has usually played a remarkably minor role in this field of research. Indeed, Wilks (2005) noted that the links between NLP and linguistics have not been either so numerous or so productive as we could imagine. Based on the authors' vast experience in this field, chapters nine and ten portray this situation from a contemporary view. In chapter nine, Brian Nolan examines the question 'What can theoretical linguistics do for NLP research?' from a number of perspectives, including linguistics, informatics and engineering. He explores the work practices and goals of contemporary linguists today and the tools they use in that work. This chapter also describes the contribution of linguistic theory (generative, constructional and functional) to linguistic realism to achieve descriptive, explanatory and computational adequacy while managing issues with linguistic and computational complexity. The functionality of many future NLP applications is not yet known; however, language-aware human cognitive technologies can point us in a very interesting direction. He concludes that the future is bright for linguists, especially those with some software skills, and that linguistic theory has a significant contribution to make to NLP. On the other hand, in chapter ten, Elke Diedrichsen supports the idea that NLP can benefit from scholars and researchers who work in theoretical linguistics, in particular functional models of grammar. The breadth of language-aware products available today is indicative of the way the IT industry is growing globally. We live in a multilingual world, and all these languages need to be properly characterized for the benefit of the customers of these products. The way to achieve this is to take on board the knowledge and insights provided by theoretical linguistics. The potential for 'next generation' IT products, arising from synergistic efforts of linguists, computer scientists and engineers working together, is huge. This chapter discusses some of the new language-aware products and applications that have recently emerged from leading IT companies.

PART 3 serves to illustrate how a linguistically aware and cognitively plausible approach to human-like processing through FunGramKB Suite can contribute to the development of enhanced knowledge engineering and NLP projects. FunGramKB Suite (Periñán-Pascual 2012, 2013) is a user-friendly environment for the semi-automatic construction of FunGramKB,

a lexical-conceptual knowledge base particularly designed for natural language understanding systems, and for the development of tools for the automatic processing of language (cf. Periñán-Pascual and Arcas-Túnez 2014b). In this regard, in chapter eleven, Carlos Periñán-Pascual and Eva M. Mestre- Mestre accurately describe a hybrid approach to the evaluation of automatic term extraction systems, which have been traditionally evaluated by means of one of two methods, i.e. gold-standard reference lists or validation based on experts' judgements (Pazienza, Pennacchiotti and Zanzotto 2005). In particular, these authors explore the way that the IATE thesaurus together with a specialized dictionary can be semi-automatically integrated with the human validation of term candidates. The experiment was performed with DEXTER, an open-access platform for data mining and terminology management that can export specialized terms to FunGramKB. In chapter twelve, Ana Díaz Galán and María del Carmen Fumero Pérez contribute to ARTEMIS, a grammar development environment that outputs the parse tree of a text based on the Layered Structure of the Clause in Role and Reference Grammar. In particular, they study the grammatical phenomenon of the insertion of the DO operator in simple sentences in English. In chapter thirteen, María José Ruiz Frutos explains how the semantic knowledge, common-sense knowledge and world knowledge in FunGramKB can help solve co-reference ambiguity. In chapter fourteen, Ángela Alameda Hernández and Ángel Felices Lago explore the design and development of specialized-knowledge ontologies in the FunGramKB framework. In particular, they describe the methodological problems encountered in the conceptualization of the superordinate concept CRIME and its configuration as an umbrella concept. Finally, the aim of chapter fifteen is to set the basis for the creation of a terminological satellite ontology of mental disorders within FunGramKB. In this work, María Beatriz Pérez Cabello de Alba and Ismael Iván Teomiro García follow Felices-Lago and Ureña Gómez-Moreno's (2012) methodological underpinnings for the construction of terminological subontologies in FunGramKB, as well as Periñán-Pascual and Arcas-Túnez's (2014a) methodology used in the design of a subontology on criminal law in FunGramKB. The authors employ a tool based on Latent Semantic Analysis (Landauer, Foltz and Laham 1998) in order to complete the phases of corpus compilation and term extraction. As can be noted in this third part of the book, linguists play a major role in those NLP systems which exploit FunGramKB as its knowledge base.

This monograph was conceived from the different perspectives of a full gamut of research projects concerned with language understanding

through the prism of theoretical linguistics, cognitive linguistics and computational linguistics. Therefore, the book will be of particular interest to scholars, researchers and postgraduate students who work in these fields of knowledge.

Finally, we would like to take this opportunity to thank Cambridge Scholars Publishing for giving us the chance to compile and publish this book, as well as acknowledging the support of the Spanish Ministry of Education and Science (grants FFI2011-29798-C02-01 and FFI2014-53788-C3-1-P) and of Generalitat Valenciana - Conselleria de Educación, Cultura y Deporte (AORG/2014/071).

The editors

References

Brennan, Susan, Marilyn Friedman and Carl Pollard. "A Centring Approach to Pronouns." In Proceedings of the 25th Annual Meeting of the Association of Computational Linguistics, 1987.

Diver, William. "Phonology as Human Behavior." In Psycholinguistic Research: Implications and Applications, edited by D. Aaronson and R.W. Reiber, 161-186. Hillside: Lawrence Erlbaum, 1979.

Dixon, Robert M. W. A typology of causatives: form, syntax and meaning. In Changing Valency: Case Studies in Transitivity, edited by R.M.W. Dixon and Alexandra Y. Aikhenvald, 30-83. Cambridge: Cambridge University Press, 2010.

Felices-Lago, Ángel and Pedro Ureña Gómez-Moreno. "Fundamentos metodológicos de la creación subontológica en FunGramKB." Onomázein 26 (2012): 49-67.

Grosz, Barbara, Aravind Joshi and Scott Weinstein. "Centring - A Framework for Modeling the Local Coherence of Discourse." Computational linguistics 21 (1995): 203-225.

Harlow, Ray. Māori: A Linguistic Introduction. Cambridge: Cambridge University Press, 2007.

Landauer, Thomas K., Peter W. Foltz and Darrell Laham. "Introduction to Latent Semantic Analysis." Discourse Processes 25 (1998): 259-284.

Nolan, Brian. The Structure of Irish: A Functional Account. Sheffield: Equinox Publishing, 2012.

Nolan, Brian. "Constructions as Grammatical Objects: A Case Study of the Prepositional Ditransitive Constructions in Modern Irish." In Linking Constructions into Functional Linguistics. The Role of Constructions in RRG Grammars, edited by Brian Nolan and Elke Diedrichsen, 143-178. Amsterdam: John Benjamins, 2013.

Pazienza, Maria Teresa, Marco Pennacchiotti and Fabio Massimo Zanzotto. "Terminology Extraction: An Analysis of Linguistic and Statistical Approaches." In Studies in Fuzziness and Soft Computing: Knowledge Mining, edited by Janusz Kacprzyk and Spiros Sirmakessis, 255-279. Berlin-Heidelberg: Springer, 2005.

Periñán-Pascual, Carlos. "The Situated Common-Sense Knowledge in FunGramKB." Review of Cognitive Linguistics 10-1 (2012): 184-214.

—. "A Knowledge-Engineering Approach to the Cognitive Categorization of Lexical Meaning." VIAL: Vigo International Journal of Applied Linguistics 10 (2013): 85-104.

Periñán-Pascual, Carlos and Francisco Arcas-Túnez. "La Ingeniería del Conocimiento en el Dominio Legal: La Construcción de una Ontología Satélite en FunGramKB." Signos 47-84 (2014a): 113-139.

—. "The Implementation of the FunGramKB CLS Constructor". In Language Processing and Grammars: The Role of Functionally Oriented Computational Models, edited by Brian Nolan and Carlos Periñán-Pascual, 165-196. Amsterdam/Philadelphia: John Benjamins, 2014b.

Tobin, Yishai. "Phonology as Human Behavior: Applying Theory to the Clinic." *Asia-Pacific Journal of Speech, Language and Hearing* 12-2 (2009): 81-100.

Tyler, Andrea and Vyvyan Evans. The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition. Cambridge: Cambridge University Press, 2003.

Van Valin, Robert D. Exploring the Syntax-Semantics Interface. Cambridge: Cambridge University Press, 2005.

Wierzbicka, Anna. Semantic Primitives. Frankfurt: Athenäum, 1972.

Wilks, Yorick. "Computational Linguistics: History". In Encyclopedia of Language and Linguistics, second edition, 761-769. Oxford: Elsevier, 2005.

PART ONE:

MEANING AND KNOWLEDGE REPRESENTATION

CHAPTER ONE

CAUSAL RELATA AND EVENT CHAINS IN THE CONCEPTS TRANSFER, LET/ALLOW AND PERMISSION IN MODERN IRISH

BRIAN NOLAN
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN, DUBLIN

1. Introduction

We know that the analysis of the causative (Nolan 2012a: 33; Nolan 2015; Nolan, Rawoens and Diedrichsen 2015) intersects with semantics, syntax and morphology and, as such, the causative construction remains one of the primary research areas for many linguists. It has been generally recognised that there are three prototypical types of lexical, morphological and syntactic causative within any consideration of a causative taxonomy. However, as well as these types, a further distinction is made along semantic lines of inquiry between direct causation and indirect causation. That is, languages are known to make a distinction between direct and indirect causation through some language specific means. For example, in order to express direct causation, a language may use a causative construction in which a higher degree of fusion is seen in the expression of cause and effect. Correspondingly, indirect causation will exhibit a lower degree of fusion of cause and effect within the expression.

Many scholars, including Van Valin (2005:42; *n*5) and Song (1996: chapter 1), realise that treating all causatives as having the same 'CAUSE' element is a gross oversimplification of the complexities involved. There is essentially a contrast among three basic types of causality, including (i) Direct (Coercive), (ii) Indirect (Non-coercive), and (iii) Permissive. Both direct and indirect causality are represented by 'CAUSE', and permissive causality can, for example, be represented by 'LET' or 'ALLOW' in logical structures. We are concerned in this chapter with causative constructions and the concepts of TRANSFER, LET/ALLOW and PERMISSION within Modern Irish, as shown in (1).

(1) x CAUSES/PERMITS/LETS/ALLOWS y to DO/MAKE/HAVE z-something

In this analysis, we will be mindful of the typology of causation in the work of Dixon (2010: 62), where he proposes the nine semantic parameters in (2), described in Table 1.1, to characterize a typology of causation.

(2) State/action, transitivity, control, volition, affectedness, directness, intention, naturalness and involvement

An important set of considerations is also found in the work of Talmy (2000), where he proposes that the physical-force model maps straightforwardly to the psychological realm, since these same predicates are used to characterise psychosocial as well as physical causal relations. This proposal develops a central theme of cognitive linguistics according to which abstract conceptual content is derived from representations of physical reality. Gärdenfors (2007) similarly extends the Talmy perspective to characterise verbal concepts as patterns of forces:

> Even though our cognition may not be built precisely for Newtonian mechanics, it appears that our brains have evolved the capacity for extracting the forces that lie behind different kinds of movements and action... In accordance with this, I submit that *the fundamental cognitive representation of an action consists of the pattern of forces that generates it.*
>
> (Gärdenfors, 2007: 254)

We argue that the appropriate way to understand dynamic events is to consider them as forces: inputs of energy. Such inputs of energy may, or may not, have an effect on the state of affairs; this inherent defeasibility provides the tools necessary to naturally accommodate the problems of the concepts of TRANSFER, LET/ALLOW and PERMISSION. This accommodates Talmy's (2000) insight that component forces are referred to in the meanings of agonist-antagonist lexical items such as *enable, prevent,* and so on. The causing event corresponds to a force that is applied to a situation where the resulting stative predicate does not hold such that this force yields a situation where the resulting stative predicate does, or may, hold.

Table 1.1. Dixon's (2010) semantic parameters of causation

VERB	1. **State/action**	Does a causative mechanism apply only to a verb describing a state, or also to a verb describing an action?
	2. **Transitivity**	What is the transitivity of the verb?
CAUSEE	3. **Control**	Is the causee lacking control of the activity or normally having control?
	4. **Volition**	Does the causee do it willingly ('let') or unwillingly ('make')?
	5. **Affectedness**	Is the causee only partially affected by the activity, or completely affected?
CAUSER	6. **Directness**	Does the causer act directly or indirectly?
	7. **Intention**	Does the causer achieve the result accidentally or intentionally?
	8. **Naturalness**	Does it happen fairly naturally (the causer just initiating a natural process) or is the result achieved only with effort?
	9. **Involvement**	Is the causer also involved in the activity (in addition to the causee) or not involved?

This chapter proceeds as follows: first we examine the concept of transfer through the verbs GIVE, TAKE and PUT and follow this with a separate discussion of the verb GET. Essentially, as encoded within these verbs and their related constructions, the theme is transferred to a recipient or location. Next, after this exploration of the verbs GIVE, TAKE, PUT and GET, we examine a construction where permission is given but where this does not lexicalise as PERMIT. This is the GIVE PERMISSION construction. In this construction, the purpose (PURP) for which permission is given is made explicit. We then examine the construction with the LET/ALLOW concepts and the PERMIT construction.

2. Transfer Constructions

In Modern Irish, the verbs GIVE, TAKE, PUT and GET can be considered as representing typical events that utilise the concepts of caused motion and transfer. These are causative ditransitive constructions (Nolan 2015, 2012) and we examine here the relationship of the event structure to the argument structure and how these are realised in the syntax, in particular, in the sense of Talmy (2000). As three-place events, we necessarily examine the role of the actor, recipient and theme and the argument realisation that occurs with these verbs in single and multi-event clauses. The general semantic representation of these verbs is given in (3), with data examples for the GIVE, TAKE and PUT

verbs of Irish in (4)-(6). GET is examined later as a construction involving the transfer of a theme, where the theme is permission.

(3)

a. *Thug* 'give':	[**do´** (x, Ø)] CAUSE [BECOME **have´** (y, z)]
causative	x: Actor GIVES z: Theme
prepositional	x causes y: Recipient to have z: Theme
ditransitive verb	Theme is transferred to recipient [actor not = recipient]
b. *Thóg* 'take':	[**do´** (x, Ø)] CAUSE [BECOME NOT **have´** (y,z) &
causative	BECOME **have´** (x, z)]
prepositional	x: Actor TAKES z: Theme
ditransitive verb	x causes x: Recipient to have z: Theme
	Theme is transferred to recipient [actor = recipient]
c. *Chuir* 'put':	[**do´** (x, Ø)] CAUSE [BECOME **be–LOC´** (y, z)]
causative	x: Actor causes z: Theme to be at y: Location
prepositional	Theme is transferred to location
ditransitive verb	

(4) Thug sé an leabhar dom.
Give-PST 3SG.M DET book to:PREP+1SG
He gave the book to me.
[**do´** (3SG.M, Ø)] CAUSE [BECOME **have´** (1SG, book)]]

(5) Thóg sí an leabhar uaidh.
Take-PST 3SG.F DET book from:PREP+3SG.M
She took the book from him.
[**do´** (3SG.F, Ø)] CAUSE [BECOME NOT **have´** (3SG.M, the book) & BECOME **have´** (3SG.F, book)]

(6) a. **Verb PUT with BE-ON**
Chuir sí cóiriughadh úr-nuaidh ar an dreisiúr.
Put-PST 3SG.F ornament fresh+new on:PREP DET dresser
She put an ornament on the dresser.
[**do´**(3SG.F , Ø) CAUSE [INGR **be-on´**(the dresser, ornament]

b. **Verb put with be-at**
Chuir sé Micheál Ó Cléirigh anall go hÉirinn.
Put-PST 3SG.M Micheál Ó Cléirigh across to:PREP Ireland
He sent Micheál Ó Cléirigh across to Ireland.
[**do´**(3SG.M, Ø) CAUSE [BECOME **be-at´**(Ireland , Micheál Ó Cléirigh)]]

c. **Verb PUT with BE-FROM**
Chuir sin Donnchadh ó obair.
Put-PST that Donnchadh from:PREP work
That put Donnchadh out of work.
[**do´**(that, Ø) CAUSE [BECOME **be-from´**(work, Donnchadh)]]

What is interesting with the verb PUT is that, at the construction level, different prepositions seem to license different meanings and directions of TRANSFER over and above the core lexical sense for the same lexical verb. These apply to the encoding of the third participant. Therefore, it seems to be necessary to distinguish the lexical meaning of the verb from the meaning it has in a particular clause in which it occurs.

3. GIVE_PERMISSION Constructions

These GIVE, TAKE, PUT and GET verbs, examined from the perspective of syntactic constructions, are concerned with the explicit transfer of the theme. These were all single-event constructions. Here, I wish to examine a number of multi-verb, multi-event GIVE_PERMISSION constructions to explore the syntactic patterns that occur including the encoding of the outer and inner event and how the various arguments are realised and shared across the verbs within each event. Specifically, the theme transferred is PERMISSION, as in (7). In these constructions, the PURPOSE, for which the permission is given, is explicitly stated in the syntax as an embedded event. As such, these are complex multi-verb constructions. In this example of a causative GIVE, the first thing we find is that the *giving of permission* is not lexicalised as a PERMIT verb. Therefore, this has a particular syntactic pattern to signal the construction signature, as in (8).

(7) a. *Ní thug sí cead teach a scuabadh, coirce a mheilt, adhmad a scoilteadh, bia a bhruith, arán nó maistriú nó abhras a dhéanamh, aghaidh a bhearradh nó a ní, capall nó asal a mharcaíocht, nó curach a chur ar linn*
She did not give permission for (the) house to be brushed, corn to be ground, wood to be split, food to be cooked, bread or churning or yarn to be made, faces to be shaved or washed, a horse or donkey to be ridden, or a currach [= a small boat] to be put on the lake.

b. Ní thug sí cead teach a scuabadh.
NEG give-PST 3SG.F permission:N house:N REL brush:VN.
She did not give permission for (the) house to be swept.

(8) a. *Thug* xNP *cead* (zNP) yNP *a* VN :**construction signature**
V1 _____V2
[[Event1] _____ [Event2]]
[T1 _____ T2]

b. [**do´** (3SG.F) CAUSE **be-at´** (z[+ANIMATE, +SPEC, -DEF]$_1$, permission)]
PURP [**do´** (z$_1$, [**brush´** (z$_1$, house)])]

The ‘x’ argument is an NP, specifically here a 3SG.F PN that acts as the **controller** of the causative force. This argument is also [+AGENT], [+ANIMATE]. As the controller of the causative force, the referent is the possessor of the permission that is to be given. That is, the transfer of the permission is within the gift of the referent of the ‘x’ argument, as denoted in the construction signature. The ‘x’ NP argument has the characteristics in (9). The nominal *cead,* ‘permission’, is the permission that is transferred. As such, it is the theme of the construction. The recipient of the permission is not encoded in the clausal syntax in this construction example. We represent this in the signature as an optional ‘z’ NP argument. Implicit in the causative event is that we expect this ‘z’ NP recipient argument to have at least the characteristics specified in (10). That is, the ‘z’ argument is a referent which is human, animate, specific and indefinite with the capability to undertake the purposive action specified by the verbal noun on the ‘y’ argument.

(9) x: NP, 3SG.F
[+ AGENT]
[+ ANIMATE]
[+ CONTROLLER]
[+ POSSESSOR_of_permission]

(10) z: NP, the optional ‘z’ NP argument
[+ ANIMATE]
[+ HUMAN]
[+ SPEC]
[– DEF]

The verbal noun encodes the explicitly permitted action, the purposive act for which permission is intentionally and volitionally given. This is the inner-caused event, i.e. V2, a bounded active accomplishment. The ‘y’ argument is the undergoer of the VN, but this argument is shared with the *thug,* ‘give’, matrix verb under the nexus-juncture relation of cosubordination (see Nolan 2012: 165 for a discussion of nexus-juncture relations in Irish). That is, some of the three explicit arguments of the ‘give’ matrix verb, i.e. V1, are in a sharing relation under cosubordination with the V2 verb, which is realised as a VN in the construction. The optional ‘z’ argument, i.e. the recipient of the permission, is the actor of the inner event.

In so far as the attributes of the argument are [+ANIMATE, +SPEC, -DEF], i.e. an animate specific indefinite entity, this encoding is reminiscent of the coding of the impersonal actor in the impersonal passive of Modern Irish (see Nolan 2012a: 107-131 for a discussion of specific indefinites in

Modern Irish). It is important to note that, in Irish, a verbal noun is used with a matrix verb to form the complex predicate, and thereby acting as a verb-verb complex. The verbal-noun morphosyntactic form retains its valence within a multi-verb and multi-event construction. We argue for an account of the complex predications within the theory of Role and Reference Grammar (RRG) (Van Valin 2005). The example (11) is a rather complex but useful one to consider, as it illustrates very well the argument realisation in the nexus-juncture relations over requesting permission from a source with a specific purposive intended action, once the permission is obtained.

(11) *Tamall ina dhiaidh sin d'iarr sí cead ar an rí cuairt a thabhairt ar a máthair, agus thug sé an cead sin di.*
Some time later she requested permission from the king to give a visit to her mother, and he gave that permission to her.
Tamall ina dhiaidh sin
Time:N after:PREP that:DEIC
d'iarr sí cead ar an rí cuairt a thabhairt
request-PST 3SG.F permission:N on:PREP DET king:N visit:N REL give:VN
ar a máthair, agus thug sé an cead sin di.
on:PREP her:possPN mother and give-PST 3SG.M DET permission that to:PREP+3SG.F.DAT

We start with identifying the various sub-clauses. In this example, *d'iarr* is V1, 'request', and the actor of this requesting event is 3SG.F. The thing requested is *cead*, 'permission', and the possessor of the permission is 'the king'. The purpose of requesting permission is to give a visit to her mother. The construction signature of this example is shown in (12). Uniquely, this complex sentence with many clauses provides multiple perspectives on the causative purposive event. The 'x' argument, i.e. the requestor of the permission, has the characteristics in (13). The thing requested is *cead*, 'permission'. The source and possessor of this permission, and therefore the controller, is 'the king' in (14).

(12) *D'iarr* xNP CEAD$_1$ yPP$_{\textbf{source}}$ zNP *a* VN PP *agus thug* yNP *an CEAD*$_1$ *sin* xNP
V1 _____ V2 _____V3
[[Event1] _____ ALLOW ___ [Event2] and [Event3]]
T1 _____ T2_T3 … T4

(13) x: NP, 3SG.F
[+ ACTOR]
[+ ANIMATE]
[+ HUMAN]
[+ INTENTIONAL]

[– POSSESSOR_of_permission]
[+ REQUESTOR_of_permission]
[– CONTROLLER]

(14) y: NP 'the king'
[+ ANIMATE]
[+ HUMAN]
[+ INTENTIONAL]
[+ POSSESSOR_of_permission]
[+ CONTROLLER]

The V2 is realised as a VN. This verb *give* has three arguments. Here, it shares the arguments with the matrix V1 *d'iarr*, 'request'. The embedded actor of the VN is 3SG.F, who is also the 'x' argument of the matrix verb, i.e. the requestor of the permission. We have a clause conjunction with *agus*, 'and'. In the conjoined clause, we encode the act of GIVING_PERMISSION to the requesting NP. In this clause, the permission is transferred from its source, the king, to the 3SG.F requestor in order to permit a particularly-identified purposive event. The logical structure underlying this clause is shown in (15).

(15) [**do´** (3SG.F_1) **request´** (3SG.F_1, $king_2$) CAUSE **be-at´** (3SG.F_1, $permission_3$)]
PURP [**do´** (3SG.F_1) **visit´** (3SG.F_1, mother)]
AND [**do´** ($king_2$) CAUSE **be-at´** (3SG.F_1, $permission_3$)]

In the example (16), we have the verb *thug*, 'give', with three arguments in a construction that encapsulates an outer event with a resulting inner event encoded by the complex conjoined pair of VNs. The VNs share some of the arguments of the V1 in a complex nexus-juncture relation. The construction signature for this example is given in (17):

(16) *Thug sé cead daobhtha fir a chruin-niughadh*
Give-PST 3SG.M permission:N to:PREP+3PL men:N REL meet:VN
agus breith ar Cheallaigh Mhór.
and judge:VN on:PREP Cheallaigh Mhór:N
He gave permission to them (for) men to meet and judge on *Cheallaigh Mhór*

(17) *Thug* xNP *CEAD* yNP zNP *a* VN *agus* VN PPlocation
V1 _____V2 ____V3
[Event1] _ ALLOW _____ [Event2 & Event3]]
[T1 _____ <**later**> T2]

We outline the characteristics of each argument starting with the 'x' argument of the matrix verb *thug*, 'give', as shown in (18). The theme is *cead*, 'permission', i.e. the thing transferred. The 'y' argument is the recipient of the permission and has the characteristics in (19). This argument is the actor of the embedded VN. We can note that, while Modern Irish has the VSOX canonical word order, the embedded event has the SOV word order as a consequence of the nexus-juncture relation of the shared arguments found here.

(18) x: NP, 3SG.M
[+ ANIMATE]
[+ HUMAN]
[+ POSSESSOR_of_permission]
[+ CONTROLLER]

(19) y: NP coded as PPN *do*:PREP 'to'+3PL
[+ RECIPIENT_of_permission]
[+ ANIMATE]
[+ HUMAN]
[– CONTROLLER]

(20) z: NP, undergoer of embedded VN
[+ ANIMATE]
[+ HUMAN]
[+ CONTROLLER]

A similar example to the previous one is given in (21) but, in this case, there is no embedded inner event: the transfer of the permission to a recipient is all that is indicated. As GIVE is a three-place verb, this is represented as a prepositional ditransitive construction in Modern Irish. In this example, the recipient, which is coded as a prepositional pronoun, is clause-final and results in an ACTOR-THEME-RECIPIENT alignment order in the clause. This clause has the construction signature in (21b) and the logical structure in (21c). The purpose for which the permission was given is unspecified.

(21) a. Thug m' athair cead dom.
Give:V-pst my:possPN+father permission:N to:PREP+1sg
My father gave permission to me.
b. Thug xNP CEAD yPPN
c. [**do´** (my father) CAUSE **be-at´** (1sg, permission)] PURP [□_**UNSPECIFIED**]

4. GET_PERMISSION to Achieve a Particular PURPOSE

In the example (22), we again have a different view on the transfer of the permission. Here, the perspective is to do with GET_PERMISSION in this multi-clause sentence that encodes three events in a causal purposive multi-event chain.

(22) Cheól sé trí h-amhráin i ndiaidh a chéile sul a bhfuair sé cead a anál a tharraingt.
Cheól sé trí h-amhráin i ndiaidh a chéile
Sing-PST 3sg.M 3:QTY songs:N in:PREP after together:PN
sul a bhfuair sé cead a anál a tharraingt
BEFORE REL get-PST 3SG.M permission:N his:possPN breadth:N REL catch:VN
He sang three songs one after the other before he got permission to catch his breadth.

This construction has the construction signature indicated in (23), where Event1 occurs BEFORE Event2, and showing the verbs within each event. Event2 has the purpose of allowing Event3. Argument sharing occurs across the three events and the three verbs. We outline the characteristics of each argument starting with the 'x' argument in (24). The 'y' argument is the NP *trí h-amhráin i ndiaidh a chéile*, 'three songs one after the other', while *sul* encodes the BEFORE.UNTIL condition that terminates the first event and acts as a trigger for the second event. After the trigger fires, Event2 unfolds, facilitating Event3 as a result. Then, the 'x' argument is [RECIPIENT_of_permission+]. The [POSSESSOR_of_permission] is not specified. The underlying logical structure of the event chain can be represented as (25). This is represented in detail in (26).

(23) *Ceol* xNP yNP ADV *sul* '**before'** *fuair* xNP.recipient *CEAD* zNP *a* VN
V1 _____V2_____V3
[Event1] **BEFORE** [Event2] PURP [Event3]
T1 _____ <**trigger**> __ T2 _____ < **immediate**> T3

(24) x: NP, 3SG.M
[+ ANIMATE]
[+ HUMAN]
[+ ACTOR]
[– CONTROLLER]

(25) [[**LS1**] before.until [[**LS2**] purp [**LS3**]]]

(26) LS1: [**one_after_the_other´** [**do´** (3SG.M) **sing´** (3SG.M, three songs)]]
BEFORE.UNTIL
LS2: [[**do´** (3SG.M) **get´** (3SG.M, permission)]
PURP
LS3: [**do´** (3SG.M) **collect´** (3SG.M, breadth)]]

The primitive PURP implies that the action is not forced but taken willingly by the 'x' 3SG.M participant. Once the permission was received, then the barrier to action, i.e. the Event1 condition trigger, was no longer an impediment, facilitating the occurrence of Event2 and the achievement of Event3 as a consequence. Therefore, this example shows the GET PERMISSION perspective of the willing actor of the inner event, once the barrier to the purposive PURP action is removed. The primitive PURP shows (a) an inner participant that is a willing agent and (b) a lack of coercion on that participant.

5. LET_ALLOW Constructions

Next, we discuss the primitive LET as lexicalised by the verb *lig*, 'let', in Modern Irish. We first explore the typical example (27) and note the characteristics of its arguments, syntactic representation and clause structure. This construction has the construction signature indicated in (28).

(27) *Lig Moulin di an comhrá a stiúradh.*
Let-PST Moulin:N to:PREP+3SG.F DET conversation:N REL steering:VN
Moulin let her steer the conversation.

(28) *Lig* xNP yPPN zNP *a* VN :**construction signature**

The 'x' argument is an NP with the characteristics shown in (29), i.e. animate and human, letting the embedded event proceed with a willing embedded inner actor. The 'y' argument, shown in (30), is willing to do the inner event and is not coerced.

(29) x: NP 'Moulin'
[+ animate]
[+ human]
[+ actor]
[+ controller]

(30) y: NP, 3sg.F.dative
[+ undergoer of external event]
[+ actor of inner event]
[+ animate]

[+ human]
[– controller]
[+ volitional]

The 'z' argument *an comhrá*, 'the conversation', which is in unmarked accusative function, is also the direct object of the embedded verb which is realised as a verbal noun. Note that the undergoer of the primary verb, i.e. the 'y' argument, is dative-marked as a PPN [to:PREP+3SG.F]. The clause with the verb *lig*, 'let', has the [VSXO REL VN] order, where X is the dative-marked subject of the embedded VN and O is the object of the primary and embedded verb. The arguments are shared across the verbs in a nexus-junction relation with VSO as the primary word order and SOV as the secondary embedded word order. The VN encodes the embedded event as transitive, requiring two arguments. This clause has the underlying logical structure in (31). That is, *lig* has lexicalised the concept ALLOW within the construction.

(31) [**do´** (Moulin) CAUSE.ALLOW **guide´** (3SG.F, conversation)]

We now look at an example of *lig* in (32), which has multiple clauses representing three events, two of which occur in a degree of parallelism. As with the earlier examples, the arguments are shared across the multiple verbs. The construction has the construction signature in (33).

(32) Thit deora ar mo lámha ach ní ligfeadh an scanradh dom iad a mhothú.
Thit deora ar mo lámha ach
Fall-PST tears:N on:PREP my:possPN hands:N but:PART
ní ligfeadh an scanradh dom iad a mhothú.
NEG let-FUT det fright:N to:PREP+1SG 3PL REL feel:VN
Tears fell on my hands but the fright didn't let me feel them.

(33) thit xNP PP ach (NEG) lig yNP zPPN xNP a VN **:construction signature**
V1 _____V2 _____V3
[[EVENT1] **BUT** ___ [EVENT2] _____ [EVENT3]]
T1 _____ _____ T2 _____ T3

This has the logical structure representation of [**LS1**] **BUT** [**LS2** ALLOW **LS3**], as shown in (34). The 'y' argument of the verb *lig* is shared across the verbs; it has the characteristics indicated in (35). The 'z' argument of the verb *lig* is a prepositional pronoun that is considered as dative-marked, as shown in (36). The VN is the embedded V of Event3.

(34) **LS1:** [fall´ (tears1, [on_my_hands]2)
BUT
LS2, LS3: <neg [**do´** (the fright) CAUSE.ALLOW **feel´** (1sg2, 3pl1)]]

(35) y: NP 'an scanradh'
[+ def]
[– animate]
[– human]
[– controller]
[– volitional]
[+ force]

(36) z: PPN do:prep 'to'+1sg
[dative marked argument and actor of embedded VN]
[+ animate]
[+ human]
[+ volitional]

Worthy of note is that the verb *lig*, 'let', collocates frequently with a set of prepositions to license some different constructional senses of ALLOW, as shown in (37).

(37) The verb *lig* collocating with prepositions

Lig 'let, allow'	Let, allow: Base form and core sense of ALLOW
Lig ***amach*** *'*Let out'	Free, discharge, release
Lig ***anuas*** *'*Let down'	Loosen, lower down
Lig ***ar*** 'let on'	Allow to rest, allow to fall
Lig ***as*** 'let from'	Release, allow to escape
Lig ***chuig, chun*** 'let go, come'	Let go, let come
Lig ***de*** 'let from'	Release from, lay aside
lig ***do*** 'let to'	Allow, let (someone) be
Lig ***faoi*** 'let under'	Let underneath
Lig ***i*** 'let in'	Let into, allow to enter
Lig ***isteach*** 'let inside'	Let in, admit
Lig ***le*** 'let with'	Let out, lengthen, extend
Lig ***ó*** 'let from'	Let go, relinquish
Lig ***siar*** 'let back'	Let back, swallow, let (something) fall
Lig ***síos*** 'let down'	Let down
Lig ***thar*** 'let pass'	Let pass
Lig ***trí*** 'let through'	Let through

6. PERMIT Constructions

We now examine the *cheadaigh* constructions, which can be contrasted with the GIVE_PERMISSION construction. The verb *cheadaigh*, 'permit', lexicalizes the concept PERMIT within its lexical semantics. There are essentially two construction forms representing this type of constructions. The first variant of the construction, as in (38), consists of a single event encoded within a single verb in the construction. The second variant of the construction, as in (40), consists of a complex multi-event with two verbs and argument sharing. The construction signatures of each variant are shown in (39) and (41) respectively. We show first the single-event single-verb construction. Within the verb *cheadaigh*, the concept PERMIT lexicalises the notion of TRANSFER OF AUTHORITY from the 'x' actor to the 'z' recipient.

(38) *Cheadaigh an chúirt an dlí um íosphá i Washington*
Permit-PST DET court DET law about:PREP low-pay:N in:PREP Washington:N
The court permitted the minimum-wage law in Washington.

(39) *Cheadaigh* xNP yNP zPP (z1PP) :**construction signature**

We show next the complex multi-event with two verbs and argument sharing. This illustrates the event permitted according to the construction signature in (41). The 'z' argument in the construction signature, i.e. the actor of the inner event, is optional, so it doesn't need to be represented or realised within the construction, as indicated in (43) and (44) showing the respective construction signatures.

(40) *Cheadaigh Kennedy don CIA ionradh náireach*
permit-PST Kennedy to:PREP+DET CIA:N invasion:N shameful:ADJ
a dhéanamh ar Chúba
REL do:VN on:PREP Cuba
LIT: Kennedy permitted to the CIA to do.make the infamous invasion on Cuba
Kennedy permitted the infamous CIA invasion of Cuba

(41) *Cheadaigh* xNP zPP yNP a VN :**construction signature**
V1 _____ V2
[[Event1] _____ PERMIT [Event2]]
T1 _____<**later**> T2

(42) *Ní ceadóidh Alice Walker 'The Colour Purple' a aistriú go hEabhrais*.
NEG permit-PRS Alice Walker 'The Colour Purple' REL translate:VN to:PREP

Hebrew:N
Alice Walker did not permit 'The Colour Purple' be translated to Hebrew.

(43) (NEG) *ceadóidh* xNP yNP a VN PPgoal :**construction signature**

(44) *Cheadaigh* xNP (zPP) yNP a VN :**construction signature**
V1 _____ V2
[[Event1] _____ PERMIT [Event2]]
T1 _____ <**later**> T2

(45) <NEG [**do´** (AW) CAUSE.PERMIT [**do´** (Ø) **translate_to´** (TCP, Hebrew)]

(46) x: NP 'Alice Walker'
[+ ANIMATE]
[+ HUMAN]
[+ ACTOR]
[+ CONTROLLER]
[+ VOLITIONAL]

(47) y: NP 'The Colour Purple'
[– ANIMATE]
[– HUMAN]
[+ UNDERGOER]
[– CONTROLLER]
[– VOLITIONAL]

7. Discussion

In this chapter, we have discussed the less direct causation involving the role that the concepts of TRANSFER, LET/ALLOW, GIVE, PERMISSION and PERMIT play. Many interesting and significant facts about the syntactic realisation of these were characterised, namely, the resolution of argument-structure relations when multi-verb and multi-event clauses are involved. In many instances, volition (or the lack of it) was a factor of the characterisation of an argument participant, especially the non-actor who was often [– VOLITION]. Similarly, control was also seen to be a factor where the actor of the outer event was [+ CONTROLLER] while a downstream argument was [– CONTROLLER]. Along with coercion, the willingness to undertake an action was also seen. It would be incorrect to state that all causee arguments were unwilling to undertake an act as many were [+ WILLING] and the syntactic realisation indicated this, especially with the choice of the verb used to denote the inner event.

Talmy (2000: 407*ff*) has characterised causation in an account that describes in semantic terms the impact of force dynamics on event chains.

Force dynamics is considered to play a structuring role across a range of language levels and has direct grammatical representation. Force dynamics is seen to emerge as systematic. As such, it structures conceptual material pertaining to force interaction in a common way across an event frame. In this perspective, the referent situation is considered as an event frame that refers to a generic unitary conceptual category resulting from the systematic segmentation of the occurrence of phenomena by human cognition. This event frame evokes a set of conceptual elements and inter-relationships, which are felt to be central and to constitute a coherent unit. One of the universal types of the event frame is a causal chain that refers to a conceptualised sequence of linked sub-events over some duration that may be immediate or have a longer temporal extent. The sequence of linked sub-events results from conceptually chunking a 'causal continuum' into relatively discrete packets. The causal chain of the event frame is demarcated by the initiating volitional act of an agent and by the final goal that the agent intends as a result of this act. In other words, the agent's volitional act and the goal mark the beginning and the end of the causal chain of the event frame. This causal chain is initiated by the agent's act of volition. It then progresses through a sequence of intermediate causally chained sub-events leading to the penultimate sub-event. This is the situation that we have found in Modern Irish with the concepts of TRANSFER, LET/ALLOW and PERMISSION, and this is how they are realised. The factors of control and volition in the sense of Dixon (2010) are of central importance, as referred in the introduction.

As is argued in this chapter, highlighting different aspects of the event is important to our analysis as languages differ in whether, for example, they conceive of 'GIVE' or 'PUT' as TRANSFER events or as 'cause to HAVE' events, that can include possession of permission and authority. We examined these constructions with the GIVE, PUT, TAKE and GET verbs and their arguments across Dixon's semantic parameters of causation. We examined GIVE_PERMISSION constructions as consisting of multi-verbs and multi-events with particular syntactic patterns that occur while encoding both the outer and inner event. Specifically, the transferred theme is the NP PERMISSION. Within these constructions, the PURPOSE for which the permission is given is explicitly stated in the syntax as an embedded event. As such, these are complex multi-verb constructions. In the causative GIVE, we found that the *giving of permission* is not lexicalised as a PERMIT verb. The GET_PERMISSION construction encapsulates the perspective of the willing actor of the inner event once the barrier to the purposive PURP action is removed. The primitive PURP signals (a) an inner participant that is a willing agent and (b) a lack of coercion on that participant.

With regard to LET_ALLOW, the verb *lig*, 'let', lexicalises the concept of ALLOW within the construction. In *lig* constructions, we also have multiple verbs and multiple events, with *lig* as the matrix verb and the allowed action realised as a verbal noun. The word order of the matrix clause is VSO but the word order of the embedded clause is SOV with the dative-marked PPN also acting as the subject of the embedded VN. The verb *lig* frequently collocates with a set of prepositions to license some different senses of ALLOW. In relation to PERMIT, the verb *cheadaigh*, 'permit', lexicalizes the concept PERMIT within its lexical semantics along with the transfer of authority from the 'x' actor to the 'z' recipient. We can summarise these dimensions as (48) for the GIVE, PUT, TAKE and GET verbs.

Additionally, in somewhat more detail, we can indicate these TRANSFER, LET/ALLOW, GIVE, PERMISSION and PERMIT concepts as (49) with their respective event chains and time-extent indicators.

It is clear, in looking at the findings of this brief analysis of indirect causation across the concepts of TRANSFER, LET/ALLOW and PERMISSION with the GIVE, PUT, TAKE, GET, LET and PERMIT verbs, that there is a set of causal relata between two events, as described in (50), and also a dependency relationship between eventualities, as shown in (51). Notwithstanding the dependency relation, we cannot assume that the dependent event will actually occur if the primary event happened. This is the nature of the concepts of TRANSFER, LET/ALLOW and PERMISSION across these verbs.

(48) The concepts TRANSFER, GIVE, ALLOW and PERMIT in constructions

TRANSFER	Transfer of theme to a recipient or location			
GIVE PERMISSION	Transfer of permission			Purpose of permission is explicit
GET PERMISSION	Transfer of permission to recipient			Purpose of permission may be explicit
LET/ALLOW		Removal of a barrier to action	Event is not impeded	
PERMIT	Transfer of authorisation			Barrier to action is not placed

(49) Event chains of TRANSFER, GIVE, ALLOW and PERMIT in constructions

a. **TRANSFER:** TRANSFER of theme to a recipient or location
[Event 1] _____ **ALLOW** ____ [Event 2 (& Event 3)]]
[T1 _____ <**later**> _____ T2]

b. **Give permission:** Transfer of permission and purpose of permission explicit
[Event] **before** [Event 2] purp [Event 3]
T1 _____ <**trigger**> __ T2 _____ < **immediate**> T3

c. **GET PERMISSION:** TRANSFER of PERMISSION and PURPose of permission (may be) explicit
[Event 1] PURP [Event 2]
T1 _____<**trigger**> __ T2

d. **ALLOW:** REMOVAL of a barrier to action and event not impeded
[[Event 1] **BUT** _____ [Event 2] _____ [Event 3]]
T1 _____ T2 _____ T3

e. **PERMIT:** TRANSFER OF AUTHORISATION and a barrier to action not placed
[[Event 1] _____ **PERMIT** [Event 2]]
T1 _____ <**later**> T2

(50) The causal relata between two events

a. Causation is a relation between two events: a causing event and a caused event.
b. Causation has a temporal dimension such that the causing event must precede the caused event.
c. Causation has counterfactual dimension:
IF the causing event had not occurred,
THEN the caused event would also not have occurred.

(51) The dependency relation between caused eventualities

a. Dependence is a relation between the two eventualities of an independently-existing causing eventuality (the 'outer' event) and a dependent eventuality (the 'inner' event).
b. Dependence has a counterfactual dimension:
IF the independently-existing causing eventuality was absent,
THEN the dependent eventuality would not occur.

(52) The nature of the commitment to caused event and caused result

a. e_1 [**CAUSE**] e_2 :Direct causation has an actual causal commitment to the caused result
b. e_1 [**TRANSFER**] (e_2) :Indirect causation with no actual causal commitment to a caused event
c. e_1 [**LET**] (e_2) :Indirect causation with no actual causal commitment to a caused event

d. e_1 [**ALLOW**] (e_2) :Indirect causation with no actual causal commitment to a caused event

e.

f. e_1 [**PERMIT**] (e_2) :Indirect causation with no actual causal commitment to a caused event

This chapter has briefly explored how Modern Irish encodes the concepts of TRANSFER, LET/ALLOW and PERMISSION with the GIVE, PUT, TAKE, GET, LET and PERMIT verbs and how their construction patterns impact on the syntax while raising important issues for understanding argument realisations across causal event chains with complex clauses that contain multiple verbs. These concepts of TRANSFER, LET/ALLOW and PERMISSION with the GIVE, PUT, TAKE, GET, LET and PERMIT verbs give us important insights into the nature of indirect causation and the causal relata between two events understood as causal forces in the sense of Talmy. These events have an intrinsic dependency relationship between the eventualities, such that we cannot assume that the dependent event will actually occur if the primary event happened. This is the core sense of the meaning of these concepts of TRANSFER, LET/ALLOW and PERMISSION across these verbs.

8. References

Dixon R. M. W. A typology of causatives: form, syntax and meaning. In Dixon R. M. W and Alexandra Y. Aikhenvald 2010. *Changing Valency: Case Studies in Transitivity*. Cambridge: Cambridge University Press. 30–83. 2010.

Gärdenfors, Peter. 2007. "Representing actions and functional properties in conceptual spaces." In Tom Ziemke, Jordan Zlatev, Roslyn M. Frank, eds., *Body, Language, and Mind*, Volume 1, Embodiment. Mouton de Gruyter.

Nolan, Brian. The Layered Structure of the Modern Irish Word: An RRG Account of Derivational Morphology Based on Lexeme Constructional Schemata. In Wataru Nakamura (ed). *Proceedings of the 10th International Conference on Role and Reference Grammar (RRG2009).* 2010. [Accessed in January 2015]: http://wings.buffalo.edu/linguistics//people/faculty/vanvalin/rrg/ProceedingsofRRG2009_02.pdf.

—. Meaning Construction and Grammatical Inflection in the Layered Structure of the Irish Word: An RRG Account of Morphological Constructions. In Watara Nakamura (ed.). *New*

perspectives in Role and Reference Grammar. London: Cambridge Scholars Publishing. 2011.

—. *The structure of Irish: A functional account*. Sheffield: Equinox Publishing Co. 2012a.

—. The GET constructions of Modern Irish and Irish English: GET passive and GET-recipient variations. In Lenz, Alexandra N. and Gudrun Rawoens. *The Art of Getting: GET verbs in European languages from a synchronic and diachronic point of view*. SPECIAL ISSUE OF LINGUISTICS. 50-6: 1111–1162. 2012b.

—. Constructional polysemy and argument realisation with the Irish GET verb. In Johanna Barðdal, Michaela Cennamo, Elly van Gelderen (eds.): *Argument Structures in Flux: The Naples/Capri papers*. [STUDIES IN LANGUAGE COMPANION SERIES 133]. Amsterdam: John Benjamins Publishing Company. 2013a.

—. Constructions as grammatical objects: A case study of the prepositional ditransitive constructions in Modern Irish. In Nolan, Brian and Elke Diedrichsen. 2013. *Linking Constructions into functional linguistics – The role of constructions in RRG grammars*. [STUDIES IN LANGUAGE COMPANION SERIES 145]. Amsterdam: John Benjamins Publishing Company. Pages 143–178. 2013b.

—. Encoding transfer, let/allow and permission in Modern Irish. In Nolan, Brian, Gudrun Rawoens and Elke Diedrichsen. *Causation, transfer and permission: Argument realisation in GET, TAKE, PUT, GIVE and LET verbs.* [STUDIES IN LANGUAGE COMPANION SERIES 167]. Amsterdam: John Benjamins Publishing Company. Pages 13 – 51. 2015

Nolan, Brian and Elke Diedrichsen. *Linking Constructions into functional linguistics – The role of constructions in RRG grammars*. [STUDIES IN LANGUAGE COMPANION SERIES 145]. Amsterdam: John Benjamins Publishing Company. 2013.

Nolan, Brian and Carlos Periñán. *Language processing and grammars: The role of functionally oriented computational models*. [STUDIES IN LANGUAGE COMPANION SERIES 150]. Amsterdam: John Benjamins Publishing Company. 2014.

Nolan, Brian, Gudrun Rawoens and Elke Diedrichsen. *Causation, transfer and permission: Argument realisation in GET, TAKE, PUT, GIVE and LET verbs.* [STUDIES IN LANGUAGE COMPANION SERIES 167]. Amsterdam: John Benjamins Publishing Company. 2015.

Song, Jae Jung. *Causatives and Causation: A Universal-typological perspective* (Longman Linguistics Library). London: Addison Wesley Longman. 1996.

Talmy, Leonard. *Toward a cognitive semantics volume I: Concept structuring systems.* [Language, Speech, and Communication Series]. Cambridge, Massachusetts and London, England: The MIT Press. 2000.

Van Valin, Robert D.; *Exploring the Syntax-Semantics Interface*. Cambridge: CUP. 2005.

CHAPTER TWO

CAUSATIVE *LASSEN* CONSTRUCTIONS IN GERMAN: SYNTAX, ARGUMENT STRUCTURE, MEANING VARIANTS AND THE IMPACT OF CULTURAL KNOWLEDGE IN DISAMBIGUATION

ELKE DIEDRICHSEN
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN, DUBLIN

1. Introduction

In this chapter, I will investigate the semantics of the German *lassen* construction, which expresses gradients between causation, permission and non-intervention. The verb meaning of *lassen* is 'let, allow'. The causativity that may be expressed in the *lassen* construction is in many cases ambiguous or a matter of degree. The sense of mere permission is almost always possible, as shown in (1).

Across languages, the ambiguity with causative constructions is a well-known fact (Kulikov 2001). In the absence of formal indicators, language users often rely on contextual factors in order to disambiguate the structure (Rawoens and Egan 2013). I will argue later in this paper that cultural knowledge plays a role as well.

(1) *Er ließ den Hund das Kissen zerfetzen*
3MsgNOM let.PAST3sg DEFMsgACC dog.sg DEFNsgACC pillow tear.apart.INF
He let the dog tear the pillow apart.
Sense 1: 'zulassen': He did not intervene, even though he did not like it
Sense 2: He allowed the dog to tear the pillow apart
Sense 3: He made the dog tear the pillow apart.

The discussion in this chapter will be limited to German *lassen* constructions with a complete structure, where a causer as the agent argument of *lassen* is added to a structure that includes a causee (the agent or undergoer of the original structure) and possibly some other arguments. In the resulting structure, the causee will have accusative case. If the original structure is transitive, the sentence will have two arguments marked in the accusative.

Formal deviations from this pattern occur, and they have an impact on the semantics of the structure. For example, the causee may be lacking in the *lassen* construction (see Nedjalkov 1976 for examples), and this has an effect on the overall interpretation of the sentence, as Rawoens and Egan (2013) show for Swedish. Moreover, the causee may appear with a *von* phrase, which resembles the appearance of the agent in passive constructions in German. *Lassen* constructions with a causee in the *von* phrase will be called 'passive' here, following a recent publication on German causatives by Enzinger (2010). The status of these constructions as 'passive' is, however, disputable (cf. Kemmer and Verhagen 1994:132).

2. Other Construction Types with *Lassen* in German

There is a variety of construction types with *lassen* which are similar to the causative *lassen* construction, but belong to different construction types. They will be introduced with the examples (2)-(9) in this section (cf. Duden online for a recent description of the uses of *lassen*).

2.1. Collocations with *Lassen*

(2) *Das* *lässt* *sich* *hören!*
DEFNsgACC let.PRES3sg 3sgREFL hear.INF
Good/impressive to hear that!

(3) *es* *krachen lassen*
3NsgACC bang.INF let.INF
have a big party

(4) *Ich* *lasse* *mir* *das* *nicht* *bieten.*
1sgNOM let.INF 1sgDAT DEFNsgACC not present.INF
I will not accept that.

(5) *Das* *muss* *man* *ihm* *lassen.*
DEFNsgACC must.PRES3sg one 3MsgDAT let.INF

You have to appreciate that (even though you generally don't think too highly of him).

2.2. *Lassen* in Middle Constructions (Fagan 1992)

(6) *Die* *Dose lässt* *sich* *leicht* *öffnen*
DEFFsgNOM can.sg let.PRES3sg 3sgREFL easily open.INF
= Die Dose öffnet sich leicht
DEFFsgNOM can.sg open.PRES3sg 3sgREFL easily
The can opens easily.

2.3. Full Verb Uses of *Lassen*

(7) *Lass* *das!*
Let.IMPsg DEFNsgACC

Stop it! / Don't do it!

(8) *Im* *Casino* *hat* *er* *viel*
PREP.DEF3sgDAT casino.sg have.PRES3sg 3MsgNOM much
Geld *gelassen*
money let.PSTP
He has spent/lost a lot of money in the casino.

2.4. Adhortative Constructions

(9) *Lasst* *uns* *froh* *und* *munter* *sein*
Let.IMPpl 1plACC happy and cheerful be.INF
(start of a song for the Christmas season)
'Let's cheer and be happy'

3. Syntax and Semantics

3.1. The Argument Structure of the *Lassen* Construction

In the causative *lassen* construction in German, *lassen* introduces an A argument that bears some responsibility with respect to the event expressed in the main verb. The responsibility value of the A argument may range from force or power to make the event happen, via permission (i.e. legal causation) to mere non-intervention, which is still some kind of causation, as one could say that the event happened because A did not intervene and finally something happened, but A did not care/did not notice.

For transitive sentences, this leads to a structure with one A argument (the causer) and two O arguments.

(10) Transitive construction

a) *Er backt den Kuchen.*
3MsgNOM bake.PRES3sg DEFMsgACC cake
He bakes the cake.

b) *Ich lasse ihn den Kuchen backen.*
1sgNOM let.PRES1sg 3MsgACC DEFMsgACC cake bake.INF
I let him bake the cake.

c) *Der Kuchen wird von ihm gebacken.*
DEFMsgACC cake become.PRES3sg PART(von) 3MsgDAT bake.PSTP
The cake is baked by him.

d) *Ich lasse den Kuchen von ihm backen.*
1sgNOM let.PRES1sg DEFMsgACC cake PART(von) 3MsgDAT bake.INF
I have the cake baked by him.

In (10), the person who bakes the cake appears as an A argument of the transitive construction in (10a). This is the causee in (10b). The causee appears as an O argument in the *lassen* causative. The causative transitive construction in (10b) has therefore two O arguments, and both are marked with the accusative case. The causee *ihn* ('him') is the O argument of *lassen* and the object *den Kuchen* ('the (ACC) cake') is the O argument of *backen*.

(10d) shows that the *lassen* construction allows a passive variant in which the object of the baking retains its O status and the causee (the person who bakes the cake) appears as a peripheral *von* phrase. Compare (10c): here, the active construction from (10a) is passivized. Note that, for the passive in the *lassen* construction in (10d), the passive auxiliary *werden* is not applied, and the full verb appears in the infinitive, and not in the past participle. Furthermore, the O of the active construction appears as S in the plain passive, but as O in the *lassen* passive. For the *lassen* construction, be it active or passive, *lassen* is the finite auxiliary. The passive *lassen* construction can only be formally identified as a passive by the *von* phrase —see Enzinger (2010: 26) for further discussion on the *lassen* passive, and Kemmer and Verhagen (1994) for arguments against the classification as passive.

(11) Ditransitive construction

a) *Mein Sohn gibt dem Obdachlosen das Geldstück.*
POSSMsgNOM son.sg give.PRES3sg DEFMsgDAT homeless.person.sgDAT

DEFNsgACC coin.sg
My son gives the coin to the homeless person.

b) *Ich lasse meinen Sohn dem Obdachlosen*
1sgNOM let.PRES1sg POSSMsgACC son.sg DEFMsgDAT homeless.person.sgDAT
das Geldstück geben.
DEFNsgACC coin.sg give.INF
I let my son give a coin to the homeless person.

c) *Dem Obdachlosen wird das Geldstück*
DEFMsgDAT homeless.person.sgDAT become.PRES3sg DEFNsgACC coin.sg
von meinem Sohn gegeben.
PART(*von*) POSSMsgDAT son.sg give.PSTP
The coin is given to the homeless person by my son.

d) *Ich lasse dem Obdachlosen*
1sgNOM let.PRES1sg DEFMsgDAT homeless.person.sgDAT
das Geldstück von meinem Sohn geben.
DEFNsgACC coin.sg PART(*von*) POSSMsgDAT son.sg ive.INF
I have the coin given to the homeless person by my son.

In the ditransitive *lassen* construction, the indirect object is dative-marked. It retains its dative-marking in both the *lassen* and the passive constructions. The other arguments of the ditransitive *lassen* construction behave as described in (10).

3.2. The Syntax of the *Lassen* Construction

In his typological approach to causativity, Dixon (2000) describes a construction type like the *lassen* causative as a periphrastic construction (his type iii). Vogel (2009), Enzinger (2010) and Bausewein (1991) regard the *lassen* construction as an ACI (*accusativus cum infinitivo*) construction, as it changes a complete sentence structure such that its subject becomes an accusative argument and the full verb appears in the infinitive. This is done by adding a new verb (*lassen*) plus its A argument to that structure. The causative *lassen* construction and the ACI construction behave very much alike in a number of construction types.

(12) The position of *lassen* and *sehen* in a *lassen* causative (b) and an ACI construction (c).

a) Transitive construction
Er hebt den Geldschein auf.
3MsgNOM pick.PRES3sg DEFMsgACC banknote.sg up
He picks up the banknote.

b) Causative *lassen* construction
Der *Polizist* *lässt* *ihn*
DEFMsgNOM policeman.sg let.PRES3sg 3MsgACC
den *Geldschein* *aufheben.*
DEFMsgACC banknote.sg pick.up.INF
The policeman lets him pick up the banknote.
c) ACI construction
Der *Polizist* *sieht* *ihn*
DEFMsgNOM policeman.sg see.PRES3sg 3MsgACC
den *Geldschein* *aufheben.*
DEFMsgACC banknote.sg pick.up.INF
The policeman sees him picking up the banknote.

(13) Word order in the periphrastic perfect: the causative *lassen* construction.
a) *Der* *Polizist* *hat* *ihn* *den* *Geldschein*
DEFMsgNOM policeman.sg have.PRES3sg 3MsgACC
DEFMsgACC banknote.sg
aufheben *lassen.*
pick.up.INF let.INF
The policeman let him pick up the banknote.
b) Word order in the periphrastic perfect: ACI construction
Der *Polizist* *hat* *ihn* *den* *Geldschein*
DEFMsgNOM policeman.sg have.PRES3sg 3MsgACC DEFMsgACC banknote.sg
aufheben *sehen.*
pick.up.INF see.INF
The policeman saw him pick up the banknote.

However, even though the causative *lassen* constructions have a very similar linear structure as the ACI constructions with perceptive verbs, there are important differences in the syntactic behaviour of these two construction types.

First, the *lassen* construction allows a passive variant in which the causee appears as a *von* phrase or may be omitted —see Kemmer and Verhagen (1994) for a different approach to these *lassen* structures with a *von* phrase. No such variant with the agent in the *von* phrase can be formed from ACI constructions with a verb of perception, as in (14b). Moreover, the *lassen* construction allows the omission of the causee, while the agent argument cannot be left out of the perceptive ACI construction, as in (15b).

(14)
a) *Er* *lässt* *den* *Geldschein* *von* *ihm*
3MsgNOM let.PRES3sg DEFMsgACC banknote.sg PART(von) 3MsgDAT
aufheben.
pick.up.INF
He lets the banknote be picked up by him.

b) **Er sieht den Geldschein von ihm*
3MsgNOM see.PRES3sg DEFMsgACC banknote.sg PART(von) 3MsgDAT
aufheben.
pick.up.INF
He sees the banknote being picked up by him.

(15)
a) *Er lässt den Geldschein aufheben.*
3MsgNOM let.PRES3sg DEFMsgACC banknote.sg pick.up.INF
He arranges for the banknote to be picked up.
b) **Er sieht den Geldschein aufheben.*
3MsgNOM see.PRES3sg DEFMsgACC banknote.sg pick.up.INF
He sees the banknote being picked up.

The *lassen* causative is a complex predicate in which two predicational elements combine to predicate as a single element. These complex predications are sometimes structurally similar with the constructions where two syntactically separate domains of predication apply and the arguments are shared across the domains. Control and raising constructions are examples of the latter (Butt 2003).

This syntactic difference can be demonstrated for the *lassen* construction and the ACI construction in German as well. The *lassen* construction as a complex mono-clausal predicate allows "clitic climbing" (Butt 2003:4), where a clitic or a weak personal pronoun can be used anaphorically in the construction, as in (16b). The same is not acceptable in an ACI construction, as shown in (17b), which shows that the constructions must be syntactically different. The ACI construction is a bi-clausal construction, where the two verbs are effective in two separate domains of predication. This can also be illustrated by showing that the ACI construction is indeed a shortened version of a complex construction with a complement clause, as in (18b), whereas such a complex construction cannot be formed from a causative *lassen* construction, as in (18a) —see Butt (2003:4-5) and Rosen (1989) for more syntactic tests that facilitate the syntactic differentiation of these two construction types.

(16)
a) *Sie lässt ihn das Auto einparken*
3FsgNOM let.PRES3sg 3MsgACC DEFNsgACC car.pl park.INF
She lets him park the car.
b) *Sie lässt es ihn einparken*
3FsgNOM let.PRES3sg 3NsgACC 3MsgACC park.INF
She lets him park it.

(17)
a) *Sie sieht ihn das Auto einparken*
3FsgNOM see.PRES3sg 3MsgACC DEFNsgACC car.pl park.INF
She sees him park the car.
b) **Sie sieht es ihn einparken*
3FsgNOM see.PRES3sg 3NsgACC 3MsgACC park.INF
She sees him park it.

(18)
a) **Sie lässt, dass er das Auto einparkt.*
3FsgNOM let.PRES3sg that 3MsgNOM DEFNsgACC car.pl park.PRES3sg
She lets = allows that he parks the car.
b) *Sie sieht, dass er das Auto einparkt.*
3FsgNOM see.PRES3sg that 3MsgNOM DEFNsgACC car.pl park.PRES3sg
She sees that he is parking the car.

The *lassen* causative as a complex predicate construction allows a variety of meanings. It is generally ambiguous between a causation meaning and a permission/non-intervention meaning. The meaning variants discussed in the following section are evaluated against Dixon's (2000) set of parameters of causativity. After that, I will suggest a constructional schema for the German *lassen* construction, which will display its semantic and syntactic features.

3.3. Meaning Variants and Disambiguation by Cultural Knowledge

This section describes the meaning variants of the causative *lassen* construction, which can express degrees of causation or responsibility on the side of the causer. Note that the examples discussed here are not supposed to cover the full range of causative *lassen* constructions in German. They are used in order to exemplify meaning differences in formally identical structures. The next section introduces a causativity scale for *lassen* constructions in German. For convenience and for a better readability of the scale, I assign the constructions a letter-number combination starting from C1, where 'C' stands for causativity. The notations C1, C2, etc. can be found in the scale.

It is argued that many of the causative *lassen* constructions can be disambiguated on the basis of cultural knowledge. This may be, for example, knowledge about the distribution of roles between people in terms of authority and responsibility. This cultural knowledge results from the experience with typical situations of the kind expressed in the respective sentence.

(19) The *lassen* constructions and their notation for placement on the scale

C1: *Hans ließ den Mantel am Haken hängen.*
Hans let.PAST3sg DEFMsgACC coat.sg PREP.DEFMsgDAT hook hang.INF
Hans left the coat hanging on the hook.

C2: *Er ließ die Torte stehen.*
3MsgNOM let.PAST3sg DEFFsgACC cake.sg stand.INF
He left the cake standing where it was.

C3: *Hans ließ mir den Mantel hängen.*
Hans let.PAST3sg 1sgDAT DEFMsgACC coat.sg hang.INF
Hans left the coat hanging for me to have it.

C4: *Sie ließen die Kinder länger schlafen.*
3plNOM let.PAST3pl DEFplACC child.pl longer sleep.INF
They let the kids sleep longer.

C5: *Er ließ den Hund das Kissen zerfetzen.*
3MsgNOM let.PAST3sg DEFMsgACC dog.sg DEFNsgACC pillow tear.apart.INF
He let the dog tear the pillow apart.

C6: *Er lässt den Einbrecher laufen.*
3MsgNOM let.PRES3sg DEFMsgACC intruder run.INF
He lets the intruder escape.

C7: *Er lässt die Kinder die Bonbons essen.*
3MsgNOM let.PRES3sg DEFplACC child.pl DEFplACC sweet.pl eat.INF
He lets the children eat the sweets.

C8: *Er lässt die Kinder den Film sehen.*
3MsgNOM t.PRES3sg DEFplACC child.pl DEFMsgACC film.sg watch.INF
He lets the children watch the movie.

C9: *Er lässt die Kinder das Gedicht lernen.*
3MsgNOM let.PRES3sg DEFplACC child.pl DEFNsgACC poem.sg learn.INF
He lets the children learn the poem = He makes the children learn the poem

C10: *Er lässt den Gärtner den Rasen mähen.*
3MsgNOM let.PRES3sg DEFMsgACC gardener DEFMsgACC lawn.sg mow.INF
He lets the gardener mow the lawn.

C11: *Er lässt den Wasserhahn laufen.*
3MsgNOM let.PRES3sg DEFMsgACC tap.sg run.INF
He lets the tap run = he turns it on

C12: *Er lässt den Plattenspieler laufen.*
3MsgNOM let.PRES3sg DEFMsgACC record.player run.INF
He lets the record player play = he turns it on

C13: *Der Wind lässt das Haus zittern.*
DEFMsgNOM wind.sg let.PRES3sg DEFNsgACC house.sg tremble.INF
The wind makes the house shake.

C14: *Blauer Lidstrich lässt Ihre Augen strahlen.*
Blue.MsgNOM eyeliner let.PRES3sg 2(pol)POSSplACC eye.pl sparkle.INF
Blue eyeliner makes your eyes sparkle.

The examples allow a reading in which the causer, i.e. the A argument of *lassen*, lets something happen without intervening. In the first two examples, this may happen by accident, while in (C4-C6) it is more plausible to assume that the causer does not intervene purposefully. Note that in (C3), the person leaves the coat on purpose, in order for another person to have it. The dative argument *mir* makes clear that there is an element of transfer involved in the non-intervention (i.e. he lets the other person have it).

Not intervening purposefully is a pre-stage of permission. I would, however, not describe the causer's role in these events as that of someone who gives permission. Even though that person has control and a certain degree of authority in the situation (e.g. he could wake the kids and stop the dog from tearing up the pillow), his role remains passive, and there is no explicit permission stated.

In (C7-C8), a reading in terms of non-intervention is possible (i.e. the kids are eating or watching the movie, and the father does not intervene). However, a more plausible scenario is that the kids want to do something that requires the permission of an adult, and the adult grants permission. Therefore, the permitting person acts purposefully, and, instead of just letting something happen, he explicitly grants permission in virtue of his authority and control over the situation. It has also to be noted that the act of giving permission entails that something happens or may be done for the benefit of the causee. Giving permission means that, measured against causation, the causer causes the causee to do something that the causee wants and that is for the benefit of the causee.

Due to their cultural experience, speakers know that, in a relationship that involves responsibility and authority, but also love and care, these acts of giving permission are part of the interaction between, for example, parents and their children. The 'plausibility' that this chapter deals with for the semantic interpretation of the examples rests on the experience with the situations described in the sentences. It is culturally acquired

knowledge. Semantically, these examples are highly ambiguous, and the use of the *lassen* construction generally does not default to one of the readings, unless the sentence evokes a situation known to the language user, or the context provides the necessary hints.

In (C9), the degree of control that the causer exhibits is even higher. Here, it is about a situation in which the causee is asked to do something as an order from the authority of the causer. The benefit may still be with the causee, but the authority of the causer is stronger in that he demands something, rather than giving permission to the causer to do something that the causee wants. Note also that this construction could be translated into English by the *make* construction which expresses direct causation.

It is possible to read this in terms of: 'He permits the children to learn the poem, provided that this is what they want and they ask for it'. Let's assume the sentence expresses the most plausible variant, which is that the children do not want to learn the poem, even though it is, after all, for their own benefit, and that they do learn it on request, or following an order, by the teacher. Thus, in this case, the causer has full control and authority, as he causes the causee to do something that the causee does not want to do (even though it is for the benefit of the causee). This kind of causation in terms of direction that involves some benefit for the causee can be found in situations of tuition, instruction and guidance, e.g. in school or other learning situations. Again, the knowledge about these situations is based on cultural experience.

In (C10-C12), the benefit to be expected from the activity expressed in the full verb is no longer with the causee.

The sentence in (C10) has the sense of control and authority in its most plausible reading: the causer causes the gardener to mow the lawn by asking (=ordering) him to do it. He has the authority to ask this action from the gardener, as the gardener, in the reading suggested here, is his employee and will receive money for obeying the orders. Therefore, a true causative sense is expressed here, even though it is not the only possible reading. From the mere form and choice of words, it is possible that the scenario describes a man who finds a strange gardener mowing his lawn and permits him to do it. However, on the basis of cultural knowledge about professional roles and employer-employee relationships, the causative reading is the most plausible one.

The sense of permission is always possible, as long as the causee is animate. A human group or individual as well as a domesticized animal would be subject to permission, while this is certainly not possible for an inanimate causee. Of course, the *lassen* construction is also possible with inanimate causees, as in (C11) and (C12). Here, one possible sense is

causation: permission is excluded, as an inanimate object would not ask for or respond to permission. However, an interpretation in terms of non-intervention in an ongoing event is possible in a context-free utterance like (C11) or (C12).

If we assume that in (C11) and (C12), causation is the intended reading, it is direct and immediate causation. The beneficiary of the whole event (or its result) is clearly the causer, and not the causee. For (C10-C12), it is true that the power exhibited by the causer also entails that the causee is committed to producing a result that can be evaluated by the causer.

Regarding all the examples and degrees of causation discussed for the *lassen* construction so far, it seems that the ambiguity between the causation, the permission and the non-intervention reading is always there. Is it? If the causee is not animate, the permission reading cannot be construed, as you could not grant permission to something inanimate. The reading in terms of non-intervention, however, seems to be always possible, as even a non-animate causee could generally be involved in an ongoing event. Non-intervention entails that the causer could intervene. How about inanimate causers, as they are found, for example, in descriptions of natural forces, but also in advertisements or articles that explain the effect of a product, as in (C13) and (C14)?

Even though the causer is not animate in those examples, they are the only examples discussed so far where the only possible interpretation is the one in terms of causation. Neither permission nor non-intervention is possible, as an inanimate force could neither grant permission nor intervene (or decide not to intervene) in an ongoing action. If one were to put up a scale of "degrees of causativity with *lassen* constructions in German", which will indeed be done in the next section, the outermost extreme in terms of direct causation would be occupied by constructions in which the causer is not animate.

Note, however, that in (C13) and (C14), the benefit of the state-of-affairs caused is not for the causer. Moreover, *lassen* constructions with an inanimate causer and a transitive sentence are hardly possible, as in (20). For my intuition, the structure with *lassen* and a transitive sentence carries a sense of intentionality that is certainly not given with an inanimate causer.

(20) ?Der Wind lässt die Fahnenstange
DEFMsgNOM wind.sg let.PRES3sg DEFFsgACC flagpole.sg
das Dach zerstören.
DEFNsgACC roof.sg destroy.INF
The wind causes the flagpole to destroy the roof.

The parameters in the description of the degrees of causativity with the *lassen* construction involve some human-only characteristics like non-intervention and permission, which express a low degree of causation. A permitting causer may be aware and in control, but the causation is not direct, so there may be a time lag and there is not necessarily a result. Most importantly, the benefit of the caused event is not for the causer, but for the causee. Therefore, the sense of permission, which outrules both inanimate causers and causees, is low in causativity because of the lack of an immediate result and benefit for the causer.

Dixon (2000) proposes nine semantic parameters, on the basis of which the form-function correlation of causative constructions across languages can be evaluated. As there are several formal expressions for causation in German, the parameters, which are introduced in (21), are relevant for our purposes.

(21) The nine semantic parameters for causative constructions (Dixon 2000: 62):
(a) Relating to the verb
    1. State vs. action: Can a causative mechanism apply to a state only or also to a verb describing an action?
    2. Transitivity: Does the causative mechanism apply to intransitive, transitive, and ditransitive verbs, or only to a limited set of these?
(b) Relating to the causee (original S or A)
    3. Control: Is the causee in control of the activity or not?
    4. Volition: Does the causee do the activity willingly? In English, *let* would be a case of wilful activity, whereas *make* would signal that the causee does not want to do the activity.
    5. Affectedness. Is the causee completely or partially affected by the activity?
(c) Relating to causer (in A function in the causative construction)
    6. Directness. Does the causer act directly or indirectly?
    7. Intention. Does the causer achieve the result by accident or is it the planned result of the activity?
    8. Naturalness: Does the activity happen naturally or is effort involved?
    9. Involvement: In how far is the causer involved in the activity?

So what would a scale of CAUSATIVITY for the German *lassen* construction look like?

The binary features I am suggesting are oriented at the causativity features that Dixon proposes for cross-linguistic comparison. They also make reference to the degree of causative force and to the relationship between the causer and the causee. The features were assigned to the constructions discussed above. The + versus – values of these features give analytical insight into the causativity value of the various

constructions and the nature of the gradience of causativity with *lassen* constructions in German.

(22) The binary causativity features
a) [+/- CONTROL]
b) [+/- AUTHORITY]
c) [+/- ORDER]
d) [+/- BENEFIT CAUSER]
e) [+/- PERMISSION]
f) [+/- DIRECT CAUSATION]
g)[+/- RESULT ORIENTED]

The features are not mutually exclusive, with two exceptions: [ORDER] and [PERMISSION] exclude each other.

Note as well that the following non-binary features play a role in the characterisation of the constructions. These are features that may be stated, so not each of the constructions should have a feature value. For some constructions, these features do not even play any role. Therefore, these are additional features, which may help the qualification of the construction or explain its place on the scale.

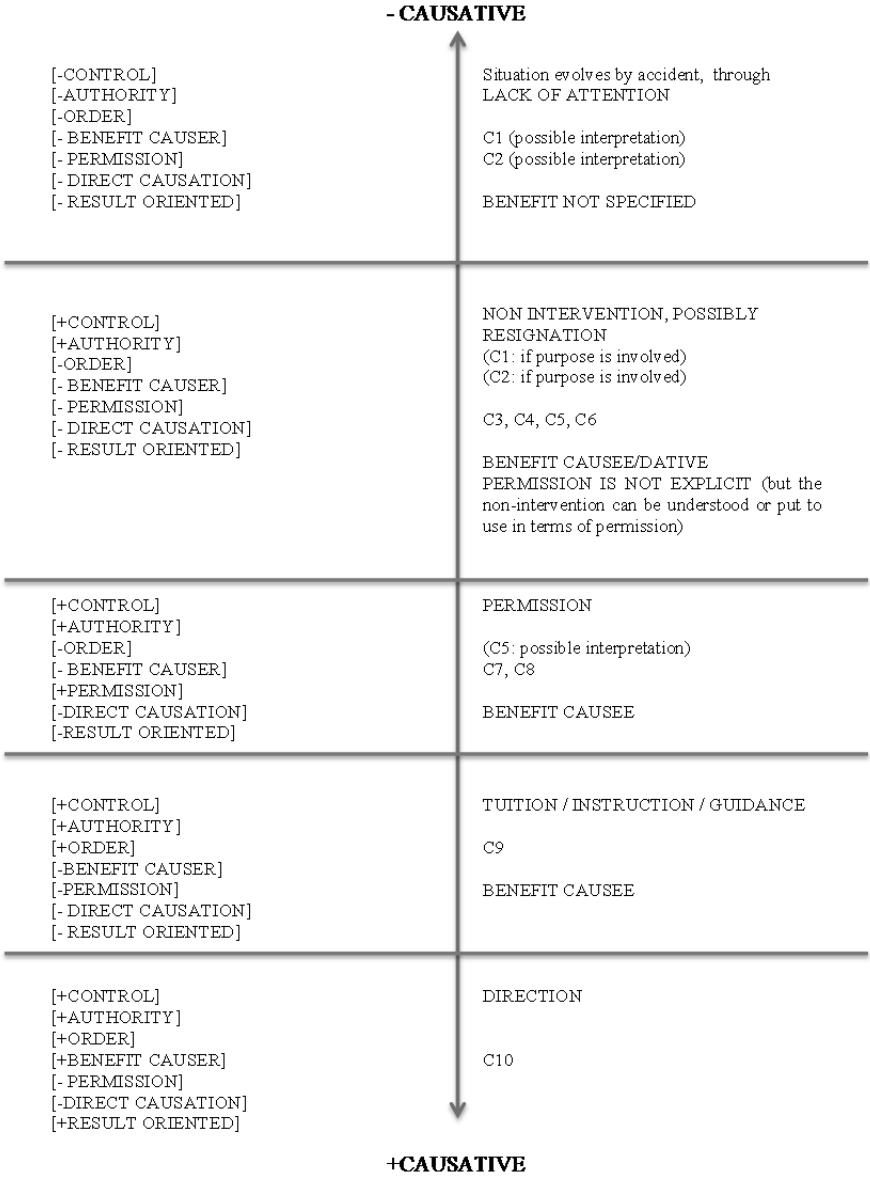
(23) Possible features of *lassen* constructions

- BENEFIT NOT SPECIFIED
- BENEFIT CAUSEE
- PERMISSION IS NOT EXPLICIT
- CAUSER IS EFFECTOR (not animate)
- IMMEDIATE RESULT

3.4. A Scale of Causativity for the German *Lassen* Construction

Figure 1 represents the positions on a scale of causativity for the German *lassen* construction. The examples (C1-C10), which have animate causers, can be ordered with respect to their degree of causativity. To measure the degree, you count the number of '+' values in the binary-feature set on the left side.

Figure 1. Scale of causativity with causative *lassen* constructions in German



The feature [DIRECT CAUSATION] is not fulfilled by any of the constructions in this scale. For causative constructions with human causers and causees, it is always thinkable that the start of the caused activity is delayed. [DIRECT CAUSATION], which is an important causativity factor in Dixon's causativity parameters (Dixon 2000: 62), plays a role in the scale in Figure 2 and its relevance for a German causativity scale is discussed accordingly.

Note that the semantics in the last three cells at the +CAUSATIVE end can only be assigned on the basis of cultural factors. Whether a given construction expresses an act of permission or direction is not denoted in the semantics of the construction nor signalled by the choice of the verb. The construction may also express a request that is intermediate between permission and direction, in that it gives an order that has to be obeyed, but the result of the action is for the benefit of the causee, not for the causer. The distinction between these meaning variants depends on the roles that can be assigned to the people on the basis of the situation described, and on the basis of the language user's previous experience with these roles in these situations. These cultural factors are part of the knowledge that a language user employs in order to properly deal with *lassen* constructions in the linguistic interaction. Therefore, they will be part of the Constructional Schema presented in the next section.

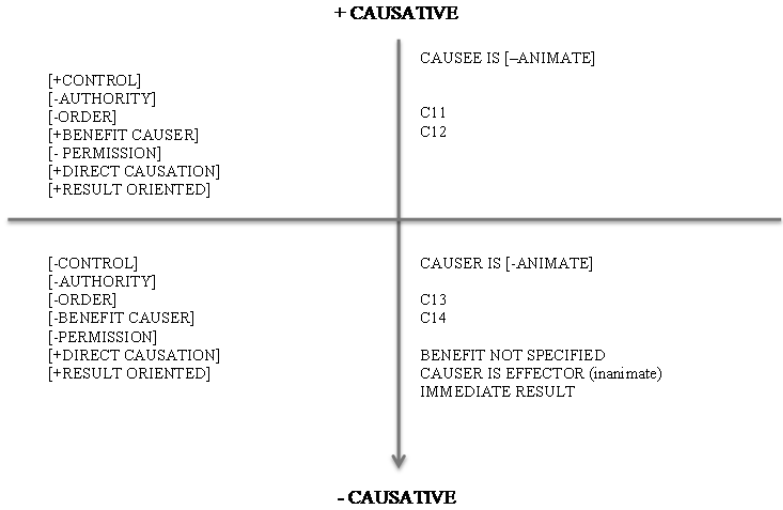
In Figure 1, the degree of causativity measured in terms of involvement, responsibility, control and benefit on the part of the causer is represented for the *lassen* constructions in (C1-C10).

There are two other construction types which are represented by the constructions in (C11-C14). These cannot be represented on the same scale as the previous types, as in (C11-C14) either the causee or the causer (or both) are inanimate entities. The parameters in terms of authority, control and benefit do not work with these. However, as pointed out before, these constructions deserve to be ranked with respect to causativity, as it seems they are the only constructions where the causativity reading is unambiguous, i.e. they could not be misunderstood in terms of permission or non-intervention. Moreover, with these construction types, the causativity effect is immediate, so there is no time lag implied between the impact of the causer and the effect on the causee. Therefore, these construction types get '+' values for the criteria [DIRECT CAUSATION] and [RESULT ORIENTED].

According to these considerations, and also for space reasons, I let these two construction types appear in a new causativity scale, which is based on the same feature set as the previous one. The constructions in Figure 2 cannot be considered to be "more causative" than the

construction involving direction in (C10), as there is a human purpose behind the latter. However, if one takes into account the directness and immediacy of the causation, and also the fact that they cannot be mistaken for permission or non-intervention, the construction types in (C11-C14) have a strong degree of causativity.

Figure 2. Causative construction with inanimate causee/causer



The considerations around a causativity scale for *lassen* constructions in German lead to a number of observations. It seems that direct causation on the one hand and authority/control on the other hand are not part of the same scale after all. Direct causation may be given when the causer is inanimate. Also, the construction is only unambiguous if the causer is inanimate. The causative *lassen* construction is generally very open and ambiguous with respect to the degree of causativity it expresses. In many cases, however, the degree of causativity can be resolved by cultural knowledge.

In Figure 3, the findings about the syntax and semantics of the causative *lassen* construction in German are summarized in a Constructional Schema, which is a means of representing constructional knowledge in Role and Reference Grammar (RRG) (Van Valin 2005, Diedrichsen 2011, 2013c, 2014, Nolan 2013). It accounts for the syntactic, morphological, semantic and pragmatic features of a construction. The construction is conceptualised as a grammatical object (cf. Nolan 2012,

2013). It has a signature, which is a morphosyntactic pattern that makes it recognisable. It also has a workspace, in which the processing of the structure takes place in real time.

Figure 3. Constructional Schema for the German causative *lassen* construction

CONSTRUCTION: German causative *lassen* construction (as exemplified in the constructions C1-C14)
SIGNATURE:
ARG_{+1} + lassen + [ARG_1 (+ ARG_2) (+ ARG_3) … (ADJ) … (PART) + V_{INFIN}]
CONSTRAINT (for the proper recognition of the construction):
Context must make clear that the construction with *lassen* is NOT one of the following:

1. Collocation with *lassen* (with idiomatic meaning)
2. Middle construction with *lassen* (these are mostly recognizable by the reflexive pronoun *sich*)
3. *Lassen* used as full verb in the sense of 'stop, stay away from' or 'release, leave behind'
4. Adhortative construction (these are mostly recognizable by the 1st person plural pronoun *uns*).

WORKSPACE: *Real-time processing according to the following construction-specific rules:*

SYNTAX:
A. Sentence Structure (cf. Dixon 2000:48-55, Enzinger 2010)

1. Complex predicate construction: Add *lassen* with its A-Argument to a complete sentence structure.
2. *lassen* and the existing full verb form a complex predicate (Butt 2003, Nübling 2008)
   A (causer) *lassen* + [x (y) (z) full verb]
   Where
   a) Word order is subject to clause type
   b) Periphrastic tense forms may be applied, additional modal verbs may occur
   c) *Lassen* is finite only if b) does not apply
   d) Passive version is formed with *lassen* as auxiliary and potential *von*-phrase for the agent causee

B. Argument Structure

1. A-Argument („causer") may be added to any sentence structure: atransitive, intransitive, transitive, ditransitive
2. The causer is always A (nominative)
3. The „subject" (S or A) of the original structure will be O of the *lassen* causative (marked by accusative). This holds for originally atransitive structures with a dummy subject as well.
4. The O of the original transitive structure (if there is one) will retain its O-marking (accusative).
5. The dative argument of the original ditransitive structure (if there is one) will retain its dative marking.

MORPHOLOGY:
Lassen is used to form a complex predicate structure with the full verb from the original structure
Perfect formation:
-Auxiliary in the perfect: *haben.*
-*lassen* appears as plain infinitive, no *ge---en* circumfix

PHONOLOGY:
Not specified.

- Constructional Schema to be continued on the next page -

SEMANTICS:
Added „causer" only vaguely specified for role in the event (Hans-Bianchi 2011, Enzinger 2010).
Causativity is a matter of degree.
Generally, [+] values in the binary feature set below are associated with a stronger degree of causativity in terms of Dixon's parameters (2000:62).

a) [+/- CONTROL]
b) [+/- AUTHORITY]
c) [+/- ORDER]
d) [+/- BENEFIT CAUSER]
e) [+/- PERMISSION]
f) [+/- DIRECT CAUSATION]
g)[+/- RESULT ORIENTED]

The features [except c) and e)] are not mutually exclusive.
Feature f) only applies if causer or causee is [-animate].

Causer may be [+/- animate]
Causee may be [+/- animate]

CULTURAL FACTORS FOR DISAMBIGUATION:
Nature of the relationship between causer and causee in terms of roles in society:

1. If relationship causer > causee = authority, responsibility, loving care
{[+PERMISSION], [+BENEFIT CAUSEE]};

2. If relationship causer > causee = tuition, instruction, guidance
{[+ORDER], [+BENEFIT CAUSEE]};

3. If relationship causer > causee = employer > employee
{[+ORDER], [+BENEFIT CAUSER]};

Correlations 1-3 are true by default.
They may be overridden by specified shared knowledge or situational factors.

→ THE LIST MAY BE CONTINUED AS RESEARCH PROGRESSES

PRAGMATICS:
Illocutionary force: not specified
Focus structure: No restrictions; PSA (causer) = topic (default)

The Constructional Schema displays the signature of the construction. It also informs about the "constraints" that can be applied for the disambiguation of the construction in case this morphosyntactic pattern is found with other construction types as well. Therefore, the "constraint" in the Constructional Schema introduces a context-based condition for the identification of a construction (Diedrichsen 2011, 2013c). In the Constructional Schema in Figure 3, the constraint states that, in order to recognize the causative *lassen* construction as the one described in this schema, the language user must make sure by context information that this construction is not one of the other expressions involving *lassen*, that we have discussed in Section 2.

The Constructional Schema in Figure 3 also lists the cultural factors that help identify the meaning of the construction, given its semantic ambiguity, whose gradient character was described by binary features in Section 3.3. As discussed in Diedrichsen (2013a, 2013c), the knowledge base applied for the use of linguistic constructions generally involves cultural knowledge to a certain degree. There are constructions whose meaning is completely based on the knowledge of the situations in which they would be properly used; on the basis of their semantics and syntax alone, they would not make sense at all. However, other constructions involve more grammatical/semantic knowledge, and the cultural aspect only plays a minor part. In the construction set examined here, we have a new phenomenon in that the semantics is there, but highly ambiguous, and the cultural aspects come into play in order to disambiguate the constructional meaning.

4. Summary and Conclusion

This paper has discussed the German *lassen* construction from the perspective of the functional syntax and its constructional properties.

The German causative construction is a complex predicate construction. Just like many complex causative constructions across languages, it is ambiguous between causative and permissive readings. For German, the description is further complicated by the fact that the verb *lassen*, which is a necessary part of the complex predicate construction, appears in many other uses and constructions as well.

As for the causative *lassen* construction, its ambiguity consists of a spectrum of meaning variants.

On a scale of causativity from mere 'non-intervention', via 'conscious forbearance' and 'explicit permission', to 'true causation', all meanings are principally attributable to the *lassen* construction, and it seems they are

disambiguated only by the verb meaning (e.g. stative verbs tend to invoke a reading in terms of 'non-intervention' and 'ignorance') and by context factors. Moreover, the animacy of the causer and the causee plays an important role.

However, some form-meaning correlations with the *lassen* construction turn out to be interpretable on the basis of cultural knowledge: for example, in a *lassen* sentence where the causer and the causee are in an employer-employee relationship, the reading will assume a notion of power and therefore entail true causation in terms of direction rather than permission.

If the situation described happens in an environment where tuition, instruction or guidance is involved, the *lassen* construction expresses direction, but it is a different kind of direction in that the benefit of the caused action is for the causee, rather than for the causer.

If the relationship between the causer and the causee is one of responsibility, loving care and authority, as it would hold between a parent and a child, it is highly probable that the *lassen* construction has to be read in terms of permission. Empirical studies would have to confirm that these culture-based form-meaning correlations apply in a wider range of constructions than the sample discussed here. From these observations, it can be concluded that cultural convention and experience are not the only basis for constructional meaning in many cases (Diedrichsen 2013a), but cultural knowledge also plays a role in the disambiguation of constructional meaning. For speakers, therefore, the semantic ambiguity of the *lassen* construction is not as big a problem as it seems, as long as the cultural background is reliable enough for the interpretation.

Syntactically, it has been argued that the *lassen* construction has similarities with the ACI construction in terms of linear structure. There are, however, important differences in the syntactic behaviour of the two constructions, and therefore the causative *lassen* construction is treated as a construction in its own right here.

The argument structure of the *lassen* causative consists of one A argument for the causer, which is added to a complete n-transitive structure. The causee argument appears as an O argument. There are variations where the causee is not mentioned. If a causer is added to a transitive structure, this results in a structure with two O arguments. In the *lassen* passive, the causee is expressed in a *von* phrase and the auxiliary is *lassen.* It is not possible to form a *werden* passive with the causer in the *von* phrase.

The syntactic and semantic properties of the *lassen* causative have been summarized in a Constructional Schema. This is an RRG

representation, in which the speaker's knowledge about a construction, its function and its usage conditions are displayed. The Constructional Schema is the representation of this knowledge associated with a construction for the linguistic theory. It has recently been suggested (Diedrichsen 2013a, 2013c) to replenish the knowledge, upon which constructions operate, with cultural knowledge, which often acts as a basis for the semantic interpretation. In the case discussed, it can be a factor in resolving semantic ambiguities, together with other factors such as contextual and common knowledge.

5. References

Duteil and Karl Heinz Wagner (eds.), Betriebslinguistik und Linguistikbetrieb. Akten des 24. Linguistischen Kolloquiums, Universität Bremen, 4-6 September Tübingen: Niemeyer, 245-251 (Linguistische Arbeiten 260). 1989.

Butt, Miriam. 2003. The Light Verb Jungle. In Gulsat Aygen, Claire Bowern, and Conour Mc Donough Quinn (eds.): *Harvard Working Papers in Linguistics,* Volume 9, Papers from the GSAS/Dudley House Workshop on Light Verbs, 1–49.

Diedrichsen, Elke. 2011. The theoretical importance of constructional schemas in RRG. In Nakamura, Wataru (ed.): *Proceedings of the RRG 2009 conference*, Newcastle upon Tyne: Cambridge Scholars Publishing, 168-198.

—.What you give is what you GET? On reanalysis, semantic extension and functional motivation with the German bekommen-passive construction. In *The Art of Getting: GET Verbs in European Languages from a Synchronic and Diachronic Point of View,* Alexandra N. Lenz & Gudrun Rawoens (eds.). Special Issue of *Linguistics* 50(6) (2012): 1163–1204.

—.Constructions as memes – Interactional function as cultural convention beyond the words. In Frank Liedtke & Cornelia Schulze (eds.): *Beyond Words*. Berlin: De Gruyter. (2013a). 283–305.

—. From idioms to sentence structures and beyond: The theoretical scope of the concept "Construction". In Nolan, Brian and Elke Diedrichsen (eds.): *Linking constructions into functional linguistics – The role of constructions in grammars.* Amsterdam: John Benjamins. (2013c) 295-330.

—.A Role and Reference Grammar parser for German. In Brian Nolan and Carlos Periñán-Pascual (eds.): *Language processing and*

grammars: The role of functionally oriented computational models. Amsterdam: John Benjamins. (2014). 105-142.

Dixon, R. M. W. A typology of causatives: form, syntax and meaning. In: Dixon, R. M. W. and Alexandra Y. Aikhenvald (eds.): *Changing Valency: Case studies in transitivity.* Cambridge: Cambridge University Press. (2000). 30-83.

Duden. Online-resource. Last accessed 21 January 2014. http://www.duden.de/rechtschreibung/lassen

Enzinger, Stefan. *Kausative und perzeptive Infinitivkonstruktionen. Syntaktische Variation und semantischer Aspekt.* Berlin: Akademie. 2010.

Fagan, Sarah M. B. *The syntax and semantics of middle constructions.* Cambridge: Cambridge University Press. 1992.

Hans-Bianchi, Barbara. *Die kausative Verbalperiphrase zwischen Grammatikalisierung und Sprachkontakt. (= Daf-Werkstatt Beiheft 1).* Arezzo: Bibliotheca Aretina. 2011.

Kemmer, Suzanne and Arie Verhagen. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5(4): 115-156. 1994.

Kulikov, Leonid I. Causatives. In Haspelmath, Martin, Ekkehard König, Wulf Oesterreicher, Wolfgang Raible (eds.): *Language Typology and Language Universals. An International Handbook.* Berlin: De Gruyter, 886-898. 2001.

Nedjalkov, Vladimir P. *Kausativkonstruktionen.* German translation from Russian by V. Kuchler und H. Vater. Tübingen: Narr. 1976.

Nolan, Brian. *The structure of Modern Irish. A functional account.* Sheffield: Equinox Publishing. 2012.

—. Constructions as grammatical objects: A case study of the prepositional ditransitive construction in Modern Irish. In Nolan, Brian and Elke Diedrichsen (eds.): *Linking constructions into functional linguistics – The role of constructions in grammars.* Amsterdam: John Benjamins, 143-178. 2013.

Nübling, Damaris. *Historische Sprachwissenschaft des Deutschen.* 2nd edition. Tübingen: Narr. 2008.

Rawoens, Gudrun and Thomas Egan. Distinguishing causative and permissive readings of the Swedish verb *låta.* In *Functions of Language* 20:1, 64-89. 2013.

Rosen, Sara. *Argument Structure and Complex Predicates. Doctoral dissertation, Brandeis University.* 1989.

Van Valin, Robert D. *Exploring the syntax-semantics interface.* Cambridge: CUP. 2005.

Vogel, Ralf. Skandal im Verbkomplex. Betrachtungen zur scheinbar inkorrekten Morphologie in infiniten Verbkomplexen des Deutschen. *Zeitschrift für Sprachwissenschaft* 28, 307-346. 2009.

CHAPTER THREE

TOWARDS THE MEANING AND REALIZATION OF MĀORI NEUTER VERBS

AOIFE FINN
TRINITY COLLEGE DUBLIN

1. Introduction

This chapter examines Māori neuter verbs. Neuter verbs are one of the five traditional Māori verbal classes. The most canonical Māori verbal class is transitive verbs. Neuter verbs exhibit distinctive syntactic behaviour from transitive verbs and the other verbal classes. Argument realization in neuter verbs is quite distinct from the transitive verb class. This is in spite of the fact that their meaning is often similar, if not the same. Unlike transitive verbs, neuter verbs have an undergoer "subject". This has led to suggestions of ergativity in Māori. Furthermore, the passivization and nominalization of neuter verbs differs from that of other verbal classes. This chapter principally examines neuter verbs in consideration with the predicate class tests from Role and Reference Grammar. Transitive verbs are considered congruently, if only to highlight the perceived differences and restrictions of the neuter verbs.

2. A Brief Introduction to Māori

Māori is the indigenous language of New Zealand. It is a member of the Austronesian language family. This is a vast family, in terms of both size and geography. Harlow (2007: 10) states that the Austronesian family has 1200 members and "stretches from Madagascar in the West to Rapanui (Easter Island) in the East". Du Feu (1996: 2) informs us that Māori is an Eastern Polynesian language along with Rapanui, Rarotongan, Tahiatian, Tuamotuan, Marquesan, Hawai'ian and Mangarevan. Sourced from the 2013 census, according to the Statistics New Zealand government website, there are approximately 148,395 people who can speak Māori.

Māori is a VSO, head-first, dependent-marking language. An introduction to the transitive verb class and neuter verb class follows.

2.1. Transitive Verbs

In Māori clauses the immediately post-verbal "subject", a privileged syntactic argument in Role and Reference Grammar terms, is marked non-overtly. This is most clearly demonstrated in the intransitive clauses below in (1) and (2). Both examples are taken from Bauer (1993: 66). The marking of the privileged syntactic argument and its order of occurrence is irrespective of the semantic role of the argument or the voice being used. The so-called passive voice (cf. (3)) marks the undergoer privileged syntactic argument in the same fashion.

(1)	Kua	tae	mai	ngā	manuhiri	ACTOR
	PRF	arrive	DIR	DET.PL	visitor	PSA
	"The visitors have arrived"					

(2)	Kua	mutu	te	hui	UNDERGOER
	PRF	finish	DET.SG	meeting	PSA
	"The meeting has finished"				

(3)	Kua	karang-tia	ngā	manuhiri	"PASSIVE
	PRF	welcome-PASS	DET.PL	visitor	VOICE"
	"The visitors have been welcomed"				

In a typical transitive clause the privileged syntactic argument is the actor. As expected of a dependent-marking language, the predicate is not marked for the number, gender, semantic roles, case or any other properties of the arguments within the clause. Rather, Māori depends greatly on the use of what Harlow (2007: 24) terms "particles" and to an extent word order. Tense and aspect are marked by particles, which typically occur before the predicate. Arguments other than the privileged syntactic argument are also marked by means of particles. Should a transitive clause occur, a second argument, the undergoer, will be marked by either *i* or *ki*. The selection of either particle for the traditional direct object is based on the sense of the verb and the semantic role of the argument. Biggs (1969: 23, 25) provides respective examples in (4) and (5). The single most important factor in distinguishing transitive verbs from other verbal classes is that they can occur in the so-called passive voice. An active clause and its passive counterpart are shown in (6) and (7) (Bauer 1997: 477). A further distinguishing characteristic of transitive verbs is that they are nominalized using <ā>; this reflects that their PSA is

an actor. There is an example of this in (8), taken from The Reports of Native Affairs Committee in 1888. A selection of canonical transitive verbs is provided in Table 1, taken from Bauer (1997: 13).

(4) Ka here a Paka i ngā kurī
tns tie art Paka obj det.pl dog
"Paka ties up the dogs"

(5) Ka whawhai te taniwha ki
tns fight det.sg monster obj
a Tamahae
art Tamahae
"The monster fights Tamahae"

(6) E here ana a Huia i ngā
prog tie.up prog art Huia obj det.pl
kurī
dog
"Huia tied up the dogs"

(7) E here-a ana ngā kurī e
prog tie.up-pass prog det.pl dog agt
Huia
Huia
"The dogs were being tied up by Huia"

(8) Kihai ia i whiwhi ki
pst.neg 3sg pst acquire acc
tetahi whenua i runga i
det land p basis p
te tango-hanga a te Matera i
det.sg take-nmz p det Mantell p
ngā whenua i te tau 1853
det land p det.sg year 1853
"She got no land when it was taken by Mr. Mantell in the year 1853"

Table 1. Typical transitive verbs

āwhina	help, assist
here	tie up
kawe	carry
patu	hit, beat
pupuhi	blow, shoot
whāngai	nourish

In ditransitive constructions, Bauer (1993: 271) explains that a third argument, the traditional indirect object, is typically marked with either *ki* or *mō/mā*. These could be referred to as non-macrorole core arguments in Role and Reference Grammar. Again, the selection of either particle is founded on the sense of the verb and the semantic role of the argument. Bauer (1993: 272) provides examples of these markings in (9) and (10). Although the English translation is the same for both (9) and (10), the sense is entirely different. In (9), *Mere* has given *te keke* to *tana tama*, but only temporarily. *Te keke* is located with him but he is not ultimately the intended recipient of the *te keke*. In (10), *tana tama* is the intended permanent recipient of *te keke*. The choice of preposition allows Māori to make subtle distinctions about the semantic role of arguments and the precise sense of the clause. It ought to be borne in mind that the examples here illustrate general tendencies; the particles can indicate alternate semantic roles. The example (11) provides a clause where all three markings are present, concurrently having a patient *māripi*, a goal *tana hoa* and a recipient *Hone*.

(9) I hoatu a Mere i te keke
pst give art Mary obj det cake
ki tana tama
in.obj poss son
"Mary gave the cake to her son"

(10) I hoatu a Mere i te keke
pst give art Mary obj det cake
mā tana tama
in.obj poss son
"Mary gave the cake to her son"

(11) I hoatu ahau i te māripi ki
pst give 1sg obj det.sg knife in.obj
tana hoa mā Hone
poss friend in.obj John
"I gave the knife to his friend for John"
or
"I gave the knife to John's friend for John"

2.2. Neuter Verbs in Māori

As well as the transitive verbs examined above, Māori scholars have traditionally identified a verbal class called "neuter verbs". Such a construction has already been seen in the example (2). In a neuter verb

construction, the undergoer is the privileged syntactic argument. It is post-verbal and non-overtly marked. Some examples are seen in (12) and (13), taken from Bauer (1993: 409) and Harlow (2001: 176) respectively. The realisation of the actor is not obligatory, as is the case in (12). If the actor is expressed syntactically, it immediately follows a particle *i*, as seen in (13). The status of the actor is disputed. Harlow (2007: 27) and Bauer (1993: 413) state that it is an oblique argument as indicated by the particle *i*. At odds with this is Harlow's (2001: 31) assertion that the actor argument is within a prepositional phrase marked with *i*. The exact status of the actor argument is dubious and has led to disagreement surrounding its proper categorisation. As constructions with neuter verbs give prominent marking to the undergoer, this has led to suggestions that neuter verb constructions are marked ergatively. Unlike transitive verbs, neuter verbs are usually nominalized using the particle <ō>; this reflects that their "subject" is an undergoer, as shown in (14) from *Kōrero: Te Māori i te ōhanga* in *Te Ara*.

(12) Kua mutu te hui
prf be.finished det.sg meeting
"The meeting has been ended"

(13) Kua oti ngā mahi i a rātou
prf be.completed det.pl work agt art 3pl
"The work has been completed by them"

(14) Ka papahoro te ōhanga Māori
tns fall.down det economy Māori
i te riro-nga o ngā whenua
p det.sg take-nmz p det.pl land
haumako i te
fertile p det.sg
rautau atu i 1800
century away p 1800
"Māori tribes lively participation in the economy fell away as the fertile land was alienated in the 19th century"

Bauer (1997: 490) has identified three classes of intransitive verbs which have been variously called neuter verbs at some point in time. Bauer named each of these classes after a prototypical member. Hence, there is the *mutu* class, the *ora* class and the *haere* class. These translate as "end", "well" and "go" respectively. Bauer follows Hooper (1982, 1984) by stating that only the *mutu* class ought to be called neuter verbs. This is

reasoned as "the three groups of intransitives follow different rules in many areas of the grammar" (Bauer 1997: 37). While the *mutu, ora* and *haere* classes will each be briefly examined, the convention that only the *mutu* verbs are true neuter verbs will be adhered to in this paper.

2.2.1. *Mutu* class

Given the description in Section 2.2, the pattern of occurrence for *mutu* verbs is exemplified in (15) (Bauer 1997: 39). A tense or aspect marker precedes the predicate, and the subject follows. The subject, an undergoer in RRG, is immediately post-verbal and non-overtly marked. If an actor is expressed, it follows the particle *i* occurring after the subject. Bauer (1997: 493) also provides the example (16). Neuter verbs are said not to be notionally stative (Bauer 1993: 413). However, Harlow (2001: 31) asserts that they "refer not much to an activity as to a state". Most importantly, unlike the transitive verbs in Section 2.1, *mutu* verbs cannot be passivized. Some prototypical *mutu* class verbs are listed in Table 2.

(15) Kua mutu te hui
pft complete det meeting
"The meeting is over"

(16) Ka mau te ika nei i a
tns catch det.sg fish deic act art
Hine
Hine
"Hine caught this fish"

Table 2. Typical mutu verbs

ea	avenge, pay for
mahue	leave behind
mau	fix, catch
mutu	finish
oti	finish, complete
riro	take

2.2.2. *Ora* Class

Bauer (1997, et passim) maintains that *mutu* class verbs are separate and distinct from the *ora* class. As shall be briefly discussed in Section 5, *ora* class verbs do indeed exhibit some different behaviour from *mutu* class verbs when they are considered under the rubric of adjectives. The *ora* class takes the form of the example in (17), taken from Bauer (1997: 38). Yet again, the subject is an undergoer. It is immediately post-verbal and non-overtly marked. Bauer refers to these *ora* class verbs as state

intransitives, since they are said to refer to states as opposed to actions. Like *mutu* class verbs, *ora* class verbs cannot be passivized. A list of some *ora* class verbs taken from Bauer (1997: 38) can be seen in Table 3.

(17) Ka reka ngā kōura
tns sweet det.pl crayfish
"The crayfish are yummy"

Table 3. Typical ora verbs

hohonu	be deep
iti	be small
kaha	be strong, able, courageous
nui	be large
ora	be well, alive
pai	be good, easy
pakaru	be smashed, broken
reka	be sweet, pleasant
roa	be long, tall
taumaha	be heavy, ill
whero	be red

2.2.3. *Haere* Class

Bauer's (1997: 490) third intransitive is the *haere* class, whose verbs are referred to as action intransitives. An example is seen in (18) (Bauer 1997: 37). They take the same form as the *mutu* and the *ora* class verbs. Their privileged syntactic argument is immediately post-verbal. However, unlike the two previous intransitive verb classes, in the *haere* class the single argument is an actor in RRG terms. Furthermore, *haere* verbs can be passivized, unlike the *mutu* and *ora* classes seen above. The passivized *haere* verb *noho* in (19) is taken from Bauer (1997: 492). A list of some *haere* class verbs, also called action intransitives, can be seen in Table 4, taken from Bauer (1997: 37).

(18) Ka oma rātou
tns run 3pl
"They ran"

(19) E noho-ia ana taua whare e
prog stay-pass prog dem house agt
te pahī manuwhiri
det.sg company visitor
"The house was occupied by a company of visitors"

Table 4. Typical haere verbs

haere	to go
hoe	to paddle, row
hoki	to go back return
noho	to sit, stay
oma	to run, move quickly
peke	to jump, leap over
rere	to fly, flee, escape
tangi	to cry, mourn, sing, chime

3. Aktionsart Classes in Role and Reference Grammar

In Role and Reference Grammar, real-life events and happenings are referred to as states-of-affairs (Van Valin and LaPolla 1997: 83). Role and Reference Grammar will hereinafter be acronymized as RRG. There are four basic states-of-affairs with their own particular characteristics (cf. (20) for examples and characteristics). These basic states-of-affairs are described as spontaneous, in the sense that they occur unprompted. However, there are four further states-of-affairs which do not occur unprompted. Although they correspond to the spontaneous states-of-affairs in (20), these are said to be induced: something causes them to happen. Some examples of the induced states-of-affairs and their spontaneous counterparts can be seen in (21).

(20)	State-of-affairs	Characteristics
	Situation Eileen is tired	Static, non-dynamic, does not involve change, may involve the constant location, the constant state, the constant condition or the constant internal experience of a participant
	Events The balloon burst	Instantaneous happenings, that may involve a change in the location, the state, the condition or the internal experience of participant
	Processes Tadhg learned German	Happenings that occur over time, that may involve a change in the location,

			the state, the condition or the internal experience of a participant
	Actions Deirdre drinks whiskey		Dynamic happening, involve a participant doing something

(21)

State of Affairs	Spontaneous	Induced
Situation	A cat being afraid	A dog scaring a cat
Event	A balloon bursting	A cat bursting a balloon
Process	Frost melting	The sun melting the frost
Action	A ball rolling	A girl rolling the ball

The real-life states-of-affairs are expressed in language, often with verbal predicates. RRG proposes six basic predicate classes, known as Aktionsart classes. These Aktionsart classes express real-life states-of-affairs. Furthermore, there are six corresponding causative Aktionsart classes. Van Valin (2005: 33) defines the six basic Aktionsart categories in terms of their binary values with respect to six features. A description of the Aktionsart classes, in terms of the six features, follows in Section 3.1.

3.1. Aktionsart Classes and Their Features

The Aktionsart classes and some examples are listed in Table 5. A discussion of the Aktionsart classes apropos of their features will follow.

Table 5. Aktionsart classes and English examples

Aktionsart class	Example
State	Danilo is afraid
Activity	The ball bounced around the room
Accomplishment	The steak defrosted
Achievement	The firework exploded
Active Accomplishment	Enda ran to the quay

Semelfactive	The torchlight flashed
Causative State	Beppe frightens Danilo
Causative Activity	Sonia bounced the ball around the room
Causative Accomplishment	Deirdre defrosted the steak
Causative Achievement	Brendan set off the firework
Causative Active Accomplishment	Brendan ran Enda to the quay
Causative Semelfactive	John flashed the torchlight

3.1.1. Static

This is the feature which conveys that a predicate is static. More precisely, if a predicate is static, it means that nothing is actually happening. The predicate is describing a constant and unchanging situation. For example, in (22) the verbal predicate is static.

(22) Tadhg is handsome STATE

(23) David is dancing ACTIVITY

The handsomeness of *Tadhg* in (22) is an ongoing attributive characteristic of *Tadhg*; nothing is truly happening or going on. States, like the predicate in (22), are static by nature. On the other hand, something is happening in achievements, accomplishments, activities, semelfactives and active accomplishments. Therefore, they are not static. The example (23) shows an activity in which there is a non-static act of dancing.

3.1.2. Telic

If a situation has an inherent endpoint, it is deemed to be telic. In the examples (24), an endpoint is not innate within the situation. For instance, the example (24) describes an unchanging character trait of *May*; it does not have a beginning or an endpoint. In (25) *Brendan* could continue to run for five more minutes or five hours; a conclusion is not given. The lights in (26) can flash on and off without implicit reference to an endpoint. Contrastively, the situations described in (27) entail an endpoint. That is, the balloon bursting signals the end of the achievement, whereas *Ciarán*'s recovery is an endpoint to the situation. In (29) *Brendan* is

running towards a definite endpoint, in this case the park. By considering (25) and (29) it should be evident that active accomplishments occur when an activity is given an endpoint. It can be deduced from the examples (24) that achievements, accomplishments and active accomplishments are telic. Conversely, states, activities and semelfactives are atelic.

(24)	May is nice	STATE
(25)	Brendan is running	ACTIVITY
(26)	The lights flash	SEMELFACTIVE
(27)	The balloon burst	ACHIEVEMENT
(28)	Ciarán recovered	ACCOMPLISHMENT
(29)	Brendan is running to the park	ACTIVE ACCOMPLISHMENT

3.1.3. Punctual

If a predicate is labelled as punctual, it imparts that the predicate has no internal duration. Trask (1996: 224) explains that a punctual event is confined to a single instant in time. If a predicate is not punctual, it has internal duration. An event with internal duration takes place over some period of time. The bomb in (30) explodes instantly; it does not explode gradually throughout a period of time. Therefore, like all achievements, having no internal duration, it is punctual.

(30)	The bomb exploded	ACHIEVEMENT
(31)	The ice melted	ACCOMPLISHMENT

On the contrary, the ice in (31) will melt over a period of time. This is typical of an accomplishment which takes place over a period of time and is not punctual. The same can be said of active accomplishments. As mentioned in 3.1.2, states and activities are atelic. Having no beginning or endpoint, they must involve internal duration and therefore are not punctual by default. On the other hand, semelfactives are also atelic; they can continue without end. However, they describe events with duration of a very short time. This short duration renders them akin to instantaneous achievements. Therefore, semelfactives are punctual.

3.1.4. Dynamic

If a predicate is dynamic, then a dynamic action is innate within the activity. Pavey (2010: 356) states that a dynamic event is one that involves action. Not unsurprisingly, states such as those in (32) are not dynamic. No action is involved in being generous, or in any other state. It could also be foreseen that activities involve dynamic actions. The activity in (33), encompassing the action of listening, is dynamic. There is also a dynamic action involved in the drinking of whiskey in (34).

(32)	Brendan is generous	STATE
(33)	Deirdre is listening	ACTIVITY
(34)	Deirdre drank the tumbler of whiskey	ACTIVE ACCOMPLISHMENT
(35)	The wine glass shattered	ACHIEVEMENT
(36)	The Castletown River froze	ACCOMPLISHMENT
(37)	Eileen glimpsed Nuala	SEMELFACTIVE
(38)	The audience clapped	SEMELFACTIVE

It might be assumed that achievements and accomplishments would be dynamic, but this would be incorrect. Although achievements and accomplishments, as in (35) and (36), describe a happening, there is no dynamic action involved as such. Nor is dynamic the opposite of static. Static describes if an event does or does not involve a happening. Even if the event is not static and involves something happening, it will only be dynamic if it has a dynamic action. Achievements illustrate this in that they are both non-static and non-dynamic. Semelfactives are somewhat anomalous in that they can be dynamic or non-dynamic. This depends on the inherent semantic character of the verb. This is demonstrated by the examples (37) and (38). The act of glimpsing in (37) is non-dynamic, whereas the clapping in (38) is clearly dynamic. The features attributed to each Aktionsart category are listed in Table 6, inspired by Van Valin and LaPolla (1997: 93).

Table 6. Basic Aktionsart classes and their features

State	+ static	- telic	- punctual	n/a
Activity	- static	- telic	- punctual	+ dynamic
Accomplishment	- static	+ telic	- punctual	- dynamic
Achievement	- static	+ telic	+ punctual	- dynamic
Active Accomplishment	- static	+ telic	- punctual	+ dynamic
Semelfactive	- static	- telic	+ punctual	± dynamic

3.2. Testing for Aktionsart Classes

RRG prescribes a series of seven tests designed to isolate certain features of the Aktionsart classes (cf. Van Valin 2005: 34-41). A predicate undergoes the Aktionsart tests. Afterwards, by consideration of the collective results, it is possible to identify to which Aktionsart class the predicate belongs. A description of each test follows.

3.2.1. Test 1 – Progressive aspect

Trask (1996: 219) defines the progressive as the aspect "which refers specifically to an action or event which is in progress at the moment of time serving as the reference point for the utterance". The progressive describes an event that is happening at that moment. As something is happening, the progressive predicate is not static. In light of this, the progressive ought not to occur with static predicates. The progressive aspect is also in progress; it is taking place over time. It therefore has internal duration; it is not punctual. Therefore, as well as static predicates, the progressive aspect ought not to occur with punctual predicates.

As seen in Table 6, which summarised Section 3.1, states are inherently static. Even though states are non-punctual, being static ensures that they will not comply with the progressive test, as seen in (39). Being punctual ensures that achievements do not occur with the progressive aspect, as shown in (40). Accomplishments, activities and active accomplishments satisfy both the non-static criteria and the non-punctual criteria; they permit the progressive aspect as in (41).

Semelfactives describe events that occur over a very short amount time. For that reason, as explained in 3.1.3, they are punctual. Accordingly, the progressive aspect cannot occur with semelfactives to produce a progressive sense. However, the progressive aspect may occur with semelfactives and give a rather different sense. As semelfactives

often involve repetition, the progressive aspect used with a semelfactive predicate can produce an iterative interpretation, as in (44). It ought to be remembered, of course, that this test is only relevant in a language that has a progressive aspect.

(39)	*Ezekiel is being a syntactician	STATE
(40)	*The balloon is popping	achievement
(41)	My hair is drying	accomplishment
(42)	May is talking	activity
(43)	Christina is walking to the cinema	active accomplishment
(44)	The Christmas tree lights are flashing	semelfactive

3.2.2. Test 2 – Dynamic adverbs

In order to determine if a predicate involves a dynamic action, the congruence with dynamic adverbs is examined. Again, it is necessary to recall that dynamic is not the opposite of static. Static describes an event that does not involve something happening. Even if the event is non-static and it does not involve something happening, it is not inevitably dynamic. It will only be dynamic if it has a dynamic action. Of the predicates encountered, only activities and active accomplishments are dynamic. This is seen in (45) and (46), where these predicates marry well with dynamic adverbs. Consulting Table 6, it can be seen that states, accomplishments and achievements are not dynamic. Accordingly, they are not compatible with dynamic adverbs, as shown in (47). As mentioned in Section 3.1.4, semelfactives can be both dynamic and non-dynamic. For that reason, depending on the particular semantic character of the predicate, semelfactives may prove incompatible or compatible with dynamic adverbs as in (50) and (51).

(45)	Brendan is jogging vigorously	ACTIVITY
(46)	Christina walked energetically to the cinema	ACTIVE ACCOMPLISHMENT
(47)	*Stevie is vigorously a teacher	STATE
(48)	*The pizza defrosted vigorously	ACCOMPLISHMENT

(49) *The balloon popped vigorously ACHIEVEMENT

(50) *Deirdre glimpsed Breda energetically SEMELFACTIVE

(51) The audience clapped energetically SEMELFACTIVE

(52) *The cat shivered deliberately ACTIVITY

(53) The cat shivered violently ACTIVITY

Van Valin (2005: 36) cautions that it is best to avoid adverbs that require a controlling subject, e.g. *deliberately* or *carefully*. This is owing to the fact that they may read incompatible with non-agentive subjects, that is, subjects who are not in control. The example (52) shows that the subject, i.e. the cat, is not in control of its actions. Therefore, the adverb *deliberately* is not compatible. On the other hand, the adverb in (53), which does not require a controlling subject, is compatible with the activity as expected.

3.2.3. Test 3 – Pace adverbs

This test checks the ability of predicates to co-occur with pace adverbs. Pace adverbs indicate the speed at which an event occurs, be it slowly or quickly. Pace adverbs are only compatible with verbal predicates that involve internal duration, that is, predicates that are not punctual. Some examples of pace adverbs include *slowly*, *quickly*, *rapidly* and *swiftly*.

Referring back to Table 6, activity, accomplishment and active accomplishment predicates are not punctual; they have internal duration. Their compatibility with pace adverbs is borne out by the examples (54). Van Valin and LaPolla (1997: 95) issue a caution regarding achievements. Achievements are punctual and have no internal duration. In being categorized as punctual, achievements are acknowledged as occurring in an instant. This instantaneous temporal duration, however small it may be, involves that pace adverbs describing a very small duration may read as acceptable with achievements. This is exemplified in (58). Being punctual, semelfactives also carry the same, as demonstrated by (59). To offset this potential error, Van Valin (2005: 36) recommends testing achievements and semelfactives with pace adverbs that describe a slow process. This will give a correct account of whether, by compatibility or incompatibility with slow-pace adverbs, predicates are achievements or semelfactives.

(54) Ezekiel is running quickly ACTIVITY

(55)	The pizza defrosted quickly	ACCOMPLISHMENT
(56)	Christina walked slowly to the cinema	ACTIVE ACCOMPLISHMENT
(57)	* The glass shattered slowly	ACHIEVEMENT
(58)	? The glass shattered instantly	ACHIEVEMENT
(59)	?The light flashed instantly	SEMELFACTIVE
(60)	The cat is fat slowly	STATE

As we saw in Section 3.1.1, state predicates are not punctual, so they might be expected to be compatible with pace adverbs. However, states are also inherently static, describing an event in which nothing is actually happening. Therefore, a pace adverb, which describes the speed of a happening, would not be compatible with state predicates. This can be plainly seen in the state predicate in (60).

3.2.4. Test 4 – Duration adposition

This test verifies whether or not a verbal predicate occurs over duration in time. To check this, the compatibility with adverbs that indicate duration in time is tested. In English, duration is indicated with the preposition "for". Van Valin and LaPolla (1997: 96) explain that "the for-phrase indicates that... an event went on for a certain amount of time, without any information about when it began or ended".

Since Table 6 shows that they are not punctual, one could deduce that states, accomplishments, activities and active accomplishments are compatible with duration adpositions. This is shown by the respective examples (61). As discussed by Van Valin and LaPolla, this is an extraneous test for accomplishments and active accomplishments, as they are inherently not punctual. As such, they will be compatible with duration adverbs. Furthermore, Van Valin (2005: 37) warns that states that describe inherent properties will not generally accept duration adpositions (cf. (62)). By contrast, states which give an account of attributive properties such as (61) will accept duration adpositions.

(61)	Ciarán was ill for six months	STATE
(62)	*Tadhg was artistic for five years	STATE

(63)	The snow was melting for two days	ACCOMPLISHMENT
(64)	Brendan ran for one hour	ACTIVITY
(65)	Deirdre drank the whiskey for one hour	ACTIVE ACCOMPLISHMENT
(66)	* The glass shattered for an hour	ACHIEVEMENT
(67)	The Christmas lights flashed for an hour	SEMELFACTIVE

On the other hand, achievements are inherently punctual and therefore will be incompatible with duration adpositions, as demonstrated in (66). Having no inherent endpoint but being instantaneous, the agreeable addition of duration adpositions to semelfactives will deliver an iterative interpretation. This is the case in (67), where the Christmas lights flashed repeatedly for an hour.

3.2.5. Test 5 – Completion adposition

Test 5 verifies if a verbal predicate explicitly references the endpoint of the event, i.e. the completion of the event. It is in that respect that this test differs from Test 4, which indicates only the length of time without explicit reference to a precise beginning point or endpoint. In English, the completion of an event is indicated with an in-phrase, e.g. *in ten minutes, in five hours* or *in two days*.

Accomplishments and active accomplishments are telic and non-punctual (cf. Table 6). Their end result that occurs over time means that they alone are compatible with completion adpositions; this is demonstrated by (68). Once more, punctual achievements flout the rules somewhat because they may be compatible with completion adpositions that denote a very short period of time (cf. (70)). Furthermore, in English at least, some activities, achievements and semelfactives may appear to agreeably occur with some in-phrases. However, these in-phrases refer not to the completion of the event itself but to the time when the event will begin. Therefore, they could not be considered as evidence for the compatibility with completion adpositions. Some examples of these "false friends" are provided in (71).

(68)	The snow melted in two days	ACCOMPLISHMENT
(69)	Breda drank the bottle of wine in two hours	ACTIVE ACCOMPLISHMENT

(70)	The bomb exploded in a fraction of a second	ACHIEVEMENT
(71)	John will golf in an hour	ACTIVITY
(72)	The bomb will detonate in two hours	ACHIEVEMENT
(73)	We will glimpse the eclipse in five minutes	SEMELFACTIVE

3.2.6. Test 6 – Stative Modifier

This test is designed primarily to distinguish between the two punctual predicates, i.e. achievements and semelfactives. Pavey (2010: 105-106) explains it very well. The main difference between achievements and semelfactives is that the former are telic while the latter are atelic. Achievements have an inherent endpoint; there is a resultant state upon completion of the event. Consequently, a stative adjective or modifier can be derived from an achievement predicate. This can be seen in the examples (74). Semelfactives do not have an inherent endpoint; the event simply occurs again and again. For that reason, a stative modifier cannot be derived from a semelfactive, as shown in (76).

(74)	Stephen shattered the window	ACHIEVEMENT
(75)	The shattered window	STATIVE MODIFIER DERIVED FROM ACHIEVEMENT
(76)	The Christmas lights flash	SEMELFACTIVE
(77)	The flashed light	STATIVE MODIFIER DERIVED FROM SEMELFACTIVE

3.2.7. Test 7 – Causative Paraphrase

In Section 3, it was stated that the states-of-affairs may be spontaneous or induced. The states-of-affairs are real-world events, while the Aktionsart predicate classes are linguistic events. Moreover, there is another correspondence between the states-of-affairs and the predicate classes. The Aktionsart classes may also be induced; they are referred to as causative classes, some examples of which can be seen in (78).

(78)

Aktionsart class	Example
State	Helen is ill
Causative State	The prawns made Helen ill
Achievement	The building collapsed
Causative Achievement	The bulldozer made the building collapse
Accomplishment	The ice-cream melted
Causative Accomplishment	The heat of the sun melted the ice cream
Activity	Stevie walks fast
Causative Activity	Sarah made Stevie walk fast
Active Accomplishment	Christina walked to the cinema
Causative Active Accomplishment	David made Christina walk to the cinema
Semelfactive	The twig tapped on the window
Causative Semelfactive	Mary tapped her pen on the desk

There is no definitive test for distinguishing non-causative Aktionsart classes from causative Aktionsart classes. However, the causative paraphrase test is helpful. If a predicate is causative, there should be an analogous causative paraphrase, as in (79). Crucially, the original causative predicate and the causative paraphrase should have the same number of arguments. Van Valin (2005: 38-39) points out that a disadvantage of this restriction is that the test is ineffective for single argument verbs.

(79) The sun melted the ice-cream — CAUSATIVE ACCOMPLISHMENT
1 2 — 2 ARGUMENTS

(80) The sun caused the ice-cream to melt — CAUSATIVE ACCOMPLISHMENT PARAPHRASE
1 2 — 2 ARGUMENTS

The causative Aktionsart predicates will have mostly the same results and cautions in Tests 1-6 as their non-causative counterparts. However,

Van Valin (2005: 39-40) does highlight some issues that ought to be taken into account when dealing with causatives. There are some instances where the non-causative predicate and the causative counterpart do not have the same results.

Test 1 for progressive adverbs may be compatible with causative state predicates if the causing state-of-affairs is comparatively active. The causing state-of-affairs in the example (81) is the less-active *your attitude*. Thus, the predicate *upset* may be deemed less compatible in the progressive to some English speakers. This is contrasted with the more active predicate *amuse* in (82), which is perfectly acceptable in the progressive aspect.

(81) Your attitude upsets/?is upsetting me

(82) The clown's zany antics amuse/are amusing the children

(83) ? Your attitude actively upsets me

(84) The clown's zany antics actively amuse the children.

This links with the dynamic adverbs involved in Test 2. Again, if the causing state-of-affairs in a causative state predicate is more active, it may be deemed more acceptable with dynamic adverbs. This is illustrated in (83). Test 2 can also potentially cause problems for causative achievements, causative semelfactives and causative accomplishments. Dynamic adverbs may sometimes be acceptable with these predicates because they may be taken as modifying the causing event, rather than the caused resultant event.

The typical results of each Aktionsart class under each test are seen in Table 7.

Table 7. Expected results of each Aktionsart test

	State	Achievem.	Accomplish.	Activity	Active Accomplish.	Semelfact.
Test 1 Is the predicate compatible with the progressive aspect?	No	No	Yes	Yes	Yes	No
Test 2 Is the predicate compatible with dynamic adverbs?	No	No	No	Yes	Yes	No

Test 3 **Is the predicate compatible with pace adverbs?**	No	No	Yes	Yes	Yes	No
Test 4 **Is the predicate compatible with duration adpositions?**	Yes	No	Yes but extraneous	Yes	Yes but extraneous	Yes
Test 5 **Is the predicate compatible with completion adpositions?**	No	No	Yes	No	Yes	No
Test 6 **Can the predicate be used as a stative modifier?**	Yes	Yes	Yes	No	Yes	No
	Caus. State	**Caus. Achievem.**	**Caus. Accomplish.**	**Caus. Activity**	**Caus. Active Accomplish.**	**Caus. Semelfact.**
Test 7 **Does the predicate have an analogous causative paraphrase?**	Yes	Yes	Yes	Yes	Yes	Yes

4. Neuter verbs, Transitive verbs and Aktionsart tests in Role and Reference Grammar

Section 49 introduced the basics of the Māori grammar. In that section transitive verbs and neuter verbs were introduced. The RRG Aktionsart classes and their features were presented in Section 3.1. This was followed by an introduction to the tests that determine to which verbal class a predicate belongs in Section 3.2. The current section considers Māori transitive and neuter verbs in terms of the RRG Aktionsart tests. An examination of the verbs under each test follows.

4.1. Neuter verbs, Transitive verbs and Test 1 – Progressive aspect

As noted in Section 2.1, tense and aspect are marked by means of particles. They typically precede the predicate. As the examples (85) show, the particles may variously mark just tense, just aspect, or both

tense and aspect. Bauer (1997: 107, 89, 106) provides the examples (85), (86) and (87) respectively.

(85)	I	pupuhi	te	hau		only	tense
	pst	blow	det.sg	wind		marked	
	"The wind blew"						
(86)	E	haere	ana	mātou	ki	only	aspect
	prog	go	prog	1pl.ex	loc	marked	
	te	one	āpōpō				
	det.sg	beach	tomorrow				
	"We are going to the beach tomorrow"						
(87)	I te		pupuhi	te	hau	both	
	pst.prog		blow	det.sg	wind	tense	and
	"The wind was blowing"					aspect marked	
(88)	I	tūtuki	tō	mātou	pahi	absolute	
	pst	crash	poss.sg	1pl.ex	bus	tense	marker
	"Our bus crashed"						
(89)	Ka	haere	ia	ki		relative	
	tns	go	3sg	loc		tense	marker
	Amerika						
	America						
	"He will go/went to America"						

Tense markers in Māori bifurcate into two types. There are absolute tense markers and relative tense markers. Absolute tense markers specify the tense in which the action took place. This is shown in the example (88) from Bauer (1997: 84), wherein a past tense context is definitive. Relative tense markers do not specify the tense in which the action took place on their own terms. They take an interpretation of tense from the most recent preceding tense particle, adverbials or other syntactic means. Bauer (1997: 87) provides the example (89), wherein *ka* does not clearly indicate a tense. In fact, without a context *ka* does not provide a definitive reading for tense. However, like many relative markers, it has a default reading when it appears out of context. As it happens, the default reading of *ka* is present.

Māori principally uses three particles to mark the progressive aspect. They are *kei te, i te* and *e... ana*. The progressive particles and their tense and aspect values, along with how they are viewed in RRG, can be seen in Table 8.

Table 8. The progressive particles

Particle	Tense value	Aspect value	Operator(s) in RRG	Layer(s) of Clause modified in RRG
Kei te	Absolute tense marker - past	Progressive	Tense and aspect	Clausal – tense Nuclear - aspect
I te	Absolute tense marker – non-past	Progressive	Tense and aspect	Clausal – tense Nuclear - aspect
E... ana	Relative tense marker	Imperfective – progressive or habitual or continuous	Aspect	Nuclear - aspect

Bauer (1997: 127) cautions that *kei te, i te* and *e... ana* can be used in imperfective contexts other than the progressive. For this reason, either or both accurate translations or/and consultation with a native speaker were sought to verify the intended sense of the tense/aspect particle. *Kei te* and *i te* are quite compatible with all verbal classes, while *e...ana* is a matter of some dispute. As seen in the example (90), taken from Bauer (1997:126), *e... ana* is quite accordant with the transitive verb class.

(90)	E	haka	ana	te	iwi	whenua
	prog	dance	prog	det.sg	tribe	land
	rā	i	te	haka		
	deic	obj	det.sg	haka		

"The local people were dancing the haka"

Williams (1862: 49, from Bauer 1997) noted that neuter verbs were incompatible with *e... ana*. Yet, Bauer (1997: 90) explicitly states that some occasional counter-examples have been found. This study has found numerous examples of *e... ana* with *mutu* neuter verbs. It would seem that in modern Māori the bar on *mutu* neuter verbs occurring with *e... ana* does not hold. Some examples of *mutu* verbs with *e...ana* can be found in both Shane Jones's speech from the Parliamentary debates and the biography of John Grace in Te Ara, as shown in (91) and (92) respectively. *E... ana* also readily occurs with the other intransitive verbal classes, i.e. *ora* verbs and *haere* verbs. Bauer (1997: 90, 89) provides respective examples in (93) and (94).

(91) E mutu ana koe i ō mahi
prog finish prog 2sg obj 2nt work
"As you end your work"

(92) E ea ana tana utu i tana noho i
prog pay.for prog 3sg fee obj 3sg stay p
te Kāreti o Te Aute
det.sg Te Aute College
"He saved enough to attend Te Aute College as a boarder"

(93) Kāore ngā tāngata o reira e
neg det.pl person p there prog
mokemoke ana
lonely prog
"The people there are not lonely"

(94) E haere ana mātou ki te one
prog go prog 1pl loc det.sg beach
"We are going to the beach"

This test is particularly important with regard to neuter verbs, as it has often been claimed that they are stative. In contrast with Bauer (1993: 413), Biggs (1969: 79) explicitly refers to neuter verbs as "stative verbs". Moreover, Harlow (2001: 31) asserts that neuter verbs "refer not much to an activity as to a state". However, as demonstrated in Section 3.2.1, this test is performed with the intention of isolating static predicates. More precisely, static predicates should not be compatible with the progressive aspect test in RRG.

Van Valin (2005: 35) cautions that certain state predicates may be compatible with the progressive aspect. State predicates that describe a situation that is not necessarily permanent may be compatible with the progressive aspect, as shown in (95). Conversely, state predicates which describe a permanent situation will not be compatible with the progressive, as shown in (96). Such judgements about the perceived permanence of the situation are best judged by a Māori speaker.

(95) Peepo the cat is lying on the bed

(96) Florence lies beneath the hills of Fiesole

4.2. Neuter verbs, Transitive verbs and Test 2 – Dynamic adverbs *and* Test 3 – Pace adverbs

The dynamic adverb test and the pace adverb test have been conflated for reasons which will subsequently be made clear.

Van Valin (2005: 36) recommends various adverbs that ought to be employed for the Aktionsart Tests 2 and Test 3. One particular Māori adverb, the modifier *ātā*, is quite frequent in its usage. The issue lies with the fact that *ātā* has many different senses. *Ātā* employs meanings of both the dynamic and pace adverbs. It can variously be translated as *gently, carefully, deliberately* and *slowly*. Recall from Section 3.2.2 that it is imperative not to use adverbs that require a controlling subject. Unfortunately, *ātā* falls under the banner of both dynamic and pace adverbs, as well as having senses that require a controlling subject.

The issue of the dynamic and pace adverb tests is further complicated by the fact that there are two means of realizing such adverbs in Māori. An adverb may modify a verb by means of apposition, wherein the adverb follows the verb, as in (97). Alternatively, the adverb can modify the verb by nominalization, as in (98). In this case, the adverb is the predicate and the verb becomes the subject or privileged syntactic argument. The verb can alternate from predicate to privileged syntactic argument without any morphological change. The preposing particles, such as tense or classifier, help to clarify the role of the verb. Both (97) and (98) are from Harlow (2001: 213).

(97) I mahi pai ia
pst work well 3sg
"He works well" or "His working is good"

(98) He pai tana mahi
cls well poss work
"His working is good"

Some examples of transitive and *haere* verbs being modified by the pace adverb *tere* are shown in (99). The example (99) is from Harlow (2001: 43), and the example (100) is from the Māori Online Dictionary. Harlow (2001: 50) states that, along with *ātā, tere* is a pre-head adverb. That being so, there are many examples of these adverbs occurring with the *mutu* classes in the corpus. Harlow (2001: 50) also conveniently provides an example in (101).

(99) Oma tere
run quick
"Run fast"

(100) Kua kite-a e te Pākehā he
pft find-pass agt det.sg Pākehā det
huarahi patu tere i te wēra
method kill quick obj det.sg whale
"The Pākehā have found a way of killing whales quickly"

(101) I tere oti te mahi
pst quick finish det.sg work
"The work was completed quickly"

The nominalization method of signifying pace or dynamics is very productive. Harlow (2007: 149) even suggests that it is perhaps more idiomatic and intuitive for native speakers. In fact, this nominalization is essential when a Maori imperative includes an adjective. This can be seen in (102), where the *haere* verb *oma* is modified by a pace adverb. The example (103) shows another *haere* verb being modified by a dynamic adverb. Finally, the transitive verb *tunu* is modified by a manner adverb in (104). The examples (102) are all taken from Harlow (2001: 219).

(102) Kia tere te oma
imp fast det.sg run
"Run quickly" or "Let the running be quick"

(103) Kia tūpato te haere
imp careful det go
"Go carefully" or "Let the going be careful"

(104) Kia pai te tunu i ngā kai
imp well det.sg cook obj det.pl food
"Prepare the food well" or "Let the cooking of the food be good"

The issues surrounding the dynamic and pace adverbs raise the question of how the Aktionsart tests should be approached in a language like Māori. It begs the question as to what exactly should qualify as a pace or dynamic adverb. Should the apposition or nominalization be considered valid means of verbal modification for the Aktionsart test? Indeed, is there a split when certain types of verb can be modified by one or both methods?

4.3. Neuter verbs, Transitive verbs and Test 4 – Duration adposition

Duration adpositions in Māori take the form of (105), from Bauer (1997: 252). They begin with the particle *mō* which is itself followed by the time phrase. This seems to be a test that all verbs from the corpus passed. This would render that there are no Aktionsart achievements in Māori, which is an unlikely result. Bearing that in mind, counsel must be sought from a Māori consultant. Moreover, examples of the *mutu* verbs with the duration adpositions were not abundant. Only one example from the corpus presented itself. *Mutu* occurs with the duration adposition in (106), an example from the biography of Te Ruki Kawiti.

(105) Mō te rua rā
for det.sg two day
"for two days"

(106) Ka mutu te pakanga mō te rima marama
tns finish det.sg war for det.sg five month
"For five months fighting ceased"

Duration adpositions seem to occur readily with transitive verbs. Bauer (1997: 252) provides two examples where *hui* and *whakatā* appear with a duration adposition, as shown in (107) and (108). The *haere* class also occurs with duration adpositions, as shown in (109) and (110) from Harlow (2001: 294) and (2007: 286) respectively. The planting advice in (111), from Pūtaiao online, gives an example of a duration adposition with *ora*, a self-evident *ora* class verb.

(107) E hui ana te kōmihana ia
prog meet prog det.sg committee 3sg
marama mō te rua rā
month for det.sg two day
"The committee meets every month for two days"

(108) Ka whakatā ahau mō te kōtahi wiki
tns breathe 1sg for det.sg one week
"I'll take a holiday for a week"

(109) Me mātua haere ki te Whare Wānanga
deon go loc det.sg university
mō te toru tau, neke atu rānei
for det.sg three year, or.more
"You will have to go to university for three years or more"

(110) I haere a Pita ki tāwāhi mō
pst go art Pita loc abroad for
te whā tau
det.sg four year
"Pita went overseas for four years"

(111) Ka ora ētahi tipu mō toru tau
tns well det plant for three year
"Some plants will live for three years"

4.4. Neuter verbs, Transitive verbs and Test 5 – Completion adposition

Bauer (1997: 243) explains that the "phrases expressing location in time generally use the same prepositions as location in space". Māori completion adposition phrases are structured similarly to the duration adpositions seen in Section 4.3. A preposition opens the temporal adjunct adpositional phrase. However, the choice of preposition varies according to the tense of the event being referenced. The prepositions, along with the tenses they are used for, are shown in (112). Some examples of the prepositions for the past, present and future tenses are shown in (113), (114) and (114) respectively (cf. Bauer 1997: 244-249).

(112)

Preposition	Tense
i	past
nō	sometimes past
kei	present
a	future
hei	sometimes future

(113) I te tau 1976 i whānau
p.pst det.sg year 1976 pst born
ai ia
prt 3sg
"She was born in 1976"

(114) Kei te tau 1996 ka tuhi ahau
p.prs det.sg year 1996 tns write 1sg
i tētahi pukupuka
obj det book
"In 1996, I'm writing a book"

(115) Ka hoki mai anō ahau a te
tns return dir again 1sg p.fut det.sg
rima karaka
five clock
"I'll be back again at 5 o'clock"

Mutu verbs have not yet easily presented themselves with completion adpositions. Along with the other provisional results shown here, all this suggests that the *mutu* class may fall under the activity Aktionsart class. However, the corpus results must be verified by a Māori correspondent.

4.5. Neuter verbs, Transitive verbs and Test 6 – Stative Modifier

Modification in Māori is not straightforward. There are two ways that a verb may modify a noun. Firstly, the verb can modify the noun by apposition, that is, by simply following the noun. Alternatively, a nominalized verb can follow the head noun. The issue of having two possible methods, apposition and nominalization, for testing the predicate has echoes of the issues surrounding the pace and dynamic adverbs in Section 0.

Ora class verbs readily occur as modifiers of a noun by apposition. An example using the *ora* verb *pai* is provided in (116), taken from Bauer (1997: 303). Harlow (2007: 108) asserts that the transitive and *haere* classes can modify. An example from Harlow (2007: 104) shows the transitive *waiata* acting as a modifier.

(116) He wāhi pai anō
cls place good again
"Is it a nice place?"

(117) He rōpū waiata
cls group sing
"A singing group"

Bauer (1997: 306) states that "it is common for verbal constructions other than state intransitives to modify nouns". However, it seems that the

normative method of modification for other verbal categories is stem nominalization. Bauer (1997: 306) gives the example (118), wherein the noun *tāima* is modified by the *haere* class verb *hoki* and its accompanying locative phrase.

(118)	Te	tāima	hoki	ki	te	kāinga
	det.sg	time	return	loc	det.sg	home
	"return-home time"					

While state intransitives, or *ora* class verbs, modify verbs by means of apposition, transitive and *haere* verbs seem to modify by stem nominalization. By contrast, Hooper (1984: 44) says that neuter verbs cannot modify nouns. This is backed up by Bauer (1997: 306), who explicitly states that "neuter verbs are not found as in texts as noun modifiers"; this seems to be true of both methods of modification. Yet, she then states that some Māori consultants may sometimes accept examples of both the apposition modification and the stem nominalization modifications. This is potentially very significant. Should my Māori consultant verify that *mutu* verbs are not acceptable modifiers, it would narrow the Aktionsart class of *mutu* verbs to only activity or semelfactive.

The issues surrounding the stative modifier tests are almost indistinguishable from those raised in Section 0 with the dynamic and pace adverbs. Here the question is raised as what exactly should qualify as a stative modifier. In a language like Māori, wherein there are two methods by which verbs may modify a noun, which method is valid for the stative modifier test? What is more, should one class of verbs pass the test via apposition while the other passes the test via nominalization? Which should be considered valid? This is particularly pertinent if both methods of modification are deemed equally intuitive.

4.6. Neuter verbs, Transitive verbs and Test 7 – Causative Paraphrase

As referenced in Section 3.2.7, Van Valin (2005: 38) cautions that the causative paraphrase test is not useful for "verbs that have only one argument in their basic form". While the status of the actor in *mutu* verbs is unclear, it is certain that the actor argument is not an intraclausal necessity for the *mutu, ora* and *haere* verbs. In the interim, until the interclausal status of the actor in the *mutu, ora* and *haere* verbs has been examined, the causative test shall not be applied to any of the neuter verbs.

5. Conclusion

Transitive verbs have been tested as a cohesive class here. However, it is unlikely and unrealistic that all transitive verbs would fall under the same Aktionsart class. The main aim of this chapter was to test neuter verbs and to compare and contrast them with transitive verbs. The transitive verbs used here are principally employed as benchmarks for the neuter verbs. A more thorough examination of the Aktionsart classes of transitive verbs will take place should it prove relevant to this study. Furthermore, this is a preliminary study, so a more exhaustive examination of the neuter verbs under the Aktionsart tests, individually and as a group, will also come to pass.

While perhaps this study begs more questions than it answers, it is constructive in that it makes ready the correct framework for testing the Aktionsart classes of the other verbal classes. This Aktionsart framework can equally be applied to other ambiguous constructions such as the actor-emphatic and the passive voice.

6. List of Abbreviations

1	first person	loc	locative
2	Second person	neg	negative
3	third person	nom	nominalization
acc	accusative	nt	neutral
act	actor	obj	direct object
agt	agentive	p	preposition
art	personal article	pass	passive
cls	classifier	pl	plural
deic	deictic	poss	possessive
dem	demonstrative	prf	perfect
deon	deontic	prog	progressive
det	determiner	prs	present
dir	directional	prt	particle
dl	dual	pst	past
ex	exclusive	sg	singular
imp	imperative	tns	tense
in.obj	indirect object		

7. References

Bauer, W., W. Parker and T. K. Evans. Māori. London: Routledge. 1993.
Bauer, W. The Reed Reference Grammar of Māori. Auckland: Reed Publishing. 1997.
Biggs, Bruce. Let's Learn Māori (Revised ed). Auckland: Auckland University Press. 1969.
Du Feu, V. Rapanui. London: Routledge. 1996.
Harlow, R. A Māori reference grammar. Auckland: N.Z., Longman. 2001.
Harlow, R. Māori: a linguistic introduction. Cambridge: Cambridge University Press. 2007.
Hooper, R. Neuter verbs and stative aspect in Polynesian, unpublished MA thesis, University of Auckland. 1982.
Hooper, R. "Neuter verbs, stative aspect, and the expression of agency in Polynesia" The Journal of the Polynesian Society. 93 (1984) 39 – 73.
Pavey, Emma L. The structure of language: an introduction to grammatical analysis. Cambridge: Cambridge University Press. 2010.
Trask, Robert Lawrence. A Dictionary of Grammatical Terms in Linguistics. London: Routledge. 1992.
Van Valin, R. D. Exploring the syntax-semantics interface. Cambridge: Cambridge University Press. 2005.
Van Valin, Robert D. & Randy LaPolla. Syntax. Cambridge: Cambridge University Press. 1997.
Williams, W.L. & R.W. First Lessons in Māori. (13th Ed). Auckland: Whitcombe & Tombs. 1862.
"2013 QuickStats about Culture and Identity", retrieved 25th January 2015 from http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-culture-identity/languages.aspx
"Biography of Grace, John Te Herekiekie", retrieved 28th June 2014 from http://www.teara.govt.nz/mi/biographies/5g14/grace-john-te-herekiekie
"Biography of Te Ruki Kawiti", retrieved 28th June 2014 from http://www.nzhistory.net.nz/people/te-ruki-kawiti
"Entry for *tere* in Te Ara, Māori Online Dictionary", retrieved 28th June 2014 from http://www.maoridictionary.co.nz/search?idiom=&phrase=&proverb=&loan=&keywords=quickly&search=
"Entry for *tipu* in Pūtaiao Online", retrieved 28th June 2014 from http://putaiao.tki.org.nz/Papakupu-Putaiao/Nga-Whakamarama-Kupu/tipu
"Kōrero: Te Māori i te ōhanga", retrieved 28th April 2013 from

http://www.teara.govt.nz/en/te-maori-i-te-ohanga-maori-in-the-economy

Reports of the Native Affairs Committee in 1888, NZETC, retrieved 28th April 2013 from http://nzetc.victoria.ac.nz/tm/scholarly/tei-Nat1888Repo-t1-g1-g2-t93.html

"Shane Jones speech from the New Zealand Parliament", retrieved 28th June 2014 from http://www.parliament.nz/en-nz/pb/debates/debates/speeches/49HansS_20111006_00001082/harawira-hone-adjournment

CHAPTER FOUR

CREATION OF THE SUBSTANTIVE CORE OF THE POLYSEMANTIC VERB OF PARTIAL RELATIONS "PART->WHOLE"

SVETLANA KISELEVA
HIGHER SCHOOL OF ECONOMICS, ST. PETERSBURG, RUSSIA
AND NELLY TROFIMOVA
HIGHER SCHOOL OF ECONOMICS, ST. PETERSBURG, RUSSIA

1. Preliminaries

The main assumption of the theory of the substantive core of a polysemantic word is that the meanings expressed by a given word have some abstract minimum notion uniting them (values) or some lexical-semantic variants (LSV) of the particular word that prevent it from becoming two homonyms. This abstract minimum notion is the average denominator of all the derived meanings (indirect values) of the word. The theory began as a method of lexical-semantic and cognitive analysis based on identifying the meaningful kernel (substantive kernel or substantive core); that is, it claimed that the meaning of any semantically complex and non-complex word can be explained by means of an exact paraphrase composed of simpler, more intelligible words than the original (Wierzbicka 1972). In our investigation, we follow these assumptions.

This research combines not only elements of the traditional semantic syntax approach but also some cognitive elements. The cognitive approach considers the process of perception and formation of the word based on a conceptual picture of it which is projected on the lexical system, where the principle of anthropocentrism allows representing it by means of human perception.

This chapter aims to show a cognitive approach to understand the semantic structure of a polysemantic verb. The main statement is that

every complex word has a substantial core, i.e. its essential functioning basis, which provides its semantic integrity.

The present study is based on the understanding of the content and form of the sign. It considers that the value (meaning) is formed first in the mind of the speaker and only afterwards in the mind of the listener (Maturana, 1978; Arkhipov, 1998, 2001). The cognitive approach to this study is that the entire role in the formation of linguistic values belongs to a person as a participant of an act of communication.

As verbs usually have several meanings, it is important to examine their polysemy, which is reflected as an objective reality or as part of the imaginary world of a person. On the one hand, the present research is conducted to find some significant evidence of differences between the values in a multivalued verb (polysemic verb). On the other hand, we aim to find the existence of a common semantic core (substantive core) able to link all the values of a given verb (in particular, using the semantics "part and whole").

Thus, the investigation, which is implemented from a cognitive approach, deals with the formation of the substantive core of a given polysemic verb, namely the verbs of partial relations in Modern English. The cognitive approach supposes the image support/schemas as well as the third meaning, which is the meaning of the speech patterns produced by the speaker in the proper communicative situation or in a context based on the nominative and primitive terms, included in the words. The substantive core of a polysemantic word is dedicated to integrate all its LSVs.

2. Prototype Theory: The Underlying Theory for Prototype Structures

For decades, linguists have searched a meaningful explanation of the content of polysemic words. One of the most popular conceptions is the prototype theory, within the field of prototype semantics, originated in the mid-1970s with the psycholinguistic research of Eleanor Rosch about the internal structure of categories (Rosch 1975, 1978, 1988). Its revolutionary character marked a new era for the discussions on the lexical meaning and brought existing theories into question.

The prototype theory is a mode of graded categorization in cognitive science, where people create categories of things and assign the same name (or label) to things that are not exactly the same but similar, and where some members of a category are more central than others. For instance, when asked to give an example of the concept *furniture*, *chair* is more frequently cited than, say, *stool*. Rosch stated that human cognition

is the primary element for any categorization process (linguistic categorization included). Rosch argued that an object is assigned to a category through comparison with its prototype object rather than through a set of criteria features. This prototype object consists of a mental entity in the human mind.

The followers of the prototypical approach argue that the categories are blurred and do not have clear boundaries. However, there is a gradation in "belonging to the category". There is an opinion that the word is naming a thing only to a certain extent, not absolutely (Kay 1990: 27). People form a concrete or abstract mental image of the object belonging to a given category. This image is termed a prototype if it can help a person perceive reality: the member of the category closest to the image will be recognized as the best example of its class or as the most prototypical one. Prototypes are tools that help people cope with an infinite number of incentives provided by reality. Based on the above, one can conclude that, not only in the assimilation of language evolution and language changes but also in the individual language flexibility in the use of categories, this is manifested (i.e. organized around prototypes) and implemented in different environments (e.g. personal, social, cultural, communicative or biological).

The prototypical approach is an attempt to establish the foundation of the formation of meanings of an LSV in the structure of polysemic words. As a starting point, it should be recalled that the lexical-semantic representations of a word are united by a semantic invariant or substantial core, i.e. the lexical prototype (LP). The traditional approach holds the view that the correlation of primary and derivative values is based on the principle of semantic derivation from the first nominative-derivative meaning. However, in the present research, it is often difficult to fit an indirect meaning to a direct (nominative-derivative and central) one. Sometimes, it takes significant effort and time to understand which of the values is the primary and which the secondary one. One can only assume that there are many of these meanings. Additionally, other difficulties arise when dealing with the definition of the sources of the secondary and tertiary shifts in the meaning of a polysemantic word (Williams 1976).

LPs have been studied deeply and successfully by linguists from different countries. The idea of the prototype is used in traditional studies of figurative language when a property, which is considered typical in an object, is transferred to another object, which is then called by the name of the first. Some examples are, for instance, fuzzy concepts—for the framework of language categorization (Zadeh 1965; Lakoff 1972; Labov 1973), anthropological linguistics—for the research of the category of

colour (Berlin 1992; Kay 1990), or linguistic semantics—for the direction associated with the works of K. Baldinger and G. Lakoff (Baldinger 1980; Lakoff 1972), in which the Aristotelian approach to meaning was criticized but adopted in the concept of semantic rules by J. Katz and P. Postal (Postal 1970; Katz 1972). Cognitive semantics postulates the internal structure of categories and allows the linguist to stay on the level of the detailed description of meanings available at that moment without sacrificing the coherence and integrity of the categories which are used. Following a similar description, a network of semantic relationships can be more complicated: "the fact that was postulated before can come into conflict later with the specified details" (Janda 1988: 6).

The concept of prototype has been explained from several perspectives. For instance, Givon (1984: 1516) believed that there are objects which are grouped around each of the prototypes, most of which may have nothing in common with the prototype. According to Givon, prototypes can vary over time, namely, in the metaphorical use of terms, when the category has new representatives. Such changes are associated with an overriding characteristic set of properties and relative ranks (Givon 1984: 19).

Taylor (1995: 19) stated that prototypical features are manifested in the unity of thought with which native speakers characterize the meanings of linguistic units in isolation from the context, where the meaning shows the best example of the category. The semantics of prototypes for Fillmore determines the meaning of a certain lexeme. According to him (1975: 123), a prototype is inherent in the human mind from birth; it is not analysed, but just "given from birth" (presented or demonstrated), so it cannot be manipulated.

Lakoff (1987: 31) believed that the manifestation of the prototypical effect lies in the fact that the central members of the category, closer to the prototype than the rest, show other cognitive characteristics than the non-central ones. They are quickly recognized and assimilated, often used in speech, and accelerate the solution of all problems related to the identification. They are used in understanding the category as a whole. Prototypical effects are particularly bright when, in the process of reasoning and recognition, some element is used instead of the category. These include, for example, the metonymic models of categorization (Lakoff 1987: 33). The sources of these effects differ. On the one hand, they are sheaves of sensations, the basic model of which is the model of the common non-metonymical type. On the other hand, they are radial patterns: there is a central case and its conventionalization variations.

Cliff (2002: 13) searched for the shared semantic core of all languages. He proposed and investigated primes in historical sequences: they were arranged in groups of comparable elements. His inventory of primes looked like a natural language in miniature. Largely, the kind of meanings included within is the kind of semantic parameters that topologists and descriptive linguists tend to assume in general language description. He identified the term "semantic prime" as a linguistic expression whose meaning cannot be paraphrased in any simpler terms. A secondary criterion is that a semantic prime should have a lexical equivalent in all languages. There is a third consideration: the meta-language of semantic primes is intended to enable reductive paraphrasing of the entire vocabulary and grammar of the language at large, i.e. it is intended to be comprehensive. Thus, semantic primes, which should have lexical equivalents in all languages, can be compared with the substantive core, which should have minimum abstract features to unite all the values of a polysemantic word.

Thus, the prototype (the original) is part of the human lexicon and will have all the updated values in the speech. Coleman and Kay (1981: 27) associated the term "prototype" with the pre-lingual, cognitive vision of the world and believed that speakers have the ability to make a conclusion about how a particular object/subject/concept corresponds to the prototypical image. Therefore, there is every reason to use the term "prototype" in relation to the lexical structure of a word. This concept is true for the Moscow semantic school, where linguists believe that the value of the full meaning of the word is formed from two elements: the nuclear set of all direct referents and the prototypical, indicating the typical referent. This prototypical element is the result of numerous direct references undertaken by it. This prototype corresponds to the sense; it has a communicative function, which is based on the hypothesis that people can understand a statement when they have something of a generalized, conceptual representation of the referential situation described in that statement. If listeners do not know a particular situation, they have to reconstruct it based on their knowledge and the prototypical values of the words included in the utterance (Koshelev 1996: 98, 200). Thus, the prototype helps to understand the perceived utterance.

Gak (1977: 17) asserted that, being stereotypes within ordinary consciousness, prototypes are created by members of a language community on the basis of a uniform division of reality, i.e. "stereotypes simplify the process of communication: people are unable to process all situations; in everyday speech for the identification it is sufficient to point to objects in general". Arkhipov (2001: 39-57) defined a multivalued word

as the word functioning at the system-level language as an LP, i.e. "the best representative of a polysemantic word, while actualization of some values occurs at the level of speech".

Plurality and diversity of opinions prove that the theory of prototypes can be seen to be a usage-based approach to language categorization. Certain areas of prototype structures may appear problematic in their theoretical base but satisfy the basic desiderata for adequately serving as word meanings. Therefore, they can help with the problem of the semantic core of a word, where prototype structures can help to render an adequate lexical prototype and thus the compositional aspect of word meaning.

3. The Substantive Core of a Word

Our interest concerns the semantic core of a polysemantic word. Based on previous research experience, we can assume that no matter how many values a word (concrete form) possesses, it always has a substantial core. At the language level, it is identified as the value of the form that can serve as a definition of the actual meaning of the word, taking into consideration the speech context.

One value should be considered as one unit. To determine the system of values of a lexical unit, it may be useful to refer to the first, main, nominative-derivative meaning on the basis that native speakers usually use these values upon presentation of appropriate language forms.

It is worth noticing that the initial understanding of the substantive core as a lexical prototype relates to a new understanding of the substantive core as a further lexical prototype, i.e. a nominative-derivative meaning and a lexical prototype. Narrowing the scope of the concept, the smallest sheaf of characteristics necessary for the identification of the subject of thought is represented in the formal concepts of Katsnelson (1986), the narrow concepts of Serebrennikov (1976) and the semantic primitives of Wierzbicka (1972). The term "the nearest meaning" belongs to Potebnya (1999). It corresponds to the "intensional" of Nikitin (1996). Under the "further meaning", Potebnya (1999) understands "meaningful value". It covers new aspects of the object, i.e. its properties and relationships with other objects.

In the structure of the lexical meaning of a word, there are two parts: "intensional" and "implicational" (Nikitin 1996: 109-110). The intensional part refers to the substantive core of the lexical meaning of a word; the implicational part refers to the periphery of the semantic characteristics surrounding this core. The intensional is a structured set of semantic features, constituting the class of denotata. For entities of this class, their

presence is mandatory while taking into account the stochastic nature of the world and its entities. In addition, the intensional is the same as the concept of class in logic. According to Nikitin, the intensional is the basis of thinking and speech operations for the classification and naming of denotata.

As the features of a word do not exist separately, because they are connected through various relationships and dependencies, some features make others consider them to a greater or lesser extent. Similarly, intensional characteristics may imply the presence or absence of other signs of the denotata of this class. Regarding the intensional (the core of the value), a set of such implicational signs forms the implicational of the lexical meaning, i.e. the periphery of its information capacity. The information about the denotatum consists of two parts: indispensable intensional characteristics and some implicational signs that are actualized by the context (Nikitin 1996: 110).

Although these terms play an important role in the scientific lexicon, there is no consensus as to which is the closest to the notion of a meaningful kernel. As already mentioned, there is one more term on the matter of a lexical prototype, described by Arkhipov (1998: 16; 2001: 50). He singles out the following identifying characteristics of the LP:

- The LP is the sheaf of meanings which are usual, abstract, important and communicative.
- The LP is a minimum sheaf of integral and differential characteristics which are needed for the identification of objects (concepts).
- The features of the LP cannot be deduced from one another.
- The LP is a semantic invariant of all the LSVs of a polysemantic word.
- The content of the LP is determined at the level of ordinary consciousness.
- The LP includes a program for all (or almost all) private LSVs of a word and vice versa: in any case, there is a hint on the model, on its distinctive features.
- The LP manages the process of semiosis of metaphorical meanings.
- The LP includes the value of the general meaning “something resembling” or “something related” in the structure of a polysemantic word.

In the beginning, when studying the substantial core of a polysemantic word, a lexical prototype was conceived as a sheaf of components only of

abstract nature (Arkhipov 1998; Pesina 2005). However, recent studies in this area show that, to determine the semantics of the whole polysemantic word, the primary value (nominative-derivative, important, direct, central and basic value) should be taken into account by all means, because it occurs first in the mind of a native speaker while understanding the semantics of a polysemantic word. The indirect value of a word can only be identified through the implicit presence of the nominative-derivative value of a word, so that the semantic structure of a word is "a set of forms and meanings inherent in one and the same word in all its usages and implementations" (LES 1990: 257). It becomes obvious only when they are compared with the motivating nominative-derivative meaning. According to Arkhipov (1998: 16; 2001: 50-51), the nominative-derivative (primary, central) meaning is called "the nearest" lexical prototype (NLP), as the value that first comes to the mind of a native speaker as an element of the basic level of understanding of the concepts of some entity. The whole meaningful core of a polysemantic word, i.e. the NLP and an abstract part which is the comprehension of all the LSVs of a word at a higher level of generalization, is then called a "further" lexical prototype (FLP). Based on the studies outlined above, the identification of the lexical prototype of a polysemantic word was complemented by Pesina (2005: 78) with some important points:

- The conceptual basis for the formation of the LP is a nominative-derivative value based on the ordinary native speaker.
- The LP does not have a declarative but a dynamic nature: the process of actualization of the speaker's knowledge is represented as a sequential composition of more complex structures on the basis of the integral and differential components of the nominative-derivative meaning of a word.

As a result of further studies on examples of verbs of partitive semantics, additional principles for identifying an LP or the meaningful kernel (substantial core) have been identified. The basis of the values of the sheaf of abstract features is formed by components such as *like, as if, so* or *as to*, conceptualizing the abstract (metaphorical, metonymical) values of a polysemantic word (Kiseleva, 2006).

Under the substantial core, we assume an invariant associative complex inherent in the word and appearing in the mind of the communicants; this complex is based on the semantic structure of the word, the grammatical organization, the patterns of word formation, motivational relations, and certain traditions of the use of lexical units.

Following the spirit of modern research in the field of cognitive science, this theory continues to develop and is the focus of attention of linguists.

Therefore, according to Arkhipov (1998: 16; 2001: 50-51), a substantive core or "lexical prototype" contains the smallest sheaf of features and is the basis of the entire structure of a polysemantic word. The study of the formation of the lexemes of a word and its functioning in speech using methods of prototypical semantics assumes that, to find the systematic value of a polysemantic word, covering all the meanings, it is necessary to make the role of the first (nominative-derivative) meaning wider. The main nominative value is an integral part of the substantial core of a multiple-meaning word, as it first comes to mind when thinking of a particular lexeme. Therefore, the invariant, including elements of abstract nature, is the summation of all values of a multiple-meaning word.

In the end, the substantive core is the product of the mapping and comparison of all lexicographic data of a word from which the total value is extracted, which is repeated in all of the presented values of a verb in the same semantic structure. The basis of the substantial core is the direct nominative meaning of a word. It can be assumed that a meaningful value represents the level of the language system as the best representative of its category. It is rational for the general classification of verbs but cannot be used to describe the semantics of an individual verb in the lexicographical representation, i.e. it cannot be used to differentiate the verbs of a separate group. However, a substantial core contributes to the classification of verbs (or any group of words) and their organization in synonymic rows or lexical-semantic groups.

Thus, if a substantial core is uniform for the whole group of verbs (words), it excludes its use in the description of the subtle semantic differences between the verbs (words) of the same group and the individual representative of the semantic structure of a verb (word). Therefore, we can say that the concept of a substantial kernel in this form has its limitations. For lexicographic works, this concept may be used only when compiling ideographic dictionaries.

4. The Invariant as a Meaningful Core of a Polysemantic Word

Now we will refer to the anecdotal evidence, according to which the lexical entry (definition) of a word should include all elements of the content of a linguistic unit. With this opinion, the research has been held in the sphere of "semantic compactness", aiming to find out the substantive

core of a word, both at the level of a single meaning and at the level of a polysemic verb.

For this chapter, the semantic values of the meanings of verbs are restricted to the meaning of "combination of parts into the whole" until there is a minimum necessary of components (semes). As language is constantly striving to economize, it may suggest that the unit-level language is stored in the mental space of the vocabulary of a person, not in the form of detailed dictionary definitions but in a more compact form.

The substantive core (intensional) at the level of the individual value of a partitive verb should involve narrowing the semantic components to the minimum required. Moreover, the meaning "the whole and its parts" should be the most recognizable. Thus, it is necessary to study and take into account all the meanings of a word.

As already mentioned above, Potebnya (1959: 19-20) developed the concept about the nearest and the furthest meaning. He considered that the meaning of the word includes "two different issues": one of them is called "the nearest meaning", belonging to the study of language, and the other one – "the furthest meaning" – is the subject of other studies, but only the nearest meaning has the real content of thought while it is pronouncing. When saying a word, the human mind does not focus on its entire set of characteristics, because it takes time to perform definite mental operations. Potebnya believed that a word out of context does not express its full content, but only one essential feature. He called it the nearest meaning, which is, together with the flight of imagination, what makes the understanding between a speaker and a hearer intelligible. The nearest (or objective) meaning was called "public" by Potebnya, because he considered that the speakers of a linguistic community have similar thoughts. The further meaning is considered subjective, i.e.

> "from personal understanding higher objectivity of thought occurs, the scientific one, but only with a help of people's understanding, i.e., language and tools, which are caused by the existence of language" (Potebnya 1999: 120-124).

Thus, the nearest meaning represents the form in which its content is represented to our ordinary consciousness. The internal form of the word implies the content of thoughts in consciousness. It gives the opportunity to see how his/her "own idea is represented to a person". According to Potebnya (1999: 124), this issue can explain why language may have many words to define the same thing and vice versa; one word can denote different things. In addition, he believed that sometimes it is impossible to distinguish the most significant feature in a given notion. Thus, perhaps

the meaning is not one, but there is instead a minimum number of possible features for a meaning. The author considered that the semantics of a word, recorded in a dictionary as the nearest meaning of the word, is secondary and derivative in relation to the knowledge about the world, which is the further meaning.

Additionally, the idea of the status of the "nation" of the meaning is nowadays alive (Apresyan 1995). Apresyan differentiated between formal and substantive concepts. The formal ones are defined as a general minimum and at the same time as more discriminatory criteria necessary to define and distinguish a thing. From Katsnelson's (1986: 20) point of view, "the formal concept sums up the most important what we should know about the issue, that is why any new step in learning of issue is out of such meaning". With regard to substantive concepts, they differ from formal concepts not only in content but also in form. This covers new aspects of the object, namely its properties and relationships with other objects. People may understand concepts differently because of their own individual experiences, education level, talent, etc. In other words, the comprehensive meaning of this author coincides with the further meaning of Potebnya.

According to Karaulov (1987: 168-170), first there is a change in the further meanings of the word, and then the components of these enriched values penetrate to the level of semantics and are partly reflected in the change of the nearest value, i.e. "the information registered by the linguistic semantics reflects a small part of the knowledge about the world and in some cases it can reflect them perversely".

There is still an open question about limiting the number of semantic components in the value. Some linguists try to narrow down the number of components by reducing their values. Other researchers consider that this number cannot be limited; this implies that the meaning cannot be described by an exhaustive set of semantic components when the semantic component itself is the kernel, capable of further partitioning the elements of meanings. For example, many linguists point to the extraordinary complexity of such operations within the lexical-semantic variants.

Gak (1977: 42-50) singled out the core of the meaning of the word as semanteme, semantic category, and semantic component or characteristic (seme), i.e. a reflection of the distinctive features. He distinguished archisemes, differential semes and potential semes in the semantic structure of the word. For example, the archiseme "means of transport" names characters peculiar to a class of objects: bus, train, airplane, etc. The archiseme may become a differential seme against the semes of the higher level; for instance, "to go" and "to talk" have achieved the "to act"

archiseme. The core of the meaning of the word corresponds to the differential semes, significantly distinguishing the semanteme of one word from another. Potential semes reflect minor, sometimes irrelevant, characteristics of the subject, as well as different associations with which the given subject is associated in the speakers' minds. In the ordinary usage of the word, their functioning is associated with the indirect meaning of the word (Gak 1977: 13-14).

Nikitin (1996: 109) highlighted the notion of intensional features close to the minimum required substantial core. The intensional is regarded to be a substantial core of the lexical meaning, that is:

> "a structured part of semantic features, constituted the class of denotata. The intensional is the same content concept about the class in logic. It is an intensional that is the base of thinking and speaking operations on classifications and denotations naming".

For instance, all mothers are women-mothers; these two characters (i.e. parent and female) make the intensional of the word "mother" in its direct meaning.

Various relations and dependencies exist among the characteristics. Therefore, "some characteristics make to think about others with greater or lesser necessity" (Nikitin 1996: 110). The author believed that, with respect to the intensional, i.e. the core of values, a set of implications forms the implicational of the lexical meaning, i.e. the periphery of its information potential. The information about the denotatum, which the word bears in the text, consists of two parts: indispensable intensional characteristics and some implicational signs, actualizing in the context.

In Nikitin's opinion, the semantic features in the intensional consist of two parts: hyperseme and hyposeme; they are connected by sexuated-specific (genus-species) or hyper-hyponymic relations. The sexuated part of the intensional is called hyperseme (archiseme), where the specific part is the hyposeme, since it has differential features. Thus, he presented an example connected with a girl, the intensional feature of whom is a female child, a hyperseme is a child concept, and the hyposeme is the concept of female.

According to Nikitin, the implication of features can be strict (mandatory, necessary), highly probable, faint (free) and negative. In the first two cases, the intensional of the meaning implies features with probability equal or close to 1. They convey a strong meaning implication. This variety forms a combination of features of the especially strong implication with probability equal to 1, i.e. with the necessity of implication from the intensional. The features of a strong implication are

close to the intensional core, the integral part of the lexical meaning. Therefore, rather often they appear as part of meanings in defining dictionaries. Nevertheless, they are beyond the intensional limits because the "theoretically possible absence of such feature in a denotatum does not exclude a denotatum from the class where it is supposed to be" (Nikitin 1996: 110). For instance, the intensional of the word *winter* is a season from December to February (in the Northern hemisphere). The strong implication meaning includes such features as the *coldest season, it is snowing, water is frozen, the sun is shining lightly, people wear warm clothes*, etc. This example shows us the difference between the intensional and the implicational features: if winter is warmer than another season, it is still winter, because the determining feature relates to time duration. Moreover, Nikitin (1996: 111) stated that the implication of some features relative to the intensional seems impossible or hardly probable (i.e. features that are incompatible):

> "in addition to strong and negative implications there is extensive area of features, the joint occurrence with this concept can be judged only conjecturally: their presence and lack are equally possible and problematic, they can be or cannot be".

This area towards the intensional of meaning expresses its weak or free implication.

Thus, according to Nikitin, the lexical meaning is a complex entity, directly woven into the cognitive system of consciousness. The structure of the lexical meaning is formed primarily by subject-logical relationships, and then extended by its intensional kernel and some implication features included in the periphery of the content. The structure of the intensional is formed by the logical dependences of its semantic features and first of all by sexuated-specific (hyper-hyponymic) relations. The implication features are also structurally ordered by their probabilistic characteristics and objective-lexical dependences (Nikitin 1996: 115). Nikitin considered the intensional to be in the nearest view (or equal) to the substantial core of meaning.

Wierzbicka (1997, 1999) believed that, to compose dictionary definitions, it is necessary to use a method of "reductive analysis", which assumes that all concepts should be defined by a set of the most indefinite semantic features. Wierzbicka quoted Leibniz's words when she stated that a great number of concepts can be obtained by combining several elements "because the nature tends to achieve the maximum effect with a minimum number of elements, i.e. to act in the simplest way" (Wierzbicka 1997: 296). The author considered that the main point is to postulate a

limited set of semantic primitives, the outer contours of which are interpretations of all lexical and grammatical meanings of natural language, that is:

> "if there is some number of conceptual primitives, directly understood (not through other concepts), so these primitives can be a very good base for all other concepts; an infinite number of new concepts can be derived from a small number of semantic primitives" (Wierzbicka 1997: 296-297).

Semantic primitives were revealed and investigated by the author on words denoting parts of the body.

It should be noted that some linguists have doubts about recognising those or other semantic units as "elements". According to the general theory of language, at the level of the basic units of meaning there is competition for the right of this or that unit to be qualified as an element to be part of the metalanguage (Kobozeva 2000). If we strictly adhere to the rules of usage of only atom-primitives in the interpretation, such descriptions seem to be very complex. Therefore, the examples of interpretation by Wierzbicka cause difficulties in perception, because there are no well-defined syntax rules of forming metalanguage manifestations. Although the author believed that every meaning of any word could be defined with maximum precision, the lexicographical value of the "universal metalanguage" is in great doubt.

Geeraerts supported the analysis developed by Wierzbicka; he believed that prototypical concepts are encyclopedic formations that require detailed description (Geeraerts 1975). Wierzbicka has undoubtedly contributed to lexical semantics, and her work can serve as an example of samples of epistemology of lexical semantics.

Our belief is that the usage of semantic primitives by A. Wierzbicka undoubtedly brings scientists closer to the minimum comprehensive meanings, which explain the sense of functioning of lexical items. Her experience in the interpretation of the first (nominative-derivative) values of polysemic words deserves special attention in this study. The explanations of the first meanings of words in dictionaries are not always available to the consciousness of an ordinary person. Therefore, the author suggested taking into account the proposed definitions without using a special metalanguage.

In the works of Apresyan, metalanguage and semantic primitives are defined in such a way that the metalanguage dictionary is reduced to two types of words: (1) semantic primitives (i.e. undetectable words, preventing further semantic reduction), and (2) more complex words from

a semantic perspective, which can be reduced to primitives in one or several steps. Apresyan (1995: 486-481) stated that:

> "the words of the natural language which are chosen to be primitives are always words of the "first plan", more implanted in language and culture. They service the more number of pragmatic situations".

Although these primitives are existing meanings, they never materialize in the words of natural languages: the physical perception (sight, hearing, etc.) denoting "to perceive", the physiological state (hunger, thirst, etc.) that means "to feel", physical actions and activities (work, rest, etc.) suggesting "to do", etc.

Thus, there are several definitions of the substantive core of the meaning of a word: "intensional" (Nikitin), "differential seme" (Gak), "semantic primitive" (Wierzbicka, Apresyan) and others. In his works, Potebnya noticed that the substantive core ("the nearest meaning") should be inherent "nation" and "singleness". It must be typical to "nationality" and "singleness". All these authors are united by the way of understanding the meaning of a word like singleness; its complex character presents a specified cognitive reflection of reality. Thus, the overall picture reflects the linguists' basic opinion regarding the notional alliance inside the meaning that is its substantive core.

This survey helps to sum up that the substantive core of the meaning of the word has the following features:

- narrowing of the semantic elements until minimum necessary;
- stable components, where the meaning should be maximally recognizable; and
- subjectivity of the opinion is admissible from the point of view of cognitive linguistics.

To sum up, it can be argued that the problem of the definition of basic cognitive mechanisms, which underlie the formation of the meanings of the polysemic word and determine its substantive core connecting all lexical-semantic variants of the word, can be solved with the help of lexical prototypes. Based on the nominative-derivative value, the realization of all indirect values is implemented, so it is formed first. The primary value is derived from the definition of dictionaries using component analysis based on the principle of frequency.

5. The Definition of the Meaningful Kernel of the Partitive Verb "to compose"

In view of the above, it is possible to define the main cognitive mechanisms that are the base of the meaning of the verb "compose" and to determine the invariant as the substantive core, connecting lexical-semantic variants of this verb. The purpose of this is to prove the function of the representative of the verb lexeme, expressing the relations between "the part and the whole" at the level of the language system, and the updating of the indirect meanings at the speech level. The cognition of the indirect meanings is realized on the basis of the nominative-primitive meaning, so it is formed first. The primary meaning is produced with the definitions of dictionaries and the usage of the component analysis on the basis of the severalty principle.

The dictionary articles of the verb ***compose (1)*** begin with the description of its values as follows:

1)
   a) be composed of smth. - to be formed of a number of substances, parts or people = consist of;
   b) to combine together to form something = to make up (LDCE 2: 314); if something is composed of particular things or people, it has those things or people as its parts or members (CCELD: 285);

2)
   a) consist of, be composed of = be made up of (something, things, or people) (LDPV: 103);
   be composed of smth. = to be made of a particular substance or substances (LPVD: 99);

3) to write a piece of music (MED: 283);
   a) to form by putting together;
   b) to form the substance of;
   c) to produce (as columns or pages of type) by composition (BDE);

4) to form by uniting two or more things;
   to form, frame or fashion;
   to form by being combined or united;
   to constitute;
   to make (NWD: 326);

5) to write or produce (music, poetry, etc.) (LDELC: 257);

6) to constitute, form, to make up, be a constituent or ingredient or component or element of, be a part of (OT: 67);

7) to be the constituent parts of; to create by combining parts or elements (RNT: 185);

8) to put together, make up: to make by putting together parts or elements: make up, form, frame, fashion, construct, produce (OED 3: 621). Examples:
Water is composed of hydrogen and oxygen (LDCE 1: 272);
The committee was evenly composed of men and women (CCELD: 285);
He sat down and composed a letter of resignation (MED: 283).

A component analysis of the above definitions gives the following results, in which dictionaries unanimously name the main identifying and most frequent items taken for analysis. For example, six dictionaries (LDCE2, LDPV, LPVD, NWD, OT and OED3) mention the element *to make* (*up/of*). Five dictionaries (LDCE 2, BDE, NWD, OT and OED3) fix the component *to form*. Four of these dictionaries contain both of these semes (LDCE2, NWD, OT and OED 3). Four definitions include the element *to be composed of*. Three dictionaries (BDE, LDELC and OED3) contain the seme *to produce*. In some of the above-mentioned dictionaries the components *to constitute* and *to construct* are found in the articles. Two dictionaries note the elements *to frame*, *to fashion* (OED3 and NWD) and *to consist of* (LDCE2 and LDPV). In two dictionaries (MED and LDELC), the first position of the article is occupied by the elements *to write or produce (music or poetry)* and *to write a piece of music*, which in other dictionaries these semes occupy the second or the third positions, i.e. they are considered by the compilers of the dictionaries as derivative values. The partitive semes *substance, part, people, member, thing, constituent, ingredient, component* and *element* are presented explicitly in many dictionaries, while in others they are given implicitly. Three definitions are specified with quantitative parameters of the whole (e.g. *a number of substances* or *to form by uniting two or more things*), and also three definitions are specified with qualitative peculiarities (e.g. *particular things or people, to produce* (*as columns or pages of type*) *by composition* and *to be made of a particular substance or substances*).

Thus, the analysis of the nominative-derivative values of the partitive verb *to compose* on the basis of the data of eleven defining dictionaries enables us to identify the most frequent of the selected semes in order to continue to formulate a nominative-derivative meaning of the analysed verb or to find a common semantic denominator (substantive core) on the basis of the initial values, resulting in the following form: ***compose (1)*** "*to form the substance/to make something by combining two or more parts or to form things or people together as a whole"*. This definition, in our

opinion, includes all necessary and full components for the purpose this word could be immediately recognized at the level of the ordinary consciousness of a person.

The first performance of the verb "compose" was associated with the image of planks or logs, put or knocked together to create a sort of construction similar to a house, wall, raft, etc. Eventually, when the life of people became more complicated and came into the world of artefacts, more complex issues and things appeared, and the relations between them also became different.

In the capacity of the lexical-semantic variant of the analysed verb, motivated by the nominative-primitive meaning, it is reasonable to make an example of a secondary meaning of this verb. Its analysis is based on comparison or identification, as a traditional way of interpretation. The purpose of this part of the investigation is to find out what items are the bases when forming and decoding metaphorical statements. Along with this, the analysis is based on the principles of the cognitive approach, supported by cognition and nomination for the appropriate images of perception. We have to prove whether cognitive images, namely, of partitive relations, underlying nominative-derivative values to comprehend metaphorical statements, are kept.

If the lexeme ***compose (1)*** (the definition of the first meaning is mentioned above by several dictionaries) gives the general concept associated with the value of formation of the partitive relationship between any two objects (things, concepts), where the composition of the whole is not specified, then the meaning of ***compose (2)*** is defined as a number of people or things composing something, i.e. they are the parts or members that form it. Example:

The elements which compose his individuality (CCELD: 285).

The sentence *There, a large part of the Felding-Roth sales force was composed of women* (Hailey: 156) is associated with a clear clarification of the content of the parts (people and things) that become integral parts of the whole "part > whole", that is, an extension of the nominative-derivative (initial) value as if/so as ***compose (1)*** (to form things or people together as a whole) or to acquire integrity.

compose (3) - (music) to invent and put into proper form. Examples:

The piece of music was composed by a famous Italian master (OED 3: 621).
Mozart composed his first symphony in 1764 (CCELD: 285).

The semantics of this metaphor is built on the basis of the analogy of composing music and building a house. Written music (i.e. composing

music) resembles the structure of a building (house) made of bricks or logs. The meaning of this "compose" involves the following components: to write, to create or to make up. In this definition, there are no parts. The components "part" and "form together" are eliminated; a nominative-derivative meaning is not seen. Thus, it means that there is no partitivity. This case can be seen as precategorical: from the class of partitive relations the verb "to compose" has moved to another with a value of creative activity as if/so as ***compose (1)*** (to form things or people together as a whole) or to acquire integrity.

compose (4) - to put together (types) so as to form words and blocks of words; to set up (type); to set up (an article, a page) in type. Examples:

> *The compositor was Mr. Manning who had composed about one half of his "Dictionary"* (OED 3: 621);
> *If you compose a short piece of writing such as a poem or speech, you write it; used especially when this requires skill or effort* (CCELD: 285);
> *Among the sales force, enthusiasm about the prospects for Montayne was running high, and someone at head office had composed a song to be sung to the tune of "America the Beautiful"* (Hailey: 250-251);
> *She composed a letter to Clarence Desmond, and two days later she received a reply from Mae* (Sheldon: 221);
> *Audrey could feel her fear pound in her chest the next morning as they went to the hotel and dropped off the note she and Charlie had composed* (Steel: 453).

Let us consider a few examples. First, it is necessary to remember that the cognitive approach in our study supposes the image support as well as the third meaning, that is, the meaning of speech patterns that is produced by the speaker in a particular communicative situation or within the appropriate context based on the nominative-derivative value included in the semantic structure of the word.

Thus, the basis of the metaphor in sentences such as "She composed satirical poems for the New Statesman; It can't be too difficult to compose a nice negative reply (if you compose a short piece of writing such as a poem or a speech, you write it; used especially when this requires skill or effort)" (CCELD: 285) is associated similarly with the formation of the parts into the whole. There is the reflection of the comparison with the partial relation between the specified subjects in the metaphorical meaning. Just as planks are the parts of some building, this LSV, which is grounded in the primary meaning, can be explained as the composition of the satirical meaning or the negative answer, taking into account a creation of the parts into the whole. This lexeme is based on the semantic elements "to write", "to create", "piece of writing", "ability" and "to form words".

Therefore, the semantics of "compose" supposes the existence of these elements. It is obvious that the nominative-primitive issue is here like the seme "to form words", where the component "words" represents the component "parts", and where the other components are eliminated. Consequently, it is possible to say that in the given specific meaning some elements of the partitive relations of the verb "to compose" are lost, because the verb acquires creative features. Consequently, the examples "She composed satirical poems for the New Statesman" and "It can't be too difficult to compose a nice negative reply" mean as if/so as ***compose (1)*** (to form things or people together as a whole) or to acquire integrity.

compose (5) - The metaphor "*Look at the way Hoyland composes his picture*" (CCELD: 285) is hypothetically motivated by the comparison, connected with the specific style of painting, with the creation of masterpieces, and with the integrity formation. This image is associated with the deviation from the conventional usage of the verb "to compose". In this metaphor, the meaning "to compose" is rethought incompletely; its semantics includes not only some additional components but the core of the nominative-primitive meaning: "if you compose a painting, a garden, or a piece of architecture, you arrange its different parts in a deliberate and usually attractive or artistic way", not given in the definition of the primary meaning. It should be noted that in this example the lexeme "compose" keeps the meaning of establishment of the partial relations between "the whole" and its "parts": the whole and the parts are represented explicitly. Therefore, the nominative-primitive meaning is seen through the substantive core of this verb. In this metaphor, there is an adaptation of the meaning of the verb "to compose", implied of the direct meaning, but keeping the idea of "combining the parts into the whole" as if/so as ***compose (1)*** (to form things or people together as a whole) or to acquire integrity. At such abstract level the conceptual integration takes place.

compose (6) - The content of this metaphor also has the abstract character: "*In the second case, I will give you some tea to compose your spirits, and do all a woman can to hold my tongue"* (Collins: 60). As it was noticed, the word "compose" represents the image of combination of wooden parts into the whole (construction), and the analysed verb means only this. When such image (frame) is over the other image in this context with the specific lexical filling, it does not keep its systematic meaning. A person understands that "to compose one's spirits" is to settle one's nerves. Here is a mechanism for searching a more appropriate sense. The components in the basis of this meaning (i.e. *to make an effort not being angry, to be calm, to pull yourself together* and *to be concentrated*) cover

all the concepts of the formation of the whole from the parts. The presence of the meaning of the general nature suggests that the semiosis occurs towards the actualization of abstract concepts and the following introductions of this lexeme. The components of the nominative-primitive meaning are eliminated and the verb “to compose” is categorized into a class with the meaning “to settle one's nerves”.

It should be noted that this meaning clearly arises from the fact that human memory keeps memories of the cases of actual implementations of indirect meanings in the semantic structures of the sentences. We must assume that this information proves that real people have real imagination of the relation between abstract things and real people/things as the basis for the metaphoric understanding with the usage of the lexeme “like” and the constructions “as if/so as”. Therefore, it is supposed to put the meaning into the value of the substantive core that has the following form: nominative-primitive meaning “***compose (1)*** (to form things or people together as a whole) or as if/so as/like ***compose (1)*** (to acquire integrity)”.

We believe that the main nominative meaning is not completely included in the substantive core of the verb; it can be altered or modified in real speech. Moreover, some part of the original definition of the verb is repeated, its status changes, and the word is used in its reduced value. The primary meaning of the word alludes to the situation, and it is quite possible that it goes through all the derivative values.

6. Conclusions

In summary, the analysis of results allows to form the following core of the word “compose” as “to acquire integrity”. This invariant as the substantive core is an item of the lexical language system and all meanings are actualized against the communicative set of participants of the communication. Thus, it should be noted that the abstract meaning “to acquire integrity” is a strong reason of the real existence of the substantive core connecting all the verbs of one class but not showing the different features of a separately taken verb. Every verb has its own substantive core, containing some specific components. The substantive core is the perception of all common things that characterize all the LSVs of a polysemic word. The nature of such meanings is so extensive that it grows out of relations between real things.

The use of the proposed prototypical approach to the study of the lexical meaning seems to be possible and relevant for the study of alternative mechanisms for the formation of the semantic structure of a word. The definition of the invariant of the polysemic word, namely of the

partitive verb, allows to identify the cognitive structures underlying semantic changes and to model the processes leading to the development of the multiple meanings of words. A research in the framework of the cognitive semantics of the problems of forming, learning and storing multivalued units of verbs with the meaning of partitive relations between the parts and the whole in the mental space of the lexicon of a person can lead to unravelling the mysteries of the development and functioning of the entire cognitive system.

7. References

Apresyan, Yuri. Leksicheskaya semantica. Sinonimicheskiye sredstva yasyka (Lexical Semantics. Synonymous Means of Language). In Apresyan Yuri The Major Works, edited by Alexey Koshelev. Vol. 1. Moscow: Yasyki russkoy kultury, 1995.

Arkhipov, Igor. "Problema yasyka i rechi v svete prototipicheskoy semantiki (Speech and Language Problems in the Light of the Prototypical Semantics)". In Studia Linguistica: Problemy lingvistiki i metodiki prepodavaniya inostrannyh yasykov (Problems of linguistics and methodology of teaching of a foreign language), edited by Novella Kobrina. St. Petersburg: Test-Print, 1998. 5-22.

—. Chelovecheskiy faktor v yasyke (The Human Factor in Language). St. Petersburg: NIYAK, 2001.

Baldinger Kurt. Semantic Theory. Oxford: Blackwell, 1980.

Berlin, Brent. Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies. Princeton, New Jersey: Princeton University Press, 1992.

Cliff, Goddard. "The Search for the Shared Semantic Core of all Languages". In Meaning and Universal Grammar. Theory and Empirical Findings, edited by Goddard Cliff and Anna Wierzbicka. Vol. I. Amsterdam: John Benjamins. 5-40.

Coleman, Linda, Kay, Paul. "Prototype Semantics: The English Word Lie". Language. Vol. 57. No. 1. 1981. 26-44.

Fillmore Charles. "An alternative to checklist theories of meaning". In Proceedings of the Berkley Linguistic Society, edited by Cathy Cogen, Henry Thompson. Berkley, 1975. 123-131.

Gak, Vladimir. "O semanticheskom invariante i sinonimii predlozheniya (On the Semantic Invariant and Synonymy of the Sentence)". In Sbornik nauchnyh trudov (Collection of scientific papers), edited by Maria Borodulina. Vol. 112. Moscow: MGPIIYa, 1977. 42-50.

Geeraerts, Dirk. “Where does Prototypicality Come from?” In Topics in Cognitive Linguistics, edited by Brygida Rudzka-Ostyn. Amsterdam/ Philadelphia: John Benjamins Publishing House “Vischa Shkola”, 1975.

Givon, Tom. “Direct Objects and Dative Shifting: Semantic and Pragmatic Case”. In Objects: Towards a Theory of Grammatical Relations, edited by Frans Plank. London and New York: Academic Press, 1984. 151-182.

Janda, Laura. “The Mapping of Elements of Cognitive Space into Grammatical Relations: An Example from Russian Verbal Prefixation”. In Topics in Cognitive Linguistics, edited by Brygida Rudzka-Ostyn. Amsterdam/ Philadelphia: John Benjamins Publishing House “Vischa Shkola”, 1988. 327-344.

Karaulov, Yuri. Russkiy yasyk i yasykovaya lichnost (Russian Language and Language Personality). Moscow: Nauka, 1987.

Kats, Jerrold. Semantic Theory. New York: Harper & Row, 1972.

Katsnelson, Solomon. Obshcheye i tipologicheskoye yasykosnaniye (General and Typological Linguistics). Leningrad: Nauka, 1986.

Kay, Paul. “Even”. In Linguistics and Philosophy, edited by Thomas Ede Zimmermann, Graeme Forbes. 1990. 59-112.

Kiseleva, Svetlana. “Soderzhatelnoye yadro i mnogosnachnost' (The Substantive Core and Polysemy)”. In Aktualnye problemy germanistiki i romanistiki (Actual Problems of Germanic and Romance Studies), edited by Larisa Nyubina. Vol. 10. Smolensk: SmolGU, 2006.

Kobozeva, Irina. Lingvisticheskaya semantica (Linguistic Semantics). Moscow: Editorial, 2000.

Koshelev, Aleksey. “Referentialny podhod k analisu yasykovyh znacheniy (Referential Approach to the Analysis of Linguistic Meanings)”. In Moskovskiy lingvisticheskiy almanah (Moscow Linguistic Almanac), edited by Nikolay Pertsov. Vol. 1. Moscow: Yasyki russkoy kultury, 1996. 82-195.

Labov, William. “The Boundaries of Words and their Meanings”. In New Ways of Analysing Variation in English, edited by Charles Bailey and Roger W. Shuy. Washington: Georgetown University Press, 1973. 67-90/

Lakoff, George. “Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts”. In Papers from the Eighth Regional Meeting of the Chicago Linguistic Society, edited by Paul M. Peranteau, Judith N. Levi, Gloria C. Phares. Chicago: Chicago Linguistic Society, 1972. 183–228.

Lakoff, George. “Prototype Theory and Cognitive Models”. In The Intellectual and Ecological Bases of Concepts, edited by Ulric Neisser. Cambridge: Cambridge University Press. 1987. 63-100.

Maturana, Humberto R. “Cognition”. In Wahrnehmung und Kommunikation, edited by Peter. M. Hejl, Wolfram. K. Köck, Gerhard Roth. Frankfurt: Peter Lang, 1978. 29-49.

Mervis, Carolyn B., Rosch, Eleanor. Categorization of Natural Objects. In Annual Review of Psychology, 1981: 32. 89–115.

Nikitin, Mikhail. Kurs lingvisticheskoy semantiki (The Course of Linguistic Semantics). St. Petersburg: Nauchny centr problem dialoga, 1996.

Pesina, Svetlana. “Prototip i semantika mnogoznachnyh slov (Prototype Semantics of Polysemantic Words)”. In Teoriya i praktika rechevogo obshcheniya (Theory and practice of speech communication), edited by Alla Bukina. Vol. 1. Magnitogorsk: MGTU, 1998: 43-52.

—. “Obraz kak ishodnyi element kognitivnogo analisa mnogoznachnogo slova (Image as the Initial Element of Cognitive Analysis of a Polysemantic Word)”. In Gertsenovskiye chteniya 2005 (Gertsenovskie reading), edited by Alexey Vorontsov. St. Petersburg: RGPU, 2005: 73-74.

Postal, Paul M. “On the Surface Verb “Remind”. In Linguistic Inquiry, edited by Samuel Jay Keyser. Vol. 1. N. 1. 1970: 37-120.

Potebnya, Aleksandr. Iz zapisok po russkoy grammatike (From the Notes on Russian Grammar). Vol. 1-2. Moscow: Nauka, 1959.

—. Mysl' i yasyk (Thought and Language). Moscow: Labirint, 1999.

Rosch, Eleanor. “Cognitive Representations of Semantic Categories”. In Journal of Experimental Psychology: General. 1975. Vol. 104: 192-233.

—. “Principles of categorization”. In Cognition and Categorization, edited by Eleanor Rosch and Barbara B. Looyd. Hillsdale, NJ: Erlbaum, 1987. 27-48.

—. Coherences and Categorization: a Historical View. In The Development of Language and Language Researches: Essay in Honor of Roger Brown, edited by Frank S. Kessel. Hillsdale, NJ: Erlbaum, 1988. 373-392.

Serebryennikov, Boris. “Svodimost yazikov mira, uchet spetsifiki konkretnogo yazika, prednaznachennost' opisaniya (The Reducibility of the World's Languages, the Specific Language, the Relevance of the Description)”. In Principy opisaniya yazykov mira (The Principles of Description of Languages of the World), edited by Victoriya Yarceva. Moscow, 1976. 7-52.

Taylor, John R. Linguistic Categorization. Prototypes in Linguistic Theory. Oxford: Clarendon, 1995.
Wierzbicka, Anna. Semantic Primitives. Frankfurt: Athenäum, 1972.
—. Lexicography and Conceptual Analysis. Ann Arbor: Karoma, 1985.
—. Understanding Cultures through Their Key Words. NY, Oxford: Oxford University Press, 1997.
Williams, Joseph. "Synaesthenic Adjectives: a Possible Law of Semantic Change". In Language, 1976: 461-478.
Zadeh, Lotfi A. "Fuzzy Sets". In Information and Control. 1965: 338-353.

8. Sources and Authorized Abbreviations

LES – Lingvisticheskiy entsiklopedicheskiy slovar' Лингвистический энциклопедический словарь. М.: Sovetskaya entsiklopediya, 1990.
BDE - Britannica Deluxe Edition, 2001, CD-Rom.
CCELD – Collins Cobuild English Language Dictionary. Collins London and Glasgow, 1990.
LDCE - Longman Dictionary of Contemporary English, New Edition. – London: Longman Group Ltd., 2004.
LDELC - Longman Dictionary of English Language and Culture. – London: Longman Group UK Limited, 1996.
LDPV – Longman Dictionary of Phrasal Verbs. – London: Longman Group UK Limited, 1996.
MED – Macmillan English Dictionary (for advanced learners). International Students Edition. London, 2002.
NWD – New Webster's Dictionary of the English Language. – USA: Surjeet Publications, 1988.
OED - The Oxford English Dictionary, Second Edition. - Oxford University: Clerendon Press, 1989. - Vol. 3.
OT – The Oxford Thesaurus. An-A-Z Dictionary of Synonyms. Oxford, Clarendon Press, 1991.
Collins - W. Collins. The Woman in White. - England: Penguin Books, 1985.
Hailey – A. Hailey. Strong Medicine. - New York, 1986.
Sheldon – S. Sheldon. Memories of Midnight. – New York: Warner Books, 2005.
Steel – D. Steele. Wanderlust. – New York, 1989.

CHAPTER FIVE

A POLYSEMY ACCOUNT OF TURKISH SPATIAL NOUN 'ÜST' IN DATIVE CASE MARKER

AYSUN BALKAN
BOGAZICI UNIVERSITY, ISTANBUL

1. Introduction

In recent years, there has been growing interest in a cognitive semantic-oriented theory of polysemy to account for the multiple distinct meanings represented in a single lexical form. The traditional views including behaviouristic and generative approaches (cf. Bloomfield, 1933; Chomsky, 1995) provide highly arbitrary and idiosyncratic semantic accounts of the lexicon. However, more recent cognitive approaches argue that lexical structures are organized and represented in the mind not in a random fashion but in a highly systematic way (cf. Lakoff, 1987; Brugman, 1988; Kreitzer, 1997; in Tyler & Evans, 2003, 2004b). Following from this, a wide range of approaches such as homonymy and monosemy attempt to account for the multiple distinct senses of a single lexical form. Among all these approaches, the polysemy account has received extensive attention from various disciplines such as cognitive linguistics or psycholinguistics by using corpus analysis, which attempts to explore the polysemy of spatial forms in English and other languages (cf. Brugman, 1988; Vandeloise, 1991; Dewell, 1996; Türker, 2005; Liamkina, 2007).

The earliest accounts of the polysemy of the English spatial preposition 'over' date back to Brugman's findings (1988) and Lakoff's studies (1987), which constitute the pioneering works for cognitive semantic analyses of spatial forms across world languages. However, these early studies remain insufficient to explain systematicity in the meaning extensions of spatial particles since they often rely on researchers' intuitions rather than on a principled model. Most recently, Tyler and Evans (2001a, 2003, 2004b) have filled the gap in the literature by

providing a comprehensive account of a polysemy network for 15 English spatial prepositions and developing a systematic methodology to determine the primary sense and distinct senses within a model termed the Principled Polysemy model.

Taking Tyler and Evans' (2001a) study on 'over' as a starting point, this chapter aims to analyse the polysemous scope of the Turkish spatial noun 'üst'. More specifically, I will follow a methodology to analyse the polysemous structure of 'üst' and demonstrate how rich and wide the scope of its semantic network is compared to 'on' and 'over' in English. Before addressing the applicability of the Principled Polysemy model, which was originally proposed for the polysemous structures of English prepositions, to a spatial form of an unrelated language such as Turkish, it is important to understand the main tenets of this model. The following section provides a review of the literature on polysemy and the studies related to child language acquisition, lending support to the argument that human spatial cognition is shaped by the human perceptual systems and the experiential interactions with the spatial-physical world. Thus, human language emerges as a complex system reflecting the cognition, i.e. conceptual categories and representations, derived from the perceptual experience with the Trajector (figure of focus) and Landmark (figure of background) spatial configurations.

2. Literature Review

The theoretical basis known as the 'Principled Polysemy model' originated in the work by Tyler and Evans (2001a, 2003), which aimed to investigate how a given lexical form is processed and interpreted in multiple distinct senses. In Tyler and Evans (2003), the Principled Polysemy model was defined as a cognitive semantic theory in which the form of a lexical unit is connected to its meanings through a motivated mapping and a pragmatic strengthening in situated contexts.

The Principled Polysemy model assumes a connectionist learning mechanism in which extended senses are derived from the primary sense through a process of *pragmatic strengthening*, a term introduced by Traugott (1989). In this process, connections between the forms and their recurring uses are established in a principled manner. Thus, they form a polysemy network of related but distinct senses. In other words, additional meanings are associated with a particular lexical form through the repeated use of that form in situated contexts. Thus, the implicatures of situated contexts are conventionalized through routinization and entrenchment of

usage patterns, which give rise to new senses in a systematic way (Tyler & Evans, 2001a, p. 745).

A great number of studies on spatial markers in English have indicated the polysemous nature of lexical forms. The most prominent research on this comes from Tyler and Evans (2001a, 2003, 2004b), in which they specified fifteen extended meanings of the spatial preposition 'over' (e.g. covering, examining, reflexive, etc.), being stored in the long-term memory and originating from a primary spatial sense termed the 'protoscene'. Tyler and Evans (2001a) demonstrated that the conventionalized distinct senses can derive from another conventionalized sense such as 'focus-of-attention'←'examining', 'repetition'←'reflexive', or become part of a cluster such as ABC trajectory (on-the-other-side-of, above-and-excess, completion, transfer) and up (over-and-above←more, control, preference) clusters by arising from the another conventionalized sense rather than directly from the primary sense.

In an in-depth semantic analysis of 'over', Tyler and Evans (2001a) described the 'protoscene' as a foundational conceptual representation derived from the human perceptual experience with the spatial-physical world:

> At the conceptual level, the primary sense is represented in terms of abstracting away from specific spatial scenes, that is, real world scenarios resulting in an idealized spatio-functional configuration. We call this abstracted mental representation of the primary sense the PROTOSCENE (p. 735).

According to this model, the semantic construction of each preposition develops from a primary sense to a set of additional distinct senses through constrained cognitive principles and constitutes a motivated semantic network. Thus, the 'protoscene' consists of a conceptual configurational-functional relation mediating between the schematic trajector (TR), namely the focus element, and the schematic landmark (LM), namely the background figure. Thus, configuration elements such as TR and LM and functional elements embody a spatial scene which is perceived as an abstract representation of recurring spatial-physical configurations mediated by the human conceptual processing (Tyler & Evans, 2003: 51). The functional element refers to the interactive relationship between the TR and the LM in a particular spatial configuration (e.g. Vandeloise, 1991; in Tyler & Evans, 2003). For instance, a spatial relationship described by the spatial particles 'on' and 'in' designates two different relations: in the former, the TR is 'supported' by the LM, while in the latter it is 'contained'. Tyler and Evans (2003), in their analysis of 15 different

English prepositions, posited one functional element for each 'protoscene'. However, some researchers have suggested that more than one functional element can be associated with a prototypical sense. Thus, such an account leaves more room for other languages to be successfully analysed within the 'Principled Polysemy' model. For instance, Turkish spatial constructions are expressed by spatial nouns in combination with spatial case markers. The evidence suggests that the meaning of each spatial noun and each case marker involves its own functional element, so these Turkish constructions include more than one functional element.

Mandler (1988, 1992, 1996; in Tyler and Evans, 2001a) used the term *image-schema* to refer to the conceptual representation of spatial relations such as containment, support, proximity, motion, and so forth, which allows us to understand the functional aspects of spatial configurations (cf. Evans & Tyler, 2004a). Thus, within the capacity of spatial cognition, we can talk about space, refer to objects in space and express motions from a source to a goal through a medium (cf. Evans & Tyler, 2004c). In the spatial domain of conceptual structures, Talmy (1985) argued for the mapping of the spatial language to the elements of spatial cognition.

Similarly, Tversky (2003) suggested the reflection of spatial knowledge in linguistic habits and conventions. In other words, the outer world interactions lead to encoding spatial concepts and categorizing their mental representations in various ways. Therefore, in the conceptualization of spatial relations in a cross-linguistic setting, it is quite likely to draw polysemy networks of spatial constructions in different semantic patterns, different degrees of semantic extensions and different lexico-semantic structures. This also affects the way speakers perceive, conceptualize and express spatial relations in the outer world.[1]

Child language acquisition studies provide additional empirical support to the Principled Polysemy model. In a most recent study, Slobin (2008) examined the representations of transitive motion events inter- and intra-typologically between verb-framed (e.g. Turkish) and satellite-framed (e.g. English) languages to determine the role of linguistic typology in the children's conceptualization of transitive motion events. In this study,

[1] Levinson (1996; 2003) suggested three different frames to encode spatial relations: *absolute frame*, which is free from speech participants and TR-LM scenes such as north, south, center, etc.; *intrinsic frame*, which is based on the intrinsic properties of an LM such as back, front, top, etc.; and *relative frame*, which is based on speech participants' points of view such as right, left, etc. Levinson reported that the relative frame is the most commonly used frame by speakers of Turkish, along with the speakers of other languages such as Dutch and Japanese.

unlike children learning English spatial markers, infants learning Turkish were found to ignore the *containment* and *support* distinction when learning the spatial case markers. Instead, Turkish children attended the *static* versus *dynamic character* of spatial references. This stems from different conceptualizations of spatiality across speakers of different languages from very early ages. For instance, when describing the action of *putting a pencil on a table,* a Turkish speaker does not always need to express a support relationship between the TR (pencil) and the LM (table) through the overt use of the spatial noun *üst* 'on top of', unless it has some contrastive or emphatic function. The listener can infer the goal-oriented spatial relation from the dative case –(y)E attached to the LM. Thus, '*put table-Dat*' is interpreted as *'put on table'* thanks to the dative case –(y)E indicating a Proto-Goal Sense. In my analysis of the lexeme 'üst', I will refer to the spatial relations prompted by combinations of linguistic elements such as spatial case markers and verbs.

The strongest evidence about the effect of typological distinction in concept formation comes from a series of studies by Choi and Bowerman (1991) and Bowerman and Choi (2001, 2003). They found that 9-month old English and Korean children could distinguish spatial relationships between tight fit, described by the verb *kkita*, versus loose fit, described by the verb *nehta* in Korean; and *containment* versus *support* described in English by the prepositions *in* versus *on* respectively. However, such cross-linguistic sensitivity to spatial relationships was no longer available to infants of both languages by the age of 18 months (cf. Slobin, 2008; Jarvis and Pavlenko, 2008). These findings shed further light on the development of language-specific spatial categorizations and representations in the human mind through active perceptual interaction with the spatio-physical world where attention is directed by language.

Language acquisition studies provide evidence for human beings' active interaction with the real-world force dynamics and various spatio-physical configurations from early stages of life. Their findings suggest that conceptual categories and representations of spatial forms derive from the interaction with the physical world through human perceptual systems in conjunction with the language they speak rather than pre-existing spatial representations in the human cognitive system. More importantly, these studies lend support to the argument that languages differ in their structuring of conceptual domains, although the human experience with the physical world is the same (cf. Slobin, 1994, 2001). Accordingly, the way each language highlights certain aspects of the spatial-physical world differs. In such an interaction, language plays an important role by focusing the speaker's attention on different aspects of the spatial scene.

Therefore, each language is expected to present a unique polysemy network arising from the interactions of different linguistic and functional elements in prompting for spatio-geometric configurations. As a result, all human languages seem to differ in the exact ways they encode spatiality.

3. Objective of Study

The first dictionary entry of 'üst' as a spatial noun means 'top', which is followed by these entries: 1. Outside surface; upper surface; 2. Over, above; 3. Clothing; 4. Superior; 5. Body; 6. (of money) Remainder. 'Üst' is also used as an adjective meaning upper, uppermost, superior (cf. Golden Dictionary: Turkish-English). However, none of these descriptive lists of the meaning of 'üst' present any motivated semantic analysis nor indicate the prototypical sense of 'üst', its extended senses and how all these related senses are conventionalized as distinct senses in the semantic memory.

The preliminary analysis of 'üst' in Turkish indicated that it encodes both contact and non-contact situations between a TR and an LM depending on its spatial case suffix, context of usage, and the verb[2] in a given sentence, which will be explained in detail in Section 4.1. Despite the fact that Turkish and English are typologically distinct languages and express spatial relations using very different linguistic elements, the analysis showed a surprising amount of overlapping with the extended meanings. Beyond its protoscene, 'üst' seems to overlap with nine distinct senses associated with 'over' in English: 'above-and-beyond (excess I), more, transfer, covering, focus-of-attention, control, preference, temporal, reflexive' in the polysemy network of the English preposition 'over' developed by Tyler and Evans (2001a). However, an in-depth semantic analysis of 'üst' through a corpus analysis is expected to show the Turkish spatial noun 'üst' with a number of additional senses not found in the English polysemy network of 'over' and a different pattern of semantic organization.

In this corpus study, I will explore further semantic extensions of the constructional polysemy of 'üst + Dat' using a methodology which follows from the 'Principled Polysemy' model by Tyler and Evans (2001, 2003). Earlier descriptive accounts of the Turkish spatial noun 'üst' failed to provide a semantic model explaining the systematic motivation of meaning extensions from its central meaning (cf. Kornfilt, 1997; Lewis,

[2] In Turkish, spatial case markers are selected by verbs.

2000; Göksel & Kerslake, 2005). Therefore, these studies display an arbitrary account of spatial nouns.

The present study will provide a detailed polysemous account of concept-form mappings for the Turkish spatial noun 'üst' and its paired spatial particle 'Proto-Goal Dative', which has not been studied so far. It will also indicate the degree of semantic polysemy displayed in the Turkish spatial noun 'üst' in comparison with the English spatial prepositions 'on' and 'over' from a cross-linguistic perspective. Overall, the results will shed light on whether the polysemy model developed for English spatial prepositions is applicable to the spatial nouns of a historically unrelated language such as Turkish.

4. Research Hypothesis and Questions

Following from Tyler and Evans (2001a) and Evans (2010), my hypothesis is that the Turkish spatial lexeme 'üst' in combination with the spatial particle 'Proto-Goal Dative' presents a highly polysemous structure with distinct non-spatial uses in a similar way to the English spatial lexemes 'on' and 'over'. However, considering the typological distance between Turkish and English, a polysemy network of 'üst' offers a different organizational pattern and depth (variety) from the polysemy networks drawn for 'on' and 'over'. In addition to the functional element of proximity which can subsume both control and support, I predict that the static versus dynamic scene will play a significant role in determining the spatial relations realized through spatial case suffixes and the context of usage in Turkish. Accordingly, the current study investigates two main questions:

1. Is the polysemy model developed for the spatial prepositions in English applicable to other unrelated languages such as Turkish, which has different linguistic resources to express spatiality?

2. If applicable, how rich is the semantic polysemy of the spatial noun 'üst + Dat' in Turkish in comparison with the spatial prepositions 'on' and 'over' in English?

4.1. Typological Sketch of Turkish

Historically, Turkish is a Turkic language, closely related to Uzbek, Turkmen and Azeri among other languages, while English is a Germanic language, closely related to German, Danish, Dutch and Swedish.

Therefore, a typological sketch of Turkish outlined in a comparative manner to English ensures a thorough understanding of the present study.

Turkish, as an agglutinative language, has six nominal case suffixes, which establish grammatical relations in the sentence: agent/subject (nominative-null case), definiteness/direct object (accusative), spatiality (dative, locative, ablative) and possession (genitive) (cf. Göksel & Kerkale, 2005; Lewis, 2000; Kornfilt, 1997).

Table 1. Turkish Nominal Case Markers

Turkish Case System	
Nominative	Ø
Accusative	-I
Dative	-(y)E
Locative	-dE
Ablative	-dEn
Genitive	-In

With its rich morphological structure, Turkish is a verb-framed language (Talmy, 1985, 1991), in which path of motion is preferably encoded in the main verb (Özçalışkan and Slobin, 1999, 2003). As an SOV language, it also has a postpositional system, where spatiality is expressed through spatial nouns paired with 'Dative, Locative, Ablative' case markers assigned by verbs. On the other hand, English has only the nominative and genitive cases for nouns, and the nominative, accusative and genitive cases for pronouns owing to its rich prepositional system to express spatial relationships.

The Turkish spatial case system displays a three-way distinction for *goal*, *location* and *source* expressed through dative [–(y)E], locative [–dE] and ablative [–dEn] particles respectively, with a combination of spatial nouns which establish (non-)spatial relationships between a TR and an LM. However, as indicated in Türker (2005), each spatial particle displays multiple distinct senses in addition to their Proto-Goal, Proto-Location and Proto-Source senses. In her thorough analysis of spatial particles, Türker demonstrated the systematic and principled meaning extensions of the prototypical senses associated with each case marker within the polysemy networks. She attributed the polysemous diversity of the particles to their grammaticalization. In other words, the frequency and variety of grammatical constructions associated with each particle determine the diversity of distinct senses derived from their Proto-senses. For example:

(a) Kitab-ı masa-nın üst-ü-**ne** koy-du-m. (Proto-Goal Sense)
*Book-ACC table-gen top-poss3sg-**DAT** put-past-1sg.*
*I put the book on top of / on**to** the table.*

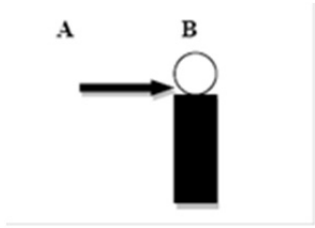


Figure 1. I put the book on top of / on**to** the table.

(b) Kitap masa-nın üst-ün-de. (Proto-Location Sense)
Book table-gen top-poss3sg-LOC3sg.
The book is on / on top of the table.

Figure 2. The book is **on** / on top of the table

(c) Kitab-ı masa-nın üst-ün-den al-dı-m. (Proto-Source Sense)
Book-ACC table-gen top-poss3sg-ABL take-past-1sg.
I took the book from top of the table.

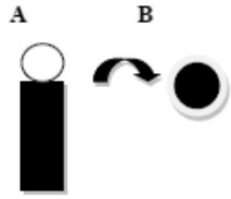


Figure 3. I took the book **from** top of the table.

The example (b) illustrates the spatial noun *üst* 'top/on' indicating a static spatial relation because of the locative case marker [-dE], while (a) and (c) illustrate *üst* indicating a dynamic spatial relation through the direction to a goal and the departure from a source due to the dative [-

(y)E] and ablative [-dEn] case markers respectively. These indicate that the spatial case markers encode different spatial relations, conceptual representations and mental imagery derived from the perceptual experience with the spatio-physical world. Accordingly, the conceptual distinction between Turkish spatial case markers is pertinent to the static versus dynamic parameters as functional elements. The following table displays the grammaticized spatial case system in Turkish and the spatial prepositions in English.

Table 2. Grammaticized Spatial Systems of Turkish and English

Turkish Spatial Case Suffixes				English Spatial Prepositions			
Category	**Goal**	**Location**	**Source**	**Category**	**Goal**	**Location**	**Source**
Dynamic	-(y)E		-dEn	**Containment**	into	in	out of
Static		-dE		**Support**	onto	on	of
				Direction	to	at	from

In Turkish, spatial relations are prompted by a combination of spatial nouns and nominal case markers assigned by verbs, unlike the English spatial system. For instance, *üst* + Dative (DAT) and *üst* + Ablative (ABL) with the verb 'atla-mak' meaning 'jump' illustrate a contact situation and require a dynamic spatial relationship between the TR and the LM in the form of a motion directed to a goal and a motion originated from a source in a downward direction respectively. However, the same verb does not allow a static spatial relationship with *üst* + Locative (LOC). In an upward direction of the TR, *üst* + ABL indicates a non-contact situation with an *above-across* sense through the completion of the motion 'jump' as in the *ABC trajectory cluster sense* drawn in Tyler and Evans (2001a). Thus, we infer that, despite being interrelated, these constructions are independent and associated with different spatial configurations and hence with different semantic interpretations in line with Goldberg's 'Constructional Approach' (1995), where a constraint is associated with the construction as a whole rather than being lexically governed. In the case of a constructional polysemy, various scenes that are not predictable from context, or from lexical or morphological items alone, are conventionally associated with a construction.

As a result, the analyses of the spatial constructions 'üst + Proto-Goal DAT', 'üst + Proto-Location LOC' and 'üst + Proto-Source ABL' within the Principled Polysemy model suggest a distinct family of extended senses within the same polysemy network originating from the protoscene of 'üst'. This allows us to account for differences and capture the relations between senses in a natural way without having to posit a collection of lexical rules pertaining to a spatial case marker.

5. Methodology

This study examines the distinctive spatio-physical configurations of the Turkish spatial noun 'üst + Dat' to form a polysemy network with a cognitive usage-based approach, using a similar methodology to that of Tyler and Evan (2001a) in the analysis of 'over'. From the Principled Polysemy model of the English spatial word 'over' proposed by Tyler and Evans (2001a, 2003, 2004b), I argue that spatial nouns in Turkish display a highly polysemous structure and thus form a polysemy network centred on a protoscene. I will draw on a corpus analysis to help determine the extended distinct senses of 'üst + Dat' in addition to its primary spatial sense.

6. Data

The data came from the METU Turkish corpus, which is a collection of 2 million words of post-1990 written Turkish samples in ten different genres: memory, narrative, research, essay, travel log, diary, news, newspaper column, article, story, novel and interview. Each written sample consists of 2000 words. The Boolean query of the lexeme 'üst' gave rise to 2200 tokens and 44 types at the paragraph level discourse, of which 'üst + Dat' occurred in 483/2200 tokens and 6/44 types in different lexical forms presenting different grammatical relations through the combinations of person, number, case and relational suffixes.

7. Analysis

In an in-depth semantic analysis, I will examine the representative types of corpora to capture the protosense of 'üst + Dat' and its extended distinct senses. Thus, the category structures and frequency of the use types for 'üst + Dat' will be analysed through 483 sentences overall. Throughout the analysis, a representative sentence will be included for the primary and extended senses along with a diagrammatic representation. In addition to the functional elements of proximity, control and support, static versus dynamic scenes will also be taken into account in the polysemy analysis of 'üst + Dat'.

a. Primary Sense

First, I will determine the protoscene of 'üst + Dat' which gives rise to extended distinct senses in a polysemy network based on the five criteria suggested by Tyler and Evans (2001a, 2003):

1. Diachronic evidence (historically the earliest meaning).
2. Predominance in the network (the most frequent spatial configuration in other distinct senses).
3. Use in composite forms (dividing the space in particular ways through spatial particles, which partially co-determine each other's meaning- e.g. Turkish 'üst' (on/over) and 'alt' (under) in the vertical dimension, and the TR 'higher-than-the-LM' vs. 'lower-than-the-LM' reading respectively).
4. Relation to other prepositions (contrasting with other members of the compositional set based on its distinctive spatio-physical configuration).
5. Grammatical predictions (predictability of additional senses directly from the primary senses or even if derived from another distinct sense, and the traceability of that other sense to the primary sense).

b. Extended Distinct Senses

Next, I will determine the additional distinct senses derived from the primary sense and instantiated in the semantic memory by applying two assessment criteria suggested by Tyler and Evans (2001a). For a sense to count as distinct:

1. First, it must contain additional meaning in a different spatial or non-spatial configuration not apparent in any other senses.
2. Second, it must derive from a primary sense or another extended sense through pragmatic strengthening. In other words, it must have a conventionalized meaning, which is not inferable from another sense or the context of occurrence.

8. Results

The protoscene of 'Üst': Analysis

1. Diachronic evidence: Turkish is from the Altaic branch of the Ural-Altaic language family of Turkic languages. Etymologically, 'üst' is an original Turkish word **'üŕ-t'** and a common Turkic derivative with secondary vowel shortening: in Old Turkish 'üst', in Turkmen 'üst', in Tatar **'ös'**, in Khal **'ist'**, etc. meaning 'top, upper part'. Its Proto-Turkic form is **'ǖŕ (/*ǒŕ)'** in Altaic etymology, which means 'on top, high, above'. Based on this diachronic evidence, the TR-being-higher-than-the-LM reading of 'üst' seems to be historically the earliest meaning.

2. Predominance in the network: The corpus analysis indicates that the sense of the TR being higher than and proximal to the LM is the most frequent spatial configuration with 18% in relation to other 21 distinct senses in the network. However, 18% only reflects the frequency of the primary sense occurrence in 'üst + DAT', since the current study addresses the interaction of 'üst' with the dative case particle [-(y)E] involving a dynamic Proto-Goal sense. Based on this, the sense that the TR is located 'higher-than-the-LM' predominates in the network and thus constitutes the primary sense.

3. Use in composite forms: This criterion postulates that spatial particles divide the space in particular ways, which partially co-determines each other's meaning. As in German and English, the compositional set dividing the vertical dimension in Turkish includes 'üst (on/over)' and 'alt (under/underneath)', and the sense distinguishing 'üst' from 'alt' is the TR being 'higher-than-the-LM' reading as opposed to 'lower-than-the-LM'. In addition, depending on the spatial case of construction, the TR's movement refers to a *static* or *dynamic* higher-than-the-LM sense. For instance, a fly's action of 'flying' can take place in a *static* manner *'over/under'* the lamp owing to the Proto-Location sense denoted by 'üst/alt + LOC' respectively. In contrast, with the Proto-Goal sense denoted by 'üst/alt + DAT', flying can be in a *dynamic* manner *'onto/to the underneath of'* the lamp. However, in the former the lamp itself constitutes the targeted LM, while in the latter it is the space underneath the lamp (see German *über* vs. *unter* in Liamkina, 2007).

4. Relation to other prepositions: This criterion postulates that the primary sense of 'üst' contrasts with other members of the compositional set such as 'alt', based on its distinctive higher-than-the-LM spatio-physical configuration. For instance, a cat's action of 'climbing' '*onto/to the top of*' a wall in a spatial configuration through 'üst + DAT' denotes

the TR being higher-than-the-LM, while the TR being 'lower-than-the-LM' constitutes a semantically ill-formed spatial configuration through 'alt + DAT' since the act of climbing requires a higher-than-the-LM reading. Thus, we can infer that each spatial particle divides the space in particular ways. The particles of the same compositional set such as 'üst' and 'alt' of the vertical dimension are distinguished in their primary sense and the semantic constraint is assigned by the verb as in the examples.

5. Grammatical predictions: This criterion postulates that the additional senses need to be directly derived from the primary sense or, if derived from another distinct sense, that other sense needs to be traceable to the primary sense. As explained and illustrated in my analysis of extended senses, the higher-than-the-LM reading of 'üst + Dat' construction has given rise to three clusters of senses (Up, Forward, and Part-Whole) besides several distinct senses. For instance, the Additive Sense giving rise to the Successive Sense directly derives from the primary sense. In a similar way, all distinct senses in the network connect each other through the higher-than-the-LM reading of the primary sense irrespective of their cluster and conventional meaning.

The following diagram adopted from Tyler & Evans' (2001a: 736) analysis of 'over' displays the protoscene analysis of the Turkish spatial noun 'üst' in accordance with the principles listed above.

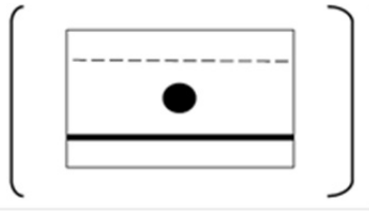


Figure 4. Protoscene of Turkish Spatial Noun 'üst'
(dark circle: TR; bold line: the LM in a bold line; dashed line: proximal relation of TR to LM)

As illustrated in Figure 4, the spatial configuration of 'üst' in its primary sense involves the TR being in a proximity relation to the surface of the LM and located higher than the LM. Thus, the TR and the LM are conceptualized within each other's sphere of influence in the vertical dimension.

Example: Yaĝmur bulut-lar-ı köy-ün tam üst-ün-DE.
Rain cloud-pl-poss3sg village-gen precisely top-poss3sg-LOC3sg.
(Lit: The rain clouds are precisely over the village.)
The rain clouds are precisely over the village.

In this scene, the TR 'rain clouds' are located above but proximal to the LM 'village'. Thus, both configuration elements are within each other's sphere of influence. For instance, the influence of rain clouds can be experienced in forms such as the blockage of the sun, having a grey sky, rain drops and so on. From this example, we infer that the spatial noun 'üst' denotes a TR-LM configuration where both elements are within each other's sphere of influence to varying extents.

Based on the protoscene analysis of 'üst', I predict a semantic network of three different constructions: PROTO-Goal, PROTO-Location and PROTO-Source senses, of which the Proto-Goal Sense of 'üst + Dat' is the main concern of this study. As proposed by Goldberg (1995: 31), constructions are associated with a family of closely related senses rather than a single, fixed abstract sense in a network of constructional polysemy. In their common primary sense, the TR is conceptualized as being higher than the LM. Under certain circumstances, the spatial relation designated by 'üst' involves the TR being in a relation of contact with the LM; other instances do not involve contact. Thus, the functional element proximity reflects the interactive relationship between the focus and the background elements.

The Spatial Construction of 'Üst + Dative'

The following diagram displays the construction of 'üst + Dat', in which the dative spatial particle adds the *Proto-Goal Sense* and the functional element of *dynamic scene* to the TR-LM spatial configurations. In this configuration, the LM serves as the goal of the action.

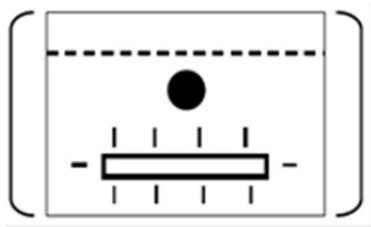


Figure 5. PROTO-Goal Sense of Turkish 'üst + DATIVE'
(dark circle: TR; bold line: the LM in a bold line; dashed line: proximal relation of TR to LM; clear arrow: dynamic scene of action directed at the LM, namely the goal)

Based on the account of Construction Grammar (Goldberg, 1995: 4), a construction is regarded as distinct if one or more of its properties are not strictly predictable from the knowledge of other constructions in the

grammar. Thus, various constructions associated with a family of distinct but related senses may give rise to subtle semantic constraints in their constructional polysemy such as the functional element of *dynamic scene* as in 'üst + Dat'.

RESULTS FOR 'ÜST + PROTO-GOAL DAT'

1. PRIMARY SENSE of 'Üst + Proto-Goal DAT' (88 Tokens/483)

In 'üst + Dat', the TR is conceptualized as being higher than the LM with a *forward trajectory* directed at the goal, namely the schematic LM as the locator, which makes the conceptual spatial relations between a TR and an LM dynamic. The functional element of *dynamic scene* is denoted by the Proto-Goal Sense of the dative particle [(y)E]. When the dative particle is attached to a spatial noun, the GOAL sense denoted by this marker combines with the PROTO sense (Türker, 2005)[3]. In the corpus analysed, 88 sentential tokens out of 483 were categorized as the primary sense of the 'üst + DAT' construction. The protoscene involves both '*proximity* and *static versus dynamic scene*' as functional parameters. Thus, both primary and extended distinct senses in the 'üst + Dat'[4] polysemy network are realized in a *dynamic* relation with the LM through the Proto-Goal Sense denoting a *forward direction* and thus a *dynamic scene*.

(a) El-ler-i-ni masa-nın **üst-ü-ne** koy-du. (Proto-Goal Sense)
*Hand-pl-poss3sg-ACC table-gen top-poss3sg-**DAT** put-past-3sg.*
*(Lit: (S)he put his/her hands on top of / on**to** the table.)*
*(S)he put his/her hands on / on**to** the table.*

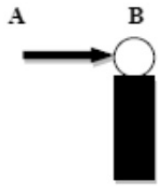


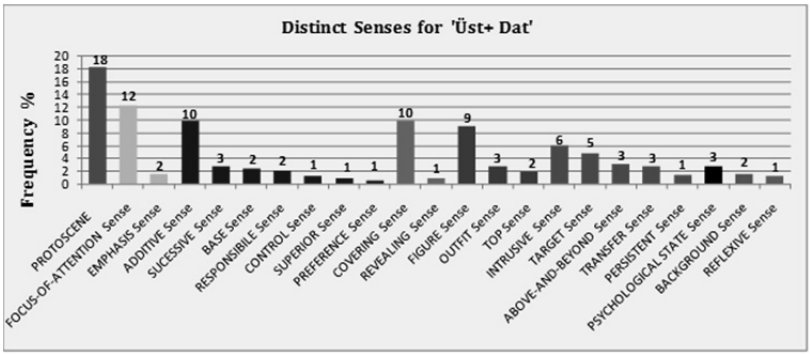
Figure 6. (S)he put his/her hands on top of / onto the table.

[3] For a thorough cognitive semantic analysis of the Dative particle [-(y)E] with other spatial case markers (Locative [-dE] and Ablative [-dEn] particles), see the polysemous semantic network in Türker (2005).
[4] Throughout the analysis of the extended distinct senses, the examples discussed and listed under the figures are given as the literal translations of the Turkish sentences for a better understanding of the conceptual representations.

EXTENDED SENSES of 'Üst + Proto-Goal DAT'

The analysis of the METU corpus indicated 22 distinct senses for 'üst + DAT' based on 483 sentential tokens. The additional distinct senses are either non-spatial in meaning or in a different spatial configuration, which derive context-independently from a primary sense or another extended sense through a number of cognitive processes such as schematization, metaphor and metonymy. With the dative case, the verb denotes varying degrees and extents of *directive actions* through a predominantly forward trajectory where the LM functions as the locator and the goal of the action. Table 3 displays the frequency percentages of the 22 extended distinct senses[5] for the 'üst + DAT' construction derived from a primary sense.

Table 3. Frequency Percentages of the Protoscene and Distinct Senses for 'Üst + DAT'



The table is organized based on the frequency occurrences of distinct senses grouped under a cluster. Accordingly, the senses within a cluster and the clusters within the network are sequenced from the most frequent to the least frequent: Protoscene 18%, Focus-of-Attention 12%, Additive 10%, Covering 10%, Figure 9%, Intrusive 6%, Target 5%, Successive 3%, Outfit 3%, Above-and-Beyond 3%, Transfer 3%, Psychological State 3%, Base 2%, Responsible 2%, Emphasis 2%, Background 2%, Top 2%, Control 1 %, Persistent 1%, Superior 1%, Preference 1%, Reflexive 1%, Revealing 1%.

A set of extended senses form complex clusters such as Up Cluster (Additive, Successive, Base, Responsible, Superior, Control, Responsible,

[5] The bars in the same colour indicate the extended senses grouped in the same cluster due to semantically related but distinct meanings.

Preference senses), Forward Cluster (Intrusive, Target, Above-and-Beyond, Transfer, Persistent senses) and Surface Cluster (Figure, Top, Outfit senses). In some instances, a distinct sense arises from the conceptualization prompted by another distinct sense rather than directly from the protoscene, such as Emphasis Sense from Focus-of Attention Sense, Revealing Sense from Covering Sense, Target Sense from Intrusive Sense, Responsible Sense from Control Sense, and Control Sense from Superior Sense. As suggested by Tyler and Evans (2003: 79), this type of conceptual construction stems from the reanalysis of spatial scenes recursively; thus, the emergence of a distinct sense results from the multiple instances of reanalysis.

The following figure displays the constructional polysemy analysis of the Turkish 'üst + Dat' construction involving *Proto-Goal Sense*, where the TR serves as the focal object and the LM is interpreted as the goal.

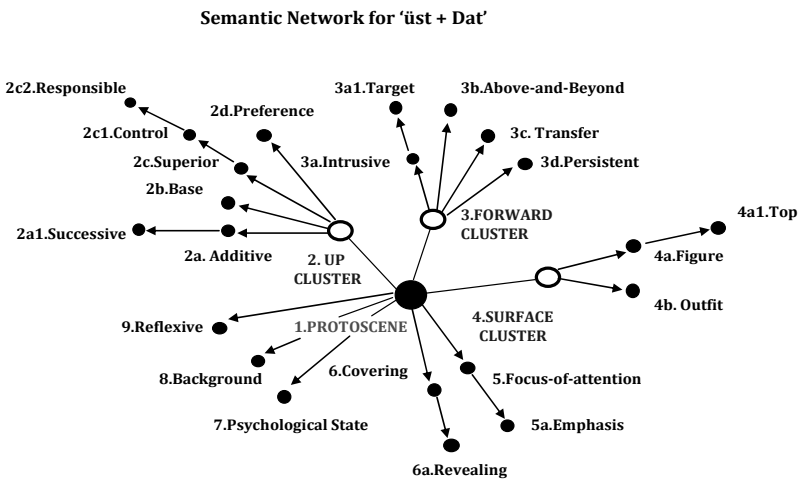


Figure 7. The Semantic Network of 'üst + Dat' within the Principled Polysemy' Model

The above figure is a display of the analysis of the extended senses. It indicates the proposed semantic network for 'üst + DAT', suggesting a total of 22 extended distinct senses organized around the protoscene. Distinct senses are represented by a dark node in the network, while the clusters by a clear node. The protoscene is placed in the central position to indicate its status as the primary sense from which the other senses ultimately derived. As also illustrated in the figure, the frequency occurrences of distinct senses positively correlate with their proximity to

the protoscene when being organized in the semantic network. A distinct sense immediately deriving from the conceptualization prompted by another conventionalized distinct sense is represented with a dark node connecting the two senses. For instance, the Revealing Sense is represented in 6a to indicate that it arises directly from the conceptualization associated with the Covering Sense in 6 rather than the protoscene. Thus, as claimed by Tyler and Evans (2003), a distinct sense can be derived from multiple instances of reanalysis; this suggests that the reanalysis of conceptualizations is potentially recursive. Moreover, a complex conceptualization such as 'Up Cluster' in 2 can give rise to several distinct senses as a result of multiple instances of reanalysis.

2. UP CLUSTER

Up Cluster involves seven distinct senses (i.e. Additive, Successive, Base, Control, Responsible, Superior and Preference). Each sense arises from construing a TR located physically higher than the LM as being vertically elevated or up relative to the LM. In this configuration, the TR is assigned to upward orientation, as illustrated in Figure 8. Accordingly, in Up Cluster the TR is vertically elevated relative to the LM.

Figure 8. Up Cluster
(dark bar: LM in upright position; dark circle: TR in a vertically elevated position; arrow: upward orientation – adopted from Tyler & Evans, 2003: 96)

The real-world experiences associated with the conceptual spatial relation of 'üst + DAT' often result in the construal of an upward orientation where the TR is vertically elevated. In other words, the TR moves from a physically lower to a higher position than the LM's. 'Up' is also associated with states such as being positive, more, superior, favourable, etc. in real-world experiences, which are not entailed in the construal of the protoscene, that is, the TR is higher than the LM. For instance, in the scene described by the sentence Kitab-ı masa-nın **üst-ün-e** koy-du-m "*I put the book on top of / onto the table"*, the book is not construed as being in a superior, more, positive, or favourable position

than the table. The relation denoted in such a spatial configuration is the TR being located higher than the LM.

2a. ADDITIVE Sense (48 Tokens/483)

Additive Sense is a non-spatial use of the spatial noun 'üst + DAT' in Up Cluster and translates into 'on top of it, in addition to it'. The implicature of addition derives in the following way: vertically up implicates greater amount in experiential terms, while physically down implicates less amount. When entities are added onto each other, this experiential correlation often results in a state of vertical elevation. Thus, the scene is conceptualized as adding a physical or abstract entity to another physical or abstract entity that is interpreted as the goal by experiencing an additive process through the upward trajectory. In real-world experiences 'up' is associated with positive states such as greater amount, superiority or favourability, which are not entailed in the construal of the protoscene, that is, the TR is higher than the LM. Thus, the natural consequence of addition is privileged and reanalysed as distinct from the spatial configuration of the protoscene from which it originates, as displayed in Figure 9. In the scene Ana yemek-te ızgara somon gel-di. **Üst-ü-ne** de tiramisu ye-di-k. "*In the main course, grilled salmon came. On / on top of it, we also ate tiramisu"*, 'the tiramisu' is construed as the TR and 'the grilled salmon' as the LM. Given our general knowledge about meals, following the main course, dessert is served as an additional variety of food. The interpretation of the 'üst + Dat' construction in this context is 'in addition to'. Thus, the implicature of addition ties in the metaphor MORE IS UP. It is conventionalized as a distinct sense through pragmatic strengthening and instantiated in the semantic memory.

Example:
Ana yemek-te ızgara somon gel-di. Üst-ü-ne de tiramisu ye-di-k.
Main course-LOC grilled salmon come-past-3sg. Top-poss3sg-DAT also tiramisu eat-past-1pl.
(Lit: In the main course, grilled salmon came. On / on top of it, we also ate tiramisu.)
In the main course, grilled salmon was served. In addition to it, we ate tiramisu.

Figure 9. Additive Sense
(vertical arrow: addition; dark circle: LM; transparent circle: TR; transparent circle in dashes: the iteration of the adding action with further TR-LM configurations)

2a$_1$. SUCCESSIVE Sense (13 Tokens/483)

Successive Sense is an extension from Additive Sense and translates into 'one over another'. Succession is another experiential correlate associated with vertical elevation. The implicature of succession derives in the following way: vertically up is associated with greater quantity in experiential terms, while physically down is associated with less quantity. The sentence "Gece boyunca mesaj **üst-ü-ne** mesaj gel-di", "*All night long, message over message came*" does not illustrate a spatial scene, in which the TR 'message' is higher than but within proximal reach of the LM 'another message'. Rather, the experientially driven implicature is that one message succeeds another with a minimal temporal gap. Thus, the messages following one another accumulate, which results in a state of vertical elevation but in a successive manner, as diagrammed in Figure 10. The LM is the earlier message which is construed as the goal, to which the TR is the new message which is directed in a transferring process through the upward trajectory. Owing to the minimal temporal gap, the succeeding entities also become proximal to each other. This suggests that, in experiential terms, there are two elements associated with the concept of succession: vertically up and proximal. In other words, the linguistic usage of 'üst + Dat' accords with how we actually experience succession. Similarly, Successive Sense correlates with the metaphor MORE IS UP. Thus, the implicature of succession is conventionalized as a distinct sense through pragmatic strengthening, differing from the spatial 'higher than' reading.

Example:
Gece boyunca mesaj **üst-ü-ne** mesaj gel-di.
*Night long message top-poss3sg-**DAT** message come-past-3sg.*
(Lit: All night long, message over message came.)
All night long, one message over another came.

Figure 10. Successive Sense
(vertical arrow: succession with minimal temporal gap; dark circle: LM; transparent circle: TR; transparent circle in dashes: the iteration of the successive action with further TR-LM configurations)

2b. BASE Sense (12 Tokens/483)

Base Sense takes place in Up Cluster and translates into 'lowermost part of any construction, the bottom support of anything on which things stand'. In this scene, the TR is construed as the receiver of the support and as a physical or non-physical entity rising over the LM, which is construed as the provider of the physical or non-physical bottom support, i.e. the goal for the TR's directed motion. The presumed intention is to have the TR come into contact with the LM, being perceived as a target base. In the example "Yani sen bütün yaşa-dık-lar-ımız-ı kendi dünya-n-da yalan-lar-ın **üstü-ne** mi inşa et-ti-n?", "*That is, did you build all the things that we lived on / on top of lies in your own world?"*, the TR '*all the things that we lived*' is built on and rises up from the LM '*lies*', as illustrated in Figure 11. The conventional interpretation of this sentence is 'building all the things that we lived on lies'. Here, the TR movement prompts for a base sense at the conceptual level by abstracting away from the spatial components of the protoscene, namely vertical elevation and proximity. However, in the experiential sense, the implicature of Base Sense associated with 'üst + Dat' correlates with vertical elevation and physical proximity. Accordingly, Base Sense derives in the following way: being vertically up and proximal to the LM in a construction situation allows the TR to build on the LM. In this configuration, the TR being in contact with the LM receives base support and moves in an upward trajectory. Vertically up is associated with greater amount and an erect position, which is privileged relative to physically down associated with lower amount and a horizontal position with respect to the support received and the emergence of a construction. Owing to this experiential correlation, the implicature of Base Sense is conventionalized and instantiated in the semantic memory as a distinct sense.

Example:
Yani sen bütün yaşa-dık-lar-ımız-ı kendi dünya-n-da yalan-lar-ın üstü-ne
That is you all live-ger-pl-poss1pl-ACC your own world-poss2sg-LOC lie-pl-gen top-poss3sg-DAT
mi inşa et-ti-n?
yes/no Q word build-past-2sg
(Lit: That is, did you build all the things that we lived on / on top of lies in your own world?)
That is, did you build all the things that we lived on lies in your own world?

Figure 11. Base Sense

2c. SUPERIOR Sense (4 Tokens/483)

Superior Sense takes place in Up Cluster and translates into 'being in a higher rank, degree, or importance; being better than or above others'. In this configuration, the TR being in a higher position exerts superior performance over the LM as an instance of physical or abstract movement through the upward trajectory. In this scene, the focal object, TR, is construed as an entity of higher rank or better merit than the background object, LM. The TR outperforms the LM owing to its superior characteristics, as illustrated in Figure 12. In the example "Yani bir görün-me-yen kadın. Buharlaş-ma-da **üst-ü-ne** yok", "*That is, she is an invisible woman. There is no one else over her in evaporating",* the implicature is that the TR '*the woman*' displays a better performance than the LM '*no one else*' in evaporating, which is a metaphor for disappearing. The LM constitutes the targeted background object and thus the goal of the TR's motion. The TR movement prompts for a 'superior sense' at the conceptual level by abstracting away from the actual spatial scene of the protoscene. In this configuration, the focus is on the resultant state of the TR movement rather than on a dynamic scene. Superior Sense derives in the following way: an entity being vertically up is also superior to an entity physically down owing to its better performance and thus to a higher rank or degree in experiential terms. Given our real-world experience, the physically big and up often exert power and influence over the physically down and small (Tyler & Evans, 2003: 101). For instance, in the army the

relationships are based on soldiers' superior versus inferior ranks, which are assigned according to their physical (e.g. strength or performance during military training) or non-physical (e.g. education level or experience in the army) merits. Accordingly, physically strong and better-trained soldiers display better performance in a combat and thus are considered to be superior to the less strong and undertrained soldiers. Such experiential correlation between vertically up and the linguistic usage of 'üst + Dat' results in the implicature of Superior Sense in a non-spatial configuration which cannot be inferred from context. Thus, it is conventionalized through pragmatic strengthening and instantiated in the semantic memory as a distinct sense in Up Cluster extended directly from the protoscene.

Example:
Yani bir görün-me-yen kadın. Buharlaş-ma-da **üst-ü-ne** yok.
*That is a/an appear-neg-ger woman. Evaporate-ger-LOC top-poss3sg-**DAT** there isn't.*
(Lit: That is, she is an invisible woman. There is no one else over her in evaporating.)
That is, an invisible woman. There is no one else any better at disappearing than she is.

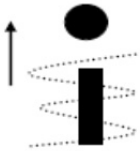


Figure 12. Superior Sense

2c$_1$. CONTROL Sense (6 Tokens/483)

Control Sense is an extension from Superior Sense in Up Cluster and translates into 'directing influence, exerting power', which is a similar finding to Tyler & Evans' (2001a, 2003) analysis of 'over'. In this scene, the TR is construed as the entity exerting power, while the LM as the entity being controlled and influenced, as displayed in Figure 13. In the example "*We wish Hadgi father would die and we would sit **onto** / on top of the property",* the figurative use of 'sit onto/on top of the property' refers to 'taking over the property; taking control of the property'. Here, the implicature is that the TR '*we*' exercises power over the LM '*Hadgi father's property*' and thus controls it. Such conventional interpretation

cannot be inferred from context, since the TR movement prompts for a control sense at the conceptual level by abstracting away from the spatial components of the protoscene, namely vertical elevation and proximity. However, in the experiential sense, the implicature of Control Sense associated with 'üst + Dat' correlates with the vertical elevation and physical proximity. Control Sense derives in the following way: being vertically up and proximal to the LM allows the TR to exert control and direct influence over the LM. On the other hand, being physically down and far from the LM is associated with absence of power and control (Tyler & Evans, 2003: 101). For instance, in wrestling, the wrestler being physically higher than and proximal to the other wrestler is associated with power and control. In the end of the competition, the wrestler finishing in the up position becomes the winner owing to his physical control over the other wrestler. Accordingly, 'üst + Dat' indicates the implicature of a control state in which the entities being in a vertically elevated and proximal position exercise a better sense of control and influence over those in a lower position. Thus, this implicature is conventionalized through pragmatic strengthening and instantiated in the semantic memory as a distinct sense.

Example:
Hacı baba-mız Allah'a selamet ol-sa da mal-ın üst-ü-ne otur-sa-k.
Hadgi father-poss1pl God-DAT willing be-if and property-gen top-poss3sg-DAT sit-if-1pl
(Lit: We wish Hadgi father would die and we would sit onto / on top of the property.)
We wish Hadgi father would die and we would take over the property.

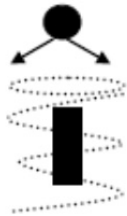


Figure 13. Control Sense
(dark sphere: TR; dark bar in upright position: LM; the spiral shape: the control exerted by the TR; downward arrows: direction of power/ influence exerted - adopted from Tyler & Evans, 2003: 102)

2c$_2$. RESPONSIBLE Sense (10 Tokens/483)

Responsible Sense is an extension from Control Sense in Up Cluster and translates into 'being accountable within one's power and control, chargeable with being cause of something'. In this scene, the focal object TR is construed as the entity of responsibility holder, while the background object LM as the entity of the responsibility directed at. The TR displays capacity for control and power over the LM and thus becomes the responsible figure, as shown in Figure 14. In the example "Biz **üst-ümüz-e** düş-en-i yap-tı-k", "*We did the thing falling on / onto us",* the conventional reading is that the TR '*we*' holds responsibility for the LM 'the thing in our share/our part' within its power and control. In this non-spatial configuration, the TR movement prompts for a responsible sense at the conceptual level. There is an independently-motivated experiential correlation between the implicature of responsibility and vertical elevation. The concept here is accountability within the TR's capacity of control over the LM owing to its better skills or higher rank, not vertical elevation. In this sense, dynamism is de-emphasized as focus on the resultant state of the TR movement. Responsible Sense derives in the following way: an entity in charge of another entity is experienced when the entity of higher capacity and control is situated in a higher position and a proximal range to the entity of lower capacity. Physically big and vertically elevated entities are often associated with power and influence over physically small and vertically low entities (Tyler & Evans, 2003: 101). Such experiential correlation between vertically up and the linguistic usage of 'üst + Dat' results in the implicature of Responsible Sense in a non-spatial configuration. As a result, it is conventionalized through pragmatic strengthening and instantiated in the semantic memory as a distinct sense directly extended from another distinct sense.

Example:
Biz üst-ümüz-e düş-en-i yap-tı-k.
We top-poss1pl-DAT fall-ger-ACC do-past-1pl.
(Lit: We did the thing falling on / onto us.)
We did our part. (We did what we were responsible for.)

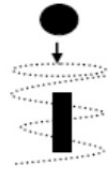


Figure 14. Responsible Sense

2d. PREFERENCE Sense (3 Tokens/483)

Preference Sense is in Up Cluster and translates into 'giving advantages/priorities to some over others, liking some better than others' as in Tyler and Evans (2001a, 2003). In this scene, the TR is construed as the entity of primary importance, while the LM as the entity of secondary importance, as displayed in Figure 15. In the example "Düşünce-ler-imiz-i ve inanç-lar-ımız-ı ne kolay dostluk ilişki-ler-imiz-in **üst-ü-ne** çıkar-dı-k", "*What easy we have raised our opinions and beliefs over / above our friendship relations",* the implicature is that 'our opinions and beliefs' are preferred over 'our friendship' rather than the protoscene reading of vertical elevation, namely the TR being higher than the LM. Thus, a preference sense is prompted at the conceptual level by abstracting away from the spatial component of the protoscene. However, dynamism is de-emphasized owing to the resultant state of the TR movement. Preference Sense derives in the following way: being vertically up is experientially associated with greater amount or intensity (e.g. prices going up), happy mood (e.g. feeling up), or a better or more advanced state. Thus, vertically up is privileged relative to being physically down associated with lower amount or intensity (e.g. calm down), unhappy mood (e.g. feeling down), or a state of less prominence. Such experiential correlation results in the implicature of preference where the entities in a vertically elevated position are preferred to those in a lower position. The reanalysis of the vertical elevation reading results in an additional meaning of primary importance owing to liking or an advantageous state. Thus, the implicature of preference is conventionalized through pragmatic strengthening and instantiated in the semantic memory as a distinct sense extended directly from the protoscene.

Example:
Düşünce-ler-imiz-i ve inanç-lar-ımız-ı ne kolay dostluk ilişki-ler-imiz-in
Opinion-pl-poss1pl-ACC and belief-pl-our-ACC what easy friendship relationship-pl-poss1pl-gen

üst-ü-ne çıkar-dı-k.
top-poss3pl-DAT raise-past-1pl.
(Lit: What easy we have raised our opinions and beliefs over / above our friendship relations.)
How easily we have put our opinions and beliefs above our friendship.

Figure 15. Preference Sense
(adopted from Tyler & Evans, 2003: 103)

3. FORWARD CLUSTER

Five distinct senses (i.e. Intrusion, Target, Above-and-Beyond, Transfer and Persistent) fall in this category. Each sense arises from the construal of a forward-oriented TR positioned in the point A moving toward a targeted LM in the point B. In this complex conceptualization, the points of reference come into existence in a sequential manner through the forward motion directed at an intended target point, as illustrated in Figure 16.

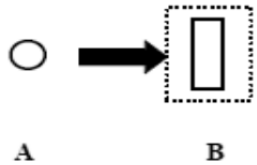


Figure 16. Forward Cluster

Although such a complex conceptualization arises from a sequentially evolving process, during the reanalysis it is subject to conceptualization in summary format (Tyler and Evans, 2003: 80). It means that each point comes into existence in a sequential manner through the TR's forward motion directed at the intended point. In real time, the object of focus cannot occupy more than one position at a time. When viewing a spatial scene in summary format, the entire trajectory that the object followed is conceptualized and represented in mind. Accordingly, the focal object initiates a forward-like trajectory starting in the point A and then prompts for and ends in the point B, where the LM is located. Thus, the points are

tied and related to each other with the directed forward motion of the TR. In this configuration, the TR is construed as the moving focal object, while the LM as the goal of the TR movement. The moving TR exerts influence on the stationary LM as a result of the forward motion.

'Forward' is associated with *directed motion, moving ahead* and *progress* in real-world experiences that are not entailed in the construal of the protoscene, namely the 'higher-than-the-LM' reading. It is also associated with *being extreme and bold*. Thus, the forward trajectory derives from real-life spatial experiences. For example:

Example: Beşiktaş-tan geç-erek Boğaz köprü-sü **üst-ün-e** var-dı-k.
*Beşiktaş-ABL pass-ger Bosphorus bridge-poss3sg top-poss3sg-**DAT** arrive-past-1pl.*
*(Lit: We arrived **onto / on top of the Bosphorus bridge** passing through Beşiktaş.)*
We arrived at the Bosphorus bridge passing through Beşiktaş.

In this scene, the TR 'we' moves from its original location in the point A to the targeted LM 'top of the Bosphorus bridge' in the point B in an attempt to reach it. A forward trajectory emerges upon the TR's completion of the 'passing' action directed at the targeted LM through Beşiktaş. Here, the conventional interpretation is 'We arrived at the Bosphorus bridge', which is prompted by the 'üst + Dat' construction. Based on our knowledge of real world-force dynamics, we know that at the point B, the TR reaches the LM in a state of potential contact. As a result of such experiential correlation, the implicature of forward is conventionalized through pragmatic strengthening and instantiated as a complex trajectory cluster from which distinct senses extend.

3a. INTRUSIVE Sense (29 Tokens/483)

Intrusive Sense is in Forward Cluster and translates into 'invasion of privacy, unwelcome visit'. In this configuration, the TR reaches the LM in an intrusive manner through a physical or abstract movement along the forward trajectory. At the conceptual level, the TR is perceived to be more powerful than the LM. In the experiential sense, power implicates control and ability to influence the targeted entity owing to the focal object's physically higher and proximal position to it. Thus, the TR movement results in a state of spatial invasion. We conceive this conventionalized sense through the TR's forceful forward motion realized in a threatening or insistent manner, which ends at the spatial boundary of the LM, as displayed in Figure 17. In the scene "Bir kez el-in-de-ki kağıt-lar-ı fırlat-ıp

üst-üm-e yürü-müştü", "*One time, (s)he had thrown the papers in his/her hand and walked onto me",* the LM 'me' is conceptualized as the goal and thus the entity of intrusion, while the TR '(s)he' as the initiator of the intrusion with its forceful forward motion. In this configuration, 'üst' refers to the body figure as the LM, and the TR's manner of motion leads to the invasion of the LM's actual or abstract spatial boundary. The conventional interpretation is '(S)he came at me' rather than the spatial components of the protoscene: vertically up and proximal. A forceful forward motion implicates a dynamic scene and an immediate directive action toward the entity of intrusion in experiential terms. Thus, the implicature of intrusion is conventionalized and instantiated in the semantic memory through a metaphorical extension: 'FORCEFUL FORWARD MOTION IS INTRUSION'.

Example:
Bir kez el-in-de-ki kağıt-lar-ı fırlat-ıp üst-üm-e yürü-müş-tü.
One time hand-poss3sg-LOC-rel.sfx paper-pl-ACC throw-ger top-poss1sg-DAT walk-hearsaypast-past-3sg.
(Lit: One time, (s)he had thrown the papers in his/her hand and walked onto me.)
Once, (s)he threw the paper in his/her hand and came at me.

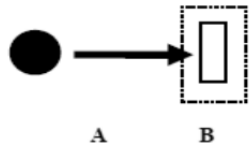


Figure 17. Intrusive Sense
(dark circle: TR; clear bar: LM; dotted rectangle: LM's sphere of influence; dark arrow: forward motion and trajectory; the tip of the arrow in LM's sphere: intrusion)

3a1. TARGET Sense (23 Tokens/483)

Target Sense is an extension from Intrusive Sense and translates into 'a fixed goal or objective'. In this configuration, the TR exerts a focused direction toward the LM as an instance of physical or abstract movement through the forward trajectory. We conceive this type of goal-oriented motion in a focused manner, which may or may not end at the spatial boundary of the LM, as shown in Figure 18. In the scene "Piraye Cavide Hanım-ın **üst-ü-ne** saldır-mış", "*It is said that Piraye attacked onto Ms. Cavide",* the TR 'Piraye' is construed as the focal object and thus the

source of the focused forward motion toward the target, while the LM 'Ms. Cavide' as the target of the TR's motion. Here, 'üst' refers to Ms. Cavide's body figure as the LM. In experiential terms, vertically up and proximal implicate an upright position and a better sense of control and influence respectively. For instance, when aiming, the archers in upright position and close to the target display a better performance in meeting the target compared to the ones in down position and far from the target, who lack power and control. In real-world experiences, 'up' is often associated with positive states such as strength, increased level of activity (e.g. Crime rates went up by 10%), or moving toward an entity for proximity (e.g. He walked up to me), which are not entailed in the construal of the spatial reading of the protoscene. In this non-spatial configuration, the focal object reaches and exerts influence on the targeted background object through the forward trajectory owing to its more advantageous positioning as higher than and proximal to the LM. As a result, a focused forward motion implicates a dynamic scene and direction toward the target entity. This type of use ties in the metaphor of GOALS ARE DESTINATIONS.

Example:
Piraye, Cavide Hanım-ın üst-ü-ne saldır-mış.
Piraye, Cavide Ms.-gen top-poss3sg-DAT attack-hearsay past-3sg
(Lit: It is said that Piraye attacked onto Ms. Cavide.)
It is said that Piraye attacked Ms. Cavide.

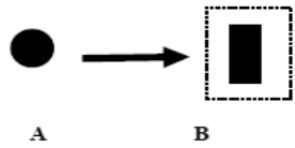


Figure 18. Target Sense
(circle: TR; dark bar: LM; dotted rectangle: LM's sphere of influence; dark arrow: forward motion and trajectory)

3b. ABOVE-AND-BEYOND Sense (15 Tokens/483)

Turkish 'üst +DAT' has an Above-and-Beyond Sense in Forward Cluster, which translates into '(go) beyond, exceed', being a similar finding to Tyler and Evans' (2001a, 2003) analysis of 'over' but in a different trajectory cluster. The relation denoted between the TR and the LM in this configuration is that the TR passes beyond the targeted LM in an excessive motion as an instance of abstract movement through the forward trajectory. In this scene, the LM is construed as the targeted entity and the goal of the

TR's motion, while the TR as the entity of focus moving toward the goal through the forward trajectory in an attempt to come into contact with the LM. However, the TR might go above and beyond the intended point as a result of the excessive forward motion, as illustrated in Figure 19. In this use, the protoscene spatial component of vertical elevation is not implicated. For instance, in the example "Piyasa-da kur 1 milyon 650 bin-in **üst-ü-ne** çık-tığ-ın-da satış gözle-n-iyor", "*When the rate of exchange in the market rises over 1 million 650 thousand, sale is monitored",* the implicature is that the TR 'the actual rate of exchange' exceeds the LM 'the established or targeted cut-off point for the rate of exchange, namely 1 million 650 thousand'. Thus, the TR movement prompts for an excess sense at the conceptual level by abstracting away from the spatial component of the protoscene. In other words, 'üst + DAT' is no longer interpreted as a spatial configuration where the TR is physically higher than the LM. Although the forward trajectory is the predicted use of the 'üst + DAT' protoscene, such a configuration brings an additional implicature to the scene. In the experiential sense, the forward motion of the TR is similar to the trajectory depicted by the protoscene in which the TR intentionally moves above and beyond the LM and completes its motion in a further point than the targeted point of the LM. Accordingly, the Above-and-Beyond Sense derives in the following way: in experiential terms, vertical elevation is associated with the concept of 'more'. For instance, when the entities increase in number, they pile up and look more than the entities of lesser amount such as half a glass of milk versus a quarter glass of milk. When we keep adding milk to the half a glass in an attempt to reach a full glass, there is a chance that the milk might overflow the glass due to carelessness or some other distraction. This results in the excessive amount of milk that is above-and-beyond the intended amount and level. Thus, the implicature of above-and-beyond is conventionalized as a distinct sense through pragmatic strengthening and instantiated in the semantic memory with a metaphorical extension: MORE IS UP.

Example:
Piyasa-da kur 1 milyon 650 bin-in **üst-ü-ne** çık-tığında satış
*Market-LOC rate of exchange 1 million 650 thousand-gen top-poss3sg-**DAT** rise-when sale*
gözle-n-iyor.
monitor-passv-pres.prog-3sg.
(Lit: When the rate of exchange in the market rises over 1 million 650 thousand, sale is monitored.)
When the rate of exchange in the market exceeds 1 million 650 thousand, transaction is monitored.

Figure 19. Above-and-Beyond Sense
(dark bar: LM in upright position; clear circle: TR; dark arrow: direction of the above-and-beyond trajectory– adopted from Tyler & Evans, 2003: 84)

3c. TRANSFER Sense (13 Tokens/483)

Transfer Sense is also in Forward Cluster and translates into 'conveyance or passing from one place to another'. In contrast to Tyler & Evans' (2001a, 2003) analysis of English 'over', the conceptualization of Transfer Sense is reanalysed in the forward trajectory on the basis of the focused forward motion of the TR intending to reach the LM, which is construed as the target point, i.e. the goal of the TR's motion. The relation denoted between the TR and the LM in this configuration is that the TR is intended to be conveyed from the point A to the point B, namely the LM in actual or abstract movement, as diagrammed in Figure 20. In this scene, a change in location of the TR gives rise to the implicature that the transfer has taken place. In the example "**Üst-ü-ne** yık-ıl-mak iste-n-en suç-lar-ın hiçbir-i kanıtla-n-ma-dan sal-ıver-il-miş-ti", "(S)*he had been released without any of the crimes that had been intended to lay* ***onto*** *her / him being proven.",* the non-physical entity 'the crimes' is construed as the TR and '(s)he' as the LM and they are tied to each other with a conventionalized Transfer Sense in a non-spatial configuration. The conventional interpretation of this sentence is that *some other people's crimes were laid on her/him.* That is to say, (s)he was blamed for other people's crimes. In other words, the responsibility for the crimes was metaphorically transferred to the subject from the people who actually committed the crimes. We cannot derive the implicature of transfer from the context in this sentence, which indicates that the transfer has emerged as a distinct sense through metaphorical extension. Crime is a non-physical entity, so nothing is physically transferred from one place to another. In a conventionalized distinct sense, it can be carried from one point to another without a literal spatial relation configured between the TR and the LM. On the other hand, the additional example "Dino Usta bacak bacak **üst-ü-ne** at-mış müdür-le konuş-uyor", "*Master Dino has apparently thrown leg on top of leg and is speaking with the principal."* illustrates the conceptualization of a spatial configuration between two

physical entities, namely the TR 'one leg' moving from the point A (its original location) to the point B (its target location) where the LM 'the other leg' is located. As a result of the motion of 'crossing one leg over the other' through the ABC trajectory, the legs take a crossed position and the process of leg-crossing is depicted with the 'üst + DAT' construction since the dative particle implicates a directed forward motion toward the goal. Thus, the implicature of physical or non-physical transfer is conventionalized through pragmatic strengthening and instantiated in the semantic memory as a distinct meaning.

Example:
Üst-ü-ne yık-ıl-mak iste-n-en suç-lar-ın hiçbir-i kanıtla-n-ma-dan
Top-poss3sg-**DAT** lay-passv-infin. intend-passv-ger crime-pl-gen none-poss3pl prove-passv-neg-ABL
sal-ıver-il-miş-ti.
release-soon-passv-hearsay past-past-3sg
(Lit: (S)he had been released without any of the crimes that had been intended to lay on**to** her / him being proven.)
(S)he had been released before none of the crimes that had been intended to lay on her/ him were proven.

Figure 20. Transfer Sense
(dark bar: LM in upright position; dark circle: TR; arc-like trajectory: transfer of the TR from the point A to the point C – adopted from Tyler & Evans, 2003: 87)

3d. PERSISTENT Sense (7 Tokens/483)

The Persistent Sense is in Forward Cluster and translates into 'insistence by holding firmly to a state or undertaking despite obstacles, being obstinately repetitious'. In this scene, the LM is conceptualized as an obstacle or impediment to the TR's forward motion. More specifically, the moving TR as the focal object exerts control and influence over the stationary LM as the background object through a repeated forward motion lacking restraint, which may or may not end in a potential contact with the targeted LM. Thus, the TR movement that is ardently inclined to reach the target LM ends up a state of persistence. We conceive this type of persistence based on the repeated forward motion of the TR realized in an insistent manner until it reaches the LM, as diagrammed in Figure 21.

The goal of the TR is to reach the targeted LM through the forward motion unrestrained. In the example "'Bu konu-nun **üst-ü-ne** gid-er-se-n program-ın-ı yayın-dan kaldır-ır-ız' di-yor-lar", "*If you go **onto** this issue, we will remove your program from broadcast." they say",* the TR 'you' is construed as the focal object and thus the source of persistence, while the LM 'this issue' as the entity being persisted and thus the goal of the TR's insistent forward motion. Here, the TR's manner of motion in the form of persistence leads to a state of invasion of the LM's actual or abstract spatial boundary. The conventional interpretation of this sentence is 'your persistence on a given matter', which cannot be inferred from context. In the experiential sense, the physically proximal TR implicates influence. Thus, the implicature of persistence arises from the independently motivated experiential correlation between proximity and influence. 'Forward' is experientially associated with directed motion and being close, as well as ardently inclined in real world, which is not entailed in the construal of the protoscene, namely the spatially higher-than and proximal reading. It is also associated with deviating from convention and being extreme and bold. For instance, a persistent lover keeps sending flowers, chocolates, love letters, poems, etc. to the beloved despite being rejected by her many times. Here the lover's intention is to be close to the beloved by sending gifts, which is interpreted as a persistent forward motion directed at the target. In addition, the gifts reaching and coming into contact with the beloved implicates physical closeness to her. Overall, the lover's manner of 'coming on**to** the beloved' is interpreted as persistence due to his insistence on the beloved by holding firmly to a state of sending gifts despite the obstacles such as her rejections. In these examples, 'going onto an issue / coming onto a beloved' are not used in a literal sense but in a metaphorical sense and implicate 'persistence in the matter / winning the beloved's heart'. Thus, the implicature of persistence is conventionalized as a distinct sense through metaphorical extension and instantiated in the semantic memory.

Example:
"Bu konu-nun **üst-ü-ne** gid-er-se-n program-ın-ı yayın-dan kaldır-ır-ız."
"*This issue-gen top-poss3sg-**DAT** go-aor-cond-2sg program-poss2sg-ACC broadcast-ABL remove-aor-1pl."*
di-yor-lar.
say-pres.prog-3pl
(Lit: "If you go onto this issue, we will remove your program from broadcast." they say.)
"If you persist with this issue, we will terminate the broadcast of your show." they say.

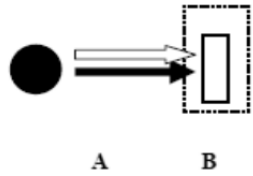


Figure 21. Persistent Sense
(dark circle: TR; clear bar: LM; dotted rectangle: LM's sphere of influence; dark arrow: forward motion and trajectory; clear arrow: repeated forward motion; the tip of the arrow in LM's sphere: invasion of the LM's space)

4. SURFACE CLUSTER

Surface Cluster involves three distinct senses (i.e. Figure, Top and Outfit). Each sense arises from the non-spatial construal of a TR-LM configuration through an independently-motivated experiential correlation between the surface implicature and the protoscene components of vertical elevation and physical proximity. In this configuration, the TR in a proximal relation to the LM interacts and influences the targeted LM through its directed motion, as illustrated in Figure 22.

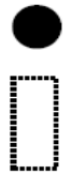


Figure 22. Surface Cluster

In Surface Cluster, the presumed intention is to have the TR come into contact with the targeted LM, which is construed as the surface of a figure. For instance, in Figure and Outfit Senses, the LM is construed as the surface of a full form or figure, while in Top Sense it is construed as the surface of a top form or figure. Here, the TR's influence can be on the whole or partial surface of the LM figure depending on the extent of the TR movement. In this cluster, the TR-LM configuration prompts for the surface of a full or partial form/figure at the conceptual level by abstracting away from the spatial construal of the protoscene through metonymy. For instance, in the sentence "**Üst-üm-e** gömlek geç-ir-di-m", "I put a shirt on**to** me", the TR 'shirt' is intended to constitute the uppermost surface of the LM 'top part of my body'. In this use of the 'üst

+ Dat' construction, the prototypical understanding of the topmost surface extends to the 'surface of the top part of a human body'. In a similar sentence, "**Üst-üm-e** elbise geç-ir-di-m", "I put a dress on**to** me", the TR 'dress' is directed at the LM of 'top and bottom parts of my body' and it is expressed by the same 'üst + Dat' construction as the above example. However, in this scene, the focus is on the outermost layer of the full figure/body rather than only the surface of the top part of the body. The second example is also a metonymic use. Thus, we infer that 'üst' can be used for either the surface of the whole body or only the top part of it.

Now let's examine the example Ev-in **üst-ün-e** de bakmak ister misin? "Would you also like to see the top of the house?". In this sentence, the TR 'you' directs its act of seeing onto the LM 'the top of the house'. Here, 'üst + Dat' actually refers to 'the top floor', i.e. 'the surface of the top level of the house' through metonymic extension rather than its roof. In contrast, in the example "Bebek **üst-üm-e** kus-tu", "The baby vomited on**to** me", the TR 'baby' directs its act of vomiting onto the LM 'me', which is expressed in the 'üst + Dat' construction. In this metonymic use, the actual meaning is 'my outfit', not 'my body'.

All these examples indicate the TR and the LM are connected to each other in a surface relationship. Accordingly, the TR, positioned higher than and proximal to the targeted LM, interacts with and influences the outermost surface of the LM in varying extents depending on the extent of the TR movement. However, the movement is de-emphasized in all surface senses since they only involve the result of the TR movement. The real-world experiences associated with the conceptual spatial relation of 'üst + DAT' often result in the implicature of surface, in which the LM is construed as the uppermost surface. 'Up' implicates vertical elevation and thus a higher position, which is associated with an erect, upright position in experiential terms. For instance, we can conceptualize the full form or figure of the objects owing to their volume if they extend in space both horizontally and vertically. Thus, the vertical dimension seems to be essential in the construal of the LM as a partial or full surface figure, at which the TR's motion is directed. The implicature of surface systematically derives from the protoscene and the basic meaning of top, in which the prototypical conceptualization of topmost surface extends to the surface of human and object figures that will be further illustrated in the analysis of surface senses. The independently motivated experiential correlation with the protoscene components of 'vertical elevation' and 'physical proximity' results in a non-spatial TR-LM configuration, in which the TR is no longer vertically elevated in relation to the LM. As a

result, the implicature of surface is conventionalized and instantiated in the semantic memory and gives rise to Figure, Top and Outfit Senses.

4a. FIGURE Sense (44 Tokens/483)

Figure Sense immediately extends from the protoscene and translates into 'bodily shape/form of a person/thing'. In this scene, the TR's focus is on the outermost surface of the targeted LM figure. Thus, the background object LM is construed as the form or figure being interacted with and influenced by the focal object TR, as illustrated in Figure 23. In the example "Koca sepet **üst-üm-e** devr-il-di", "*The huge basket fell onto / on top of me",* the implicature is that the TR 'the huge basket' fell on the LM 'my body'. The impact of the fall can be on the whole surface of the body or some part of it such as head, arm, etc. Here 'üst + Dat' is used for 'me' as a full figure; however, it actually refers to some surface of my body through metonymy. Similarly, in the second example, "Heykel-in **üst-ü-ne** de merdiven daya-yacak-lar", "*They will lean a ladder onto the statue, too",* the TR 'a ladder' influences the LM 'the statue' by leaning. In this example, the surface of the statue figure constitutes the source of support for the ladder and thus the goal of the TR's leaning-motion despite the fact that the ladder can only be based on some part of the statue surface rather than the whole surface. Thus, the TR movement prompts for a figure sense at the conceptual level by abstracting away from the spatial construal of the protoscene through metonymy. However, the movement is de-emphasized in Figure Sense since it only involves the resultant state of the TR movement. The Figure Sense derives systematically from the protoscene and the basic meaning of top. 'Up' implicates vertical elevation and thus a higher position, which is associated with an erect, upright position in experiential terms. For instance, any entity occupying space in both vertical and horizontal dimensions is regarded as a figure, such as body, statue, building, etc., owing to their volume. However, the implicature of figure is not literally spatial. In this configuration, the prototypical understanding of the topmost surface gets extended to the surface of forms and figures. Thus, the TR is no longer vertically elevated in relation to the LM. In such a non-spatial configuration, the implicature of figure is conventionalized through the metonymic use of 'üst + Dat' and pragmatic strengthening, which is eventually instantiated in the semantic memory as a distinct sense.

Example:
Koca sepet **üst-üm-e** devr-il-di.
Huge basket top-poss1sg-**DAT** fall-passv-past-3sg

(Lit: The huge basket fell on**to** / on top of me.)
The huge basket fell on me.

Figure 23. Figure Sense
(dark circle: TR; dark bar: LM; dashes around the dark bar: LM's sphere of influence)

4a1. TOP Sense (9 Tokens/483)

Top Sense is an extension from Figure Sense in Surface Cluster and translates into 'upper part / upper portion of a form or figure'. In this configuration, the targeted LM is construed as the surface of the top part of a full form or figure, which is interacted and influenced by the TR, as illustrated in Figure 24. In the example "Blucin-im-in **üst-ü-ne** yeni bir mor bluz al-ır o-nu giy-er-im", "*I would buy a new purple blouse on**to** / on top of my blue jeans and wear it",* the implicature is that the TR 'a new purple blouse' completes the LM 'my blue jeans' by filling the outer layer of my top body. The purple blouse also influences the blue jeans by matching or mismatching with them. Here, the 'üst + Dat' construction is used for a top garment to complete the blue jeans, but it actually refers to the surface of the top part of a body figure through metonymy. Thus, the top part of the body is occupied fully or partially depending on the volume and shape of the purple blouse, which completes the blue jeans into a full figure. Similarly, in the second example, "Vücud-unuz-un alt kısm-ı **üst-ü-ne** nazaran daha mı sıkı?" *"Is the bottom part of your body more firm relative to its top?",* the TR 'the bottom part of your body' is tied to the LM 'its top' on the comparative grounds of body features. Also, in this example, the LM 'its top' refers to the outermost top part, namely surface, of a body/figure. In these non-spatial configurations, there is no direct correlation between the spatial construal of the protoscene and the 'top' implicature associated with 'üst + Dat'. As in the case of Figure Sense, the movement is de-emphasized in Top Sense owing to the involvement of only the resultant state of the TR movement.

The TR-LM configuration in this scene arises from an independently-motivated experiential correlation between the implicature of top and the

functional elements of vertical elevation and proximity denoted by the protoscene. Accordingly, Top Sense develops in the following way: up implicates vertical elevation and thus a higher position, which is associated with an erect, upright position, superiority, or a more favourable or advanced state in experiential terms. For instance, the top floor of a two-storey house is more favourable than the bottom floor since the range of view is better. Despite the fact that both top and bottom floors are easily distinguished as two separate levels and thus two separate parts, they still belong to the same whole, namely the house figure which is analogous to the body figure discussed earlier. Moreover, the top floor stands in an upright position and constitutes the top portion of the house based on the support it receives from the bottom floor. Thus, the TR-LM configuration in this scene prompts for Top Sense at the conceptual level by abstracting away from the spatial construal of the protoscene through metonymy. Such experiential correlation results in a non-spatial configuration of the TR-LM in a surface relationship and the association of the 'üst + Dat' construction with the conventionalized implicature of 'the outermost surface of a top figure' via pragmatic strengthening. Accordingly, Top Sense derives from Figure Sense and its basic meaning of top through the extension of the prototypical sense of 'outermost surface' to the implicature of 'the surface of the top part' through metonymy. Thus, the conventional meaning of 'top' figure is instantiated in the semantic memory as a distinct sense in Surface Cluster.

Example:
Blucin-im-in **üst-ü-ne** yeni bir mor bluz al-ır o-nu giy-er-im.
*Blue jeans-poss1sg-gen top-poss3sg-**DAT** new a/an purple blouse buy-aor-(1sg) it-ACC wear-aor-1sg*
*(Lit: I would buy a new purple blouse **onto** / on top of my blue jeans and wear it.)*
I would buy a new purple blouse over / on top of /on top with my blue jeans and wear it.

Figure 24. Top Sense
(dark circle: TR; bar: LM; dashes around the top bar: surface of LM's top layer within sphere of influence)

4b. OUTFIT Sense (13 Tokens/483)

Outfit Sense immediately extends from the protoscene and translates into 'clothes, garments for the body'. In this scene, the TR's motion is directed at the part or whole of the LM surface in a similar way to Figure Sense. Thus, the focal object TR interacts with and influences the background object LM, which is construed as the goal of the TR movement, as illustrated in Figure 25. In this scene, the LM 'outfit' is represented by a clear bar framed by dark lines, unlike the dark bar illustrating the Figure Sense in Figure 26, because clothing is conceptualized as the most outer layer of a human figure. In the example "Leon çorba-yı **üst-ü-ne** dök-me-den dikkatle iç-iyor", "*Leon is drinking the soup carefully without spilling (it) on**to** himself*", the implicature is that the TR 'the soup' has been spilled on the LM 'his clothes'. Here, the 'üst + Dat' construction is used for an outfit implicature through metonymy. However, the focus can be on some part of the outfit surface rather than its whole surface. The result of such influence is the distortion of the colour and texture of the fabric to varying extents depending on the amount and proximal range of the TR. Similarly, in the second example, "**Üst-üm-e** bir çekidüzen ver-me-m şart", "*It is necessary that I give neatness on**to** myself",* the TR 'I' influences the LM 'on**to** myself' referring to *'my outfit'* by fixing it up. In this use, the act of fixing oneself up applies to the adjustment of clothes for better overall looks. In the same way, the impact of the TR involves either the partial or the full portion of the LM. Thus, the TR movement prompts for Outfit Sense at the conceptual level by abstracting away from the spatial construal of the protoscene through metonymy. However, Outfit Sense only involves the resultant state of the movement due to the de-emphasized TR movement.

Accordingly, Outfit Sense derives in the following way: up implicates vertical elevation and thus a higher position, which is associated with an erect, upright position or a more favourable or advanced state in experiential terms. For instance, we can see the full form and shape of the clothes only when we put them on since they can stand in an upright position and be visible in a full figure owing to their volume. In such a scene, the outfit is construed as the uppermost layer of a figure in which the TR 'clothes' constitutes the topmost surface of the LM 'human body'. In experiential terms, the skin constitutes the outermost layer of the human body and is thus positioned upmost. However, clothes are used to cover our body, so they are positioned higher than the human body and conceptualized as the uppermost layer. Moreover, regardless of the physical positions humans take (e.g. lying down on one's back, lying on

one's stomach, sitting, standing, etc.), the clothing on human bodies is still conceptualized in an upward position in relation to our bodies. As a result, through recurring experiences of clothing, the 'outermost' layer is associated with the 'uppermost' layer in a human figure. For instance, when we use 'üst + Dat' for a targeted LM in the context of *'spilling soup'* as the below example, it refers to the clothes, not to the body figure which is a metonymic use. Thus, the implicature of outfit arises from the independently-motivated experiential correlation with a human figure in a surface relationship in which the TR 'outfit' is construed as the 'upper surface', while the LM 'body figure' as the 'lower surface'. Such experiential correlation results in the association of the 'üst + Dat' construction with the conventionalized implicature of outfit via pragmatic strengthening. Accordingly, Outfit Sense extends systematically from the protoscene and the basic meaning of top. The prototypical conceptualization of the outermost surface gets extended to the implicature of outfit through metonymy in which the TR is no longer vertically elevated in relation to the LM. As a result, Outfit Sense is conventionalized in Surface Cluster and instantiated in the semantic memory as a separate sense.

Example:
Leon çorba-yı üst-ü-ne dök-me-den dikkatle iç-iyor.
Leon soup-ACC top-poss3sg-DAT spill-neg-ABL carefully drink-pres.prog-3sg
(Lit: Leon is drinking the soup carefully without spilling (it) onto / on himself.)
Leon is drinking the soup carefully without spilling it on his outfit.

Figure 25. Outfit Sense
(dark circle: TR; clear bar framed by dark lines: LM; dashes around the clear bar: LM's sphere of influence)

5. FOCUS-OF-ATTENTION Sense (57 Tokens/483)

The Focus-of-Attention Sense directly derives from the protoscene and translates into 'about'. The implicature of focused attention derives in the

following way: physical closeness implicates clear vision and thus greater amount of attention which enables focus on the background object in experiential terms, while physical distance implicates obscure vision and thus less attentive action. Given the fact that entities closer to us are more easily seen, this experiential correlation is a state of focused attention reanalysed as distinct from the spatial configuration that it originates, as displayed in Figure 26. The functional element of dynamism is de-emphasized in Focus-of-Attention Sense. Instead, the resultant state of the TR movement is involved. In the example "Kimse 'siyasetçi yalan-lar-ı' **üst-ü-ne** bir incele-me-ye giriş-mi-yor", *"Nobody initiates an investigation on politician lies."*, the LM 'politicians' lies' is conceptualized as the focus of attention and the TR 'investigation' as the source of attention. Every spatial scene is viewed from a particular vantage point that refers to how a particular scene is viewed, and in large part this determines the functional configuration of a scene and the ways it is meaningful. Thus, the construal of a scene is not restricted to the default vantage point represented by the conceptualizer. Rather, the same scene can be construed from many different vantage points. In this scene, the vantage point shifts to the TR's location, which is higher than the LM. The LM is construed as the goal of the attention directed from the TR being positioned within a proximal range so that the line of vision and attention can be conveyed to the background object (Tyler & Evans, 2003). However, the TR is not always construed as higher than the LM. In some instances, such as watching the stars with a telescope, the vantage point is located lower than the LM and the interpretation still remains the focus-of-attention rather than the protoscene reading, i.e. the TR higher than the LM. For instance, when focusing on a star (LM), one simply adjusts the lens of the telescope to a focal vision so that the image is enlarged and thus seems closer. In such a non-spatial configuration, proximity stands as a crucial element of how we actually experience focus-of-attention. Thus, this implicature derives from an independently motivated experiential correlation between focus-of-attention and vertical elevation through pragmatic strengthening, which is eventually conventionalized as a distinct sense and instantiated in the semantic memory.

Example:
Kimse 'siyasetçi yalan-lar-ı' **üst-ü-ne** bir incele-me-ye giriş-mi-yor.
*Nobody politician lie-pl-poss3pl top-poss3pl-**DAT** a/an investigate-ger-DAT initiate-neg-pres.prog-3sg*
(Lit: Nobody initiates an investigation on politician lies.)
Nobody initiates an investigation about 'politicians' lies'.

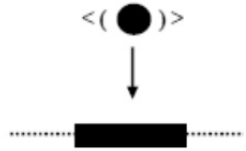


Figure 26. Focus-of-Attention Sense
(dark circle: TR; dark bar: LM; eye icon: vantage point; arrow: direction of attention - adopted from Tyler & Evans, 2003: 96)

5a. EMPHASIS Sense (8 Tokens/483)

Emphasis Sense is an extension from Focus-of-Attention Sense and translates into 'special attention and importance'. The implicature of emphasis derives in the following way: vertically up implicates greater amount and control; physical closeness implicates greater attention and thus focuses on the background object in experiential terms. Given our real-world experiences, we know that the entities of greater importance are placed higher and closer than the entities of less importance, which results in a state of emphasis as diagrammed in Figure 27. However, Emphasis Sense only involves the resultant state of the TR movement in which the dynamism is de-emphasized. In the example "Muzaffer Bozok sözcük-ler-in **üst-ü-ne** bas-arak konuş-ma-ya başla-dı", "*Muzaffer Bozok began to talk pressing* ***onto*** */ on top of the words",* the TR 'Muzaffer Bozok' is construed as the source of emphasis and the LM 'the words' as the goal and entity receiving emphasis. In other words, the TR determines the amount and extent of attention given to the LM, namely the words when producing them. Thus, the scene is conceptualized as such: one entity, namely the TR being in 'up' position puts stress on the LM in the metaphorical sense. The entity receiving stress is interpreted as the goal by virtue of the dative case marking at which the focused attention and importance is directed. Stress also implicates physical closeness and thus a better control of the background object in experiential terms. In addition, it might bring about a noticeable structural change in the object due to the stress applied. In this specific instance, the vocalic force applied to words when speaking leads to an auditory change such as emphasized words in higher pitch and volume. Thus, the emphasized background object attracts more attention sources from conceptualizers since it is highlighted through emphatic stress. However, lack of stress implicates physical distance and a weak control of the background object. Thus, an LM devoid of emphasis attracts less attention from conceptualizers. This suggests that, in experiential terms, there are two elements associated with the concept of

emphasis: vertically up and proximal. As a result, 'üst + Dat' develops a conventional emphasis sense and is instantiated in the semantic memory.

Example:
Muzaffer Bozok sözcük-ler-in üst-ü-ne bas-arak konuş-ma-ya başla-dı.
Muzaffer Bozok word-pl-gen top-poss3plr-DAT press-ger speak-ger-DAT begin-past-3sg
(Lit: Muzaffer Bozok began to talk pressing onto / on top of the words.)
Muzaffer Bozok began to talk stressing the words.

Figure 27. Emphasis Sense
(dark circle: TR; dark bar: LM; eye icon: vantage point; arrow: direction of emphasis; rectangle border: containment)

6. COVERING (Occlusion) Sense (48 Tokens/483)

Covering Sense immediately extends from the protoscene and translates into 'covering/occlusion'. In this scene, the vantage point is on-stage since both the TR and the LM are lower than our line of vision. As a result, the TR stretching over the LM is conceptualized as occluding the LM. In experiential terms, the construer's line of vision is directed at a TR-LM thanks to its being positioned higher than the TR. Thus, the TR in a motion toward the locating object can be larger or perceived to be larger than the LM such as a sheet over a bed, a skirt over tights, etc., as diagrammed in Figure 28. The dynamism is de-emphasized in Covering Sense. Instead, the resultant state of the TR movement is focused. In the example "Siyah, kalın bir tayt giy-miş-ti. **Üst-ü-ne** kısa kırmızı bir etek…", "*She had worn black thick tights. On / on top of it, a short red skirt…*", the TR 'a red short skirt' covers the LM 'thick black tights' in a stretching manner through the vertical and horizontal trajectory. As in Tyler and Evans (2003), the scene is conceptualized as the TR occupying the whole surface or some significant portion of the LM due to the perception of the TR as larger than the LM. The LM is construed as a goal at which the TR's motion is directed. In addition, the TR is not necessarily construed as being in a higher position than the LM. The following example illustrates a TR-LM configuration in which the TR 'skirt' and the

LM 'tights' are parallel to each other in a vertical position. The skirt is shorter than the tights, but it still covers some significant portion of them. As a result, the interpretation of covering is distinguished from the default protoscene reading based on two essential changes: the vantage point shifting from off-stage to on-stage and the TR being perceived as larger than the LM (Tyler & Evans, 2003: 91). Thus, this implicature of covering is conventionalized through pragmatic strengthening, leading to the instantiation of Covering Sense independently in the semantic memory.

Example:
Siyah, kalın bir tayt giy-miş-ti. **Üst-ü-ne** kısa kırmızı bir etek…
*Black thick a/an tights wear-hearsay past-past-3sg. Top-poss3sg-**DAT** short red a/an skirt…*
(Lit: (S)he had worn a black thick tights. On / on top of it, a short red skirt…)
(S)he had worn thick black tights. Over them, a red mini skirt…

Figure 28. Covering Sense
(Eye icon: vantage point; elongated sphere: TR; vertical bar: LM - adopted from Tyler & Evans, 2001: 753)

6a. REVEALING Sense (4 Tokens/483)

Revealing Sense is an extension from Covering Sense and translates into 'revelation/appearance'. In this scene, the vantage point is also on-stage since the construer's line of vision is higher than the TR and the LM. The following example illustrates a TR's motion directed at an LM in a non-spatial configuration, in which the TR (e.g. Turkey's overlooked realities) is conceptualized as a revealing entity by moving above the LM (e.g. water), as diagrammed in Figure 29. As in the case of Covering Sense, Revealing Sense involves the resultant state of the TR movement, i.e. de-emphasized dynamism. In the example "Türkiye gözardı ed-il-miş gerçek-ler-i-ni su **üst-ü-ne** çıkar-ma-yA başlı-yor", "*Turkey is beginning to raise its overlooked realities on**to** / on top of water*", the figurative use of 'to raise realities onto / on top of water' implicates revelation. Thus, the conventional interpretation is that the TR 'overlooked realities' is revealed

by moving above the impeding LM 'water'. In contrast to Covering Sense, the LM is construed as an occluding entity and an impediment to the visibility of the TR in this configuration since it stretches to a larger space than the TR. Given our real-life experience, up correlates vertically with visibility, while physically down with invisibility. For instance, a cat sitting on top of a table is visible owing to its physically elevated position and a construer higher than the TR and the LM, while a cat under the table is invisible owing to its physically down position and thus the occluding table. However, in some instances, physically lower TRs can still be visible, such as the bottom part of a shirt (TR) and a skirt (LM) that are parallel to each other in a vertical position. The skirt covers the bottom of the shirt completely when the bottom is worn inside the skirt. However, when kneeling to pick something up on the floor or rising up toes to reach an object higher than us, the bottom of the shirt moves out the skirt and is revealed over the skirt through a stretching move. Therefore, as with Covering Sense, the interpretation of revealing is distinguished from the default protoscene reading based on two essential changes: the vantage point shifting from off-stage to on-stage and the TR being perceived as smaller than the LM. In all the scenes configured in this sense, the verb 'rise' is used metaphorically to implicate revelation since the focus object TR constitutes an entity that was previously hidden because of being covered and is conceptualized as revealed owing to its movement onto the background object LM and its location at the top. Thus, Revealing Sense derives from an independently-motivated experiential correlation between revelation and vertical elevation through a metaphorical extension, which is eventually conventionalized as a distinct sense and instantiated in the semantic memory.

Example:
Türkiye gözardı ed-il-miş gerçek-ler-i-ni su üst-ü-ne çıkar-ma-yA
Turkey overlook make-pass-evid. reality-pl-poss3sg-ACC water top-poss3sg-DAT raise-ing-DAT(to) out
başlı-yor.
begin-pres.prog-3sg
(Lit: Turkey is beginning to raise its overlooked realities onto / on top of water.)
Turkey is beginning to reveal its neglected realities.

Figure 29. Revealing Sense
(clear circle: TR; dark rectangle: LM; eye icon: vantage point)

7. PSYCHOLOGICAL STATE Sense (13 Tokens/483)

Psychological State Sense is extended directly from the protoscene and translates into 'emotional/internal state'. As indicated in Evans (2010), the conceptualization of the implicature Psychological State is reanalysed as a distinct lexical concept and denotes a relation between the TR and the LM in a non-spatial configuration. In this scene, the TR causes an emotional state in the LM. Thus, the background object LM is construed as the target point and the focal object TR as the source of the psychological state which is intended to be conveyed to the targeted LM through the inward trajectory, as schematized in Figure 30. In the example "**Üst-ü-ne** ezik büzük bir hal gel-miş-ti.*", "An overwhelmed state had come over her / him.",* the TR 'overwhelmed state' influences the LM '(s)he' by surrounding her/him and changing her/his ordinary mood from normal to overwhelmed. Thus, the TR's state surrounds and influences the LM through its directed motion at the LM from all angles. Such a non-spatial configuration gives rise to the implicature that the TR exerts some type of influence on the LM. Accordingly, the LM displays a new internal state as a consequence of a psychological or emotional state brought by the TR into the LM, which may or may not have observable manifestations (cf. Evans, 2010). The prototypical meaning involves dynamism; however, this sense involves the resultant state of the movement due to the de-emphasized TR movement.

Psychological State Sense derives in the following way: vertically up implicates strength, greater amount and control power; physically proximal implicates a better sense of influence. For instance, the state of love permeating the internal state of a lover is caused by the emotional attachment felt toward somebody or something. The LM 'the person in love' is conceptualized in a weaker position and down in mood because of suffering, while the TR 'the person beloved' is conceptualized in a stronger position and up in mood because of the feeling of love felt for her/him. In addition, physical proximity is important for the feeling of love

to develop and influence the overall internal state of the lover. If a relationship is experienced in the distance, the feeling of love tends to decrease in time, while it flourishes in a relationship, in which the partners are physically close. The implicature of a psychological state is not literally spatial. Rather, it is conventionalized as a distinct sense via pragmatic strengthening through independently-motivated physical correlation between the psychological state and the protoscene components of "up" and "proximal". We cannot infer the meaning of psychological state from context because the TR-LM non-spatial configuration is conventionalized through pragmatic strengthening and instantiated in the semantic memory as a distinct meaning.

Example:
Üst-ü-ne ezik büzük bir hal gel-miş-ti.
*Top-poss3sg-**DAT** overwhelmed a/an state come-hearsay past-past-3sg*
(Lit: An overwhelmed state had come over her / him.)
(S)he got into an overwhelmed mood.

Figure 30. Psychological State Sense

8. BACKGROUND Sense (8 Tokens/483)

Background Sense immediately extends from the protoscene and translates into 'the scenery or ground behind something'. In this scene, the TR is construed as the entity in the foreground such as a pattern, colour, etc., while the LM as the ground lying behind the foreground entity. Therefore, the LM is conceptualized as the targeted background object at which the TR's motion is directed. The presumed intention is to have the TR come into contact with the LM, as illustrated in Figure 31. In the example "Altın **üst-ü-ne** pırlanta bir kolye", "*A diamond necklace on gold",* the TR '*a diamond'* is placed onto the LM '*gold*'. The conventional interpretation of this sentence is '*a necklace made of diamond on a gold background'*. Notice that in this sentence the TR is not necessarily located in a vertically elevated position in relation to the LM. Here, the spatial relation denoted by the 'üst + Dat' construction is configured on a

'foreground-background' relation between the TR and LM, which differs from the protoscene spatial configuration and the 'higher-than' reading. Thus, a background sense is prompted at the conceptual level by abstracting away from the spatial construal of the protoscene. The dative particle prototypically involves movement of a TR to a position vertically elevated in relation to the LM, either with or without contact. However, the movement is de-emphasized in Background Sense since it only involves the result of the TR movement. In experiential terms, the implicature of background derives from an independently-motivated experiential correlation with vertical elevation and physical proximity. Accordingly, the TR being vertically up and proximal to the LM results in being foregrounded and thus highlighted by a spatially bigger ground figure lying or standing behind. In this configuration, the TR being in contact with the LM exerts influences on the LM by changing the colour and pattern of the background object. For instance, if "diamond" were not added on the gold necklace, the resulting pattern would be completely different from a diamond necklace, in which diamonds are highlighted more than the gold background. Owing to this experiential correlation, the implicature of Background Sense is conventionalized and instantiated in the semantic memory as a distinct sense.

Example:
Altın **üst-ü-ne** pırlanta bir kolye.
*Gold top-poss3sg-**DAT** diamond a/an necklace*
(Lit: A diamond necklace on gold.)
A diamond necklace on a gold background.

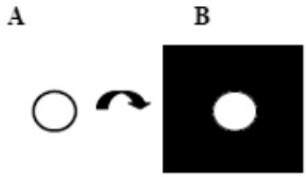


Figure 31. Background Sense

9. REFLEXIVE Sense (6 Tokens/483)

Reflexive Sense is associated with both a vertical elevation and a horizontal positioning of the TR through a movement from upright position to downward position in a reflexive manner to reach the LM (goal), which is conceptualized in a summary fashion. It translates into 'to

collapse onto / fall onto'. In a similar finding to Tyler and Evans' (2001a) analysis of 'over', the scene is conceptualized as the TR occupying multiple positions in the integration of the TR-LM configuration. Using Lindner's (1981) term "reflexivity", Tyler and Evans pointed out the dynamic character of such spatial experience. However, the actual or abstract transformation of the TR ends in a 'static' spatial configuration. As displayed in Figure 32 and illustrated in the example "Bitkince diz-ler-i **üst-ü-ne** çök-üyor İzzet", "*Exhaustedly, İzzet is collapsing on / onto his knees",* the initial upright position of the TR, the knees, is conceptualized in a horizontal and static position, being closer to the ground as a result of falling onto the LM, the knees. The ninety-degree arc-like fall down of the knees constitutes a support not only for the knees but for the whole body as well. Thus, the knees are conceptualized reflexively. In other words, the same entities can be both the TR and the LM in an abstracted spatial configuration by mediating the two temporally situated positions into a single spatial configuration.

Example:
Bitkince diz-ler-i **üst-ü-ne** çök-üyor İzzet.
*Exhaustedly knee-pl-poss3sg top-poss3pl-**DAT** collapse-pres.prog-3sg İzzet*
*(Lit: Exhaustedly, İzzet is collapsing on / **onto** his knees.)*
Exhaustedly, İzzet is kneeling down.

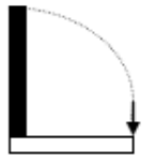


Figure 32. Reflexive Sense
(dark bar: TR in upright position; clear bar: the same TR in horizontal position constituting the LM position – adopted from Tyler & Evans, 2001a: 760)

9. Discussion

This chapter examined the spatial configurations denoted by the Turkish spatial construction 'üst + Dat' within the Principled Polysemy model. This construction was found to be polysemous with multiple senses originating from the primary sense. In parallel to Tyler and Evans' (2001a, 2003) findings, the meaning extensions identified in the semantic network of 'üst + Dat' occur in a motivated manner through cognitive semantic

processes represented in the human conceptual system and derived from human experiences with the spatio-physical world. In other words, cognitive semantic processes such as metaphor, metonymy, schematization, etc. constitute an instrument for the embodiment of human experience and the interaction with the outer world, which give rise to distinct but related senses in the conceptual system. Moreover, the empirical data from child language acquisition studies (cf. Slobin, 2008; Bowerman & Choi, 2003) lend support to the claim that the events encoded by constructions derive from human experience.

The constructional polysemy analysis of 'üst + Dat' has resulted in a semantic network, in which the spatial noun 'üst' and the dative particle '(y)E' propose a complex polysemous structure consisting of 22 distinct senses organized in a radial manner (cf. Lakoff, 1987). Eight senses (i.e. protoscene, above-and-beyond, transfer, focus-of-attention, reflexive, covering, preference and control) overlap with the senses determined in the polysemy network of 'over' in English (cf. Tyler and Evans, 2003). The different pattern of semantic organization and the additional senses not found in the polysemy network of 'over' demonstrate the cross-linguistic differences in the formalization of spatial expressions. As a result, there are multiple ways to divide space and hence multiple factors to focus on. Accordingly, languages vary in their choices of which factors are picked out.

In an analysis following from Tyler and Evans (2001a) and Evans (2010), the Turkish spatial construction 'üst + Dat' displays a highly polysemous structure with distinct non-spatial uses in a similar way to the English spatial lexemes 'on' and 'over'. However, the Turkish spatial noun 'üst' involves three different constructions in the combinations of Dative, Locative and Ablative particles which denote Proto-Goal, Proto-Location and Proto-Source senses respectively. The constructional polysemy analysis of 'üst + Dat' addressed in this study also presents a wider semantic network and a number of distinct senses not found in the polysemy network of English 'over' in Tyler and Evans (2001a, 2003). In addition to the above analyses, the functional element of the static versus dynamic scene plays a significant role in the organization, representation and interpretation of the TR-LM configurations in the system of Turkish spatial nouns.

The present study sheds light on the clear interaction of the spatial noun 'üst' and the dative particle [-(y)E] within the Principled Polysemy model (Tyler & Evans, 2001a). The prototypical meaning involves dynamism, where *üst* indicates a *dynamic spatial relation* through a motion directed at a *goal*. However, in a great number of extended senses

such as Superior, Responsible, Preference, Figure, Top, Outfit, Focus-of-Attention, Emphasis, Covering, Revealing, Psychological State, and Background, the dynamism is de-emphasized and only the resultant state of the TR movement is highlighted.

Thus, 'üst + Dat' encodes a different spatial relation, conceptual representation and mental imagery from the other spatial particles, e.g. the static locative [-dE] and the dynamic ablative [-dEn] denoting *location* and *source* respectively. As a result, the conceptual distinction between Turkish spatial case markers is pertinent to *static* versus *dynamic scenes* as functional elements depending on the verb, which assigns a particular case marker to a noun, which is different from the construal of the spatial postposition system in English.

10. Conclusion

This study constitutes an important source of reference since it provides a detailed cognitive account for the organization and representation of Turkish spatial constructions. Unlike earlier descriptive studies, the present study is grounded in the Principled Polysemy model, which sheds light on a wide array of usages and semantic systematicity in constructions with a family of distinct but related senses. Moreover, the findings of the present study provide a detailed polysemous account of the concept-form mappings for the Turkish spatial noun 'üst' and its paired spatial particle 'Proto-Goal Dative', which has not been studied so far. Overall, the results shed light on the applicability of the Principled Polysemy model developed for English spatial prepositions to the spatial nouns of a historically unrelated language such as Turkish. Evidence from child language acquisition studies and an analysis of the diachronic development of 'üst' would provide further support for the development of the extended semantic senses associated with the grammatical construction of 'üst + Dat'. As a result, further empirical research is needed for a better understanding of the polysemous structure and patterns of the spatial noun and spatial particle constructions in Turkish.

11. References

Bloomfield, Leonard. *Language.* New York: Holt, Rinehart & Winston, 1933.

Bowerman, Melissa, and Soonja Choi. "Space under Construction: Language-specific Spatial Categorization in First Language Acquisition." In *Language in Mind: Advances in the Study of*

Language and Thought, edited by Dedre Gentner, and Susan Goldin-Meadow, 387-427. Cambridge, UK: Cambridge University Press, 2003.

Bowerman, Melissa, and Soonja Choi. "Shaping Meanings for Language: Universal and Language-specific in the Acquisition of Spatial Semantic Categories." In *Language Acquisition and Conceptual Development,* edited by Melissa Bowerman, and Stephen C. Levinson, 483-511. Cambridge, UK: Cambridge University Press, 2001.

Brugman, Claudia M. *The Story of Over: Polysemy, Semantics and the Structure of the Lexicon.* New York: Garland Press, 1988.

Choi, Soonja, and Melissa Bowerman. "Learning to Express Motion Events in English and Korean: The Influence of Language-specific Lexicalization Patterns." *Cognition* 41 (1991): 83-121.

Chomsky, Noam. "Categories and Transformations." In *The minimalist Program*, edited by Noam Chomsky, 219-394. Cambridge, MA: MIT Press, 1995.

Dewell, Robert. "The Separability of German *über*-: A Cognitive Approach." In *The Construal of Space in Language and Thought,* edited by Martin Pütz, and René Dirven, 109-133. Berlin: Mouton de Gruyter, 1996.

Evans, Vyvyan. "From Spatial to the Non-spatial: The 'State' Lexical Concepts of in, on and at." In *Language, Cognition and Space,* edited by Vyvyan Evans, and Paul Chilton, 215-248. London: Equinox, 2010.

Evans, Vyvyan, and Andrea Tyler. "Spatial experience, lexical structure and motivation: the case of *in*." In *Studies in Linguistics Motivation (In the Cognitive Linguistics Research Series),* edited by Günter Radden, and Klaus-Uwe Panther, 157-192. Berlin: Mouton de Gruyter, 2004a.

—. "Rethinking English Prepositions of Movement: The Case of *To* and *Through*." In *Adpositions of Movement (*Belgian Journal of Linguistics *18),* edited by Hubert Cuyckens, Walter de Mulder, and Tanja Mortelmans, 247-270. Amsterdam: John Benjamins, 2004c.

Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press, 1995.

Göksel, Aslı, and Celia Kerslake. *Turkish: A Comprehensive Grammar.* London: Routledge, 2005.

Jarvis, Scott and Aneta Pavlenko. *Crosslinguistic Influence in Language and Cognition.* New York: Routledge, 2008.

Kreitzer, Anatol. "Multiple Levels of Schematization: A Study in the Conceptualization of Space." *Cognitive Linguistics* 8 (4) (1997): 291-325.

Kornfilt, Jacklin. *Turkish.* New York: Routledge, 1997.

Lakoff, George. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind.* Chicago, IL: Chicago University Press, 1897.

Lewis, Geoffrey. L. *Turkish Grammar (2nd Ed.).* London: Oxford University Press, 2000.

Liamkina, Olga. "Semantic Structure of the German Spatial Particle *über*." *Journal of Germanic Linguistic* 19 (2) (2007): 115-160.

Lindner, Susan. A Lexico-semantic Analysis of English Verb Particle Constructions with *out* and *up*. PhD diss., University of California at San Diego, 1981.

Mandler, Jean. "How to Build a Baby: On the Development of An Accessible Representational System." *Cognitive Development* 3 (1988): 113-136.

Mandler, Jean M. "How to Build a Baby: II Conceptual Primitives." *Psychological Review* 99 (1992): 587-604.

—. "Preverbal Representation and Language." In *Language and Space*, edited by Paul Bloom et al., 365-384. Cambridge, MA: MIT Press, 1996.

Özçalışkan, Şeyda, and Dan I. Slobin. "Codability Effects on the Expression of Manner of Motion in Turkish and English." In *Studies in Turkish Linguistics*, edited by Sumru Özsoy, Didar Akar, Mine N. Demiralp, Eser E.

Taylan, and Ayhan A. Koç, 259-270. Istanbul: Boğaziçi University Press, 2003.

Özçalışkan, Şeyda, and Dan I. Slobin. "Learning How to Search for the Frog: Expression of Manner of Motion in English, Spanish, and Turkish." *BUCLD 23 Proceedings* (1999): 541-552.

Slobin, Dan I. "Putting Things in Places: Developmental Consequences of Linguistic Typology." In *Event Representation,* edited by Jürgen Bohnemeyer, and Eric Pederson, 1-28. Cambridge: Cambridge University Press, 2008.

Slobin, Dan I. "Form-function Relations: How Do Children Find out What They are?" In *Language Acquisition and Conceptual Development*, edited by Melissa Bowermann and Stephen C. Levinson, 406-449. Cambridge: Cambridge University Press, 2001.

—. "Crosslinguistic Aspects of Child Language Acquisition." *Sophia Linguistica Working Papers in Linguistics* 35 (1994): 2-80.

Talmy, Leonard. "Lexicalization Patterns: Semantic Structure in Lexical Forms." In *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon,* edited by Timothy Shopen, 57-149. Cambridge, UK: Cambridge University Press, 1985.

—. "Path to Realization: A Typology of Event Conflation." *Proceedings of the Seventh Annual Meeting of Berkeley Linguistic Society* (1991): 480-519.

Traugott, Elizabeth C. "On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change." *Language* 65 (1989): 31-55.

Tversky, Barbara. "Places: Points, Planes, Paths, and Portions." In *Representing Direction in Language and Space,* edited by Emile Van der Zee, and Jon Slack, 132-143. Oxford: Oxford University Press, 2003.

Tyler, Andera, and Vyvyan Evans. "Applying Cognitive Linguistics to Pedagogical Grammar: the Case of *over*." In *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*, edited by Michael Achard, and Susanne Niemeier, 257-280. Berlin: Mouton de Gruyter, 2004b.

—. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition.* New York: Cambridge University Press, 2003.

—. **"**Reconsidering Prepositional Polysemy Networks: the Case of *over*." *Language* 77(4) (2001a): 724-765. Reprinted in *Polysemy: Flexible Patterns of Meaning in Mind and Language*, edited by Brigitte Nerlich, Zazie Todd, Vimala Herman, and David D. Clarke, 95-160. Berlin: Mouton de Gruyter, 2003.

Türker, Ebru. "Locative Expressions in Korean and Turkish: A Cognitive Grammar Approach." PhD diss., University of Hawaii at Manoa, 2005.

Vandeloise, Claude. *Spatial Prepositions: A Case Study in French.* Chicago: Chicago University Press, 1991.

CHAPTER SIX

INTERPRETATION OF COREFERENTIAL CHAINS IN CZECH[1]

ALENA PONCAROVÁ
CHARLES UNIVERSITY IN PRAGUE

1. Introduction

This chapter aims to present the results of my research considering the influence of the constituent structure and the information structure on coreferential chaining in Czech. Research focused on information structure has a very strong tradition in the Czech academic environment, but the Centering Theory (Grosz, Joshi and Weinstein 1995; Walker, Joshi and Prince 1998; Brennan, Friedmann and Pollard 1987, among others) could still bring new insights. That is why I decided to perform this research combining the Centering Theory with the Czech tradition of examining the information structure. My research focuses on coreferential chaining in Czech, and the tested situation is as follows: there are two noun phrases in the first utterance of the minimal context and one pronoun in the second utterance of the minimal context; the main research question is to establish the preferred noun phrase as an antecedent of the pronoun, where both information and constituent structures were manipulated. The research consists of two basic parts: the corpus research based on the Prague Dependency Treebank and the questionnaire in which almost three hundred native speakers of Czech participated.

[1] This research was performed within the subproject *Centering and Czech – syntactic analysis* (No. 26910102) at Charles University in Prague, Faculty of Arts, from the Specific University Research 2015.

2. Theoretical Background[2]

2.1. Information Structure in Czech

The Czech approach to the INFORMATION STRUCTURE was formed in the late 1930s, when Vilém Mathesius's founding paper was published in 1939. Since then, there has been massive development in this field. Vilém Mathesius presented the concept of contextual *boundness* (Mathesius 1939). He stated the initial point of the utterance (*východiště*) as "what is in the situation known or at least is evident and from the speaker starts" (Hajičová 2012, p. 60) in contrast with the nucleus of the utterance, the core (*jádro*): "what the speaker utters about or with respect to the starting point of the utterance" (Hajičová 2012, p. 60). The slightly different distinction is defined on the basis of familiarity: as what the utterance is about in the case of the topic (*základ*) and as what the utterance says about it in the case of the focus (*jádro*). It is obvious that the terminology is mixed, but the difference between the initial point of the utterance and the topic has to be distinguished, because the initial point of the utterance is anchored in the surface structure of the sentence, whereas what the utterance is about forms the topic of the utterance.

Jan Firbas (1971, 1992 etc.) introduced the term FUNCTIONAL SENTENCE PERSPECTIVE into Czech linguistics. He developed the communicative dynamism perspective covering the information structure principle. This perspective aspires to cover the dynamic character of communication. His hierarchy of degrees is a tool that captures "the extent to which the element contributes towards the development of communication" (Firbas 1971); it is considered from a conversational perspective. He distinguished the theme part of the utterance, i.e. the least dynamic element—"element or elements carrying the lowest degree of communicative dynamism within a sentence" (Hajičová 2012, p. 61), from the rheme part of the utterance, i.e. the most dynamic element.

The notion of scale can already be found in Mathesius's theory. He defined the centre and the accompanying elements (*jevy průvodní*) of both the initial point and the nucleus of the utterance. The similar scalarity perspective applies in Svoboda's concept of COMMUNICATIVE IMPORTANCE (*sdělná závažnost*) from the point of view of the intention of the speaker (Svoboda 2007).

Petr Sgall's concept of TOPIC–FOCUS ARTICULATION is another perspective considering the information structure. It represents a "formal

[2] A further-developed overview of this issue is available in Hajičová (2012).

model of functional generative description of language, namely of the representation of sentences of the underlying (so-called tecto-grammatical) sentence structure" (Hajičová 2012, p. 62). The concept of contextual boundness is incorporated into this perspective including the understanding of boundness as a characteristic of the entity that is not necessarily known from the previous context or new but "rather structured as regards the information structure" (Hajičová 2012, p. 62).

All previously introduced classifications are being applied to a single utterance. The perspective of dynamic progression across the border of two utterances is described by František Daneš's THEMATIC PROGRESSION. There are five types of progression in the discourse: (1) linear thematization of rhemes, (2) continuous theme, (3) derived themes, (4) exposition of a split rheme, and (5) thematic skip (Daneš 1974). Linear thematization of rhemes is a simple linear thematic progression: each rheme becomes the theme of the next utterance. The second type of thematic progression is the thematic progression with a continuous theme: there is a constant theme in a series of utterances. According to Daneš (1968), the first two types of thematic progression are the most frequent principles of discourse development. The third and fourth types of thematic progression are thematic progressions composed of several parts of themes or rhemes respectively. Hence, the third type is the thematic progression with derived themes: it assumes the existence of the hypertheme, from which the particular themes are derived. The fourth type is parallel to the derived themes in the thematic progression – it is the thematic progression with the split rheme, where rhemes are explicitly or implicitly doubled or multiplied; this type of thematic progression is very rare, because a discourse structure mainly based on it would be very incoherent. The last type of thematic progression is the thematic progression with thematic skip. It is based on the principle of elision; in the hierarchy of thematic development, there is one step omitted, and the listener has to be able to reconstruct the omitted link to identify the complete meaning of the text. The concept of thematic progressions can be captured in the following schemas:

Table 1. Thematic progressions

Thematic progression with...	Utterance No.1	Utterance No.2	Utterance No.3
linear thematization of rhemes	T1 – R1	T2 (R1) – R2	T3 (R2) – R3
continuous theme	T1 – R1	T1 – R2	T1 – R3
derived themes	Ta (T) – R1	Tb (T) – R2	Tc (T) – R3
split rheme	T1 – R1 (Ra + Rb)	T2 – Ra	T3 – Rb
thematic skip	T1 – R1	T2 (R1) – R2	T4 (R3) – R4

2.2. Sentence Structure in Czech

Czech is a language with a so-called free word order (with some exceptions as clitics), but the dominant and neutral word order arrangement is SVO. The Subject part of the sentence is very often omitted (under the condition of the referent's familiarity from the previous context), and the "use of personal pronoun in the function of Subject is always somehow motivated" (Zimová 1994: 59) by factors such as unclearness of the Subject of a sentence, development of the Subject phrase, emphasis of Subject, contrastive position of Subject, etc. A rich repertoire of morphological affixes offers an elaborated system to distinguish Subject from Object even if the Subject is omitted.

The Object as the second sentence function is typically pronominalised, but still present on a surface level. It is omitted significantly less often than the Subject: "[U]noccupied Subject position is usually compensated by congruent suffixes of verbum finitum in the Predicate. […] In the case of the Object […], the surface realization – it does not matter whether in the form of full expression or pronoun – seems to be the basic setting" (Zimová 1994: 92).[3] The elision of the Object as well as the repetition of full expression is what Zimová (1994) considered to be atypical.

Each type of sentence function has a typical form of expression. Based on Zimová's assumption, the typical form of the Subject is elision, while in the case of the Object it is marginal. The Object is typically expressed by a pronoun (if it is contextually bound and there is no coreferential ambiguity) or by a full nominal expression.

Table 2 shows that the Subject has a richer repertoire of forms of expression, because it could (hierarchically) be expressed by elision, pronoun and repetition of the full nominal expression, which is very rare and unnatural. In the case of the Object, the most frequent form is the pronoun, followed by elision and full nominal expression. The expression by nominal phrase of both Subject and Object is typical only in the case of the first realization of the entity in a discourse (placing on the stage); if the entity has already been present in a discourse, it is quite unnatural to use a full phrase to refer to it.[4]

[3] All citations from Zimová (1994) were translated from Czech by Alena Poncarová.
[4] Cf. Poncarová (2013) for more details.

Table 2. Subject vs. Object

	Subject	Object
Basic form of expression	elision	pronominalisation
Elision	typical	atypical
Pronominalisation	possible	typical
Full nominal expression	atypical	atypical

2.3. Coreference in Czech

In Czech, it is possible to use many variants of expressions to refer to the same entity in the discourse, e.g. a full nominal phrase, many kinds of pronouns, or by means of verbal agreement. We are able to reconstruct the coreferential chaining on the basis of agreement (i.e. grammatical coreference based mainly on morphological means) as well as on the basis of textual relations (i.e. text coreference). For the purpose of my research, I use the coreference concept to capture the respondents' approach to the development of text relations (cf. Section 3.2).

2.4. Centering Theory

The Centering Theory is a complex "local attentional state model" (Grosz, Joshi and Weinstein 1995, p. 207) based on English-language material, which uses tools such as Centres of Attention and Classification of Transitions to describe the local coherence of a text. Apart from the two mentioned tools, the Centering Theory also considers realization relationships, centre management rules and other theoretical tools. In this theory, the utterance is placed at the prominent position in the first instance: "it is an utterance and not a sentence in isolation that has centres and the same sentence when uttered in different discourse situations may have different centres, because determination of centres is related to reference resolution" (Kruijff-Korbayová and Hajičová 1997, p. 37).

2.4.1. Centres of Attention

The concept of Centres of Attention is the basic pillar of the Centering Theory. Centres of Attention are the "semantic objects" (Grosz, Joshi and Weinstein 1995, p. 208) which represent the "entities serving to link [one] utterance to other utterances in the discourse segment that contains it" (Grosz, Joshi and Weinstein 1995, p. 208). It is very important to say that

the centre is not connected with the word or any surface form – it is anchored in the cognitive concept of the entity.

There are three types of Centres of Attention: forward-looking, backward-looking and preferred. Forward-looking centres are the representation of each entity present in the discourse and are ranked according to the hierarchy of activation. These centres open the possibility of coreferential chaining, since they are possible antecedents of potential coreferential chains. The backward-looking centre is the representation of one entity that connects the current utterance with the previous context. The preferred centre is the highest-ranked element in the hierarchy of forward-looking centres; it is mainly represented by the Subject of the sentence.[5]

2.4.2. Types of Transitions

The second basic pillar of the Centering Theory is the concept of Transitions. It follows the classification of Centres of Attention, because it reflects the dynamic changes of the classification between two immediately adjacent utterances. The principle of Transitions is based on the relations between the preferred centres and the backward-looking centres of an utterance and the relations between the backward-looking centres of current and previous utterances. The typology of Transitions is shown in Table 3.

Table 3. Types of Transitions

	$Cb(U_i) = Cb(U_{i-1})$ or $Cb(U_i) = ?$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	Continue	Smooth-shift
$Cb(U_i) \neq Cp(U_i)$	Retain	Rough-shift

2.4.3. Realization relationships

Each entity can be represented in context in two variants. The first one is the direct realization; the second one is the (indirect) realization. To be directly realized, the entity has to be explicitly mentioned in the context. This is the case of the coreferential relation between two expressions referring to the same referent, typically between full phrases and pronouns (for example, *Mr. Cook* and *he*). The second type of entity representation

[5] It is very important to distinguish between *utterance* and *sentence*. Utterance is used when speaking about the information structure, whereas sentence is connected with the constituent structure.

is the realization. We can assume that the entity is realized in the discourse in case of an associative relation between two expressions in the context – for example, there is only an indirect relation between two expressions such as *house* and *door*, according to Kruijff-Korbayová and Hajičová (1997) and Grosz, Joshi and Weinstein (1995). Other types of realization are not allowed.

2.4.4. Rules and Claims

After defining the Centres of Attention and the Types of Transition, instructions as to how to treat them have to be developed. First of all, the rules and claims for the management of the Centres are presented (i.e. 1–6); then, the claim related to Transitions is reported (i.e. 7). The list is based on Grosz, Joshi and Weinstein (1995: 210).

1. Every entity from the set of forward-looking centres has to be (directly) realized in the utterance.
2. The set of forward-looking centres is partially ordered, whereas forward-looking centres are ranked according to the hierarchy of activation.
3. There is only one (unique) backward-looking centre in every utterance.
4. The backward-looking centre is strictly local. It is based on the set of forward-looking centres of the previous utterance.
5. Subject rule: the highest-ranked element of the set of forward-looking centres, i.e. preferred centre, determines the backward-looking centre of the following utterance, i.e. $Cb(U_{i+1})$.
6. Pronoun rule: if any element of $Cf(U_i)$ is realized by a pronoun in U_{i+1}, then the $Cb(U_{i+1})$ must also be realized by a pronoun; if there is no pronoun in the utterance, this rule is not applicable.
7. There are preferences among sequences of Centre Transitions: Continue > Retain > Smooth-shift > Rough-shift.

These rules and claims form the basic instructions for applying the Centering Theory in praxis. They are more useful criteria than strict conditions; sometimes they may even be contradictory: the preference for the Continue type of transition predicts the preservation of Cb from the previous utterance, but in combination with the Pronoun rule it could be problematic.

2.4.5. Discussion

There are a few complications to consider when applying the Centering Theory to authentic language material. They arise mainly from methodological and theoretical limitations. As an example, I can present (i) a limitation of a backward-looking centre in the utterance (which is not plausible considering Czech), (ii) an application of the strict locality principle, i.e. chasing only relations between immediately adjacent utterances [more specifically, "the backward-looking centre for an utterance U_i is chosen from the set of forward-looking centres of the previous utterance U_{i-1} which are claimed not to be constrained by the features of any previous utterance in that segment" (Kruijff-Korbayová and Hajičová 1997, p. 36)], and (iii) an application of the Pronoun Rule in multi-pronoun utterances.[6]

3. Survey Design

3.1. Tested Hypotheses

The Centering Theory was used as the main background theory in the survey, because it states the prediction of preference of the discourse development. For the purpose of the survey, it had to be combined with the traditional Czech approach to the information structure analysis. In spite of this fact, two hypotheses are based primarily on the Centering Theory approach:

0. Null hypotheses 1: There is no correlation between the constituent structure of the sentence and the coreferential chaining.
0. Null hypotheses 2: There is no correlation between the information structure of the utterance and the coreferential chaining.
1. There is a preference for the Subject of the previous utterance (in contrast with the Object) to be the antecedent of the pronoun in the current utterance.
2. There is a preference for the Topic of the previous utterance (in contrast with the Focus) to be the antecedent of the pronoun in the current utterance.

To examine Czech data, I used two separate methods of verification: the first is corpus research and the second is a questionnaire.

[6] Cf. Poncarová (2014) for a more sophisticated analysis or the solution proposition.

3.2. Testing Conditions

Centering Theory as a dominant background of the research set the basic testing condition: only minimal contexts of two utterances were taken into account within both the corpus research and the questionnaire. However, additional conditions had to be respected:

1. The first utterance of the minimal context must contain two noun phrases, and the second must contain a referring expression (pronoun). A setting with more than two noun phrases within the first utterance is quite common (see 3.4.1); by "two noun phrases" I mean two noun phrases serving as potential antecedents of the pronoun. If there are more than two noun phrases, the third, fourth, etc. noun phrases cannot be a part of a particular coreferential chain.
2. Both noun phrases must be explicit, and the referring expression can be omitted; in that case, it is reconstructed as a pronoun.
3. Only the coreferential relations based on text linking (i.e. only the relations not based on grammatical rules) are tested in both the corpus research and the questionnaire.
4. There is no preference (intra- or extralinguistic) for either of the noun phrases to be the antecedent of the pronoun.
5. Only personal and possessive pronouns are studied, because, in the case of other pronoun types (especially demonstrative), there is a strong preference for one of the noun phrases to be the antecedent of the pronoun. Pronouns represent a balanced set of gender and number settings.
6. According to the Centering Theory approach, coreferential relations have to be realized across the utterance border and have to be anaphoric: the pronoun has to be an anaphor, and one of the noun phrases has to be an antecedent.
7. To make the research methodologically smooth, it has to be possible to vary minimal contexts with respect to both the constituent and the information structure without the violation of the requirement of the naturalness of each expression in Czech.
8. Associative relations, i.e. the phenomena related to coreference, are ignored. Only true coreferential relations are taken into account.

3.3. Prague Dependency Treebank

For the purpose of the survey, I chose the Prague Dependency Treebank (hereinafter referred to as PDT) – approximately one million tokens, about fifty thousand sentences, as it is the only Czech-language

corpus that is fully syntactically annotated. Despite this great advantage, there are several limitations which have to be kept in mind while interpreting results, such as the dominant proportion of journalistic texts in the corpus or the methodological background of the Praguian linguistic tradition. In PDT, utterances are captured in the form of trees based on the dependency relations (valencies) and annotated within three interlinked layers (i.e. morphological layer, analytical layer and tectogrammatical layer). More information about the PDT annotation can be found in Mikulová (2005).

3.3.1. Information Structure in PDT

In PDT, there are two different ways to obtain the information structure: communicative dynamism and contextual boundness. Communicative dynamism is coded by means of nodes – i.e. the left side of the tree is dedicated to the least dynamic parts of the utterance, and the right side to the most dynamic ones. Contextual boundness is coded by assigning the attribute *tfa* (topic–focus articulation) value to every node of the utterance. The *tfa* attribute exhibits three different values: *t* for contextually bound nodes, *c* for contrastive contextually bound nodes and *f* for contextually non-bound nodes. For the purpose of this research, only *tfa* attribute values were taken into account, but both the surface and the deep position of the node were captured. The difference between *tfa* = *c* and *tfa* = *t* is not important from the Centering Theory perspective.

3.3.2. Constituent Structure in PDT

In PDT, the constituent structure is captured within the analytical layer of the annotation. Many types of constituent functions are differentiated (e.g. Subject, Object, Attribute, Adverbials, Auxiliary sentence members and Predicate), but for the purpose of this research it is important to differentiate between Subject and Object (and other constituent functions).

3.3.3. Coreference in PDT

The tectogrammatical layer of the annotation in PDT is the level where the coreferential relations are captured.[7] There are two basic types of coreferential relations: grammatical coreference and textual coreference. Grammatical coreference is devoted to the relations of coreferential

[7] Only endophoric reference was taken into account.

expressions identifiable on the basis of grammatical relations, as for example the agreement structure (i.e. reflexive pronoun, relative elements, verbal modifications with dual dependency, control constructions, quasi-control constructions, and reciprocity constructions). In contrast, textual coreference is intended to encode the relations formed within the text realization, as in *Do you think that the NATO's$_i$ decision of whether it$_i$ will expand or not will depend on Russia's attitude?*[8]

3.3.4. Corpus Research

There are 6,126 coreferential relations in PDT based on textual coreference and 2,064 of them match the set of conditions presented above. In more than 50% of these contexts, the antecedent of the pronoun is placed in a contextually bound part of the utterance; in almost 50% of the contexts, the antecedent of the pronoun functions as the Subject of the sentence:[9]

Table 4. PDT results – information structure

	to t/c	to f	sum
from t/c	1 178	853	2 031
from f	21	12	33
sum	1 199	865	2 064

Table 5. PDT results – constituent structure

	to Sub	to Obj	other	sum
from Sub	164	219	97	480
from Obj	261	203	114	578
other	517	286	203	1 006
sum	942	708	414	2 064

The measurement of coreferential relations in PDT can lead to inferring the following partial observations:

1. The data from the PDT support the hypotheses about the preference of both Subjects and contextually bound parts of the utterances to be chosen as an antecedent of the pronoun.

[8] *Myslíte, že rozhodnutí NATO$_i$, zda se {#PersPron}$_i$ rozšíří, či nikoli, bude záviset na postoji Ruska?*

[9] All results presented here are statistically significant at a 5% level of probability. In the case of the constituent structure, only the Subject-Object relationship is taken into account.

2. From the perspective of the Centering Theory, PDT shows that the Continue type of transition is the dominant one, as predicted.
3. The position of anaphors in coreferential chains seems to be more stable than the position of antecedents (comparing the *from*-parts of the tables with the *to*-parts).
4. The most frequent anaphor in coreferential chains seems to be the Object of the previous utterance.
5. Tendencies shown in PDT data are not completely convincing, although they are statistically significant.

However, it is necessary to keep in mind that claims about the antecedents' preferences are valid with respect to highly specific (restricted) arrangements of utterances as well as to specific data sets.

3.4. Questionnaire

The second part of the survey consists of an on-line questionnaire. Its purpose is to examine the interpretative part of communication with respect to coreferential chaining. Coreferential chaining interpretation was tested within minimal contexts fitting the previously presented conditions.

3.4.1. Test contexts

The contexts' design respects the set of conditions presented above (i.e. minimal contexts with two noun phrases in the first utterance, the referring expression in the second one, and a coreferential relation between the referring expression and either of the noun phrases). Due to the intended extent of the research and the lack of natural contexts fitting all the conditions, it was necessary to specifically develop artificial contexts for this survey. For this reason, the set of contexts had to be pre-tested in both grammaticality and naturalness judgments. The first proposal of contexts (i.e. the basic set) was as follows:[10]

[10] This notation, which adapts the abbreviations proposed by the Leipzig Glossing Convention, considers only the relevant grammatical meanings and categories. Subject in Czech is in all examples in Nominative. The information structure notation is not included: the pre-verbal part of the first utterance represents the Topic part, and the post-verbal part represents the Focus part.

1. *Knihovna darovala knihy fakultě.*
library(F):A;SBJ;SG donate:PSR;3SG;F book(F):OBJ;ACC;PL faculty(F):P;OBJ;DAT;SG
Bylo to pro ni výhodné.
be:PST;3SG;N it(N):SBJ;SG for she(F):OBJ;ACC;SG advantageous:NOM;SG;N
'Library$_m$ donated books to faculty$_n$. It was advantageous for it$_{m/n}$.'

2. *Pořadatel nejprve uvítal na konferenci zahraničního*
organizer(M):A;SBJ;SG first welcome:PST;3SG;M at conference(F):OBL;LOC;SG foreign:ACC;SG;M
profesora.
professor(M):P;OBJ;ACC;SG
Měl radost, že se potkali.
have:PST;3SG;M pleasure(F):OBJ;ACC;SG that each.other meet:PST;3PL;M
'The organizer$_m$ first welcomed the foreign professor$_n$ at the conference. He$_{m/n}$ was pleased that they had met.'

3. *Tomáš pozval na zápas Michala.*
Thomas(M):A;SBJ;SG invite:PST;3SG;M to match(M):OBL;ACC;SG Michael(M):P;OBJ;ACC;SG
Byl velký fanoušek fotbalu.
be:PST;3SG;M big:NOM;SG;M fan(M):SBJ;SG football(M):OBJ;GEN;SG
'Thomas$_m$ invited Michael$_n$ to the match. He$_{m/n}$ was a big football fan.'

4. *Petr řídí Volkswagen.*
Peter(M):A;SBJ;SG drive:PRS;3SG Volkswagen(M):P;OBJ;ACC;SG
Jezdí rychle.
go:PRS;3SG;M fast
'Peter$_m$ drives Volkswagen$_n$. [He/It]$_{m/n}$ goes fast.'

5. *Pan domácí srdečně přivítal Karla.*
mister(M):A;SBJ;SG landlord(M):A;SBJ;SG heartily welcome:PST;3SG;M Carl(M):P;OBJ;DAT;SG
Půjčil mu peníze.
lend:PST;3SG;M he(M):OBJ;DAT;SG money(M):OBJ;ACC;PL
'Landlord$_m$ welcomed Carl$_n$ heartily. He$_{m/n}$ had lent him money.'

6. *Dlouholetá kolegyně doporučila na volné*
longtime:NOM;SG;F colleague(F):A;SBJ;SG recommend:PST;3SG;F for free:ACC;SG;N
místo Kláru.
position(N):OBL;ACC;SG Clara(F):P;OBJ;ACC;SG
Pak jí koupila květiny.
then she(F):OBJ;DAT;SG buy:PST;3SG;F flower(F):OBJ;ACC;PL
'Longtime female colleague$_m$ had recommended Clara$_n$ for the position. Then she$_{m/n}$ bought her$_{n/m}$ flowers.'

7. *Galileo veřejně podporoval Koperníka.*
Galileo(M):A;SBJ;SG publicly support:PST;3SG;M Copernicus(M):P;OBJ;ACC;SG
Věřil, že Země se otáčí kolem Slunce.
believe:PST;3SG;M that Earth(F):SBJ;SG itself rotate:PRS;3SG around Sun(N):OBL;GEN;SG
'Galileo$_m$ supported Copernicus$_n$ publicly. He$_{m/n}$ believed that Earth rotated around Sun.'

8. *Liberální voliči podporují demokratické kandidáty.*
liberal:NOM;PL;M voter(M):A;SBJ;PL support:PRS;3PL Democratic:ACC;PL;M candidate(M):P;OBJ;ACC;PL
Bojují za občanské svobody, volný trh a
fight:PRS;3PL for civil:ACC;PL;F liberty(F):OBJ;ACC;PL free:ACC;SG;M trade(M):OBL;ACC;SG and
zmírnění chudoby.
reduction(N):OBL;ACC;SG poverty(F):GEN;SG
'Liberal voters$_m$ support Democratic candidates$_n$. They$_{m/n}$ fight for civil liberties, free trade and poverty reduction.'

9. *Policie vyšetřuje korupci.*
police(F):A;SBJ;SG investigate:PRS;3SG corruption(F):P;OBJ;ACC;SG
Její vliv ve veřejném prostoru je
she(F):POSS;NOM;SG;M influence(M):SBJ;SG in public:LOC;SG;M space(M):OBL;LOC;SG be:PRS;3SG
nezpochybnitelný.
unquestionable:NOM;SG;M
'Police$_m$ investigates corruption$_n$. Its$_{m/n}$ influence in the public space is unquestionable.'

10. *Začínající autor na veletrhu představuje*
novice:NOM;SG;M author(M):A;SBJ;SG at market(M):OBL;LOC;SG present:PRS;3SG
svůj první román.
he(M):POSS;REFL;ACC;SG;M first:ACC;SG;M novel(M):P;OBJ;ACC;SG
Velký zájem o něj nastartovala televizní
enormous:ACC;SG;M interest(M):OBJ;ACC;SG in it(N)/he(M):ACC;SG start:PST;3SG;F TV:NOM;SG;F
reklama.
advertising(F):SBJ;SG
'Novice author$_m$ presents his first novel$_n$ at the market. TV advertising started an enormous interest in [it/him]$_{m/n}$.'

11. *Známý architekt Gaudí navrhl moderní*
famous:NOM;SG;M architect(M):A;SBJ;SG Gaudí(M):A;SBJ;SG designe:PST;3SG;M modern:ACC;SG;M
dům ve Francouzské čtvrti.
house(M):P;OBJ;ACC;SG in French:LOC;SG;F Town(F):OBL;LOC;SG
Velice rychle ho proslavil.
very quickly it(N)/he(M):ACC;SG make.famous:PST;3SG;M
'Famous architect Gaudí$_m$ designed modern house$_n$ in French Town. [It/He]$_{n/m}$ made [him/it]$_{m/n}$ famous very quickly.'

12. *Personalisté minulý měsíc přijali do naší*
HR.professional(M):A;SBJ;PL last:ACC;SG;M month(M):OBL;ACC;SG hire:PST;3PL;M to we:POSS;GEN;SG;F
firmy dvacet nových zaměstnanců.
firm(F):OBL;GEN;SG twenty:ACC;PL new:GEN;PL;M employee(M):P;OBJ;GEN;PL
Jejich pracovitost je nepřekonatelná
they:POSS;NOM;SG;F hard.work(F):SBJ;SG be:PRS;3SG insuperable:NOM;SG;F
'The HR professionals$_m$ hired last month twenty new employees$_n$ to our firm. Their$_{m/n}$ hard work is insuperable.'

13. *Ronald McDonald letos jmenoval Alexandra*
Ronald(M):A;SBJ;SG McDonald(M):A;SBJ;SG this.year name:PST;3SG;M Alexander(M):P;OBJ;ACC;SG
Schramma novým generálním.ředitelem.
Schramm(M):P;OBJ;ACC;SG new:INS;SG;M CEO(M):OBJ;INS;SG
Jeho prozřetelnost zachránila firmě
he(M):POSS;NOM;SG;F providence(F):SBJ;SG preserve:PST;3SG;F company(F):OBL;DAT;SG
reputaci.
reputation(F):OBJ;ACC;SG
'Ronald McDonald$_m$ named Alexander Schramm$_n$ as the new CEO this year. His$_{m/n}$ providence preserved the company's reputation.'

14. *Palice odrazily útok seker.*
bud(F):A;SBJ;PL repulse:PST;3PL;F attack(M):ACC;PL axes(F):P;OBJ;GEN;PL
Jejich majitelé byli překvapeni.
they:POSS;NOM;PL;M owner(M):SBJ;PL be:PST;3PL surprised:NOM;PL;M
'Buds$_m$ repulsed an attack by axes$_n$. Their$_{m/n}$ owners were surprised.'

15. *Paní primářka netrpělivě očekává vrchní sestru.*
head(F):A;SBJ;SG physician(F):A;SBJ;SG eagerly await:PRS;3SG head:ACC;SG;F nurse(F):P;OBJ;ACC;SG
Její dokumenty nejsou připraveny.
she(F):POSS;NOM;PL;M document(M):SBJ;PL be:PRS;NEG;3PL prepared:NOM;PL;M
'Head physician$_m$ eagerly awaits the head nurse$_n$. Her$_{m/n}$ documents are not prepared.'

16. *Paní učitelka pochválila Aničku.*
female(F):A;SBJ;SG teacher(F):A;SBJ;SG praise:PST;3SG;F Anna(F):P;OBJ;ACC;SG
Měla radost, že se jí cvičení
have:PST;3SG;F happiness(F):OBJ;ACC;SG that * she(F):OBJ;DAT;SG exercise(N):SBJ;SG
povedlo.
*.manage.well:PST;3SG;N
'Teacher$_m$ praised Anna$_n$. She$_{m/n}$ was happy she$_{n/m}$ managed an exercise well.'

Each context is presented in four variants. The aim is to vary both Topic–Focus and Subject–Object positions of the noun phrases separately, because I did not want to mix the variables, in order to make the causality clear. The example in Table 6 shows the differences between the individual variants.

Table 6. Variants of contexts

Information structure	A	*Galileo$_{SBJ;T}$ veřejně podporoval Koperníka$_{OBJ;F}$. Věřil, že Země se otáčí kolem Slunce.* 'Galileo supported Copernicus publicly. He believed that the Earth rotated around the Sun.'
	B	*Koperníka$_{OBJ,T}$ veřejně podporoval Galileo$_{SBJ;F}$. Věřil, že Země se otáčí kolem Slunce.* 'Galileo supported Copernicus publicly. He believed that the Earth rotated around the Sun.'
Constituent structure	C	*Galileo$_{SBJ;T}$ veřejně podporoval Koperníka$_{OBJ;F}$.Věřil, že Země se otáčí kolem Slunce.* 'Galileo supported Copernicus publicly. He believed that the Earth rotated around the Sun.'
	D	*Galileem$_{OBJ;T}$ byl veřejně podporován Koperník$_{SBJ;F}$. Věřil, že Země se otáčí kolem Slunce.* 'Copernicus was publicly supported by Galileo. He believed that the Earth rotated around the Sun.'

There are two noun phrases in the first utterance: *Galileo* and *Copernicus*. Both of them have the same adequate chance to become an antecedent of the pronoun *he* in the second utterance. There are two forms of context varying the information structure (A and B) and two forms varying the constituent structure (C and D).

The A and C variants of the context are the same, but they serve as primary settings for variants B and D, which are the marked ones: A (*Galileo* – Topic, Subject; *Copernicus* – Focus, Object) for B (*Galileo* – Focus, Subject; *Copernicus* – Topic, Object) and C (*Galileo* – Subject, Topic; *Copernicus* – Object, Focus) for D (*Galileo* – Object, Topic; *Copernicus* – Subject, Focus).

3.4.1.1. Pre-testing

A basic set of contexts was presented to 20 native speakers of Czech. They had to classify the contexts in a seven-point scale of acceptability from "quite a natural Czech expression" (1) to "not Czech at all" (7); the score 4 is the central value, so it represents the neutral evaluation. Every respondent was presented all the variants of all the contexts, but the variants of the same context were not presented together. Initially, the first and third variants (A and C) of the contexts were presented; they were followed by the second variants (B), and finally the fourth variants (D) were presented.[11]

[11] The questionnaire is available at http://1url.cz/pZ7L.

It is easily predictable that the more neutral the context is, the better the evaluation results; in other words, the more natural the expression is, the lower the score is. The results considering each variant of each context were grouped on the basis of the mean and median obtained:

Table 7. Pre-testing results

group		number of variants	variants	number of contexts	contexts	action
1	both mean and median lower than 4	38	1A–4A, 6A–13A, 15A, 16A; 1B–16B; 4C, 8C–14C	7	4, 8, 9, 10, 11, 12, 13	test contexts
2	both mean and median from 4 to 4.5	3	2C, 7C, 16C	3	2, 7, 16	test contexts
3	mean or median from 4 to 4.5 & only one variant of the context received this evaluation	3	14A, 6C, 15C	3	6, 14, 15	test contexts, modified
4	mean or median above 4.5 & more than one variant of the context received this evaluation	2	5A, 5C	1	5	filling context, replaced by new context
5	both mean and median above 5	2	1C, 3C	2	1, 3	filling contexts, replaced by new contexts

Contexts belonging to the types 1 and 2 were preserved in the form in which they occurred in the first proposal, three contexts were modified (type 3), and three contexts were replaced by new ones (types 4 and 5). The contexts replaced by new ones were preserved as special kinds of fillers – they were also analysed to detect possible inconsistencies in the dataset.

3.4.1.2. Final set of test contexts

The contexts 1, 3 and 5 are new contexts developed following the example of well-rated contexts. The contexts 6, 14 and 15 were modified. The rest of the set remained the same. The contexts 17–20 are the special type of filling contexts. Although the contexts 17–19 were replaced in the set of testing contexts by new ones, they stayed as fillers. The context 20, which is an authentic context from newspapers (Lidové noviny 8-3-2013), is

a filler rather than a test context; however, because of its authenticity, I decided to include it into a list of monitored contexts. After some modifications, replacements and changes of the status of several particular contexts, the final set of contexts looks the way outlined below. Because of the simplicity of texts, I refer to this set of contexts as "test contexts", even though the last four are not literally test contexts.

1. *Pan listonoš dnes doručil objednaný*
   mister(M):A;SBJ;SG postman(M):A;SBJ;SG today deliver:PST;3SG;M ordered:ACC;SG;M
   časopis.
   magazine(M):P;OBJ;ACC;SG
   Zpozdil se.
   be.late.*:PST;3SG;M *
   'The postman_{m} today delivered ordered magazine_{n}. $[\text{It/He}]_{n/m}$ was late.'

3. *Fotbalisté Bohemians pozvali na zápas*
   football.player(M):A;SBJ.PL Bohemians(F):GEN;SG invite:PST;3PL;M to match(M):OBL;ACC;SG
   kluky.
   boy(M):P;OBJ;ACC;PL
   Holky je neobdivovaly.
   girl(F):SBJ;PL he(M):OBJ;ACC;PL admire:PST;NEG;3PL;F
   'Football players_{m} invited boys_{n} to the match. Girls did not admire $\text{the}_{m/n}$.'

5. *Pittsburští Tučňáci předstihli newyorské*
   Pittsburg:NOM;PL;M Penguin(M):A;SBJ;PL overtake:PST;3PL;M New.York:ACC;PL;M
   Rangery.
   Ranger(M):P;OBJ;ACC;PL
   Jejich fanoušci byli překvapeni.
   they:POSS;NOM;PL;M fan(M):SBJ;PL be:PST;3PL;M surprised:NOM;PL;M
   'Pittsburg Penguins_{m} overtook New York Rangers_{n}. $\text{Their}_{m/n}$ fans were surprised.'

6. *Dlouholetá kolegyně doporučila na volné*
longtime:NOM;SG;F colleague(F):A;SBJ;SG recommend:PST;3SG;F for free:ACC;SG;N
místo Kláru.
position(N):OBL;ACC;SG Clara(F):P;OBJ;ACC;SG
Pak si s ní domluvila schůzku.
then * with she(F):OBJ;INS;SG *.arrange:PST;3SG;F meeting(F):OBJ;ACC;SG
'Longtime female colleague$_m$ had recommended Clara$_n$ for the position. Then she$_{m/n}$ arranged a meeting with her$_{n/m}$.'

14. *Palice odrazily útok seker.*
bud(F):A;SBJ;PL repulse:PST;3PL;F attack(M):ACC;PL axes(F):P;OBJ;GEN;PL
Jejich zástupci bojovali statečně.
they:POSS;NOM;PL;M representative(M):SBJ;PL fight:PST;3PL;M bravely
'Buds$_m$ repulsed an attack by axes$_n$. Their$_{m/n}$ representatives fought bravely.'

15. *Paní doktorka netrpělivě očekává snaživou*
female(F):A;SBJ;SG doctor(F):A;SBJ;SG eagerly await:PRS;3SG aspiring:ACC;SG;F
kolegyni.
colleague(F):P;OBJ;ACC;SG
Její pacient není připraven na
she(F):POSS;NOM;SG;M patient(M):SBJ;SG be:PRS;NEG;3SG prepared:NOM;SG;M for
operaci.
surgery(F):OBJ;ACC;SG
'The female doctor$_m$ eagerly awaits the aspiring female colleague$_n$. Her$_{m/n}$ patient is not prepared for the surgery.'

17. *Knihovna darovala knihy fakultě.*
library(F):A;SBJ;SG donate:PSR;3SG;F book(F):OBJ;ACC;PL faculty(F):P;OBJ;DAT;SG
Bylo to pro ni výhodné.
be:PST;3SG;N it(N):SBJ;SG for she(F):OBJ;ACC;SG advantageous:NOM;SG;N
'Library$_m$ donated books to faculty$_n$. It was advantageous for it$_{m/n}$.'

18. *Tomáš pozval na zápas Michala.*
Thomas(M):A;SBJ;SG invite:PST;3SG;M to match(M):OBL;ACC;SG Michael(M):P;OBJ;ACC;SG
Byl velký fanoušek fotbalu.
be:PST;3SG;M big:NOM;SG;M fan(M):SBJ;SG football(M):OBJ;GEN;SG
'Thomas$_m$ invited Michael$_n$ to the match. He$_{m/n}$ was a big football fan.'

19. *Pan domácí srdečně přivítal Karla.*
mister(M):A;SBJ;SG landlord(M):A;SBJ;SG heartily welcome:PST;3SG;M Carl(M):P;OBJ;DAT;SG
Půjčil mu peníze.
lend:PST;3SG;M he(M):OBJ;DAT;SG money(M):OBJ;ACC;PL
'Landlord$_m$ welcomed Carl$_n$ heartily. He$_{m/n}$ had lent him$_{n/m}$ money.'

20. *Miloš Zeman po deseti letech střídá.*
Miloš(M):A;SBJ;SG Zeman(M):A;SBJ;SG after ten:LOC;PL year(M):OBL;LOC;PL replace:PRS;3SG
prezidenta Václava Klause.
president(M):P;OBJ;ACC;SG Václav(M):P;OBJ;ACC;SG Klaus(M):P;OBJ;ACC;SG
Jeho příznivci se na inauguraci sjíždějí z
he(M):POSS;NOM;PL;M supporter(M):SBJ;PL * to inauguration(F):OBL;ACC;SG *.come:PRS;3PL from
celé země.
all.over:GEN;SG;F country(F):OBL;GEN;SG
'Miloš Zeman$_m$ replaces President Václav Klaus$_n$ after ten years. His$_{m/n}$ supporters come to the inauguration from all over the country.'

3.4.2. Filling Contexts

In addition to the test contexts, the set of filling contexts forms an equally important part of the survey. The contexts used as fillers are based on authentic contexts from newspapers, magazines, Internet discussions, etc. and are modified to look like test contexts, especially with respect to the two-noun-phrase structure. I present only a few fillers here:

1. *Obraz Emila Fily pokořil rekord. Získal ho nový majitel.*
   Emil Fila's painting broke a record. A new owner acquired it.
2. *Roger Federer odehrál exhibiční zápas s Rafaelem Nadalem. Publikum mu tleskalo ve stoje.*
   Roger Federer played an exhibition match with Rafael Nadal. He received standing ovations from the crowd.
3. *Barley nabídl whiskey bratrovi. Chutnala mu.*
   Barley offered a whiskey to his brother. He liked it.
4. *Autosalon propaguje nový vůz. Jeho úspěšnost závisí na reklamě.*
   A showroom promotes a new car. Its success depends on advertising.
5. *Postřelenou dívku převezli do jiné nemocnice. Je lépe hlídaná.*
   The shot girl was transferred to another hospital. [She/It] is guarded better.

Most of the fillers are coreferentially ambiguous, but not necessarily. It is sometimes the extralinguistic influence what makes the reference clear. The filling contexts were modified to the structure of the first variant of contexts (A), because it is the most neutral one, so it does not influence the testing situation by any anomaly.

3.4.3. Variants of the Questionnaire

This part of the survey was conducted through four variants of the questionnaire. Each variant of the questionnaire, which contains 20 test contexts and 30 fillers, was designed with respect to the following rules:

- Each variant of the questionnaire contains only one variant of a particular context (A, B, C or D) to prevent the situation of interaction.
- The position of the contexts is randomly chosen, but once determined, it is preserved across the four variants of the questionnaire; they differ only in which context variant is placed in a particular position.
- Choosing the concrete variant of a particular context is not random: each variant of the questionnaire contains 5 A variants, 5 B variants, 5 C variants and 5 D variants of test contexts – there is a uniform distribution in the questionnaire.
- Filling contexts have no variants: they are represented by a unique expression following the A variant of the contexts.

3.4.4. Testing Situation

Test and filling contexts are presented individually. First of all, the respondent is shown two utterances of the minimal context and the question related to the text including some answers. The form of the question depends on the context, but it fulfils the structure of *Who/What does what?* because it respects a neutral order Agent – Predicate – Patient: for example, *Who received standing ovations from the crowd?*, *Who arranged a meeting with whom?*, *Who/What is watched better?* and so on.[12] The question is connected with a particular context, and it remains

[12] Czech is a flective language. For this reason, there have to be variants of those expressions that demonstrate different affixes by different meanings. For example, in the context *The shot girl$_m$ was transferred to another hospital$_n$. [She$_m$/It$_n$] is*

consistent in the four variants of the questionnaire. All the information contained in the second utterance of the minimal context does not have to be used in the question.

The answers are developed on the basis of František Daneš's questioning test (Daneš 1968). They are constructed as a forced choice not to draw attention to the ambiguity of coreferential chaining, which will happen if possibilities such as "both" and "none of them" are available. The order of possible answers is random, but it is preserved in all four variants of the questionnaire. In order to minimize the cognitive effort and reading complications, answers only consist of noun phrases, not complete utterances. If the noun phrase is very developed, it is not necessary to preserve all the members of the noun phrase (e.g. *famous architect Gaudí* or *modern house in French Town*). In contexts about donation, the benefactor is located on the second position in the answer; the first position is occupied by a donator (Agent). The only exception is the only authentic context *Miloš Zeman replaces President Václav Klaus after ten years. His supporters come to the inauguration from all over the country*, because there is very strong extralinguistic experience (i.e. *people know who both the new and the old president of Czech Republic are*). Therefore, this context offers the opportunity to test the respondents' attention and careful reading by reversing the order of Agent–Patient in the answers offered. Even so, the respondents chose the proper answer much more frequently than the false (261:18). The prepositions are always included in the proposed answers.

A total of 284 respondents (71 for each variant of the questionnaire) participated in the survey: both males and females, with all education levels and ages. The contexts were presented to them on-line. The questionnaire variant was assigned automatically depending on which variant exhibited the lowest number of responses. At the beginning of the testing, there were a short instruction about how to proceed through the questionnaire and three instructional items. Then each context, question and answer was presented, and the respondents had to choose an answer within 20 seconds. A time limit was included in the testing because it was not required to give an elaborate answer; the intention was to make respondents choose the most natural and first-choice answer. When respondents did not choose their answer before expiration, they were automatically directed to the next context. This situation occurred only in 74 cases from more than 5,500 measurements.

---

guarded better. – Who$_m$/What$_n$ is guarded better?, the question in Czech has to be *Kdo$_m$/Co$_n$ je lépe hlídaný$_m$/hlídané$_n$?*.

3.5. Results

Including all the measurements' conditions (e.g. information structure, constituent structure, four variants of the questionnaire, etc.), 5,680 measurements were included: 2,840 for each variable (for both information and constituent structures) and 1,420 for each variant of the questionnaire. The following tables contain the number of responses choosing a particular variant as an antecedent of the pronoun.

3.5.1. Information Structure

From the perspective of a single questionnaire, all variant differences between Topic and Focus as an antecedent of the pronoun are significant except in the questionnaire Q4. As a whole, the distinction between Topic and Focus is significant. Considering the 2,810 measurements, the coreferential chain to the Topic part of the previous utterance was dominant over the Focus part, with a difference of 274 responses. The dissimilar situation observed in the case of the questionnaire Q4 could probably be attributed to the respondents, because other possible factors such as lexical representation, text situation or context characteristics were monitored. However, it would require larger analysis to identify the crucial reason of the distinction.[13]

Table 8. Questionnaire results – information structure

Variant of questionnaire	to T	to F	difference	p-value	$p < 0.05$
Q1	321	385	64	0.016	yes
Q2	439	255	184	2.858e-12	yes
Q3	419	284	135	3.55e-07	yes
Q4	363	344	19	0.475	no
Sum	1,542	1,268	402	2.355e-07	yes

The hypothesis considering the information structure predicts that there should be a preference for Topic as an antecedent of the pronoun in the Topic setting of the context – variant A – (and implicitly for Focus in the Focus setting – variant B). Table 9 shows that this hypothesis is fully

[13] We can speculate about the reasons why the questionnaire Q4 shows a slightly different result. It could be caused by the structure of the particular contexts contained in it (but in that case it would probably be reflected in another questionnaire variant too). It could also be caused by the respondents' specificity or by many other factors. Further research would bring more evidence.

supported by the survey: the data show the preference for Topic in the Topic setting (except the Topic setting in Q2). It is also valid for Focus in the Focus setting (except the Focus setting in Q3).

Table 9. Questionnaire results – information structure hypotheses[14]

variant of questionnaire	T > T	T > F	F > F	F > T
Q1	208	145	240	113
Q2	112	237	202	143
Q3	222	131	153	197
Q4	225	129	215	138

The testing of the information structure variable showed a significant tendency towards the Topic. The Centering declaration about the preference of the Continue type of Transition was confirmed, assuming that the Subject of a particular utterance is realized in a Topic part of the utterance, which is the most neutral adjustment.

3.5.2. Constituent Structure

From the perspective of a single questionnaire, variant differences between the Subject and the Object to be an antecedent of the pronoun are significant except in the questionnaire Q4, but as a whole the distinction between Subject and Object is significant. From 2,796 measurements in total, 60% (1,701) of responses referred to the Subject. That is a confirmation of the basic Centering theory preference of Subject.

Table 10. Questionnaire results – constituent structure

Variant of questionnaire	to Sbj	to Obj	difference	p-value	p < 0.05
Q1	495	204	291	2.2e-16	yes
Q2	464	228	236	2.2e-16	yes
Q3	406	298	108	4.693e-05	yes
Q4	336	365	29	0.273	no
Sum	1 701	1 095	606	2.2e-16	yes

The hypothesis considering the constituent structure predicts that there should be a preference for the Subject to be an antecedent of the pronoun in the Subject setting of the context – variant C (and implicitly for the Object in the Object setting – variant D). Table 11 shows that this

[14] T > T stands for "Topic in the Topic setting"; T > F for "Focus in the Topic setting".

hypothesis is supported by the survey: the data show the preference for the Subject in the Subject setting (except the Subject setting in Q4). However, the Object setting exhibits an opposite tendency than I expected: there is a preference for the Subject of the previous sentence to be an antecedent of the pronoun even in the Focus setting. This conclusion supports the Centering Theory approach to the Subject of the sentence as the prominent constituent.

Table 11. Questionnaire results – constituent structure hypotheses[15]

variant of questionnaire	Sbj > Sbj	Sbj > Obj	Obj > Obj	Obj > Sbj
Q1	224	129	75	271
Q2	236	109	119	228
Q3	212	141	157	194
Q4	100	251	114	236

4. Conclusion

I have presented the traditional Czech approach to the information structure (Mathesius 1939; Firbas 1961, 1972; Daneš 1968, 1974) and the basis of the Centering Theory (i.e. Centres of Attention, Types of Transitions, Rules and Claims, and Realizations). Research considering coreferential chaining in Czech has combined the traditional Czech approach with the Centering Theory approach. My own research was briefly introduced (i.e. design, methods and results). At this point I would like to mention some of the critical points of my research. The first critical point is the crucial role of the lexical arrangement: if the minimal contexts had contained other lexical items, the results would have probably been different (with respect to some additional intra- and extralinguistic factors). The second critical point is the time limit, which could influence a respondent's answers. There are still many questions to be solved (Poncarová 2014). However, the hypotheses following the Centering Theory's predictions were mostly supported. I can assume that the data obtained supported the hypothesis of a preference of Subject and Topic of the previous utterance to be the antecedent of a pronoun in the second utterance. Therefore, there is a strong preference to develop discourse on the basis of the Continue type of transition. From the point of view of coreferential chaining, the differences between the particular variants of contexts are not as clear as, for example, in English, because Czech is a

[15] Sbj > Sbj stands for "Subject in the Subject setting"; Sbj > Obj stands for "Object in the Subject setting".

free-word-order language. In case of contradictory tendencies based on both Topic and Subject preferences (i.e. Subject in the Focus part and Object in the Topic part), there is a strong preference for the Subject to be an antecedent of the pronoun, even if it is in the Focus part. That is why I can predict that the constituent structure is a stronger factor than the information structure – but it is necessary to consider this claim to be only a basic general conclusion; a more developed analysis would be required to make a solid prediction.

5. List of Abbreviations

5.1. Text

1–16	particular context (1–20 in case of final set of contexts)
A, B, C, D	variant of particular context
Cb	backward-looking centre
Cf	forward-looking centre
Cp	preferred centre
F	focus
OBJ	object
PDT	Prague Dependency Treebank
Q1–Q4	variant of questionnaire
SBJ	subject
T	topic

5.2. Notation

()	lexical inherent category
1, 2, 3	first, second, third person
A	agent-like argument of the verb
ACC	accusative case
DAT	dative case
F	feminine gender
GEN	genitive case
INS	instrumental case
LOC	locative case
M	masculine gender
N	neuter gender
NEG	negation
NOM	nominative case
OBJ	object

OBL	oblique
P	patient-like argument of the verb
PL	plural
POSS	possessive
PRS	present
PST	past
REFL	reflexive
SBJ	subject
SG	singular

6. References

Brennan, Susan, Marilyn Friedman and Carl Pollard. A Centring Approach to Pronouns. Stanford: Proceedings of the 25th Annual Meeting of the Association of Computational Linguistics, 1987.

Daneš, František. Functional Sentence Perspective and the Organization of the Text. In Papers on Functional Sentence Perspective, 106-128. Prague: Academia, 1974.

—. "Typy tematických posloupností v textu: (na materiále českého textu odborného)." Slovo a slovesnost 29 (1968): 125-141.

Firbas, Jan. Functional sentence perspective in written and spoken communication. New York: Cambridge University Press, 1992.

Firbas, Jan. "Topic and Comment: A Study in Russian and General Transformational Grammar". Journal of Linguistics 7 (1971): 91-101.

Grosz, Barbara, Aravind Joshi and Scott Weinstein. "Centring - A Framework for Modeling the Local Coherence of Discourse." Computational linguistics 21 (1995): 203-225.

Hajičová, Eva. "On scalarity in information structure." Linguistica Pragensia 22 (2012): 60-78.

Kruijff-Korbayová, Ivana and Eva Hajičová. "Topics and Centres: A comparison of the salience-based approach and the Centring theory." The Prague Bulletin of Mathematical Linguistics 67 (1997): 25-50.

Mathesius, Vilém. "O tak zvaném aktuálním členění věty." Slovo a slovesnost 5 (1939): 171-4.

Mikulová, Marie. Anotace na tektogramatické rovině Pražského závislostního korpusu: Anotátorská příručka. Prague: Charles University in Prague, 2005.

Poncarová, Alena. Antecedent zájmena třetí osoby: tendence v současné češtině. Prague: Charles University in Prague, 2013.Unpublished thesis.

—. Centra pozornosti v teorii centeringu. In Lingvistika Praha, http://lingvistikapraha.ff.cuni.cz/node/206, 1-13. Prague: Charles University in Prague, 2014.
Svoboda, Aleš. Brněnská škola funkční větné perspektivy v pojmech a příkladech. Ostrava: University of Ostrava, 2007.
Walker, Marilyn, Aravind Joshi and Ellen Friedmann Prince (eds.) Centring Theory in Discourse. Oxford: Clarendon Press, 1998.
Zimová, Ludmila. Způsoby vyjadřování větných členů v textu: konkurence pojmenování, pronominalizace a elize. Ústí nad Labem: Jan Evangelista Purkyně University, 1994.

CHAPTER SEVEN

A HYPOTHESIS ABOUT THE ORIGIN OF META-SYMBOLS AND SUPERORDINATE CATEGORIZATION

CIRO ANTUNES DE MEDEIROS
UNIVERSIDADE ESTADUAL DE CAMPINAS, SÃO PAULO

This chapter is dedicated to Professors Ruth Vasconcellos and Edson Françozo.

1. Introduction

According to archaeological data, modern human behaviour started around 75,000 years ago (Mellars 2006). The hallmark of such behaviour is full cognitive fluidity – something never achieved by previous human behaviour. The transition to modern human behaviour would have been caused by the integration of different types of specialized human intelligence. Supposedly, that integration would have been caused by the emergence of a general intelligence capable of accessing cognitive processes from different specialized intelligences. Previous human behaviour would have resulted from a general sort of intelligence not capable of doing that. Indeed, behaviours caused by relations between different intelligences, as the use of food for establishing social relationships, would have remained very simple (Mithen 1996: 178-187). Therefore, the transition to a modern human behaviour would be associated to the appearance of modern symbolic life and general-purpose syntactic language – the Cultural Palaeolithic Revolution (Mithen 1996). However, the origins of the current human species occurred much earlier, between 200,000 and 150,000 years ago (Mellars 2006). This chapter proposes an explanation for the beginning of the Cultural Palaeolithic Revolution and for the time delay between the appearance of the species and the Palaeolithic Revolution. The main premise is the idea that the cognitive skills that support language were already present in our species

since its origin. However, such skills can only be fully developed in the presence of proper cultural stimuli, and such stimuli were not present when anatomically modern humans appeared. The chapter tries to explain the cultural catalysing event that provided such cultural stimuli and, therefore, allowed humans to mobilize their cognitive capacities to deal with symbolic issues and develop syntactic language.

The nature of the explanation built along this chapter tries to be as close to the nature of the experimental contexts of studies about language development as possible. Such experimental contexts provide the elements of the explanation, so they should be presented first.

The mentioned studies dealt with the conceptual development of langauge. They focused on the development of categorization capacities, especially on the development of basic- and superordinate-level categorization and on how language is affected by such capacities. Children from different cultures throughout the world display similarities in some fundamental aspects of language and cognition development, disregarding whether caretakers start to speak directly to them when they are a few months old or only after they begin to speak the first words. In spite of cultural differences, children start to produce words at roughly the same age, and their first words are predominantly nouns. At the end of the first year, infants can appreciate taxonomic categories (e.g. dog, cat, etc.) and property-based concepts (e.g. red, soft, etc.). However, after children begin to produce their first words, the learning of categories gets linked to the learning of new words (Waxman 1999: 273; Waxman 2002: 102-107), since exposure to a new word highlights and facilitates object categorization (Waxman & Booth 2000: 274; Klibanoff & Waxman 1998; Hall & Waxman 1993).

In the early acquisition period, children expect that a new name will highlight commonalities among objects independently of its grammatical form (Waxman & Booth 2000: 774). During this period, nouns are the most prevalent type of word in the children's lexicon, and children link nouns to object categories, especially to basic-level ones (Waxman 2002: 109). When children face an unknown word highlighting a new category, they prefer to pay attention to the shape of the object as the defining feature of the taxonomic category (Hall & Waxman 1993: 1551; Becker & Ward 1991; Ward et al. 1991). This can explain why children prefer basic-level categories. The different items of basic-level categories have shapes that are not clearly distinct from one another (Murphy 2002: 349), and a category is easily recognized according to the average shape of its items. The same does not apply to superordinate categories. Items of a

superordinate category can present a great diversity of shapes (Rosch et al. 1976: 405):

> "In general, the basic level of abstraction in a taxonomy is the level at which categories carry the most information, possess the highest cue validity, and are, thus, the most differentiated from one another" (Rosch et al. 1976: 383).

Although preschool children tend to relate new names to items of basic-level categories (Waxman 1999: 245), such tendency can be obviated. When properly stimulated, one-year-old children can include items with clearly distinct shapes from one another in the same category, forming superordinate categories. In experimental contexts, preschool children can form a superordinate category when the different category items are highlighted by a common word (Waxman & Markow, 1995; Murphy, 2002: 349). Experimental results indicate that the incapacity to form superordinate categories is correlated to the ability to produce only nouns. The children that are able to form superordinate categories are also capable of producing nouns and words from classes such as *verb* or *preposition*. However, such results are insufficient to distinguish between the factors that are causes from those that are effects (Waxman & Markow 1995: 290).

2. Superordinate Categorization and Human Language

In experimental contexts, two-month-old or older children present attention decrease when they face constant or repetitive visual patterns. In such contexts, an increase in the attention dedicated to a new pattern could be interpreted as evidence of habituation to the old one. It suggests that children learned the old pattern (Fantz 1964). That idea has been used to support several experiments that try to test the capacity of children to form taxonomic categories (Quinn & Eimas, 1996: 189-190).

In the experiments performed by Waxman and Markow (1995), children were exposed to four items of a superordinate category during the habituation phase. In the experiments, the exposure to each item lasted 30 seconds. The needed stimulus to form superordinate categories was the simultaneous exposure to a common noun - which played the role of the name of the category - and the items of the category during the habituation phase. After the end of the habituation phase, the test phase started. During the test phase, children were simultaneously exposed to two items: an item of the category used during the habituation phase and an item of the new category. Children's preference to manipulate the item of the new category

was considered evidence that children formed a representation of the old category (i.e. the one presented during the habituation phase) (Waxman & Markow 1995: 281-290).

Other similar experiments were made using instrumental music melodies, instead of words, as the superordinate category. Such experiments produced similar results to those made by Waxman & Markow (1995) (cf. Roberts & Jacob, 1991; Roberts & Cuff, 1989). These are very interesting data because instrumental music melodies are not names of superordinate categories in human language. Thus, it is possible to elicit a superordinate category without the symbols used to communicate the categories. Moreover, these data raise some questions about what else could elicit superordinate categories. Above all, the main question is: could a superordinate category be elicited by the stimulus created by the subjects that were not trying to elicit the category or do not even master more than one level of categorization?

The mastery of superordinate categorization is very important for human language development because word classes are superordinate categories. Each verb is clearly distinguishable from others. Moreover, each verb is an item of a common category (verbs). Thus, the category of verbs is a category composed by clearly distinguishable items – a superordinate category. The same applies to the remaining word classes. Therefore, the capacity to comprehend superordinate categories is used to comprehend what a grammatical class is.

Once the mastery of superordinate categorization supports the ability to include many clearly distinguishable items in the same category, it could also be related to the acquisition of a very important human-language feature. By thinking in words as categories of meanings, the ability to form superordinate categories would be a process analogous to the capacity to deal with the perspectival nature of linguistic symbols. An object can be simultaneously a flower, a rose or a gift, because the perspectival nature of linguistic symbols multiplies indefinitely the meanings of symbols (Tomasello 1999: 107). Therefore, the capacity to form superordinate categories would have been a necessary step for the emergence of the typical human language, since such capacity is required for the comprehension of grammar and for the comprehension of the perspectival nature of linguistic symbols.

However, syntactic language did not emerge at the beginning of the species (Bickerton 2007: 520) or even before the Palaeolithic Cultural Revolution (Mithen 1996). It is very probable that the Palaeolithic Cultural Revolution was correlated with the origin of modern human language and that both were in turn correlated with a great and abrupt improvement in

the human creativeness and capacity to innovate (Mithen 1996). If language is assumed prior to the Cultural Revolution as devoid of grammar and of the perspectival nature of linguistic symbols, the mastery of superordinate categorization was not necessary to learn the cited language.

3. Language before Syntax and Basic-Level Categorization

Mithen (1996) divided the history of anatomical modern humans as a species in two periods: before the Palaeolithic Cultural Revolution and after the Palaeolithic Cultural Revolution. Humans from the first period are called "Early Modern Humans"; the archaeological registers they left are very similar to the ones left by Earlier Humans (e.g. *Neanderthals* and *Homo erectus*), including the number of individuals per group and the low diversity of stone tools (Mithen 1996: 131-151). Mithen (1996) named humans from the second period "Modern Humans". The fossil register left by Modern Humans suggests they had the same modern mind and cognitive-function integration we have today. Their group size became bigger than that of their predecessors; they had symbolic artefacts, improved the design of stone tools according to the particularities of their local necessities and produced a great diversity of stone tools. All these innovations also suggest a great difference between the language of Early Modern Humans and the language of Modern Humans, indicating that the latter should have been very similar to our current human language (Mithen 1996). Bickerton, among others, hypothesized that the great difference between those two languages is the absence of syntax in the language of Early Modern Humans (Jackendoff 1999).

Some decades before the appearance of modern theories about the origin of current human language, Wittgenstein criticized St. Augustine's description of human language. According to St. Augustine, each word has a meaning, and this meaning is the object the word substitutes. Wittgenstein (1968: 2-3) defended that what St. Augustine described is a communication system that is not what we call language but fits well with a representation of a "more primitive language" than current human language.

A language in which each word has a meaning is not a language in which each word has multiple possibilities of meaning. Therefore, such language is not endowed with the perspectival nature of the linguistic symbols described by Tomasello (1999) and does not demand the mastery of superordinate categorization. If we assume that the "more primitive language" was the language produced by humans before the Palaeolithic

Cultural Revolution (and before syntax), then it can be assumed that it was supported by basic-level categorization. This suggests basic-level categorization evolved before superordinate categorization. The idea that basic-level categorization is phylogenetically more basal than other levels of categorization was explicitly defended by Rosch et al. (1976):

> "If basic categories are the level of abstraction at which it is generally most useful to refer to objects, one would expect, in the evolution of languages, that names would evolve first for basic level objects, spreading both upwards and downwards as taxonomies increased in depth" (Rosch et al. 1976: 407).

> "The same work predicted "basic object categories should be the basic classifications made during perception, the first learned and first named by children, and the most codable, most coded, and most necessary in the language of any people" (Rosch et al. 1976: 435).

The idea that each word has one meaning, and this meaning is the object the word substitutes, describes countable nouns better than any other kind of word. Moreover, the "more primitive language" evolved gradually during hundreds of thousands of years before the origin of syntax; therefore, it is possible that the cognitive capacities that supported the use of that language suffered more selective pressures than the ones that support current human language.

According to the above assumptions and the principle of *Ontogeny recapitulates Phylogeny*, it would be expected that the products of the phylogenetically most basal language would appear before the products of the syntactic language.

Studies show that the conceptual-linguistic links between countable nouns and object categories are developed earlier in the mind and are very similar across cultures. While the conceptual-linguistic connections between adjectives and object properties are developed later (Bloom 2001: 1095-1097) and are more variable across cultures (Waxman 1994: 232), other kinds of words that are used to express particular forms of meaning are also more variable across different languages (Waxman & Markow 1995: 293). Children seem to be more skilled to learn countable nouns first, even in languages in which the greater part of the input takes the form of verbs – as in Korean language (Waxman & Markow 1995: 261).

Moreover, it would be expected that the cognitive skills that supported more basal human languages were learned earlier and suffered more selective pressures than the cognitive skills that support current human language. The main uses of a language where one word has just one

meaning do not require the skill to master more than one level of categorization.

Basic-level categories are the first taxonomic categories formed by children; in other words, children start to categorize without making hierarchical structures (Murphy 2002: 327). In particular, three-year-old children show difficulties using items from superordinate categories. They avoid coordinating two categories and forming hierarchical structures, even when the superordinate category is presented and understood before the lower-level category. It seems that children have problems understanding how one thing has two different names (Markman et al. 1980: 238-239). In addition, adults are quicker and prefer to identify objects belonging to basic-level categories (Waxman 1999: 235-236; Murphy 2002; Graham et al. 1998: 105; Hall & Waxman 1993: 1567). Such problems and preferences are evidence of the differences of evolution between the capacity to categorize without making hierarchical structures and the capacity to form superordinate categories. It is probable that the neural circuitries employed to deal with basic-level categories are not entirely the same as those employed to deal with hierarchical structures of categories. Moreover, it is probable that the neural circuitries employed to deal with basic-level categories are developed earlier. From an evolutionary perspective, the two capacities seem to belong to two distinct stages.

4. The Hypothesis

Waxman proposed that (1) children start language acquisition endowed with a general expectation to treat any kind of word as a name of an object category based on data, implying that this phase of language acquisition is less variable across languages than the following phases, and that (2) in a second moment, children develop a trend to link specific forms of meaning to specific kinds of words (Waxman 1999: 253). Assuming that this proposition is correct, and following the principle of *Ontogeny recapitulates Phylogeny*, the "more primitive language" could have been learnt just at the first stage of language acquisition proposed by Waxman. The more primitive language could be supported by basic-level categories, since it served to the purpose of linking nouns to objects. Furthermore, it seems that basic-level categorization does not require a language to be developed, since children can form basic-level categories without the words eliciting the categories (Waxman & Markow 1995; Graham et a. 1998: 104). Thus, such cognitive function should have preceded syntax

and should have been incorporated by the "more primitive language" when it emerged.

The hypothesis starts from the idea that the human language prior to syntax would be like the "more primitive language" described by Wittgenstein in *Philosophical Investigations*, in which "every word has a meaning" and the ability to recognize superordinate categories is considered less important. In this more basal language, the categories predominantly represented would be those included in the basic level. However, it is from such language that the ability to form superordinate categorization emerged.

The experimental results from Robert & Jacob (1991) opened the possibility that categories could be elicited by non-linguistic factors influencing attention. This suggests that it is cognitively possible for children to form superordinate categories without the stimulus from people who already master superordinate categorization, and this is probably what happened (in the cultural catalysing event). Within each human group, each object should have just one name in the scenario of the catalysing event, in the case of the objects that were named. It is not expected from those that do not master superordinate categorization to create different names for an object or to give the same name to different objects. Moreover, stone tools – very important objects – did not present great shape variation across different human groups (Mithen 1996: 123-132; Leakey 1994). Because of the importance of stone tools[1], it is very likely that they were given a name by those people. If the tool name varied across human groups, due to geographic isolation, the meeting of different groups would have been the source for a very important stimulus: one object associated to different symbols. However, in the experimental context, what happens is that one symbol or one melody is associated to different objects. The melody is not exactly a symbol, but it can elicit a superordinate category. Probably, when a common melody is associated to different objects, it becomes a symbol, i.e. a representation of the superordinate category.

In the scenario of a meeting between different human groups, there were common objects (i.e. tools with a similar shape) and different vocalizations (i.e. variations of the tool name across groups). In the setting of the categorization experiments, there are different objects playing the role of the category item, and a common vocalization playing the role of the category elicitor. In order to use the scenario of the experiments to

[1] The mastery of stone-tool making is considered one of the two most important survival factors for humans during the Pleistocene (James 1989).

explain the catalysing event, the tools should play the role of the elicitor, while the different vocalizations should be considered as category items. Two-year-old children have the ability to use objects as symbols (Rakoczy et al. 2005: 73). Therefore, it could be possible for an object to elicit categories. If a given object can highlight a category as non-linguistic sounds can do, then the category could also be elicited by an object. In this case, the tools could have played the role of elicitor.

The hypothesis scenario is that of the gathering of different human groups. Moreover, periods of geographic isolation are needed to explain the differentiation of tool names. These two hypothesis elements could have existed if the distribution of human population had followed a pattern of geographic expansion periods interspersed with periods of geographic retraction. Archaeological data could indicate if such pattern could have happened.

Fossil evidence suggests the presence of anatomically modern humans between 198,000 and 194,000 years ago in the region of Omo River, Africa (McDougall et al. 2005), while evidence from mitochondrial DNA suggests that the origin of species is located in Africa around 150,000 years ago (Cann et al. 1987; Oppenheimer 2012; Templeton 2007). Another study with mitochondrial DNA suggests that the geographic expansion out of Africa that originated current human populations occurred around 60,000 years ago (Behar et al., 2008). However, that expansion was not the first for anatomically modern humans to leave Africa. The first dispersal out of Africa and towards the Middle East occurred around 100,000 years ago (Ambrose 1998). However, another study suggests that the occupation of the Middle East started approximately 120,000 years ago and ended around 90,000 years ago (Oppenheimer 2012). The estimated size of human groups at the time of that first expansion was 25 individuals per group (Harpending et al. 1993).

The differentiation of tool names could have occurred during the period of oxygen-isotope stage 5. During that period, human demographic expansion was part of the expansion of the Afro-Arabian biotic community (Ambrose 1998: 625). Probably, human demographic expansion followed the expansion to tropical areas during warmer periods. If such expansions occurred across the whole African continent, the probability of many episodes of tool name differentiation becomes higher. Once there was a considerable diversity of tool names across the continent, a period of contraction of the tropical areas would facilitate the conditions for the cultural catalysing event. The shrinking of the tropical areas could have led different groups to migrate to common areas – the areas remaining fertile and plentiful of resources. Such migration pattern could

have led to the aggregation and fusion of different human groups. As evidence of such aggregation, an increase in the size of human groups would be expected. Thus, as the aggregation allowed the cultural catalysing event, a correlation between the increase in the size of human groups and the appearance of modern human behaviour would be expected. After the end of oxygen-isotope stage 5, the shrinking of tropical areas did happen.

There is a considerable possibility that a volcanic winter was responsible for a reduction in resources that led to a human-population bottleneck (Ambrose 1998: 632-635). The volcano responsible for that event would have been Toba (in northern Sumatra, Indonesia) and its eruption, or multiple eruptions, would have been significant enough to change world climate (Zielinski et al. 1996). The event caused a reduction of solar transmission between 10% and 0.001% of normal sunlight (Chesner, et al. 1991: 202). "Toba massive eruption would have caused a reduction in sunlight levels for several years following the eruption" (Ramaswamy 1992), and it would have produced 200 Km^3 of pyroclastic deposits in the period between 77,000 and 69,000 years ago. A study predicted that such eruption would have led to a temperature reduction of 3.5ºC in the northern hemisphere and 5ºC in the areas covered by the smoke cloud (Rampino & Self 1992: 50-51).

> "The first 200 years of this stadial event are marked by indicating high atmospheric dust concentrations, probably from Aeolian erosion due to reduced vegetation cover and sediments exposed by a drop in sea level" (Ambrose, 1998: 633).

The scenario fits with a reduction of the geographic areas that were plentiful of resources. It could have led to migrations to the warmer areas that remained productive. The idea of group aggregation requires a scenario in which some areas suffered more impact than others. It would make the migration fluxes converge to the less impacted areas.

In the human context, Toba should have been responsible for, at least, six years of reduction of plant biomass and for a disastrous famine (Ambrose, 1998: 634). The oxygen-isotope stage 4, related to the period between 74,000 and 60,000 years ago, indicates a colder and more arid period than its predecessor during all its period, which corresponds to a period of human-population bottleneck (Ambrose 1998: 625).

Most of the estimates for the human population during the bottleneck indicate between 500 and 10,000 individuals (Harpending et al. 1993: 495; Rogers & Harpending 1992: 565), while the smallest number encountered among the estimates is 4,000 individuals for 20,000 years. It is likely that

the survivors were confined to tropical refuges (Ambrose 1998). Thus, it is very likely that migration episodes of different groups to common areas (e.g. tropical refuges) occurred.

In the same period the human-population bottleneck happened (i.e. 60,000–80,000 years ago), there was a burst of technological innovation in southern Africa: the industries of Still Bay emerged in the period between 75,000 and 68,000 years ago (Jacobs et al. 2008: 735). The sites in Still Bay show much resemblance to the sites from the Upper Palaeolithic Revolution in the period between 45,000 and 50,000 years ago, i.e. tool-design diversity, personal ornaments and abstract art (Mellars 2006: 9382). Several authors have argued that an increase in subpopulation densities and in inter-group interactions would be correlated to the appearance of the Upper Palaeolithic Revolution behaviour (Powell et al. 2009: 1299).

The hypothesis requires the size of the human groups in Still Bay to be similar to the size of the human groups from the Upper Palaeolithic Revolution period, contrasting with the size of Early Modern Human groups. The size of human groups after the Palaeolithic Revolution was larger than the size of groups before the Revolution. Such contrast is known as Puzzle 5 (Mithen 1996: 134). The growing in size combined with the reduction of usable areas makes an aggregation of different groups probable. Such aggregation would have been the cause of the cited increase of group size and the source of different names (symbols) for the same tools. As the items of the category elicited by the tools are the symbols, the use of a common tool as the symbolic representation of the superordinate category would be the use of a meta-symbol. Eventually, when this symbolic representation was renamed with a vocalization (i.e. a common name), a vocalic name would be given to a superordinate category composed by words.

After the near extinction of the species, there was another period of geographic expansion that replaced the populations of *Homo erectus* and *Homo sapiens* – according to the replacement hypothesis (Rogers & Harpending 1992: 565). There were possibly two new expansions between 80,000 and 60,000 years ago. The human groups that made such expansions (known as Clade L2 and Clade L3) overpowered all prior human groups (Behar et al. 2008: 1130). These subsequent expansions were related to modern technology and were driven by culture (Harpending et al. 1993: 495).

Therefore, the change of the size of human groups and the change of technology are very likely related to the survival in the near-extinction period, and both the survival and the changes are related to the new symbolic life of those groups. If we assume that the new symbolic life

includes syntax, the abrupt change could not have happened without the emergence of superordinate categorization and meta-symbols. The aggregation of human groups would have provided the necessary cultural stimuli to trigger the catalysing event: tools from the same type with clearly distinguishable names due to previous geographic isolation. The exposure to the same object that is named with different vocalizations together with the capacity to use this object as a symbol would have been the trigger for the creation of meta-symbols and for the inclusion of clearly distinguishable items in a given category, i.e. the emergence of superordinate categorization and meta-symbols.

5. Discussion: Targeting possible Future Experiments

The key condition of the hypothesis is the capacity to use a common object as the elicitor of superordinate categories. Traditional experiments about the learning of superordinate categories used vocalizations as elicitors and objects as category items. In order to test the key condition of the hypothesis, an inversion should be made: the items should be vocalizations and the elicitor should be a common object. For example, it is possible to verify if a child is familiarized with a category of vocalizations using the head-turn procedure.

The experiment must contain the simultaneous exposition to a common object – playing the role of the elicitor – and to the vocalizations – playing the role of the items in the category. One possibility is to divide the children into two groups; the experimental group would have a common object as the elicitor, while the control group would have different objects playing the role of the elicitor.

If the experiment started with a familiarization phase divided into four trials, the experimental group would be exposed to a set of slightly different vocalizations and to the common object, while the control group would be exposed to the same set of vocalizations and to four types of objects – one in each trial. The objective of such design would be to guarantee that the common object was the only possible cause of potential differences between the two groups. However, if each child completed the task only in one of the groups (in the control group or in the experimental group), there would exist the possibility that the result differences could be attributed to the individual differences of the children.

A child that is capable of forming the category independently of which group he or she belongs to would be a *false positive*; only children capable of forming the category in the experimental group, but not in the control group, would contribute to corroborate the hypothesis. If a child presented

the same reaction (to form or not to form the category) in both groups (i.e. experimental and control), such subject should be identified as a subject contributing to falsify the hypothesis. Therefore, one way to avert false positives would be to include each child in both experimental and control groups and to create two sets of vocalizations and objects. Each child would undergo a familiarization phase with each set, thus making it possible to contrast the two reactions. The experiment would be completed with the head-turn procedure that is used to test if the category had been formed or not.

However, even if experimental results showed that objects could elicit superordinate categories, such results would not be better than indirect evidence: they would become a small and simplified piece of a huge puzzle. They would show their real value, i.e. as a useful mistake or as a possible correct hypothesis, only when put together with the remaining pieces. Undoubtedly, there is a lot of work to be done.

This chapter integrates experimental data about the co-development of language categorization with archaeological data about human evolution in order to create a hypothesis. As a goal of the chapter, the elements of the hypothesis had to be those from an experimental context in language development. To use such elements, the hypothesis should be the source for new experiments and for the result prediction of such experiments. The described process shows how the author understands research about language evolution. It can be described as a cycle that (i) integrates experimental data from developmental studies with data from research on human evolution in order to create new evolutionary hypotheses, (ii) uses the new evolutionary hypotheses to create new developmental experiments, (iii) integrates the results of the new experiments with data from previous developmental studies and with research on human evolution in order to create new evolutionary hypothesis, and so forth.

6. References

Ambrose, Stanley. H. “Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans.” Journal of Human Evolution 34 (1998): 623–651.

Becker, Angela. H., Ward, T. B., “Children’s use of shape in extending novel labels to animate objects: Identity versus postural change.” Cognitive Development 6 (1991): 3-16.

Behar, Doron. M., Richard Villems, Himla Soodyall, Jason Blue-Smith, Luisa Pereira, Ene Metspalu, Rosaria Scozzari, Heeran Makkan, Shay Tzur, David Comas, Jaume Bertranpetit, Lluis Quintana-Murci, Chris

Tyler-Smith, R. Spencer Wells, Saharon Rosset and The Genographic Consortium. "The Dawn of Human Matrilineal Diversity." The American Journal of Human Genetics, 82 (2008): 1130–1140.

Bickerton, Derek. "Language evolution: A brief guide for linguists". Lingua 117 (2007): 510–526.

Bloom, Paul. "Précis of How Children Learn the Meanings of Words." Behavioral and brain sciences 24 (2001): 1095–1103.

Cann, Rebecca L., Mark Stoneking and Allan C. Wilson. "Mitochondrial DNA and human evolution." Nature 325 (1987): 31–36.

Chesner, C. A., W. I. Rose, A. Deino and R. Drake. Westgate, J. A. "Eruptive history of earth's largest Quaternary caldera (Toba, Indonesia) clarified." Geology 19 (1991): 200–203.

Fantz, Robert L. "Visual experience in infants: Decreased attention to familiar patterns relative to novel ones." Science, New Series 164 (1964): 668–670.Graham, Susan. A., Rachel K. Backer, Diane Poulin-Dubois. "Infant's expectations about object label reference." Canadian Journal of experimental Psychology. 52 (1998): 103-112.

Hall, D. Geoffrey, and Waxman, Sandra. R. "Assumptions about word meaning: Individuation and basic-level kinds." Child Development 64 (1993): 1550–1570.

Harpending, Henry. C., Stephen T. Sherry, Alan R. Rogers and Mark Stoneking. "The genetic structure of ancient human populations." Current Anthropology 34 (1993) 483–496.

Jackendoff, Ray. "Possible stages in the evolution of the language capacity." Trends in Cognitive Sciences 3 (1999): 272-279.

Jacobs, Zenobia, Richard G. Roberts, Rex F. Galbraith, Hilary J. Deacon, Rainer Grün, Alex Mackay, Peter Mitchell, Ralf Vogelsang, Lyn Wadley. "Ages for the Middle Stone Age of southern Africa: implications for human behavior and dispersal." Science 322 (2008): 733–735.

James, Steven. R. "Hominid use of fire in the lower and middle Pleistocene." Current Anthropology 30 (1989): 1-26.

Klibanoff, Raquel Stote, Sandra R. Waxman. "Preeschoolers'acquisition of novel adjectives and the role of basic-level kind". In A. Greenhill et al. (Eds.), Proceedings of the 22nd Boston University Conference on Language Development 442-453. Somerville, MA: Cascadilla Press, 1998.

Leakey, Richard. "The Origin of Humankind." New York, NY: BasicBooks: 1994.

Markman, E. M., M. S. Horton and A. G. Mcclanahan. "Classes and collections: Principles of organization in the learning of hierarchical relations." Cognition 8 (1980): 227–241.

Mellars, Paul. "Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model." PNAS 103 (2006): 9381– 9386.
McDougall, Ian., Francis H. Brown and John G. Fleagle. "Stratigraphic placement and age of modern humans from Kibish, Ethiopia." Nature 443 (2005): 733–736.
Mithen, Steven. "The prehistory of the mind. A search for the origins of art, religion and science." London, UK: Thames and Hudson: 1996.
Murphy, Gregory L. "The Big Book of Concepts." 1ª Ed. Cambridge, Massachusetts; London, England: A Bradford Book, The MIT Press: 2002.
Oppenheimer, S. "Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map." Philosophical Transactions of the Royal Society B (2012): 770–784.
Powell, Adam, Stephen Shennan and Mark G. Thomas. "Late Pleistocene Demography and the Appearance of Modern Human Behavior." Science 324 (2009): 1298-1301.
Quin, Paul, C. and Peter D. Eimas. "Perceptual Cues That Permit Categorical Differentiation of Animal Species by Infants". Journal of Experimental Child Psychology 63 (1996): 189-211.
Ramaswamy, V. "Explosive start to the last ice age." Nature 359 (1992): 14 -14.
Rampino, Michael R. and Stephen Self. "Volcanic winter and accelerated glaciation following the Toba super-eruption." Nature 359 (1992): 50-52.
Rakoczy, Hannes, Tricia Striano, Michael Tomasello. "How children turn objects into symbols: A cultural learning account." In Symbol use and symbol representation, 69-97. New York: Erlbaum, 2005.
Roberts, Kenneth and Martin D. Cuff. "Categorization Studies of 9- to 15-Month-Old Infants: Evidence for Superordinate Categorization?" Infant Behavior and Development 12 (1989): 265-288.
Roberts, Kenneth and Marianne Jacob. "Linguistic Versus Attentional Influences on Nonlinguistic Categorization in 15-Month-Old Infants." Cognitive Development 6 (1991): 355-375.
Rogers, Alan R. and Henry Harpending. "Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences." 9 (1992): 552 – 569.
Rosch, Eleanor., Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem "Basic objects in natural categories." Cognitive Psychology 8 (1976): 382–439.
Templeton, Alan. R. "Genetics and recent human evolution." Evolution 61 (2007): 1507-1519.

Tomasello, Michael. "The cultural origins of human cognition." Cambridge: Harvard University Press, 1999.

Ward, Thomas. B., Angela H. Becker, Sally Duffin Hass and Edward Vela. "Attribute Availability and the Shape Bias in Children's Category Generalization." Cognitive Development 6 (1991): 143-167.

Waxman, Sandra R. "The dubbing ceremony revisited: Object naming and categorization in infancy and early childhood." In D. L. Medin & S. Atran (Eds.), Folkbiology, 233-284. Cambridge, MA: MIT Press/Bradford Books, 1999.

Waxman, Sandra R. "The development of an appreciation of specific linkages between linguistic and conceptual organization." Lingua 92 (1994): 229–57.

Waxman, Sandra R. "Early word learning and conceptual development. Everything had a name, and each name gave birth to a new though." In. U. Goswami Ed. Blackwell Handbook of Childhood Cognitive Development, 102-126. Oxford UK: Blackwell Publishers, 2002.

Waxman, Sandra R. and Amy Booth. "Distinguishing count nouns from adjectives: Evidence from 14-month-olds' word extension." In Proceedings of the 24th Boston University Conference on Language Development, 773-784 Somerville, MA: Cascadilla Press, 2000.

Waxman, Sandra R. and Dana B. Markow. "Word as invitations to form categories: Evidence 12 to 13- Month-Old Infants." Cognitive Psychology 29 (1995): 257-302.

Wittgenstein, Ludwig. "Philosophical investigations." New York, NY: Macmillan, 3rd, 1968.

Zielinski, G., P. A. Mayewski, L. D. Meeker, S. Whitlow, and M. S. Twickler. "Potential impact of the Toba mega-eruption ~71,000 years ago." Geophysical Research Letters 23 (1996): 837–840.

CHAPTER EIGHT

AN ICONIC AND A SYSTEMATIC FEATURE OF "IRREGULAR" FORMS IN ENGLISH

ELENA EVEN-SIMKIN
BEN-GURION UNIVERSITY OF THE NEGEV

> A natural language is a system – it is systematic; it is structured; it is a tight, interconnecting, logical system. A system by definition cannot have exceptions, because exceptions are unsystematic, they would have to lie outside any system of language, and in the case of irregular verbs one is placing a tremendous burden on speakers' memory capacity if one simply accepts them as exceptions.
>
> (Beedham 1989: 191)

1. Introduction[1]

This chapter explains the Internal Vowel Alternation (IVA) system in Past Tense (ablaut) verbal formations, e.g. *give–gave* or *win–won* in English. These IVA forms are considered irregular in Modern English (ModE) because they are limited in number (ca. seventy-six "irregular" Past Tense verb forms which are divided into fourteen subclasses). However, in Old English (OE), Indo-European, Semitic, and other languages, the IVA was a prevalent and productive process and was part of a larger system.

In OE IVA verbs, historically referred to as "strong" (ablaut-vocalic) verbs, underwent a phonological backing process of varied degrees that has been preserved in ModE despite many diverse historical phonological changes that have occurred in the language. Furthermore, there are "irregular" ModE IVA verbs that originally were "weak—ed" (also called

[1] I gratefully acknowledge Yishai Tobin for his helpful comments and for encouraging me to write this paper. I am, of course, responsible for any and all possible flaws or errors.

consonantal verbs) which were not derived from the IVA verbs in OE, such as *ring, stick* or *wear* (Emerson 1910: 351–4). There are also borrowed verbs, such as *take–took* (Norse), *thrive–throve* (Norse) or *strive–strove* (French), which have become part of the IVA system. It is most interesting to note that these "new" IVA verbs adhere to the iconic backing process, which reflects the metaphorical movement "backwards" (in time) to the past.

Both OE and ModE have a similar and parallel system of IVA nouns (fewer in number than IVA verbs), which are also considered to be "irregular". These IVA nouns display the opposite phonological process of fronting in nominal plural declension (e.g. *man/men*, *tooth/teeth* or *mouse/mice*). This fronting process may also be considered metaphorically iconic in noun pluralisation denoting a move forward or an expansion of the number of entities. It is also interesting to note that both the verbal and the nominal IVA systems are opposed to the so-called regular systems composed of Lexical Item + Apical Suffix (V + (e)d and N + (e)s). Moreover, both the verbal and the nominal IVA systems are composed almost exclusively of monosyllabic words and can be easily distinguished from each other by the metaphoric iconic direction of their IVA forms.

Furthermore, it has been demonstrated in Even-Simkin (2012) that all the IVA verbs and nouns in both OE and ModE are semantically related and even share a common semantic denominator in the form of distinctive semantic features. All the IVA verbs have been shown to share the same distinctive semantic feature of +RESULT, where the notion of resultative is defined as "an action, state or event [that] must be viewed from the point of view of a result, goal, consequence, conclusion, destination, telic endpoint, etc., which may be explicitly stated or implicitly implied" (Tobin 1993: 17). In Even-Simkin (2012), the invariant meanings (*signifiés*) of all the phonological classes and sub-classes of the diachronic IVA verb system are presented by annotating their resultative messages. All the IVA nouns have been shown to share the distinctive feature of Semantic Integrality originally defined in Tobin (1990: 156) as "an entity or entities is/are perceived as occupying a single continuous space".

The present analysis is based on the sign-oriented linguistic theory (e.g. Saussure 1959 [1916]) and the Columbia School (CS) (e.g. Tobin 1990, 1993, 1994; Diver 1995; Contini-Morava and Tobin 2000; Reid et al. 2002; Davis et al. 2006), including the phonological component of the CS theory known as Phonology as Human Behaviour (PHB) (cf. Diver 1979, 1995; Tobin 1997, 2009). This chapter uncovers the underlying phonological system of the IVA verb system in OE and ModE. The IVA verbs historically were part of the larger rule-governed grammatical

system in OE, i.e. the period that extends "from the earliest times to the year 1100" (Emerson 1910: 44), when as Hogg (1992: 9) pointed out: "[t]here are clear linguistic indications [… in] the structure […] that it is reasonable to make […] the dividing line between Old English and Middle English". Thus, a comprehensive understanding of the ModE IVA forms lies in their etymology, which would be incomplete without considering the historical causes and explanations of IVA diachronically. Historical studies (cf. Emerson 1910: 223, Quirk and Wrenn 1955, Lass 1994, Crystal 1995, Smith 2009, Lass and Anderson 2010 [1975]) imply that IVA forms are etymological remnants of the former noun and verb systems in English.

2. "Irregular" versus "Regular" Forms

The "irregular" IVA verb forms are neither obvious nor transparent in ModE and are acquired as random and individual lexical items (Kuczaj 1977, Pinker 1999, Pinker and Ullman 2002) despite the fact that they were part of a productive and prevalent paradigmatic system in OE. OE IVA verbs were divided into seven subclasses of "strong" verbs (cf. Quirk and Wrenn 1955, Hulbert 1963, Smith 2009, Lass and Anderson 2010 [1975]).

This transformation from what may be viewed as a systematic grammatical system to an idiosyncratic lexical one has been the subject of much research, particularly, in the realm of first language acquisition. Plunkett and Marchman (1993), for example, indicated the *multi-lateral pattern of errors*: the "irregularization" of some "regular" known and novel forms (e.g. *pick-*puck*, *flow-*flew* or **bing-*bang*) and the overgeneralization or "regularization" of some "irregular" lexical and novel forms (e.g. *come-*comed*, *win-*wined*, *blow-*blowed*, *break-*breaked*, *go-*goed*, *went-*wented* or **gling-*glinged*). Similarly, Marchman's (1997: 300) findings in the study of school-age children's productivity in the "irregular" versus "regular" Past Tense in English (e.g. *glow-*glew* or **spow-*spew*, and *catch-*catched*, *shed-*shedded* or **bod-*bodded*) "are consistent with the conclusion that the mechanisms underlying the productive use of regular and irregular patterns are actually more similar than they are different". Moreover, Kuczaj (1977), who examined the spontaneous speech of pre-school-age children, indicated the absence of incorrect IVA forms. Plunkett and Juola's (1999: 486) single-process model/network that handles the "regular" and "irregular" Past Tense inflection of verbs in English also indicated the absence of error in the initial performance of "irregular" forms including the IVA – their

"model not only produces initial error-free performance on noun plurals but does so in the context of initial error-free performance on verb past tense forms". Plunkett and Juola (1999: 486) as well as Marchman (1997) concluded that the "onset of over-regularization errors on nouns tends to occur earlier than overgeneralization errors on verbs in the network", which is most probable because the number of "irregular" nouns versus verbs is smaller (seven IVA nouns in ModE) and they occur in the most frequently used lexical items (Even-Simkin and Tobin 2011: 317). Following Bybee and Slobin (1982: 266), "instances of incorrect vowel changes […] would be predicted if the child were formulating vowel-change rules, rather than learning rote forms". Indeed, the examples of such instances were reported in other studies, such as Bybee and Slobin (1982), Plunkett and Marchman (1993), Marchman (1997), and Ramscar (2002).

All of the above studies support the claim that there may be some covert mechanism underlying the so-called irregular nominal and verbal Past Tense forms in addition to the overt mechanisms of the so-called regular formations composed of lexical item plus suffix, i.e. noun + plural *(e)s* and verb + past *(e)d*. As Beedham (2005: 112) pointed out:

> [i]rregular verbs are a historical vestige, but so are the regular verbs (so is everything in language), that does not stop the irregular verbs from being rule-governed and meaningful synchronically, if we can only find the rule(s) and the meaning(s).

The etymology of the so-called regular and irregular Past Tense formations, i.e. the fact that both "[…] strong and weak verbs follow regular patterns or paradigms, called conjugations" (Smith 2009: 109) strengthens the claim that the IVA forms in Past Tense inflection are not irregular patterns but rather grammatically structured forms. Furthermore, additional support for the systematic character of the so-called irregular Past Tense formation comes from Bybee's (2001: 110) study: her findings of "the model developed in Bybee (1985, 1988, 1995), along with the connectionist models (Rumelhart and McClelland 1986) and the analogical model (Skousen 1989, 1992), would claim that both regulars and irregulars are handled by the same storage and processing mechanisms". As with the IVA verbs, the historical overview of the origin of the plural *i-mutated* form such as *foot-feet* and *mouse-mice* reveals that the IVA Plural nouns are not arbitrary or "irregular" but are the remnants of an earlier phonological system (Crystal 1995: 19):

> […] in Germanic there were many words where a vowel in a stressed syllable was immediately followed by a high front vowel ([i]) or vowel-like […] ([j]) in the next syllable. The plural of **fōt* is thought to have been **fōtiz*, with the stress on *fō*. For some reason […], the quality of this high front sound caused the preceding vowel to change (mutate). In the case of **fōt*, the ō became ē, which ultimately came to be pronounced [i:], as in modern *feet*. The *–iz* ending dropped away, for once the plural was being shown by the e vowel, it was unnecessary to have an ending as well. […] this is what happened in 7th-century Old English. All back vowels in the context described above were changed into front vowels–and all short front vowels and diphthongs were affected, too, being articulated even further forward and higher (with the exception of [i], of course, which is already as far forward and as high in the mouth as any vowel can be).

Thus, Crystal (1995) described this process as a natural and widely spread process in Germanic and other languages. Baayen and Moscoso del Prado Martín (2005: 668) also claimed that:

> It is clear that the Germanic irregular nouns of English, although formally and etymologically highly heterogeneous, pattern along lines of semantic similarity.

Likewise, the "irregularity" of the IVA verb forms has also been questioned in other studies. For instance, Hoard and Sloat (1973: 107) claimed that "irregular [… verbs] are much more regular than they seem". They have proposed a small number of very abstract rules, in which the most clearly non-assimilatory phonological processes are formulated over unmarked/marked values instead of the plus/minus features, which also partially define some vowel variations in the "strong" conjugation of verbs. However, in the opinion of most researchers, the precise phonological characteristics that exhaustively define the so-called irregular Past Tense verbs have not been proposed (cf. Pinker and Prince 1988, Plunkett and Marchman 1993). In the present analysis I provide a phonological characterization of the nominal and verbal IVA systems in OE and ModE.

3. The Nominal IVA System in Old and Modern English

The so-called irregular Plural formations of nouns in English presented in Table 1 demonstrate the remnants of the Plural formation of the nominal system in OE: i.e. seven nouns that have been retained in ModE and ten modern nominal plurals that originally were part of the IVA system but now follow the "regular" rules of adding the lexical item + apical suffix (-

s/-es) in Plural formation. From Table 1, it is evident that the so-called irregular Plural formations of nouns in ModE systematically follow various degrees of the fronting process: some of the examples go from historical long back vowels to front vowels (e.g. *foot/feet, goose/geese* or *tooth/teeth*); others go from historical long back vowels that were changed from central-back diphthongs to central-front diphthongs (e.g. *mouse/mice* or *louse*/*lice*); finally, others go from the historical low central-back vowel /a/ to the medium-low front vowel /æ/ for the singular to the medium or medium-high front lax vowels /ɛ, ɪ/ for the plural (e.g. *man/men* or *woman*/*women*).

However, regardless of the degree (from back vowel or diphthongs to different front vowels and diphthongs), each ModE Plural formation still displays a variation of the fronting process and still follows the same OE phonological fronting process of the IVA in plural declension. This kind of phonological consistency may also be drawn from the common historical background of these nominal IVA plurals, i.e. i-mutation.

In addition, these IVA forms are found in other Germanic languages and other language families (Emerson 1910, Crystal 1995). Following Quirk and Wrenn's (1955: 151) study, the IVA Noun Plural forms undergo a fronting or raising of non-front vowels to mid or high front vowels, i.e. the Noun Plural IVA nuclei are uniformly fronted**.** Lass and Anderson (2010 [1975]: 119) described this phonological process as a systematic one in their study of OE phonology: "[t]he basic effects of the umlaut may be summed up as follows: in a certain context, back vowels front [… and] [i]f the vowels undergoing umlaut are nonback and low, they raise".

Indeed, in the examples given in Table 1, we may witness different degrees of the phonological fronting process of the IVA from singular to plural declension, i.e. from back vowels or diphthongs to different front vowels or diphthongs in ModE IVA Noun Plurals as well as in their OE forms.

Table 1. Phonological *Fronting Process* of IVA

OE Singular Form of Nouns with the *Vowels*: /ō, ā, a, ū, u, ēō/	OE Plural Form of Nouns with the *Vowels*: / ē, æ, y, īē/	Phonetically-Phonological *Fronting Process* of IVA in OE Forms of the Nouns (+)	ModE Singular form with the Following Phonological Representation of the Vowels*:* , ʊu, au][æ	ModE Plural form with the Following Phonological Representation of the Vowels*:* [e, Ii, aI]	Phonetically-Phonological *Fronting Process* of IVA in ModE Forms of the Nouns (+)
Old English Forms					
mann	menn	/a/ → /ē/ = (+)	man	men	[æ] → [e] = (+)[2]
wīfmann	wīfmenn	/a/ → /ē/ = (+)	woman	women	[æ] → [e] = (+)
fōt	fēt	/ō/ → /ē/ = (+)	foot	feet	[ʊu] → [Ii] = (+)
tōð	tēð	/ō/ → /ē/ = (+)	tooth	teeth	[ʊu] → [Ii] = (+)
gōs	gēs	/ō/ → /ē/ = (+)	goose	geese	[ʊu] → [Ii] = (+)
mūs	mys	/ū/ → /y/ = (+)	mouse	mice	[au] → [aI] = (+)
lūs	lys	/ū/ → /y/ = (+)	louse	lice	[au] → [aI] = (+)
Modern English Forms					
brōc	brēc(OE)/ brēche(ME)	/ō/ → /ē/ = (+)	breeches, trousers, pants	Without IVA in ModE	----
bōc	bēc	/ō/ → /ē/ = (+)	book	Without IVA in ModE	----
fēōnd	fīēnd/fynd	/ ēō/ → /īē/ or /y/ = (+)	foe	Without IVA in ModE	---
frēōnd	frīēnd/frynd	/ ēō/ → /īē/ or /y/ = (+)	friend	Without IVA in ModE	---
hōnd	hēnd	/ō/ → /ē/ = (+)	hand	Without IVA in ModE	---
gōte	gēt	/ō/ → /ē/ = (+)	goat	Without IVA in ModE	---
hnute	hnyte	/u/ → /y/ = (+)	nut	Without IVA in ModE	---
burg	byrg	/u/ → /y/ = (+)	fortress	Without IVA in ModE	---
āc	æc	/ā/ → /æ/ = (+)	oak	Without IVA in ModE	---
cū	cy	/ū/ → /y/ = (+)	cow	Without IVA in ModE	---

[2] Although both [æ] and [e] are front vowels, the low front vowel [æ] is further back than the mid-front vowel [e]. In the current discussion I refer to any and all relative degrees of the fronting process.

In other words, despite all of the simplifications and changes from earlier Germanic to OE and finally to ModE, we still observe a consistency in the fronting process of the IVA in singular to plural nominal declension, which cannot be accidental. This apparent regularity and consistency in the IVA system evidently points to the non-arbitrariness of the IVA process. Furthermore, this phonological fronting process in the Singular to Plural declension of the IVA is iconic for metaphoric "moving forward" because it points to the fronting as a representation for the addition of plural; this additional iconic element provides a further support for viewing the nominal IVA system in English as being phonologically systematic and not irregular.

4. The IVA Past Tense System in Old and Modern English

OE verbs had two main groups of verbs: "the strong and the weak, or as we call their descendants respectively the irregular and the regular" (Emerson 1910: 344). Following Quirk and Wrenn (1955: 40), in OE about one-quarter of all the verbs were "strong" verbs. These IVA verbs originally were divided into Seven Classes based on the variable stem-vowel within the qualitative *ablaut (Abtönung)* system (Quirk and Wrenn 1955, Lass and Anderson 2010 [1975]), which can be traced back to the Germanic languages (Emerson 1910, Quirk and Wrenn 1955, Hulbert 1963, Mitchell and Robinson 1982), further "derived from Proto-Indo-European Ablaut variation" (Smith 2009: 114). For example, following Emerson (1910), Quirk and Wrenn (1955), and Lass and Anderson (2010 [1975]) studies, these IVA variations are divided into seven different classes of the OE IVA verbs, as presented in Table 2.

Table 2. Seven Classes of "Strong" Verbs in OE

CLASS	PRESENT	(PRET1/PRET2) = PAST	BACKING (+/–)
I	ii	aa i	+
II	eo	æa u	+
III	1). i	a u	+
	2). e	ea u	+
	3). eo	ea u	+
IV	e	æ ææ	+
V	e (i, ie)	æ ææ	+
VI	a	oo oo	+
VII	ea	eo eo	+

Table 2 presents the vowel-gradation (*ablaut*) forms of seven classes of OE IVA verbs. Moreover, we may point out that, besides a phonemic distinction based on the vowel length in OE, there was also a further differentiation between singular and plural forms in the Past Tense form of the OE "strong" verbs. Thus, OE had seven classes of "strong" verbs with a wide variety of stem-vowel grading and phonemic distinctions based on the vowel length: short versus long. But, in all of these classes of "strong" verbs, the vowel gradation distinguished between the Non-Past and the Past Tense forms. There were various degrees of the backing process in all of the classes of the IVA in the singular and/or the plural Past Tense forms. For example, in the first class of Table 2, the IVA process is from a long high front /i/ to a low-central-back /a/ in singular and a high front /i/ in the plural form. As Emerson (1910: 346) noted, "[o]f the two preterit stems in Old English sometimes one, sometimes the other has been retained in the modern speech, but more commonly the singular has outlasted the plural". In the second class, the IVA process is from the medium-front back diphthong /eo/ to the medium-low-front low-central-back diphthong /æa/ in singular or the high back /u/ in plural form. The third class contains more kinds of the IVA variations, but all of them display different degrees of backing: from the front /i/ or the mid-front /e/ or the medium-front back diphthong /eo/ to the low-central-back /a/ or the medium-front low-central-back diphthong /ea/ in singular and the high back /u/ in plural forms. In the fourth and fifth classes, there is another degree of backing from a medium-front /e/ to a medium-low front /æ/ in singular and a long medium-low front /ææ/ in plural that, according to spectrographic analysis (cf. Ladefoged 1993: 197), is less front or further back than [e]. In the sixth class, there is a transition from a low-central-back /a/ to a long back /oo/ in both singular and plural forms and, following Ladefoged (1993: 197), [o] is higher and further back than [*a*]. In the last class of the *reduplicating verbs*, "the changes of vowels in Cl. VII are of obscure origin, but those in Cl. I-VI have arisen by **gradation**" (Quirk and Wrenn 1955: 46). Nevertheless, we still observe a slight variation of the backing process of the IVA in class VII similar to the previous classes: from a medium-front low-central-back diphthong /ea/ to a medium-front-back diphthong /eo/ in both singular and plural forms.

Not surprisingly, in Table 3 we point out this kind of regularity in phonological processes, which clearly demonstrates the phonological backing process of the OE IVA verbal forms that is particularly apparent in the conjugation from Present second and third person singular to Past Tense singular and/or plural forms of seven "strong" classes.

Table 3. The IVA Systems of the OE "Strong" Verbs

Class No.	*Stem Vowel in Infinitival and Non-Past Tense Forms*	*Stem Vowel in 2sg. and 3sg. Present Tense Forms*	*Stem Vowel in Past Tense Form*	*Backing Process (+/-)*
I	/ī/	/y/, /ī/	/ā/ - sg. /i/ - pl.	+
II	/ēo/, /ū/	/y/, /ī/	/ēa/-sg. /u/-pl.	+
III	1). /i/ 2). /eo/,/ie/	/i/	1)/a/-sg.; /u/-pl. 2)/ea/-sg.;(/i/ or /u/-pl.)	+
IV	1). /e/ 2). /e/ (/ie/)	/i/,/y/	1)/ō/-sg./pl.; /æ/-sg.;/ǣ/-pl. 2)/ea/-sg; /ēa/-pl.	+
V	1) /ēo/ or /ie/ 2) /i/ or /e/	/i/, /y/, /ī/	1). /ea/-sg.; /ā/ or /ēa/-pl. 2). /æ/-sg.; /ǣ/-pl.	+
VI	/a/,/e/,/ea/,/æ/	/e/, /y/	/ō/-sg./pl.	+
VII	/ō/, /ā/ or/ǣ/,/ea/	/y/, /ǣ/, /ē/, /e/	(/ēo/ or /ē/-sg./pl.)	+

Historically, the OE ablaut or "strong" verbs were divided into seven classes, but the ModE IVA verbs may be divided into fourteen classes according to the phonological variation of the IVA form or pattern in the Past Tense. The groups of the IVA verbs presented in this study are similar to the earlier proposed by Bybee and Slobin (1982), Bybee and Moder (1983), Bybee (1985) and Pinker (1994, 1999), which are commonly referred to as the classes or "families" of the "irregular" verbs. As Pinker and Prince (1996: 314) note: "[i]rregular subclasses display a family resemblance structure". That is, by defining the "irregular" Past Tense forms, Pinker and Prince (1996), like Bybee and Slobin (1982), Bybee and Moder (1983) and Bybee (1985), refer to the phonological patterns or schemas, which in this study are presented as the IVA patterns.

Table 4 presents fourteen groups of seventy-six ModE IVA verbs arranged according to the phonological IVA class, following the OED (Oxford English Dictionary) system for transcribing the vowel sounds. Table 4 shows the corresponding varying degrees of the backing process found in the IVA verbs; for example, *slink, wring, swing, sting, strike, sling, cling, fling, win, hang → slunk, wrung, swung, stung, struck, slung, flung, won, hung*, which represent the same phonological backing pattern and which may be classified as a subgroup with its distinctive IVA form: from the high front lax [I] or the low front [æ] or the low-central high-front diphthong [aI] to the mid-central [ʌ]. The phonological backing process is also evident in other IVA verbal Past Tense formations, i.e. the different types of the IVA represent various degrees of the backing

process, as in *see, get, fight → saw, got, fought*, from the high-front diphthong [Ii] or the mid-front [ɛ] or the low-central high-front diphthong [aI] to the long low-back [ɔ:] or the low-back [ɔ] or [ɒ][3]; as in *find → found*, from the low-central high-front diphthong [aI] to the low-central high-back diphthong [aʊ]; as in *drink, sink → drank, sunk/sank*, from the high lax front [I] to the low front [æ] or the mid-central [ʌ]; as in *eat → ate*, from the high-front diphthong [Ii] to the mid-front [ɛ]; as in *draw, grow, fly → drew, grew, flew*, from the long low back [ɔ:] or the mid-central high-back diphthong [əʊ] or the low-central high-front diphthong [aI] to the high-back diphthong [ʊu]; as in *take, stand → took, stood*, from the mid-front high-front diphthong [eI] or the low-front [æ] to the high-back diphthong [ʊu]; as in *drive → drove*, from the low-central high-front diphthong [aI] to the mid-central high-back diphthong [əʊ]; as in *swear → swore*, from the mid-front mid-central diphthong [eə] to the low-back mid-central diphthong [ɔə]; as in *speak → spoke*, from the high-front diphthong [Ii] to the mid-central high-back diphthong [əʊ]; as in *awake → awoke*, from the mid-front high-front diphthong [eI] to the mid-central high-back diphthong [əʊ]; and as in *shoot → shot*, from the high-back diphthong [ʊu][4] to the low-back [ɔ] or [ɒ], or the mid-central high-back diphthong [əʊ].

[3] There is a dialectal variation of these IVAs of the verbs *tread* and *get* in transcribing the different vowels of British English and American English. Thus, there is some difference in the transcription of vowels (cf. Ladefoged 1993: 70). In the chart of English vowels, Ladefoged compared and presented the IPA symbols that are used by different authors, i.e. [ɒ] used by Wells (1990) in the *Longman Pronunciation Dictionary* in the transcription of the internal vowel sound of the word *bother* corresponds to the [ɔ] used by Jones (1977) in the *Everyman's English Pronunciation Dictionary* (14th ed.). For example, [ɔ:] used by Jones (1977) in this dictionary in transcribing the internal vowel sound of the verb *brought* differs from the transcription of this sound [ɔ] given in the other dictionaries, e.g. Kenyon and Knott (1953) and Prator and Robinett (1985). In the *New Collins Concise Dictionary of the English Language* (1982: xix), by McLeod and Hanks, there is also the following remark concerning the merging of the sounds [ɔ:] and [ɒ]: "[t]he old-fashioned /**ɔ:**/ in words like *off, cloth, cross* is abandoned in favour of /ɒ/".

[4] In the case of the verbs *choose* and *shoot*, the internal vowel of their Non-Past Tense forms is already a back vowel that also retains its *backing* feature in their Past Tense form, but with the addition of the feature of *lowering*. That is, the internal vowel of their Past Tense forms is marked with the more apparent [low] feature than that of the Non-Past Tense form.

Table 4. Backing Process of IVA

Group No.	Stem-Vowel in Non-Past Tense Form in OE or ModE	Stem-Vowel in Past Tense Form in OE or ModE	No. of Verbs	Backing Process in ModE or OE (+/–)	The Verbs
1	**ModE: [I], [aI],[æ]**	**ModE: [ʌ]**	10	+	slink, wring, swing, sting, strike, sling, cling, fling, win, hang
2	**ModE: [Ii], [aI],**	**ModE: [ɔ:]**	4	+	fight, see,
	[ɛ]	**[ɔ]/[ɒ]**		+	get, tread
3	**ModE: [aI]**	**ModE: [aʊ]**	4	+	find, grind, bind, wind
4	**ME: [I]**	**ME: [æ]**	4	+	drink, begin, sit, swim
	OE: /i/ ModE: [ʌ]	**OE: /o/** or **/u/ ModE: [æ]**	1	+ –	run
4 or 1	**ModE: [I]**	**ModE: [æ]** or **[ʌ]**	6	+	sink, spin, stink, sing, shrink, spring
5	**ModE: [Ii]**	**ModE: [ɛ]**	1	+	eat
	OE: /ea/ ModE: [ɔ:]	**OE: /ēo/ ModE: [ɛ]**	1	+ –	fall
	OE: /ea/ ModE: [əʊ]	**OE: /ēo/ ModE: [ɛ]**	1	+ –	hold
6	**ModE: [ɔ:], [əʊ], [aI]**	**ModE: [ʊu]**	6	+	draw, grow, know, blow, throw, fly
7	**ModE: [eI], [æ]**	**ModE: [ʊu]**	5	+	take, shake, stand, forsake, slay
8	**ModE: [aI]**	**ModE: [əʊ]**	8	+	(a)bide, rise, shine, drive, strive, dive, stride
9	**ModE: [eə]**	**ModE: [ɔə]**	4	+	swear, tear, shear, bear
10	**ModE: [Ii]**	**ModE: [əʊ]**	6	+	speak, weave, steal, heave, freeze, yield
11	**ModE: [eI]**	**ModE: [əʊ]**	3	+	awake, break, wake
12	**OE: /ī/ ModE: [aI]**	**OE: /ā/** or **/i/ ModE: [I]**	2	+ –	bite, slide
12 or 8	**OE: /ī/ ModE: [aI]**	**OE: /ā/** or **/i/ ModE: [I]** or **[əʊ]**	3	+ –/+	ride, write, smite
13	**OE: /ēo/ ModE: [ʊu]**	**OE: /ēa/** or **/u/ ModE: ([ɒ/ɔ]/[əʊ])**	2	+ +	choose, shoot
14	**ModE: [I]**	**ModE: [eI]**	3	+	bid, give, cleave
	OE: /e/ ModE: [ʌ]	**OE: /ō/ ModE: [eI]**	1	+ –	come
	OE: /i/ ModE: [aI]	**OE: /æ/** or **/ǣ/(/ā/) ModE: [eI]**	1	+ –	lie

That is, the vast majority of the ModE IVA verbs display various degrees of the backing process. However, it is worth pointing out that there are seven ModE IVA verbs (run, fall, bite, slide, come, lie and hold) that did not preserve the backing process in ModE, as well as in three alternative Past Tense forms in some dialects for the verbs ride, write and smite. Nevertheless, in the verbs ride, write, smite, which have two alternative dialectal Past Tense IVA forms in ModE, one of the dialectal variations is marked for the backing feature. It is also interesting to note that the ModE forms run, fall, bite, slide, come, lie and hold, which have not retained the backing process in ModE, followed the phonological backing process in their OE counterparts: from the medium-high front /i/ to the medium-low back /o/ in singular form or the medium-high back /u/ in plural, from the medium-front low-back diphthong /ea/ to the high-medium-front-long back diphthong /ēo/, where [o] is further back than [a] (Ladefoged 1993: 197), from the high front long /ī/ to the long low back /ā/ in singular and the medium-high front lax /i/ in plural, from the medium-front /e/ to the long medium-back /ō/, and from the medium-high front lax /i/ to the medium-low front /æ/ or the long low back /ā/.

Table 4 also indicates that the ModE IVA verbs are fundamentally systematic and not "irregular". The ModE like OE IVA forms display various degrees of the backing process, such as: from the mid-high front [I] or the low front [æ] or the low-central mid-high-front diphthong [aI] to the mid-low-central [ʌ], from the high-front diphthong [Ii] or the low-central mid-high-front diphthong [aI] to the long low back [ɔː], from the mid-low-front [ɛ] to the low back [ɒ]/[ɔ], from the low-central mid-high-front diphthong [aI] to the low-central mid-high-back diphthong [aʊ], from the mid-high lax front [I] to the mid-low-central [ʌ] or the low front [æ], which spectrographically appears as a lower and further back vowel than [I] (cf. Ladefoged 1993: 197), from the high-front diphthong [Ii] to the mid-low-front [ɛ], which is marked by a slight backing from [Ii] (cf. Ladefoged's spectrographic analysis), from the long mid-low back [ɔː] or the mid-central high-back diphthong [əʊ] or the low-central mid-high-front diphthong [aI] to the high-back diphthong [ʊu], from the mid-front mid-high-front diphthong [eI] or the low-front [æ] to the high-back diphthong [ʊu], from the low-central mid-high-front diphthong [aI] to the mid-central high-back diphthong [əʊ], from the mid-front mid-central diphthong [eə] to the mid-low-back mid-central diphthong [ɔə], from the high-front diphthong [Ii] to the mid-central high-back diphthong [əʊ], or from the mid-front mid-high-front diphthong [eI] to the mid-central high-back diphthong [əʊ]. Thus, regardless of the degree (from front or central vowels or diphthongs to different further central or back vowels and

diphthongs), these Past Tense constructions still demonstrate different degrees of the backing process.

5. Five "Weak" Verbs as a Further Evidence of the Phonological System of the IVA

In the previous section, I have presented the phonological system of the IVA in the originally "strong" verbs. In this part we will further see the evidence for the phonologically motivated nature of the IVA system that comes from another category of verbs, namely, the historically "weak" verbs that in ModE have the IVA patterns as the marker of Past Tense inflection. In addition, there is also a further category of either originally "weak" or originally "strong" verbs; as well as the IVA pattern in their ModE forms, these verbs have an additional apical suffix (-d/t) that is realized in the phonetic sound [*t*] as the marker of the Past Tense inflection, as in the originally "weak" verbs *br<u>i</u>ng–br<u>ou</u>**ght**, t<u>ea</u>ch–t<u>au</u>**ght*** and *s<u>ay</u>–s<u>ai</u>**d***,[5] or as in the originally "strong" verbs *w<u>ee</u>p–w<u>e</u>p**t*** and *sl<u>ee</u>p–sl<u>e</u>p**t***. What is common to the verbs of these remaining categories, presented in Table 5, is that all of them have the IVA patterns either in ModE or in their archaic Past Tense inflection or in both OE and ModE forms. By studying these IVA patterns, it becomes evident that these forms show the apparent regularity of the phonological backing process of the IVA verbs, just like the originally "strong" verbs, which only have the IVA backing pattern in their Past Tense inflection.

Table 5. *Backing Process* of the IVA in the Originally "Weak" and/or "Strong" Verbs with the Additional Marker of Past Tense

Stem-Vowel in Non-Past Tense Form	Stem-Vowel in Past Tense Form	No. of Verbs	Backing Process (+/–)	The Verbs	Past Tense forms with: (IVA) or (IVA + apical suffix)
[I]	**[ʌ]**	5	+	ding, sneak, dig, string, stick	dung, snuck, dug, strung, stuck
[I], [aI], [Ii], [ɜ:], [eI], [æ]	**[ɔ:]**	6 (*3 archaic forms*)	+	think, teach, seek, catch, buy, bring, *(work, freight, reach*)	thought, taught, sought, caught, bought, brought, (*wrought, fraught, raught*)

[5] According to Jember (1975: 66), the verb *say* had two different "weak" forms in OE: *secgan* and *swengan*, whereas only the former one is recorded in the OED.

Stem-Vowel in Non-Past Tense Form	Stem-Vowel in Past Tense Form	No. of Verbs	Backing Process (+/–)	The Verbs	Past Tense forms with: (IVA) or (IVA + apical suffix)
[I]	**[æ]**	1	+	spit	spat
[I]	**[æ]** or **[ʌ]**	1	+	ring	rang/rung
[Ii]	**[ɛ]**	23	+	creep, flee, sweep, leap, sleep, weep, breed, deal, feed, feel, keep, kneel, lead, lean, leave, mean, meet, plead, speed, bleed, read, dream, bereave	crept, fled, swept, leapt, slept, wept, bred, dealt, fed, felt, kept, knelt, led, leant, left, meant, met, pled, sped, bled, read, dreamt, bereft
[aI]	**[əʊ]**	1	+	climb	clomb
[eə]	**[ɔə]**	1	+	wear	wore
[Ii]	**[əʊ]**	1	+	reeve	rove
[eI]	**[əʊ]**	1	+	stave	stove
[eI]	**[ɛ]**	1	+	say	said
[e]	**[əʊ]**	2	+	sell, tell	sold, told
[Iə]	**[ɜ:]**	1	+	hear	heard

Table 5 presents various degrees of the phonological backing process of the IVA in these additional forty-seven "irregular" verbs of ModE. Indeed, the IVA verbs *ring–rang/rung, dig–dug, stick–stuck, string–strung, spit–spat, wear–wore, reeve–rove* and *stave–stove* originally were "weak—ed" verbs, i.e. historically these verbs were not derived from the "strong" *vowel-gradation*. However, these verbs still follow the backing process in the Non-Past to Past Tense conjugation of verbs: from the mid-high lax front [ɪ] to the mid-low-central [ʌ] or the low front [æ], which is articulated further back in the mouth (cf. Ladefoged 1993: 197), from the mid-front mid-central diphthong [eə] to the mid-low back mid-central diphthong [ɔə], or from the high-front diphthong [ɪi] or the mid-front high-front diphthong /eɪ/ or the low-central high-front diphthong [aɪ] to the mid-central high-back diphthong [əʊ]. That is, these transformed "weak" to "strong" IVA verbs still conform to the phonological process of backing, thus providing a further evidence not only for the systematic backing of the IVA verbs in English over time but also for the productivity of this IVA process that was addressed in Keuleers and Sandra (2003).

As is the case with these originally "weak" verbs that became "strong", the historical "strong" IVA verbs which became "weak" show the same phonological process, i.e. various degrees of the backing process: e.g. *climb–clomb [climbed], shave–shove [shaved], heave–hove [heaved], (a)bide–(a)bode [(a)bided], crow–crew [crowed], sow–sew [sowed], shear–shore [sheared], shrive–shrove [shrived], weave–wove [weaved], yield–yold [yielded], wake–woke [waked], awake–awoke [awaked]*, etc., with the IVA from the low-central high-lax-front diphthong [aɪ] to the mid-central high-lax-back diphthong [əʊ], from the high-front diphthong [Ii] or the mid-front high-lax-front diphthong [eɪ] to the mid-central high-lax-back diphthong [əʊ], from the mid-front mid-central diphthong [eə] to the low back mid-central diphthong [ɔə], or from the mid-central high-lax-back diphthong [əʊ] to the high-back diphthong [ʊu]. These examples present the different degrees of the phonological process of backing in the conjugation of the IVA verbs from Non-Past to Past Tense, thus supporting the systematic nature of the IVA process.

However, there are two verbs, i.e. *light* and *hide*, which follow the so-called irregular Past Tense conjugation in Standard English and do not show the phonological backing of the IVA process. However, these two verbs cannot be considered as being counterexamples because historically they were not originally "strong" verbs. Besides, these verbs are monosyllabic, and today they share the same diphthong /aɪ/ and end with an apical consonant, usually associated with the dominant "*–ed*" suffix system in Past Tense formation, i.e., /ɪd/, which is reserved for those verbs that end in /t/ or /d/. In other words, these two verbs historically belonged to the "weak" conjugation system, i.e. they did not follow the ablaut gradation; consequently, they further share the prevalent monosyllabic structure of the present-day IVA verbs and end with the apical consonants /t/ and /d/ [*ɪt*, *ɪd*] that are usually associated with the Past Tense. As a result, this pair of verbs does not contradict the *phonological backing process* hypothesis proposed in this study. However, it is also worth noticing that there are two other verbs that do not show a clear backing process in the conjugation of the Past Tense in ModE, one of which is originally a "strong" verb (*lose*), while the second is originally a "weak" verb (*shoe*); both of them are generally considered to be the IVA "irregular" verbs in Standard English. However, not surprisingly, the Non-Past Tense form as well as the Past Tense form of these verbs historically has the [+back] word-internal vowel: *l**o**se*[l**uː**z]–*l**o**st*[l**ɒ**st, US l**ɔː**st] and *sh**o**e*[ʃ**uː**]–*sh**o**d*[ʃ**ɒ**d]. In addition, these two verbs differ from other verbs of the main corpus presented in Table 4 because they have the additional

apical-suffix marker of the Past Tense. In other words, beside this IVA pattern with the backing feature in both forms (i.e. the Non-Past and the Past Tense form) of these two verbs, there is another additional marker of Past Tense, which is the addition of the apical suffix *–t/d* that characterizes the so-called regular Past Tense conjugation system of verbs in ModE. This extra marker of the apical suffix may have a more central role than the IVA process, which, subsequently, does not show a clear and radical backing of the IVA in these two forms. However, this may be blurred by the fact that the original vowel is a back vowel to begin with as it is in the verbs that it can only display the IVA process as its sole marker of the Past Tense form. Thus, the less salient IVA process in the verbs *shoe* and *lose* may show a slight deviation from the backing process of the IVA verbs, supporting the argument that the IVA is not a random, an exceptional or an irregular formation, but rather is a well-structured system that is also iconic. That is, the morpho-phonological IVA Past Tense process metaphorically reflects "moving backward in time". This phonological featuring of the IVA is diachronically consistent over time, i.e. the *ablaut* forms, like their ModE IVA counterparts and the novel (originally "weak") "strong" forms, follow the backing process of the IVA.

6. Discussion and Conclusions

This study proposes a new insight into the phenomenon of the so-called irregular verbs in English by answering the following questions: Is the phonological IVA system phonologically motivated? Why should the IVA be viewed as a classification system and not as an arbitrary list of exceptions to the general rule of the Past Tense conjugation of verbs and the Plural declension in nouns? First, the IVA appears to be non-random *phonologically*; even though the IVA system has changed over time, it still reveals the basic consistency in the particular phonological process it displays, i.e. the *fronting process* of various degrees for all IVA noun plurals and the *backing process* of various degrees in the overwhelming majority of cases of IVA verbal Past Tense formations.

Moreover, it is interesting to point out that the modal IVA verbs like *may/might, shall/should, will/would* and *can*/*could* also adhere to this backing process of the IVA in the conjugation from the Non-Past to the Past Tense form (Table 6).

Table 6. *Backing Process* in Modal Forms

Non-Past Tense form [ei]/[i]/[e]/[æ]	Past Tense form [ai]/[ʊu]	*Backing Process* (+/–)
may	might	+
will	would	+
can	could	+
shall	should	+

This kind of phonological backing process of the IVA in the modal verbs further supports the prediction concerning the non-arbitrary and iconic character of the backing process of the IVA in the Past Tense conjugation process. However, the specific issue of the modal IVA verb forms remains a topic for a later study.

Figure 1 presents the percentage ratios of the *+backing* versus the *–backing* ModE IVA verbal forms of the three IVA corpora which were studied and described in this chapter: (I) "Strong" (verbs) Corpus (SC), which includes the originally "strong" ModE verbal forms with the IVA process only, (II) Additional Corpus (AC), which includes the ModE IVA verbal forms that are either of "weak" origin or of "strong" origin with the additional marker of the Past Tense (beside the IVA), and (III) Modal (verbs) Corpus (MC), which includes the modal IVA verbs which were presented in Table 6.

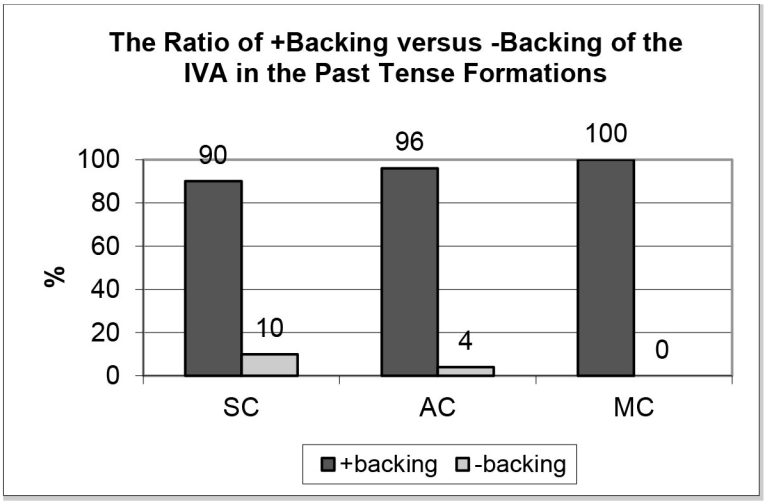


Figure 1. The Ratio of *+Backing* versus *–Backing* Modern English IVA Verbal Forms

The percentage ratios in these three corpora, i.e. SC, AC and MC, present a clear predominance of the +*backing* process in the vast majority of all of the IVA verbs of ModE. Moreover, as previously discussed, the –*backing* cases of the IVA in ModE verbal forms cannot be considered as counterexamples, because all of these cases followed the backing process in their OE forms, thus supporting the historical phonological systematization of the IVA verbal Past Tense forms.

This study provides an explanation for the essential aspects of the phenomenon known as "irregular" IVA verbs in ModE. I have shown that there are systematic phonological patterns and features that triggered and allowed the OE verbal IVA forms to be retained in ModE. This was accomplished by viewing the IVA forms as systems composed of linguistic signs, since in ModE as well as in OE the IVA preserved its phonological classification system[6].

Additional support for the systematized nature of the IVA forms comes from the area of first language acquisition. Pinker (1999: 198–9) pointed out the remarkably small number of cases when children use "irregular" forms incorrectly; in the study of L1 spontaneous speech acquisition, he reported about ninety-six percent of correct responses. Beedham (2005: 114) claimed that "[t]his result indicates that even in the language of small children the irregular verb-forms are not learnt by rote, they are rule-governed and meaningful". In a similar vein, Langstrof (2011: 137) noted the ability of children to formulate or systematize the rules on the basis of grammatical and phonological data that, in its turn, may explain the remarkable number of correct responses of the IVA forms by children at the earliest stages of L1 acquisition.

Stemberger (2001: 19) showed that "phonology affects syntax, both in grammar (e.g. Rice and Svenonius 1998; Broadwell 2000) and in language acquisition (Stemberger and Bernhardt 1997)". Thus, it is not surprising that in ModE as well as in OE the IVA preserved its phonological features as a part of a full-fledged morphosyntactic classification system in nominal and verbal systems, showing that the meta-IVA processes in English are systematic in nature. There are also the alternative non-IVA forms which have also been categorized as the so-called irregular verbs in ModE, such as *put/put*, *build/built*, *cast/cast*, *send/sent* or *spill/spilt*, but which were excluded from the present analysis and discussion. These

[6] Moreover, Beedham (2005: 168) in his phonotactic study of the "irregular" versus "regular" verbs showed that "the irregular verbs differ from the regular verbs phonologically: the English irregular verbs contain VCs and CVs which the regular verbs tend not to contain", thus once again pointing to the phonologically non-arbitrary character of the "irregular" verbs.

other classes of verbs await further investigation following the same sign-oriented approach exemplified in this analysis that is based on the assumption that human actions convey meaning and function as signs.

7. References

Baayen, H. R., and F. Moscoso del Prado Martín. "Semantic Density and Past-Tense Formation in Three Germanic Languages." Language 81/3(2005): 666–698.

Beedham, C. "Investigating Grammar Through Lexical Exceptions: Tense and Irregular Verbs in English, German and Russian." Journal of Literary Semantics 18/3(1989): 187–202.

Beedham, C. Language and Meaning. Amsterdam/Philadelphia: Benjamins, 2005.

Broadwell, G. A. On the Phonological Conditioning of Clitic Placement in Zapote*c*, [ROA 422], 2000.

Bybee, J. L. Morphology: a Study of the Relation Between Meaning and Form. Amsterdam/Philadelphia: Benjamins, 1985.

Bybee, J. L. "Morphology as Lexical Organization." In Theoretical Morphology, edited by M. Hammond and M. Noonan, 119–41. San Diego, CA: Academic Press: 1988.

Bybee, J. L. "Regular Morphology and the Lexicon." Language and Cognitive Processes 10 (1995): 425–55.

Bybee, J. L. "Phonology and Language Use." Cambridge/New York: Cambridge University Press: 2001.

Bybee, J. L. and D. I. Slobin. "Rules and Schemas in the Development and Use of the English Past Tense." Language 58 (1982): 265–289.

Bybee, J. L. and C. L. Moder. "Morphological classes as natural categories." Language 59 (1983): 251–270.

Contini-Morava, E. and Y. Tobin. "Between Grammar and Lexicon." Amsterdam/Philadelphia: John Benjamins, 2000.

Crystal, D. The Cambridge Encyclopaedia of the English Language. Cambridge: Cambridge University Press: 1995.

Davis, J., Gorup, R. J., and Stern, N. (eds.). Advances in Functional Linguistics: Columbia School Beyond Its Origins. Amsterdam/ Philadelphia: John Benjamins, 2006.

Diver, W. "Phonology as Human Behavior." In D. Aaronson and R. W. Reiber, ed., Psycholinguistic Research: Implications and Applications, 161–186. Hillside, N.J.: Lawrence Erlbaum, 1979.

Diver, W. "The Theory." In E. Contini-Morava and B. Sussman Goldberg, ed., Meaning as Explanation: Advances in Linguistic Sign Theory, 43–114. Berlin: Mouton de Gruyter, 1995.

Emerson, O. F. "The History of the English Language." New York /London: Macmillan and CO., Ltd., 1910.

Even-Simkin, E. and Y. Tobin. "Common Semantic Denominators of the Internal Vowel Alternation System in English." Poznan Studies in Contemporary Linguistic*s* 42/2 (2011): 308–330.

Even-Simkin, E. Internal Vowel Alternations in Nominal and Verbal Forms According to the Sign-Oriented Theory and the Theory of Phonology as Human Behaviour. Ph.D. dissertation, Ben-Gurion University of the Negev, 2012.

Hoard, J. E. and C. Sloat. "English Irregular Verbs." Language 49/1 (1973): 107–120.

Hogg, R. M., ed. The Cambridge History of the English Language: The beginnings to 1066, vol. 1. Cambridge: Cambridge University Press: 1992.

Hulbert, J. R. Bright's Anglo-Saxon Reader. Unites States of America: Holt, Rinehart and Winston, Inc., 1963.

Jember, G. K. English-Old English, Old English-English Dictionary. United States: Westview Press, Inc., 1975

Jones, D. Everyman's English Pronouncing Dictionary, (14th ed.). London: Dent., 1977.

Keuleers, E. and D. Sandra. "Similarity and Productivity in the English Past Tense." Computational Linguistics 234 (2003): 1–49.

Kenyon, J. S. and T. A. Knott. A Pronouncing Dictionary of American English. Springfield, Mass.: G. and C. Merriam, 1953.

Kuczaj, S. A. "The Acquisition of Regular and Irregular Past Tense Forms." Journal of Verbal Learning and Verbal Behavior 16 (1977): 589–600.

Ladefoged, P. A Course in Phonetics. (3rd ed.) Fort Worth: Harcourt Brace Jovanovich College Publishers, 1993.

Langstrof, C. "Vowel Change as Systemic Optimization: Why the New Zealand English Front Vowel Shift is Not a Good Example." English Language and Linguistics 15/1 (2011): 137–147.

Lass, R. and Anderson, J. M. Old English Phonology. New York/Cambridge: Cambridge University Press: 2010[1975].

Lass, R. Old English: A Historical Linguistic Companion. Cambridge: Cambridge University Press: 1994.

Marchman, V. A. “Children's Productivity in the English Past Tense: The Role of Frequency, Phonology, and Neighborhood Structure.” Cognitive Science 21/3 (1997): 283–304.
McLeod, W. and P. Hanks. The New Collins Concise Dictionary of the English Language. London/Glasgow: William
Collins Sons and Co Ltd, 1982.
Mitchell, B. and F. C. Robinson. A Guide to Old English. Oxford: Basil Blackwell, 1982.
Oxford English Dictionary (Second Edition).Oxford/New-York: Oxford University Press, 1991. [cited in the text as OED]
Pinker, S. The Language Instinct: The New Science of Language and Mind. London: Penguin, 1994.
Pinker, S. Words and Rules: The Ingredients of Language. London: Weidenfeld and Nicolson, 1999.
Pinker, S. and A. Prince. “On Language and Connectionism: Analyses of a Parallel Distributed Model of Language Acquisition.” Cognition 28 (1988): 59–108.
—. “The Nature of Human Concepts: Evidence From an Unusual Source.” Communication and Cognition 29(3/4) (1996): 307–362.
Pinker, S. and M. T. Ullman. “The Past and Future of the Past Tense.” Trends in Cognitive Sciences 6 (2002): 456–463.
Plunkett, K. and V. A. Marchman. “From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets.” Cognition 48 (1993): 21–69.
Plunkett, K. and P. Juola. “A Connectionist Model of English Past Tense and Plural Morphology.” Cognitive Science 23/4 (1999): 463–490.
Prator, C. H. and B. W. Robinett. Manual of American English Pronunciation, (4th ed.). New York: Holt, Rinehart and Winston, 1985.
Quirk, R. and C. L. Wrenn. An Old English Grammar. Great Britain: Methuen and Co., Ltd., London, 1955.
Ramscar, M. “The Role of Meaning in Inflection: Why the Past Tense Doesn't Require a Rule.” Cognitive Psychology 45 (2002): 45–94.
Reid, W., Ricardo, O., and N. Stern, eds. Signal, Meaning and Message: Perspectives on Sign-Based Linguistics. Amsterdam/Philadelphia: John Benjamins, 2002.
Rice, C. and P. Svenonius. “Prosodic V2 in Northern Norwegian.” Paper presented at the 17th West Coast Conference on Formal Linguistics, Vancouver, BC, (February), 1998.
Rumelhart, D. E. and J. L. McClelland. “On Learning the Past Tense of English Verbs.” In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, ed., Parallel distributed processing: Explorations in

the microstructure of cognition, Vol. 2: Psychological and biological models, 216–271. Cambridge, MA: MIT Press: 1986.
Saussure, F. de. A Course in General Linguistics/Cours de Linguistique Général. Paris: Payot, New York: Philosophical Library, 1959 [1916].
Skousen, R. Analogical Modelling of Language. Dordrecht: Kluwer, 1989.
—. Analogy and Structure. Dordrecht: Kluwer, 1992.
Smith, J. J. Old English: a Linguistic Introduction. United Kingdom: Cambridge University Press: 2009.
Stemberger, J. P. and B. H. Bernhardt. "Syntactic Development Limited by Phonological Development." Paper presented at the 38th meeting of the Psychonomic Society, Philadelphia, (November), 1997.
Stemberger, J. P. Overtensing Within Optimality Theory, 2001. [ROA nº 477,] http://ruccs.rutgers.edu/roa.html.
Tobin, Y. Semiotics and Linguistics. London/New-York: Longman, 1990.
—. Aspect in the English Verb. London/New-York: Longman, 1993.
—. Invariance, Markedness and Distinctive Feature Analysis: A Contrastive Study of Sign Systems in English and Hebrew. Amsterdam/Philadelphia: John Benjamins,1994.
—. Phonology as Human Behavior: Theoretical Implications and Clinical Applications. Durham, NC/London: Duke University Press: 1997.
—. "Phonology as Human Behavior: Applying Theory to the Clinic." Asia-Pacific Journal of Speech, Language and Hearing 12/2 (2009): 81–100.
Wells, J. C. Longman Pronunciation Dictionary. Harlow, U.K.: Longman, 1990.

PART TWO:

THEORETICAL LINGUISTICS AND NLP

CHAPTER NINE

WHAT CAN THEORETICAL LINGUISTICS DO FOR NATURAL LANGUAGE PROCESSING RESEARCH?

BRIAN NOLAN
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN, DUBLIN

1. Introduction

The term "Natural Language Processing" (NLP) is as an engineering subdomain of computer science concerned with the building of products and solutions that treat human language through information technology in some way. Essentially, NLP is seen as providing engineered products. However, in the second decade of this 21st century we can observe that NLP is giving way to the term "Human Language Technology" (HLT) as one that is more meaningful and useful to express the nature of this endeavour and the enterprise of providing solutions to do with the use, deployment and understanding of human languages in its widest sense through contemporary technology. We sometimes see the use of the term "Human Cognitive Technology" (HCT).

In this regard, HLT has become an increasingly central component of computer science over the last decade. HLT has also become increasingly pervasive in our lives, through ubiquitous applications such as Internet search and information retrieval, speech technology, business intelligence and data mining. Language Technology is a fast-growing interdisciplinary field concerned with interactions between computers and spoken and written human language. Importantly, HLT requires knowledge of both computer science and linguistics, and often of other related disciplines, but especially it requires an understanding of linguistics and how languages work to meet the levels of adequacy needed to be successful and accepted.

Combining computer science with linguistics has the potential to create many disruptive innovations in ways we can hardly imagine now in areas including machine translation of the world's languages, Internet-based

question-and-answer dialogue systems, conversational agents or avatars, natural language parsing, Internet search, automatic speech recognition, speech and text-to-speech synthesis, software localisation and internationalisation, sentiment analysis, along with computer-aided language learning.

Linguistics is the scientific study of human language as a system of human communication. Consequently, linguists are concerned with a number of particular questions about the nature of language. What properties do all human languages have in common? To what extent do languages differ? To what extent are the differences systematic with patterns? Some of the basic questions that theoretical linguistics studies are indicated in (1).

(1) Important questions for linguists:

What is language and how is it organized in the languages and cultures of the world?

How is human language to be analysed? How might we discoverer its units and tokens?

Where is language learned and then stored and processed in the brain?

What do all spoken, written and signed languages have in common and what do these properties tell us about the nature of human cognition?

What is the relationship between language, culture and thought?

How is such a knowledge system structured and acquired?

How is it used in the production and comprehension of messages?

How does it change over time?

What is the nature of the cognitive processes that activate when we produce and understand language?

Therefore, linguistics is the study of human languages as knowledge systems in all their dimensions. The part of linguistics that is concerned with the structure of language is divided into a number of subfields, as shown in (2).

(2) Subfields of linguistics

Phonetics - the study of the physical aspects of speech sounds and their classification.

Phonology - the study of how sounds are organised and used in natural languages and the rules that determine which sounds may be combined.

Morphology - the study of word formation and the internal structure of words.

Syntax - the study of rules that govern the ways in which words combine to form phrases, clauses, and sentences.

Semantics - the study of the meaning of linguistic expressions in language.

Pragmatics - the study of those aspects of meaning and language use dependent on the speaker, the addressee and other features of the context of utterance.

Each human language is a complex of knowledge and abilities enabling speakers of the language to communicate with each other and to express ideas, hypotheses, emotions and desires. An emphasis on language as a system of human communication guides many linguists towards employing functional-cognitive models of grammar such as Role and Reference Grammar, Construction Grammar, Cognitive Grammar and similar related models.

Because the use of language is such a central feature of being a human, linguistics has intellectual connections with many other disciplines in the humanities, the social sciences and the natural sciences, with close connections to philosophy, psychology (e.g. cognition and knowledge representation), physics (e.g. acoustics), biology (e.g. anatomy and neuroscience), computer science (e.g. speech synthesis and recognition and avatars) and health sciences (e.g. aphasia and speech therapy). The main purpose of the study of linguistics in a research environment is the advancement of knowledge. However, because of the centrality of language in human interaction and behaviour, the knowledge gained through the study of linguistics has many practical consequences and uses, in particular when coupled with computer science in an applied research context.

On the other hand, computer science is a discipline where theory and practice overlap and can be seen as a science of problem solving with technology, namely with hardware and software. Computer scientists model and analyse problems in order to design solutions while verifying that the proposed solutions are correct and effective. Problem solving requires precision, creativity and careful reasoning. Like theoretical linguistics, computer science also has strong connections to other disciplines. Many problems in science, engineering, health, business and other areas can be solved effectively with computing strategies. However, finding a coherent effective solution generally requires both computer-science expertise and the application of domain knowledge. Consequently, computer scientists often develop knowledge of other subjects including, for example, computer architecture, software systems and programming, graphics, artificial intelligence, computational science, mobile platforms, web development and software engineering. Indeed, some of the core areas of computer science are concerned with theory, algorithms and data

structures, programming methodology and languages, and computer elements and architecture. Other areas include computer networking and communication, database systems, parallel and distributed computation, computer-human interaction, operating systems, and numerical and symbolic computation. Computer engineering provides the techniques for integrating hardware and software and has a close relationship with many dimensions of computer science.

Considered as a profession, computer science is a discipline that involves the understanding and design of computers and computational processes. In its most general form, it is concerned with the understanding of information transfer and transformation, as well as making processes efficient and endowing them with some form of intelligence. The discipline ranges from theoretical studies of algorithms to practical problems of implementation in terms of computational hardware and software. A central focus is on processes for handling and manipulating information. Thus, the discipline is concerned with advancing the fundamental understanding of algorithms and information processes in general as well as the practical design of efficient reliable software and hardware to meet given specifications.

2. What do Contemporary Linguists use in their Work?

Linguistics is the scientific study of language and, for our purposes here, can be considered to broadly have three major dimensions of study: (1) language form, (2) language meaning, and (3) language in context and use. Strangely enough, most people outside the linguistics profession have no idea what a linguist actually does. A frequent assumption is that a linguist knows a lot of languages or is perhaps involved in language teaching in some way. It is very rare indeed that one hears that a linguist is a scientist working within the humanities on the scientific study of language or families of languages.

Specifically, a linguist analyses human language as a system for relating sounds (or signed gestures) and meaning. Phonetics studies acoustic and articulatory properties of the production and perception of speech sounds. The study of language meaning, on the other hand, deals with how languages encode relations between entities, properties and other ontological aspects of the world to convey, process and assign meaning, as well as to manage and resolve ambiguity of various kinds that may exist in an utterance. Typically, semantics concerns with word meaning and word-level lexical semantics, together with phrase- and clause-level truth conditions, whereas pragmatics deals with how context influences

meanings and how meaning is imbued through context and used in an utterance. Grammar is composed of the system of rules that governs the form of the utterances in a given language. At different levels of the grammar, it encompasses sound and meaning, phonology (how sounds function and pattern together), morphology (the formation and composition of words), and morphosyntax (the formation and composition of phrases and sentences and other constructions assembled from these words and appropriately marked). While engaged on such intellectual work on one or more areas in phonetics, semantics or syntax, linguists today make use of a number of technologies to assist their scientific investigations.

The professional linguist today typically works on, or with, any number of the tools and applications shown in (3). These are all engineered NLP products, applications and tools from the domain of computer science and computer engineering which mostly, but not exclusively, serve a linguistic purpose. Many of these have been created with linguist users and linguistic applications in mind, but some have been more generally created for a general informatics market and innovatively applied to good use in the scientific study of language by linguists.

(2) NLP products, applications and tools

Language-aware software
Online digital corpora
Digital ontologies (FunGramKB, RDF and topic maps, etc.)
Design patterns in digital ontologies
Knowledge engineering
Lexicons and their architecture
Computational semantics
Data mining for linguistic purposes
Digital collection and annotation of authentic language data
Text annotation, probably using XML
Text mining
Information visualization for typological purposes
Machine translation (Google or Bing)
Tree-banks
Dialog systems
Question-answer software systems
Digital conversational agents or avatars
Speech recognition (ASR)
Speech biometrics for security
Text-to-speech synthesis (TTS)

Clearly, considering human-language technologies as a growth area, there is potential for mutual benefit for linguists working closely with computer scientists in the specification, design and creation of these tools, applications and products. When we examine the HLT sector, we will see how large the market opportunity actually is and the huge scale of the opportunity that exists for focused and principled applied research between linguists, computer scientists and the industry.

3. The Role of Theoretical Adequacy in our Models and Linguistic Realism

Butler (2009: 1–66) has described in detail the levels of adequacy expected from contemporary linguistic models in order to be considered as fit for purpose. These levels span across the descriptive, typological, psychological and explanatory adequacy required in a viable model of language. In today's world we can consider that another level of adequacy is also required, that of computational adequacy. This is a concept well known to those scientists and professionals working within computer science and software development. A computer system and its underlying model must be fit for purpose as specified. It must deliver its results in a coherent and timely manner with efficient utilisation of available resources.

In linguistic-modelling terms, a linguist is concerned with the model of the computation in the mind of a person who actually computes the link between the concept-semantics-syntax interfaces in an utterance. In computational-linguistic terms, however, this model is a subset of the model of grammar used (for example, RRG) once implemented in software. Specifically, does the model work to deliver the correct and expected results optimally and efficiently considered as a computer system? An implication of the requirement of computational adequacy for a linguistic model implemented in software is that one must address the appropriate levels of granularity required for software specifications in order to actually model a grammar. Typically, this extra level of detail needed to specify the model of grammar adequately to be understandable by a computer system puts additional and complex demands on the linguist to reach the correct levels of granularity and precision so that the model can be programmed.

For a functional-cognitive model of grammar where syntax is not viewed as autonomous but as semantically motivated, the scales of the challenges are increased. However, as a scientific enterprise, it is worthwhile, as it has the potential to model the computation of language from semantics, morphosyntax and the lexicon. In functionally motivated models of language, a grammar-lexicon-construction continuum exists and the interfaces between these need to be rigorously expressed (Nolan and Periñán 2014; Nolan, Mairal-Uson and Periñán 2009; Nolan and Salem 2011; Nolan 2014a; Nolan 2014b). Therefore, the architecture of the lexicon must be strictly defined. Heretofore, traditionally, typical models of language have included HPSG and LFG from the generative tradition, Fluid Construction Grammar from within the world of contemporary construction grammars and RRG from the functional-linguistics tradition.

A disadvantage of most construction-based accounts (with the sole exceptions of Sign-based Construction Grammar and Fluid Construction Grammar) is that they do not pay sufficient attention to how lexical elements are combined into particular language-specific constructions and to how the constraints operating over these constructions are resolved (Nolan 2014c: 152). Overall, construction-based accounts would appear to be not computationally friendly.

As a way of treating linguistic and computational complexity in pursuit of linguistic realism with a robust cultural awareness of linguistic conventions, this scientific enterprise of creating linguistically motived language software and computational models of grammar is crucial. Internationalisation and localisation in software also needs to be guided by linguistic awareness and, in most cases, is culturally and linguistically sensitive to the world's diverse locales. Clearly, professional linguists and linguistic research have a key role to play in this enterprise which will ultimately lead to innovative language-aware products in collaboration with the industry at the SME and at larger global-scale levels.

4. Convergence

It is easy to see in today's world of rapid innovation that the domains of linguistics, computer science and (software and hardware) engineering are overlapping into a convergence to provide the basis for next-generation NLP and HLT applications and scientific insights. We are in a time of fast convergence into a new area of language-aware technologies, many of which are Internet-based and exist on mobile platforms.

Advances in the application of language-aware technologies give us, as linguists, access to very large linguistic datasets over which we do our data collection and build our corpora. We can easily make use of data mining strategies, and many linguists working in functional typology make use of tools and applications to enhance their statistical methods and the visualisation of complex data sets.

As linguists, we know we live in a multilingual world. The contribution of computer science to this multilingual world is the ubiquitous availability of UNICODE on all of our computer systems and font sets to represent the world's languages. Through the widespread availability of high-speed Internet broadband and Wi-Fi, we can make use of Google and Bing as corpus building tools to guide our investigations using n-gram analysis as needed and if required. Additionally, we can use public content as linguistic data from the explosion that is social media (e.g. Facebook or Twitter). There are many new areas of convergence

emerging, and we will soon see new 21st-century linguistically based tools applied to combat cyber-bullying, where the language form and use are a diagnostic for a bullying attack.

5. The Next Generation of Linguistically Motivated Language Software

The functionality of many future NLP and language applications is not yet known. Consider what we have emerging today:

- IBM Watson as an exemplar of HCT; Watson is an artificially intelligent computer system capable of answering questions posed in natural language.
- Internet-based question-answer systems that are multilingual
- Apple *Siri*, Microsoft *Cortana*, Google *Voice*, and their underlying speech and language models
- Spoken real-time machine translation [i.e. speech (source language) to speech (target language) translation] is now available with Microsoft's Skype
- Intelligent knowledge agents and avatars for sign languages
- Dialogue and conversational agents

The potential for language-aware software to provide enhanced services within the multilingual EU is huge. The potential for language-aware HCTs is huge. Additionally, language-aware HCTs that build on existing HLTs point us in a very interesting direction, where collaboration between professional linguists and computer scientists in research, innovation and product development is key. Clearly, linguists have a contribution to make in motivating the linguistic component of such research and development work through our university research centres with the industry.

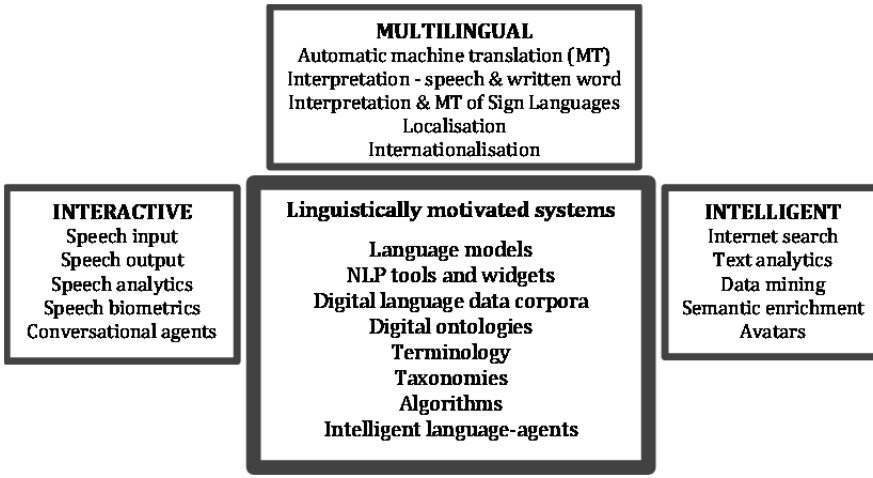


Figure 1. The scope of linguistically motivated Human Language Technology

This will lift the quality and adequacy of the emerging language-aware platforms beyond mere engineering onwards to utility and societal benefit. It would seem that the future is bright for linguists, especially those with software skills.

6. Conclusions

The Language Technology industry has emerged from the areas of Intelligent Content, Translation Technology and Speech Technology, as shown in Figure 1. Up to now, these three areas have been historically separate, in spite of their common scientific and technological roots across computer science and engineering, computational linguistics and theoretical linguistics. Intelligent Content enables search and analytics, Translation Technology enables multi- and cross-language content processing, and Speech Technology enables natural human interaction. The role that linguistics has played in convergence with computer science has been significant and has enabled these new innovations and new industries to emerge through convergence.

Research evidence from LT-Innovate (2014) suggests that Languge Technology applications will impact many different industries at the same time —see Table 1, based on the LT-Innovate report (2014).

Table 1. Industry-specific challenges and needs for language-aware technologies

Domain	DATA	APPLICATION	Language-aware PLATFORM
Automotive	Service Information	Technical Translation	Embedded Systems / Speech interfaces
Culture and Tourism, Digital Humanities	Maps, Schedules, Archives	Automatic Interpreting	Augmented Reality
e-Commerce	Product Data	Competitive Monitoring	Avatars
e-Government	PSI	Curation	Semantic Annotation / Machine translation / Text-to-speech
Education	Academic Publications	Language Learning	Voice Clones / Avatars / Intelligent agents
Financial Services and Banking	Actuarial Data	Fraud Monitoring	Speaker Verification
Games and Entertainment	Video archives	Speech Translators	Voice Banks / Avatars
Life Sciences and Healthcare	Clinical Data	Public Health Monitoring	Electronic Health Records / Intelligent Agents
Media and Publishing	Content and Rich Media	Language-aware Recommender	Multilingual Authoring
Software Publishing	Translation Data	Self-Service Support	Website Localisation / Machine translation
Telecommunications	Service Data	Service Optimisation	Speech Translation / Automatic Speech to Text conversion

A report by LT-Innovate (2012) has estimated that there are over 500 SMEs operating in the Language Technology product space in Europe alone; there are several thousand additional European SMEs that provide Language Technology services in the translation segment, and the industry enjoys a profusion of spin-offs from European university research. According to this report, the Language Technology industry has been propelled to a new level of activity and interest by the rise of the Big Data

and social media phenomena, the need for analytics of all types (but particularly to handle exploding volumes of unstructured data), the emergence of cloud platforms that overcome the computationally challenging and resource-hungry limitations for Language Technology applications as well as the explosion of the mobile market with its demand for personalisation and new, better, more human-friendly interfaces.

The most significant and important trend in the Language Technology industry is the convergence of language technologies themselves. What we see today is Speech Technology combined with Intelligent Content tools to enable personable, conversational intelligent agents or "avatars". Translation Technology is combined with Speech Technology for "talking translators" with Intelligent Content tools to enable cross-language searching for content and media. Subtitling and dubbing are both being automated. The industry must meet this challenge of convergence across computing with language-aware software and the need to move to a new generation of products and services that respond to the computing and communication of the next generation. To successfully achieve this, it requires linguists working together with computer scientists and computer engineers to produce new language-aware products, platforms and software across next-generation ICT-based solutions and services.

Linguistically motivated HLT is a core discipline for computational-linguistics research and development; it will enable innovative, pervasive and ubiquitous solutions and services that have the potential to generate revenue for industry and provide societal benefits. While the segments of Intelligent Content, Translation Technology and Speech Technology are converging, they are at different levels of maturity and implementation at this time. Intelligent Content applications are generally built on existing software environments while the growing need for Internet search, business intelligence with data mining and knowledge discovery on unstructured data have created immediate opportunities and recent strong growth.

Translation Technology is emerging quickly from its specialist status as toolsets for professional translators into mainstream multilingual application environments. Cloud platforms for sharing linguistic resources have the potential to transform this segment of the industry. We see ubiquitous Internet-based translation platforms such as Google Translate and Microsoft's Bing Translate. Speech Technology is emerging quickly on new platforms, especially mobile, giving this area vibrancy on major mobile platforms with Apple *Siri*, Microsoft *Cortana* and Google *Voice*. Strong dominance of speech recognition will eventually be balanced with other applications including speaker recognition and speech generation.

The potential for combining these language-aware technologies in innovative ways is huge. Consider, for example, combining Intelligent Content technology with speech recognition and speech analytics for biometric intelligence, information security and security systems to protect individuals and institutions.

A key finding of the LT-Innovate (2012) report is that current global market for Language Technology software and services will have a value of nearly €30bn now, in 2015, with significant growth expected especially in Europe.

As recently as 2013 (LT-Innovate 2013), it was estimated that over 64 of the top 100 vendors in the "globalisation industry" are based in Europe, where there are several thousand companies offering Translation Technology services, many of them micro-enterprises but with revenues over €50m. It was found that the domain of Translation Technology is the segment of the Language Technology market where Europe has a clear lead overall because of its strength in services. In the Language Technology industry, translation is the largest area and expected to grow to half the total industry by 2015, owing to its 'service' channel to market. On the product side, Europe and the US compete evenly in the sale of tools and products, with emerging developments in Asia that will be challenging within the Machine Translation industry. As of now, most of the world's top IT companies have significant bases in Ireland (including, for example, IBM's R&D centre and the European headquarters of Google, Microsoft and Facebook, all being based in Dublin) and a large portion of their Language Technology services are conducted in Europe. According to LT-Innovate (2014), with a growth in excess of 10%, year on year, Language Technology is the new frontier in IT and is a domain in which European researchers and companies have been pioneers for decades. Today, language-aware technology empowers us in many different ways, as shown in (4).

(4) Empowerment potential of language-aware technology:

- the streamlined optimisation of many complex time-consuming language-dependent processes throughout industry and society
- seamless conversational interaction of humans with devices, from home appliances to cars and from personal digital assistants to robots
- language-neutral processing (search, analysis, enrichment and exploitation) of "intelligence" in all formats opening up new avenues to an inclusive knowledge society

As mentioned in the introduction, the combination of computer science with linguistics has the potential to create many useful innovations that will cause a phase shift in our expectations. As it has become increasingly pervasive in our lives, HLT has become an increasingly central component of computer science too. Central to future research and developments in this area is the convergence of knowledge of computer science and linguistics and other related disciplines. Crucially, the key ingredient is an understanding of linguistics and how languages work to meet the levels of adequacy needed to be successful and accepted and to guide computer science in application development and in the engineering of solutions of value to us. Theoretical linguistics can do a lot for NLP research and assist in a significant way with the design and development of next-generation HLTs.

7. References

Butler, Christopher, S. 2009. Criteria of adequacy in functional linguistics. *Folia Linguistia: Acta Societatis Linguisticae Europaeae* 42(1): 1–66.

LT-Innovate report. 2012. *The Global Language Technology Market: Innovation for the next generation of ICT*. Report from The Forum for Europe's Language Technology Industry. Available [Jan 2015] at: http://www.lt-innovate.eu/resources/document/

—. 2013. *Landmark Report on the State of the European Language Technology Industry*. Report from The Forum for Europe's Language Technology Industry. Available [Jan 2015] at: http://www.lt-innovate.eu/resources/document/lt-20-13

—. 2014. *LT-Innovate Summit 2014 (Brussels)*. The Forum for Europe's Language Technology Industry. Available [Jan 2015] at: http://www.lt-innovate.eu/event/item/lt-innovate-summit-2014-brussels

Nolan, Brian and Elke Diedrichsen. 2013. '*Linking Constructions into functional linguistics – The role of constructions in RRG grammars*'. [Studies in language Companion Series 145]. Amsterdam and New York: John Benjamins Publishing Company.

Nolan, Brian. 2014a. Introduction - Language processing and grammars. In Nolan, Brian and Carlos Periñán. *Language processing and grammars: The role of functionally oriented computational models*. [STUDIES IN LANGUAGE COMPANION SERIES 150]. Amsterdam and New York: John Benjamins Publishing Company.

Nolan, Brian. 2014b. *Motivating a model of conversational software agents, speech acts, language constructions and an emergent common*

ground. Beyond Words - New developments in pragmatics Conference. Leipzig University, Germany. May 2014.

Nolan, Brian. 2014c. *Extending a lexicalist functional grammar through speech acts, constructions and conversational software agents*. In Nolan, Brian and Carlos Periñán. *Language processing and grammars: The role of functionally oriented computational models*. [STUDIES IN LANGUAGE COMPANION SERIES 150]. Amsterdam and New York: John Benjamins Publishing Company.

Nolan, Brian and Carlos Periñán. 2014. *Language processing and grammars: The role of functionally oriented computational models*. [STUDIES IN LANGUAGE COMPANION SERIES 150]. Amsterdam and New York: John Benjamins Publishing Company.

Nolan, Brian, Ricardo Mairal-Uson and Carlos Periñán. 2009. *Natural Language processing applications in an RRG Framework.* Proceedings of the 10th International Role and Reference Grammar Conference. University of California, Berkeley USA.

Nolan, Brian and Yasser Salem. 2011. UniArab: RRG Arabic-to-English Machine Translation. In Wataru Nakamura (ed.). *New perspectives in Role and Reference Grammar*. Newcastle upon Tyne: Cambridge Scholars Publishing. 312–346.

CHAPTER TEN

DOES NLP NEED THEORETICAL LINGUISTICS?

ELKE DIEDRICHSEN
INSTITUTE OF TECHNOLOGY BLANCHARDSTOWN, DUBLIN

1. Introduction

There is a growing need for Natural Language Processing (NLP) in Information Technology. Nowadays, there are many personal devices on the market that use language-aware software such as smartphones, tablets, and "wearables", which include smart-watches and glasses such as Google Glass and HoloLens (cf. YouTube: Google Glass; Microsoft HoloLens; TechHive). The user interface of these relies on spoken-language input to a large degree. Therefore, they require speech-related software, which uses Automatic Speech Recognition. The devices are able to "understand" natural language, talk back to the user and provide the requested service. The commands to the device still require "keywords" but are becoming more and more natural. Machine Translation using spoken language is an enhancement to this functionality, where the software translates a spoken or written input into a selected target language and outputs it in spoken or written mode. The implementation of Automatic Speech Recognition, Text-to-Speech Synthesis and Machine Translation software requires the expertise from theoretical linguists, as every language has its own challenges with respect to syntax, semantics, pragmatics, morphology and phonology (Jurafsky and Martin 2000). In the rapidly growing field of NLP, where natural language data are processed, there is a growing demand for computer systems to learn and adapt to the user's needs, and to interact with the user in a very natural manner. This is a great opportunity for linguists and an inspiring challenge for linguistic theories. As NLP systems are no longer just statistic and rule-based, the broad range of linguistic topics and modern theories can be usefully applied to NLP software and the language models they require.

This chapter introduces some of the fields of NLP, together with their most popular and promising applications. It will become apparent that the modern technology that comes with personal devices relies on linguistic processing to a large degree, and that these technological fields need professional linguistic advice.

2. Personal Devices

In the last few years, the market of Internet-capable personal devices has grown rapidly. Smartphones are ubiquitous, and people use them to make phone calls, write text messages and consume content from the Internet, but they also rely on them for the organization of their personal lives. Smartphones come with an address book, a calendar, reminder functions, an alarm clock, health apps, and many more things that people used to need in pre-smartphone times. The desire for communication and connectedness is key in the use of smartphones, and these devices facilitate communication in several ways. They connect the user to the Internet and, therefore, to social networks and communication channels. They also keep the user up-to-date with respect to political and professional developments. Moreover, and this is the fascinating and growing dimension in the use of personal devices, they are developing into a personal assistant, the interaction with which is perfectly natural. Modern technology achieves this function by equipping the device with language-aware software.

This software is being developed such that it is able to "understand" natural spoken language. If desired, the device is able to talk back to the user, in either spoken or written language, and provide the requested service. This service might be something simple and straightforward like setting an alarm or putting a notice into the calendar. However, modern smartphone software is also able to make appointments by automatically checking other people's calendars, provide information from the Internet and provide weather or traffic information for the current location of the user. It does this by automatically detecting the user's location through GPS. Further developments in personal-assistant software will go a step beyond this, being able to book not only an appointment into the calendar, but also a restaurant table or a flight. These latter services require that the booked service should be available for booking over the Internet, of course.

The commands used for the interaction with the device are becoming more and more natural. The software is able to "understand" natural, casual, everyday speech that a user would also employ to talk to a friend

or colleague. Some "keywords" are still required, like "Hey Siri" or "OK, Google" to activate the speech-recognition functionality of the device. The aim is to make the interaction with the device functional and natural by letting the device adapt to the speaker's voice and vernacular, and by offering as many languages and dialects as possible.

The NLP field that is required to provide such a service is called "Automatic Speech Recognition" (ASR). It means that natural spoken input is processed by the device, and this is done in such a way that the requested service is provided, which gives users the impression that the device has "understood" what they "mean".

The technology that is employed to talk back to the user is called "Text-to-Speech Synthesis" (TTS). TTS provides natural-sounding speech from a written source, which is either the user's own writing or the content that comes from the device itself. TTS is useful to provide screen reading to sight-impaired users, for example, but it is also very useful for situations where the user cannot keep his or her eyes on the screen during the interaction, e.g. when you are driving and require direction instructions. Every modern device for car navigation employs TTS, but it is also found in public spaces where the screen interaction in vending machines, ATMs, etc. is supported by speech output for sight-impaired people.

Another important field for NLP is Machine Translation, where the software translates a spoken or written input into a selected target language and outputs it in spoken or written mode. This functionality has recently been enhanced to work with word detection from a camera: a new Google application is able to "read" a word from a sign and provide the appropriate translation in the desired target language (Spiegel Online: Netzwelt).

What enables computer software to understand, provide and translate natural language? The software needs an up-to-date language model for the language it operates in. Ideally, the software serves a large number of languages, and even minority languages and dialects. A language model is also called a "grammar" (Jurafsky and Martin 2000). It is not a grammar in the sense that traditional linguistic theories would have it, but it certainly builds on insights resulting from the linguistic analysis.

As we know, every language has its own challenges with respect to syntax, morphology, phonology, semantics and pragmatics. These research fields of the linguistic theory are very important to understand the challenges that an NLP system faces.

For syntax, there are rules which may be straightforward to process, but there is hardly any natural language where these rules would not show ambiguities and exceptions here and there. For semantics, the aspect of

"meaning" is debatable. What is the scope of word meanings? How can we handle idioms and metaphors? Pragmatics plays a role for NLP as well, especially as screen interaction is getting more and more natural and personal. The theory of speech acts is employed to equip dialog systems, for example; for any linguistic interaction, the use of pronouns has to be understood and integrated, which is a difficult task.

For phonology, the phonemes of a language have to be identified and integrated, and it is particularly important to include dependencies between neighbouring phonemes in a word, e.g. phonetic features like final devoicing, as it appears in German. For morphology, word forms, inflections and irregularities have to be collected and represented.

Other than that, there are localization features for every language, which include, for example, formats for date and time, address or number, and other conventionalized formats of the written output, e.g. for letters or the way the user is addressed.

A language model accounts for all these language features. It has a lexicon and models for syntax, morphology and phonology. The models are based on both large linguistic corpora and linguistic reference works. They provide the totality of syntactic, morphological and phonological rules, as appropriate for automatic processing. That means they do not *explain* the rules, but provide them as a large data structure, which in turn is used by the language processing software.

Moreover, there has to be a particular language model for the mapping of phonemes to graphemes and vice versa, to ensure that the pronunciation that a device provides is in line with a native speaker's output, but also to make sure that the spoken input provided by the user is properly processed.

A language model for semantics can be represented to contain an ontology, which is a data structure that provides a representation of linguistic entities, together with their meanings and dependencies. Some examples of such a model are described in Periñán-Pascual (2013) and Ruiz de Mendoza Ibáñez (2013).

For pragmatics, the model includes pragmatic rules like speech-act rules and coreference rules for pronoun use.

It should be noted that a language model is not a model of a language in its entirety, like a reference grammar. Rather, it is provided for the linguistic features which are necessary to process the given functionality.

In the case of speech-recognition software, as it is found in services such as Siri, Google Now or Cortana, the language model is to a large degree built on phonological and morphological information.

On the other hand, language avatars, as described by Nolan (2014), need a model of pragmatics to ensure that their behaviour in interaction is natural and adequate for not only the situation but also the user and the purpose they are used for.

At the moment, we are mainly speaking of smartphones and tablet computers, when it comes to discussing personal mobile devices. The next development will be wearables, which are computer devices being attached to the user's body, so they are even more personal. These devices will be worn as a watch that can sense the heart rate of a user and communicate it, for example, or as a set of glasses that can interact with the user's actual line of sight and thus provide an experience of augmented reality: you can see the world through the glasses and, at the same time, request and view real-time information from the Internet, be it about the weather, traffic or directions. They can also record their vision, take photographs and therefore share their view interactively via the Internet.

Therefore, these wearable devices get even closer to the reality that the user experiences than a smartphone can, and, as they are missing a keyboard, the interaction with the device has to work via speech input. This again enhances the future proofness of speech-recognition software and highlights the importance of research and development in that field.

3. NLP for Research and Industry

Given that these days the Internet is generally accessible, and that, furthermore, its content is growing tremendously, there are modern efforts to organize this vast information load in such a way that it can be used easily and professionally.

There is a growing need to process and organize the content provided on the Internet and make it available to users according to their needs and use cases. Organisations, communities and individuals are requesting for content and services to be delivered in their own language, according to their own needs, preferences and context. The challenge is to invent software that is able to productively harvest global unstructured data from the Internet. One device that is able to perform such potent searches is an IBM machine called "Watson" (High 2012). Internet searches, as performed with search engines like Google, have been keyword-based. The user types in a keyword (i.e. a search request) and is provided with a list of locations where an answer might (or might not) be located. The new aim is to create a search engine that provides an intuitive, conversational means of discovering a set of confidence-ranked responses, which are tailored for the individual need of the user. How can this be achieved?

Many natural language systems have attempted to emphasize precision within the confines of specific well-formed rules. However, natural language use does not always respond to such rules. Natural language can be imprecise and idiosyncratic.

So far, many systems have been working by responding only to "word appearance", without "understanding" what was actually "meant", and this may certainly produce bad results. For example, "Find me a pizza restaurant" and "Never find me a pizza restaurant" will produce the same result, which is a list of pizza restaurants. This is because the "word appearance" search looked for the appearance of the string *pizza restaurants*, but certainly did not take into account the negative polarity expressed by the word *never*.

Now, the software definitely does not "understand" what is "meant" by the utterance. It only acts as if it does understand, and produces exactly the output required by the user. This is achieved by letting the software interpret on the basis of *context*, rather than word appearance. The context is provided by feeding huge training corpora into the software and letting it learn meanings of strings on the basis of the frequency of common appearance. The interpretation is enhanced by taking into account user information like the user's location and also previous searches in the search engine, which can give the software statistical hints towards his or her interests.

The context provided by the corpus training is the information that the computer needs to acquire the meaning of language. As we know, there is a historical and cultural component to meaning. One example is the word *tweet*. In the last few years, the meaning most frequently associated with this word is based on the social network service *Twitter*. Therefore, *Twitter* will be the most frequent context of the word *tweet,* so a computer will "learn" that *tweet* is, for the time period covered by the corpus it uses, the verb that expresses "write something on the social network service *Twitter*". This will work for other expressions as well; consider for example the meaning of verbs such as *like, connect* and *suggest*, which are nowadays frequently used in connection with other social network services like *Facebook* and *LinkedIn.*

Therefore, a computer program like Watson will not really "understand" the individual words in the language. Rather, it refers to features in the language from which it can determine whether one text passage (which we call a *question)* infers another text passage (which we call an *answer),* with a high level of accuracy and under changing circumstances.

This functionality lends itself to a whole new dimension in artificial intelligence. For example, there is a function Watson provides that is called "Debater". This application can provide an argumentative discussion with facts and arguments from the Internet. The potential for application of this complex communicative interaction system in research, politics and economy is huge.

We are just at the beginning of a major new era of computing that is less focused on precision, but much more accurate, as it takes into account real-world and up-to-date information on language meaning. It is an era of applying human-like behaviour to large-scale computing problems. It is the era of cognitive systems.

4. Summary and Outlook

Theoretical linguistics can enhance NLP in numerous ways, as NLP systems are no longer just statistic and rule-based. Rather, the broad range of linguistic topics and modern theories can be usefully applied to NLP software and the language models they require.

Thus, the main fields of the linguistic theory are a necessary companion for any production of NLP software. The syntactic theory delivers linearization rules and constructions as grammatical objects (Nolan 2013, Diedrichsen 2013, 2014). From the field of semantics, insights can be used for the provision of ontologies for natural language understanding (Luzondo-Oyón and Jiménez-Briones 2014, San Martín and Faber 2014, Ruiz de Mendoza Ibáñez 2014), but also metaphor studies and theories of semiotics and cultural knowledge can be applied.

Phonology is a very important field for the provision of ASR software. Insights from linguistic phonology are required to provide grapheme-to-phoneme mapping for any language, but also to take into account the syllable structure and intonation patterns. All of these are necessary components of a functioning ASR software that is able to process natural spoken language and to separate the linguistic input from sounds that are just noise, like coughing, background noise, etc.

The linguistic subdomain of pragmatics is used for modern NLP research and development as well. In order to ensure natural interaction of the software with its user, NLP requires insights into reference tracking, speech acts, conversation analysis and the inclusion of relevant context information.

In an NLP project, there is a full integration between computer engineers and linguists. The role of the linguists consists mainly in providing advice from the linguistic theory, which is required in any step

of the production process. An example of linguists and computer scientists working together is given in Mayer, Wälchli, Rohrdantz and Hund (2014).

Linguists provide the theoretical background for the linguistic models. They set up the training corpora, which must be representative samples of the given language, and write guidelines for annotation or transcription. These guidelines are the basis for the training of the software. They represent what is expected of the functionality of the software. Therefore, it is very important that the guidelines give a correct account and application of relevant categories and rules of the language, like part-of-speech types, linguistic rules, correct transcriptions of spoken input, or semantic connections, to name just a few. Linguists will also be needed to check the quality and document the performance of the software against the guidelines. They may identify problems and errors in the performance of the software, document these and make suggestions for solutions.

These are just a few of the many applications of theoretical linguistics for computer software. With the rising importance of Internet communication and the permanent availability of Internet access, more areas will arise that will need the expertise of linguists. One of them is sentiment analysis, where industries automatically search social networks and other communication channels for assessments of their products by users. Sentiment analysis is becoming an important component in surveying customer satisfaction. For automatized sentiment analysis to be accurate, it is necessary to identify expressions of positive or negative opinion, and this can be done by lexical-semantic and syntactic analysis.

With the rise of social media, there is also increasing danger of cyberbullying and mobbing. This is another growing area, where linguistic theory combined with computer science and NLP can be used to create applications to recognize the emergence of a cyberbullying incident, again by lexical and syntactic analysis.

5. References

Diedrichsen, Elke. A Role and Reference Grammar parser for German. In Nolan, Brian and Carlos Periñan-Pascual (eds.): *Language Processing and Grammars. The role of functionally oriented computational models*. Amsterdam: Benjamins. Pages 105-142. 2014.

—. From idioms to sentence structures and beyond: The theoretical scope of the concept “Construction”. In Nolan, Brian and Elke Diedrichsen (eds.): *Linking Constructions into Functional Linguistics. The role of constructions in grammar.* Amsterdam: Benjamins. Pages: 295-330. 2013.

High, Rob. The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works. IBM: IBM Watson, Redguides for Business Leaders. 2012. Available: www.ibm.com_smarterplanet_us_en_ibmwatson_assets_pdfs_Era_of_Cognitive_Systems-An_Inside_Look_at_Watson.pdf last accessed 1/24/15 11:00 PM

Jurafsky, Daniel and James H. Martin: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* New Jersey: Prentice Hall. 2000.

Luzondo-Oyón, Alba and Rocío Jiménez-Briones. FrameNet and FunGram KB: A comparison of two computational resources for semantic knowledge represntantion. In Nolan, Brian and Carlos Periñán-Pascual (eds.): *Language Processing and Grammars. The role of functionally oriented computational models*. Amsterdam: Benjamins. Pages 197-231. 2014.

Microsoft Holo Lens: http://www.microsoft.com/microsoft-hololens/en-us last accessed 1/24/15 11:00 PM

Nolan, Brian. Constructions as grammatical objects: A case study of the prepositional ditransitive construction in Modern Irish. In Nolan, Brian and Elke Diedrichsen (eds.): *Linking Constructions into Functional Linguistics. The role of constructions in grammar.* Amsterdam: Benjamins. Pages: 143-178. 2013.

—. Extending a lexicalist functional grammar through speech acts, constructions and conversational software agents. In Nolan, Brian and Carlos Periñán-Pascual (eds.): *Language Processing and Grammars. The role of functionally oriented computational models*. Amsterdam: Benjamins. Pages 143-163. 2014.

Periñán-Pascual, Carlos. Towards a model of constructional meaning for natural language understanding. In Nolan, Brian and Elke Diedrichsen (eds.): *Linking Constructions into Functional Linguistics. The role of constructions in grammar.* Amsterdam: Benjamins. Pages: 205-230. 2013.

Ruiz de Mendoza Ibáñez, Francisco José. Low-level situational cognitive models within the Lexical Constructional Model and their computational implementation in FunGramKB. In Nolan, Brian and Carlos Periñan-Pascual (eds.): *Language Processing and Grammars. The role of functionally oriented computational models*. Amsterdam: Benjamins. Pages 367-390. 2014.

—. Meaning construction meaning interpretation and formal expression in the Lexical Constructional Model. In Nolan, Brian and Elke Diedrichsen (eds.): *Linking Constructions into Functional Linguistics.*

The role of constructions in grammar. Amsterdam: Benjamins. Pages: 231-270. 2013.

San Martín, Antonio and Pamela Faber. Deep semantic representation in a domain-specific ontology: Linking EcoLexicon to FunGramKB. . In Nolan, Brian and Carlos Periñan-Pascual (eds.): *Language Processing and Grammars. The role of functionally oriented computational models*. Amsterdam: Benjamins. Pages 271-295. 2014.

Spiegel Online Netzwelt: *Please respect the milk foam not in the ablaufgitter runs* (on Google's new text recognition and translation software). Spiegel Online 15 January 2015, http://www.spiegel.de/netzwelt/apps/google-translate-was-taugen-die-neuen-uebersetzer-funktionen-a-1012973.html, last accessed 1/24/15 11:00 PM

TechHive: Microsoft HoloLens macht da weiter wo Google Glass aufgehört hat (on the sales stop of Google Glass and its new competitor, HoloLens from Microsoft) http://techhive.de/microsoft-hololens-macht-da-weiter-wo-google-glass-aufgehoert-hat-2210272036/ last accessed 1/24/15 11:00 PM

YouTube: Google Glass: http://www.youtube.com/user/googleglass last accessed 1/24/15 11:00 PM

PART THREE:

A LINGUISTICALLY AWARE AND COGNITIVELY PLAUSIBLE NLP PROJECT

CHAPTER ELEVEN

A HYBRID EVALUATION PROCEDURE FOR AUTOMATIC TERM EXTRACTION

CARLOS PERIÑÁN-PASCUAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
AND EVA M. MESTRE-MESTRE
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

1. Introduction

The use of corpora in terminography is currently a requirement for specialized knowledge acquisition. Identifying those lexical units which belong to a given specific domain is a complex task, where simple introspection or concordance analysis does not really become effective. For instance, the application of standard frequency criteria to a data collection tends to extract general-purpose vocabulary and is therefore of limited use in identifying technical words. Thus, the automatic extraction of specialized lexical units from text-based corpora is currently a priority field of research in the language industries. In fact, the benefits of automatic term extraction (ATE) projects are immediate, particularly in areas such as document categorization, machine translation or ontology development.

This chapter focuses on the evaluation of ATE systems, whose procedure is traditionally performed by means of one of two methods, as described by Pazienza et alii (2005: 265). On the one hand, an *a priori* reference list of terms for the specific domain is used as a gold standard against which to measure the system performance. The quality of the system is evaluated in terms of precision (i.e. the percentage of extracted terms which are in the reference list) and recall (i.e. the percentage of terms in the reference list which were extracted by the system). On the

other hand, when a reference list is not available, human experts validate the candidate terms extracted by the system. This type of evaluation usually focuses on precision, i.e. the percentage of extracted candidates which have been recognized as terms by the expert.

But which evaluation method is more adequate: gold-standard reference lists or validation based on experts' judgements? It is important to highlight that both methods present some problems. On the one hand, the system can extract terminological expressions which are not present in the reference list. In this case, although these candidates are true terms, they are tagged as false. On the other hand, validation is a time-consuming task, as well as being prone to the expert's subjectivity and personal interpretation. In fact, "little effort has been done for reaching some kind of consensus in a standard evaluation procedure" (Vivaldi & Rodríguez, 2007: 225). We propose to apply a semi-automatic hybrid approach to the evaluation of the candidate terms extracted by ATE systems, thus alleviating the problems presented by both methods. More particularly, this chapter demonstrates that (i) the use of more than one reference list as gold standard does not guarantee a correct evaluation of the candidates, and (ii) the manual validation of candidates should be grounded on an unambiguous criterion about how to decide that a given candidate can really be considered a term characteristic for the domain of the corpus.

We performed the experiment with DEXTER (Discovering and EXtracting TERminology), a multilingual platform for data mining and terminology management, whose aim is not only the search, retrieval, exploration and analysis of texts in domain-specific corpora but also the automatic extraction of terms from specialized domains.[1] The remainder of this chapter is structured as follows. Section 2 examines the components of the metric SRC. Section 3 explores the advantages and disadvantages of reference lists in ATE evaluation. Finally, Section 4 describes the evaluation procedure and discusses the results.

[1] DEXTER, which has been developed in C# with ASP.NET 4.0 by Prof. Carlos Periñán-Pascual, is intended to be freely accessible from the FunGramKB website (www.fungramkb.com), a knowledge-engineering project for natural language understanding. Although only English and Spanish are currently supported in DEXTER, French and Italian are about to be included. Indeed, the modular architecture of this terminology workbench facilitates the processing of corpora in any language and about any specialized domain.

2. SRC: A Metric for Term Extraction

The metric used in the experiment was SRC, which is employed in DEXTER for the identification and extraction of candidate terms. SRC is a user-adjustable composite metric for the extraction of unigrams, bigrams and trigrams from specialized corpora.[2] On the one hand, SRC is composite, since it is the result of combining other metrics on a rational basis. On the other hand, SRC is user-adjustable, since it contains parameters whose values can be adjusted by the user, so that it can accommodate to the configuration of the document collection. It should be noted that the smallest unit of analysis in SRC is the stemmed ngram. Stemming does not only reduce the number of candidates to inspect during validation, but it also serves to determine terminological prominence more effectively. SRC is based on the following equation:

(1)

$$SRC(g) = termhood(g) + unithood(g)$$

$$termhood(g) = S(g) * \alpha + R(g) * \beta$$

$$unithood(g) = \begin{cases} 0, & \text{iff } |g| = 1 \\ C(g) * \gamma, & \text{iff } |g| > 1 \end{cases}$$

where *g* is a stemmed ngram, and the coefficients α, β and γ are the user-adjustable parameters, where α + β = 1 for unigrams and α + β + γ = 1 for complex ngrams. Unlike for most of the research in automatic term extraction, where single metrics are usually combined indiscriminately to produce the best results, SRC is grounded in the theoretical principles of salience (i.e. *S(g)*), relevance (i.e. *R(g)*) and cohesion (i.e. *C(g)*). The metric (1) shows that SRC enables varying degrees of unification between termhood and unithood,[3] where the salience and relevance components of

[2] When the ngram is undergoing the validation process, it will be called "candidate". Only after the validation and lemmatization of the ngram, the candidate can be really called "term".

[3] Kageura & Umino (1996: 260-261) differentiated between unithood and termhood:

> 'Unithood' refers to the degree of strength or stability of syntagmatic combinations or collocations. Thus 'unithood' is not only relevant to complex terms, but to other complex units as grammatical collocations or idiomatic expressions. On the other hand, termhood refers to the degree that a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts. Thus 'termhood' is not only relevant to complex linguistic units, but also to simple units.

SRC serve to measure termhood and the unithood of complex terms is determined by cohesion. The three terminological features of SRC are briefly described in the next sections.

2.1. Termhood in SRC: Salience

One of the pillars of the DEXTER metric is the notion of salience, which is based on the termhood measure TF-IDF (cf. Salton, Wong, and Yang 1975; Salton, Yang, and Yu 1975; Salton and Buckley 1988; among many others), i.e. the weight of a term is determined by the relative frequency of the term in a certain document (or term frequency, i.e. TF) compared with the inverse proportion of that term in the entire document collection (or inverse document frequency, i.e. IDF). Indeed, TF-IDF is one of the most popular AKE (Automatic Keyword Extraction) measures, which can derive lists of keywords from a collection of documents and where each one of these keywords is typically assigned a weight representing how salient the keyword is to the selected document. Therefore, if the keyword is given a high value, then it is perceived as more related to the topic of the document than a keyword with a low value.

In DEXTER, the salience of the stemmed ngram *g* in the document *d* is calculated by applying the following formula:

(2)

$$S_d(g) = TF(g) * IDF(g) * NORM(g)$$

$$TF(g) = f_d(g)$$

$$IDF(g) = 1 + \log\left(\frac{N_T}{df(g)}\right), \text{ where } df(g) > 0$$

$$NORM(g) = \frac{1}{\sqrt{\sum_{g \in d} (TF(g) \times IDF(g))^2}}$$

where $f_d(g)$ is the number of occurrences of g in d, N_T is the number of documents in the target corpus, and $df(g)$ is the number of documents in which the ngram appears in the target corpus. Apart from treating all documents as equally important regardless of their size, the normalization factor makes the salience index range from 0 to 1. It should be noted that this cosine normalization is calculated on the basis of the type of ngram, i.e. unigram, bigram or trigram. For example, the weight of a certain

bigram in a given document is normalized by calculating the weights of all and only the bigrams in the same document.

Finally, the salience of ngrams with respect to the whole target corpus, and not just to a single document, can be calculated as follows:

(3)

$$S(g) = \frac{\sum_{d \in CP_T} S_d(g)}{\sqrt{\sum_{g_j \in CP_T} \left(S(g_j)\right)^2}}$$

Again, the normalization factor of this formula only takes into account ngrams of the same type.

2.2. Termhood in SRC: Relevance

A key issue about termhood is the difference between prevalence and tendency, as explained by Wong, Liu, and Bennamoun (2008). Salience measures the prevalence of the term in a particular target domain, but it does not reflect the tendency of term usage across different domains. In other words, TF-IDF fails to comprehend that terms are also properties of domains, and not just of documents. Consequently, the salience of TF-IDF is combined with a measure which quantifies the relevance of ngrams through the contrastive analysis between the target corpus and a reference corpus. Relevance is calculated in DEXTER as follows, which results from an adaptation of Ahmad, Gillam, and Tostevin's weirdness (2000):[4]

(4)

$$R(g)'' = \frac{P_T(g)}{P_R(g)}$$

$$P_T(g) = \frac{f_T(g)}{|CP_T|}, \text{iff } |g| = 1; \text{otherwise}, P_T(g) = \frac{\sqrt[|g|]{\prod_{k_i \in g} f_T(k_i)}}{|CP_T|}$$

$$P_R(g) = \frac{f_R(g)}{|CP_R|}, \text{iff } |g| = 1; \text{otherwise}, P_R(g) = \frac{\sqrt[|g|]{\prod_{k_i \in g} f_R(k_i)}}{|CP_R|}$$

where $f_T(g)$ and $f_R(g)$ represent the frequency of the stemmed ngram g in the target corpus and the reference corpus respectively, $f_T(k)$ and $f_R(k)$

[4] DEXTER uses the *British National Corpus* (BNC) as the corpus of reference for English.

represent the frequency of a given unigram in *g* with respect to target corpus and reference corpus respectively, $|CP_T|$ and $|CP_R|$ represent the total number of words in the target corpus and the reference corpus respectively, and $|g|$ is the number of lexical items included in the ngram. In this setting, if an ngram is used more frequently in the target corpus than in the reference corpus, then the relevance index of the ngram is greater than 1, and conversely. If the ngram does not occur in the reference corpus, then $f_R(g) = 1$. It should also be noticed that the relevance of complex candidates is calculated on the basis of the geometric mean of each lexical item within the candidate. In this way, the metric can minimize the effects of extremely small or large values in a skewed frequency distribution of the items within the multi-word candidate.

Finally, the relevance index is normalized with the following equation:

(5)

$$R(g) = 1 - \frac{1}{\log_2(2 + R(g)'')}$$

2.3. Unithood in SRC: Cohesion

The notion of cohesion was introduced to determine the unithood of complex ngrams, i.e. cohesion is aimed to quantify the degree of stability of bigrams and trigrams. Cohesion is calculated in DEXTER as follows, which results from an adaptation of Park, Byrd and Boguraev's Term Cohesion (2002):

(6)

$$C(g)'' = \frac{f_T(g)}{\sqrt[|g|]{\prod_{k_i \in g} f_T(k_i)}} \times F, \text{iff } |g| > 1$$

$$F = \begin{Bmatrix} 1, & \text{iff } f_T(g) = 1 \\ \log_2(f_T(g)), & \text{iff } f_T(g) > 1 \end{Bmatrix}$$

where $f_T(g)$ is the frequency of the stemmed ngram *g* in the target corpus, $f_T(k)$ is the frequency of a given unigram in *g* with respect to the target corpus, and $|g|$ is the number of unigrams in *g*. It can be noted that cohesion is not only proportional to the frequency of the complex ngram but also takes into account the frequency of the items which compose *g*. More particularly, cohesion is high when the items which compose the ngram are more frequently found within the ngram than alone in texts. As

with the relevance metric, geometric mean smooths the result in a frequency distribution where extreme values are present.

Finally, cohesion values are normalized in a manner similar to those of relevance:

(7)
$$C(g) = 1 - \frac{1}{\log_2(2 + C(g)'')}, \text{ iff } |g| > 1$$

3. Dictionaries and Thesauri as Reference Lists

Specialised glossaries, dictionaries and thesauri are common terminological tools used in scientific and technical documentation. These tools can help knowledge exchange from experts to professionals, so that novices can become experts. On the other hand, classificatory structures of knowledge organisation can also be used for pedagogical purposes, e.g. in LSP courses, being the means by which lexis is learnt in specific domains. In both cases, these lexical repositories help us understand the conceptual system underlying a given domain.

But what are the disadvantages of using lexical resources as reference lists in ATE validation? In other words, could they be used as reliable gold standards for rigorous terminological studies? The key issue is that gold standards in terminology can be recognised as exemplars of quality but not of perfection. It is astounding the pace at which science and technology advance and consequently require the coinage of new terms. The problem lies in the fact that neologisms cannot keep up with the slower pace at which these words are studied, agreed upon, accepted and finally incorporated into the language, as pointed out by Cabré (2007). In a nutshell, lexical resources take a snapshot of language in use at the time of compilation. As a result, the ATE system usually extracts terms which are not present in the reference list, leading us to think that a gold standard is not a golden standard, since the latter implies "a level of perfection that can never be attained" (Claassen, 2005: 1121).

Therefore, reference lists cannot be solely relied upon. However, they undoubtedly help to accelerate the process of evaluation. Then, for a multilingual platform such as DEXTER, one of the best options of gold standard would be IATE (InterActive Terminology for Europe), the multilingual term database of the European Union.[5] The main goal of IATE, which has about 8.5 million terms in all 24 official EU languages, is

[5] http://iate.europa.eu/

to ensure the quality of the terminology employed in the documentation of the EU institutions. Thus, IATE results from the compilation of all the terms used in many subject matters (e.g. politics, finance, education, applied sciences, humanities, among many others) by the translators of the various language services of the EU institutions. IATE is undoubtedly a well-reputed and useful resource, despite the several problems detected by Zorrilla-Agut (2013), e.g. duplicates, incomplete entries, misspellings, broken hyperlinks, obsolete data, and others, which result from the merger with former terminology databases, such as EURODICAUTOM or EUTERPE. Although the contents of IATE are constantly updated, the next section demonstrates that this multilingual database, or even its combination with other resources, is not sufficient for a reliable evaluation of candidate terms.

4. Evaluation Procedure

4.1. Experiment

The experiment was conducted by using a small corpus of 46 documents (78,988 tokens) about electronics. The documents were obtained from a website[6] whose aim is to provide beginners who study electronics with basic information to help them develop knowledge and understanding of this subject. In particular, documents about the following subtopics on electronics were collected:

Table 1. Composition of the corpus.

Subtopic	Documents
AC theory	17
DC circuits	10
Input-output devices	8
Miscellaneous circuits	11

DEXTER extracted 1,268 unigrams, 1,143 bigrams and 377 trigrams. Although most of these ngrams took the form of nouns or noun phrases, the system also discovered several adjectives (e.g. *capacitive, harmonic, inductive, ohmic* or *sinusoidal*) and a few verbs (e.g. *amplify* or *regulate*).

Following the argumentation in Section 3, we decided to perform a hybrid procedure of evaluation for SRC. On the one hand, the type-A evaluation was based on two reference lists: IATE and *Modern Dictionary*

[6] http://www.electronics-tutorials.ws

of Electronics (Graf, 1999). This evaluation method based on gold standards consisted of eight successive steps. The first step was to download the IATE database in XML format.[7] In the second step, and because of the huge size of the file (2.16 gigabytes), instead of managing all data from a single large XML file, the free application GSplit[8] was used to create several smaller XML files for easier and faster processing. It is important to bear in mind that terminological entries in the IATE XML are structured into three levels: concept, language and term. To illustrate some of the fields in each level, we take the sample entry in Appendix 1:[9]

(i) Concept level

- *TermEntry:id* is a unique identifier of each concept in IATE, e.g. IATE-59612.
- *Descrip:type* contains numeric identifiers which represent subject domains. For example, the code 4816001 represents the subject field of *Land transport*.

(ii) Language level

- *LangSet:lang* indicates an ISO language code. For example, the codes *de, en es, fr* and *it* refer to German, English, Spanish, French and Italian respectively.

(iii) Term level

- *Term* stores a simple or complex lexical unit which designates a concept in a particular language. For example, the lexical realizations of the concept IATE-59612 are *Gaszug* in German, *throttle cable* in English, *acelerador* in Spanish, *accélérateur* in French and *filo del gas* in Italian.

XML is a popular markup language for data exchange. In fact, it is one of the best platform-neutral technologies for interoperability between systems. The problem lies in the text-based and verbose nature of XML, so the third step was to export the XML data to a SQL-supported database (MySQL). As a result, three relational tables were constructed:

[7] The XML file was downloaded on 25 June 2014 from http://iate.europa.eu/tbxPageDownload.do.
[8] http://www.gdgsoft.com/download/gsplit.aspx
[9] The complete XML schema is shown diagrammatically in Appendix 2.

Table 2. Relational tables for IATE data.

Table (fields)	records
IATE_domains (domain_id, descrip)	671 domains
IATE_concepts (concept_id, domain_id)	1,842,937 concepts
IATE_terms (concept_id, term)	981,170 English terms

The fourth step involved the stemming of the terms stored in the *IATE_concepts* table, since the candidate terms extracted by DEXTER take the form of stems.

On the other hand, XML files are also used in DEXTER for data storage (i.e. main.xml, texts.xml, unigrams.xml, bigrams.xml and trigrams.xml). Therefore, the fifth step was to export the XML files containing the candidate terms to relational tables, which were stored in the same MySQL database. As a result, three relational tables were constructed:

Table 3. Relational tables for DEXTER ngrams.

Table (fields)	records
DEXTER_unigrams (candidate, SRC, S, R, f, selected)	1,268 unigrams
DEXTER_bigrams (candidate, SRC, S, R, C, f, selected)	1,143 bigrams
DEXTER_trigrams (candidate, SRC, S, R, C, f, selected)	377 trigrams

In the sixth step, DEXTER unigrams, bigrams and trigrams which were present in the IATE domains of Electrical Industry (6621001) Electronics and Electrical Engineering (6826), Electrical engineering (6826001) and Electronics industry (6826002) were automatically tagged as positive candidates. This was achieved through SQL queries such as (8):

(8)
```
UPDATE DEXTER_unigrams SET selected = true WHERE
(IATE_concepts.domain_id='6826' OR
IATE_concepts.domain_id='6826001' OR
IATE_concepts.domain_id='6826002') AND
IATE_concepts.concept_id=IATE_terms.concept_id AND
DEXTER_unigrams.term= IATE_terms.term;
```

In the seventh step, positive candidates were reviewed to detect those which had been ill-categorized by IATE contributors. For example, unigrams such as *component, device, diagram, digit, element* or *relationship*, bigrams such as *maximum value* or *mean value*, and trigrams such as *instant in time* or *rate of change* were erroneously tagged as terms of electronics, so they were all deselected. In the final step, the remaining

candidates extracted by the system were contrasted with the 25,000 headwords in *Modern Dictionary of Electronics*. Consequently, this type-A evaluation was performed semi-automatically: the IATE database was automatically exploited after some pre-processing, whereas the second reference list was manually browsed, because the dictionary is not available in a machine-tractable format.

On the other hand, the type-B evaluation consisted in validating manually those candidates which were not selected as terms in the previous stage. As this type of evaluation was applied to the outcome of the type-A evaluation, the problem of subjectivity was minimized.

This experiment was mainly aimed at analysing the impact of both types of evaluation procedures on the performance of SRC, S, R and C in terms of precision from the top 200 unigrams, bigrams and trigrams of our corpus.

4.2. Results

Tables 4 and 5 illustrate the results of the type-A and Type-B evaluation of unigrams respectively, where SRC was calculated with $\alpha = 0.8$ and $\beta = 0.2$ according to the equation (1).

Table 4. Precision in the Type-A evaluation of unigrams.

#candidates	S	R	f
1-40	0.82	0.45	0.75
41-80	0.57	0.65	0.37
81-120	0.45	0.70	0.42
121-160	0.45	0.62	0.27
161-200	0.47	0.55	0.50
	0.55	0.59	0.46

Table 5. Precision in the Type-B evaluation of unigrams.

#candidates	SRC	S	R	f
1-40	0.82	0.85	0.60	0.75
41-80	0.72	0.57	0.75	0.37
81-120	0.67	0.45	0.70	0.42
121-160	0.65	0.47	0.65	0.27
161-200	0.57	0.47	0.57	0.50
	0.69	0.56	0.65	0.46

Tables 6 and 7 show the results of the Type-A and Type-B evaluation of bigrams respectively, where SRC was calculated with $\alpha = 0.5$, $\beta = 0.3$ and $\gamma = 0.2$ according to the equation (1).

Table 6. Precision in the Type-A evaluation of bigrams.

#candidates	S	R	C	f
1-40	0.55	0.17	0.52	0.60
41-80	0.55	0.25	0.52	0.55
81-120	0.52	0.32	0.42	0.42
121-160	0.45	0.37	0.32	0.47
161-200	0.32	0.35	0.42	0.50
	0.48	0.29	0.44	0.51

Table 7. Precision in the Type-B evaluation of bigrams.

#candidates	SRC	S	R	C	f
1-40	0.82	0.80	0.35	0.62	0.80
41-80	0.57	0.67	0.47	0.62	0.70
81-120	0.65	0.60	0.45	0.52	0.55
121-160	0.62	0.57	0.47	0.52	0.55
161-200	0.60	0.55	0.42	0.52	0.52
	0.65	0.64	0.43	0.56	0.62

Finally, Tables 8 and 9 show the results of the Type-A and Type-B evaluation of trigrams respectively, where SRC was calculated with $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 0.8$ according to the equation (1).

Table 8. Precision in the Type-A evaluation of trigrams.

#candidates	S	R	C	f
1-40	0.07	0.00	0.07	0.07
41-80	0.05	0.02	0.07	0.10
81-120	0.05	0.05	0.05	0.00
121-160	0.07	0.07	0.02	0.02
161-200	0.00	0.00	0.10	0.00
	0.05	0.03	0.06	0.04

Table 9. Precision in the Type-B evaluation of trigrams.

#candidates	SRC	S	R	C	f
1-40	0.52	0.50	0.35	0.55	0.50
41-80	0.50	0.27	0.45	0.47	0.35
81-120	0.40	0.30	0.27	0.25	0.25
121-160	0.17	0.17	0.45	0.27	0.12
161-200	0.30	0.20	0.20	0.25	0.15
	0.38	0.29	0.34	0.36	0.27

4.3. Discussion of Results

4.3.1. Reference list vs validation

After analysing the results from the two evaluation methods, it can be noted that the integration of two large gold-standard data sets, that is, the IATE thesaurus and an electronics dictionary, is not sufficient. In fact, manual validation increased the number of recognized terms proportionally to the complexity of the ngram, as can be seen in Table 10:

Table 10. Precision increase after type-B evaluation.

	S	R	C
Unigrams	+0.01	+0.06	
Bigrams	+0.16	+0.14	+0.12
Trigrams	+0.24	+0.31	+0.30

Once the importance of a hybrid evaluation has been demonstrated, I should highlight now the criterion applied to identify false negative terms, i.e. ngrams which were true terms but ignored by our two linguistic resources. In this regard, most of the ATE research focuses just on the number and/or profile of the evaluators involved in the experiment, as shown in the following descriptions:

> Three judges inspected the 500 glossary items and marked good glossary items. (Park, Byrd & Boguraev, 2002: 6)

> We instructed three judges to manually inspect the top 300 candidate terms produced by each algorithm and mark those they believed to be terms one would expect to encounter when reading texts about animals. (Zhang, Iria, Brewster & Ciravegna, 2008: 2110)

> Then, we asked two experts from different branches of genetics to inspect first 100 terms produced by each method. Their task was to decide which terms are characteristic for the genetic domain. (Knoth et alii, 2009: 91)

But how can "good glossary items" be identified? When to believe that a given candidate is really a term? How to decide that the candidate is "characteristic for the domain"? Even in the experiment described in the last quotation, the researchers were aware of the significance of having an unambiguous criterion, but surprisingly it was not well defined:

> The task may seem simple, but the domain experts found it ill-defined. The lack of a precise definition of "the characteristic domain term" showed to

> be the major problem. […] The evaluators also found it difficult to be consistent across large set of results. In order to increase their consistency they had to evaluate the same results more than once. (Knoth et alii, 2009: 91-92)

All agree that defining this criterion of terminological consistency is not an easy task, but few researchers attempt to do it. In this evaluation of SRC, the criterion to evaluate complex terms should combine the concepts of unithood and termhood in such a way that "the candidate must be a lexicalization which serves to describe the specific domain of the corpus". Indeed, this statement was crystallized into three corollaries aimed to provide verifiable evidence for multi-word terms:

Corollary 1. Several lexical items ($item_i$, $item_{i+1}$, ..., $item_n$) can be treated as a single lexicalization (*complexItem*) as long as the definition of *complexItem* can be found in the corpus or in any other document resource.[10]

Corollary 2. The definition of *complexItem* must go beyond the simple combination of the meanings of $item_i$, $item_{i+1}$, ..., $item_n$.

Corollary 3. The definition of *complexItem* must contain at least one lexical item, different from $item_i$, $item_{i+1}$, ..., $item_n$, which is typically used to describe the corpus domain.

These corollaries, being used by the evaluator as guidelines, were particularly useful to differentiate between complex terminological units and phraseological collocations. For example, should *coil of wire* and *coil rotation* be considered terms of electronics? The case of *coil of wire* fulfils the three corollaries. First, the following definition was found:

> A coil of wire is simply a resistor as far as steady voltage is concerned, but for alternating voltages it behaves as an inductor. (Sinclair, 2011: 36)

Second, this definition clearly goes beyond the combination of the meanings of *coil* and *wire*. Third, there are three words in this definition, i.e. *resistor, voltage* and *inductor*, which undoubtedly pertain to the electronics terminology. On the other hand, *coil rotation* is only a statistically significant co-occurrence which does not fulfil any of the previous corollaries. Therefore, *coil of wire* is recognized as a term, but *coil rotation* is tagged as a collocation.

[10] In this experiment, the evaluator browsed Google Books and Wikipedia, among other electronic resources.

In this fashion, the type-B evaluation revealed 12 unigrams, 38 bigrams and 64 trigrams as false negative terms among the top 200 SRC-ranked ngrams. For the sake of the transparency of the evaluation procedure, we present these ngrams in alphabetical order:

a) Unigrams: *astable, kWhr, NPN, ohmic, photodevice, photoelectric, photojunction, PMDC, PN, PSU, RLC*, and *RTD*.[11]
b) Bigrams: *AC capacitance, AC circuit, AC inductance, AC resistance, AC waveform, analogue sensor, ATX connector, ATX PSU, capacitive circuit, common anode, common cathode, cosine wave, circuit current, DC circuit, DC voltage, electrical current, induced EMF, LC circuit, LDR photocell, lights sequencer, linear solenoid, Molex connector, PN-junction, proximity sensor, RC snubber, RLC circuit, RMS value, RMS voltage, rotary solenoid, Schmitt trigger, solar cell, solenoid coil, sound transducer, sound wave, stator field, voltage source, voltage triangle*, and *voltage tripler*.[12]
c) Trigrams: *AC capacitance circuit, ATX power supply, audio sound transducer, bench power supply, bipolar NPN transistor, brushed DC motor, brushless DC motor, circuit phase angle, coil of wire, common anode display, common cathode display, contact temperature sensor, conventional current flow, current limiting resistor, DC output voltage, DC servo motor, delta-connected network, delta star transformation, flow of current, flow of electrons, half-wave rectifier, intermittent duty cycle, Johnson Decade Counter, Kirchoffs Circuit Law, Kirchoffs Current Law, Kirchoffs Voltage Law, lagging phase difference, LC tank circuit, light dependent resistor, light emitting diode, linear solenoid construction, load resistor RL, mesh current analysis, Molex ATX connector, negative half-cycle, nodal voltage analysis, Norton equivalent circuit, Ohms Law triangle, parallel LC circuit, parallel resonance circuit, parallel RLC circuit, permanent magnet rotor, phototransistor light sensor, photovoltaic solar cell, positive half cycle, potential divider network, RC snubber network, resistance temperature detector, Schmitt trigger inverter, series LC circuit, series RC circuit, series resonance circuit, series RLC circuit, sinusoidal AC waveform, solenoid duty cycle, star delta transformation, stator field winding,*

[11] kWhr (kilowatt-hour), NPN (negative-positive-negative), PMDC (permanent-magnet direct-current), PN (positive-negative), PSU (power supply unit), RLC (i.e. the letters used for the electrical symbols for resistance, inductance and capacitance respectively), and RTD (resistive temperature detector).
[12] AC (alternating current), ATX (Advanced Technology eXtended) DC (direct current), EMF (electromotive force), LC (inductor-capacitor), LDR (light dependant resistor), RC (resistor-capacitor), and RMS (root-mean-square).

Thevenins equivalent circuit, unregulated power supply, voltage doubler circuit, voltage multiplier circuit, voltage quadrupler circuit, voltage tripler circuit, and *Wheatstone Bridge circuit*.[13]

4.3.2. Performance of SRC

As shown in Tables 5, 7 and 9, the best precision was obtained with SRC when the top 200 unigrams, bigrams and trigrams were retrieved, so it can be concluded that the combination of metrics can actually improve the performance of this ATE system. Moreover, the distribution of terms in the different cut-off points along the top 200 ngrams is also better with SRC than with the single metrics. Indeed, for ATE systems focusing on small- and medium-sized specialized corpora, SRC is one of the most efficient measures as far as precision is concerned.

4.3.3. Analysing false positive candidates

To the best of our knowledge, scientific literature on ATE is not characterized by the deep analysis of false positive candidates. However, analysing the noise is the most productive way to find out how the performance of our system can be improved. For this reason, false positive candidates found among the top 200 SRC-ranked ngrams have been classified into four main categories:

i. Common words, such as *average, maximum* or *value*.
ii. Words typically used in the description of other specialized domains, e.g. *complex number, polar form multiplication, root mean square* and *vector* (mathematics); *output device* and *star network* (computer science); *cadmium sulphide* (chemistry); *absolute position encoder* (mechanical engineering).
iii. Words nested in multi-word terms of the given domain, e.g. unigrams such as *nodal* (*nodal voltage analysis*) or *unregulated* (*unregulated power supply*), bigrams such as *parallel RLC* (*parallel RLC circuit*) or *width modulation* (*pulse width modulation*), and trigrams such as *permanent magnet DC* (*permanent magnet DC motor*), *pole double-throw* (*single-pole double-throw*) or *power transfer theorem* (*maximum power transfer theorem*). One of the most representative cases in this category is proper nouns, since they are almost always part of complex terms. Some examples include

[13] RL (resistor-inductor).

Kirchoffs (*Kirchoffs Circuit Law, Kirchoffs Current Law* and *Kirchoffs Voltage Law*), *Molex* (*Molex connector*), *Schmitt* (*Schmitt trigger*) or *Thevenin* (*Thevenin's theorem*).

iv. Mathematical variables and symbols. For example, in the following passage, the variables XL, XC and XT were converted into the ngrams XL, XC y XT:

Then in the series RLC circuit above, it can be seen that the opposition to current flow is made up of three components, X_L, X_C and R with the reactance, X_T of any series RLC circuit being defined as: $X_T = X_L - X_C$ or $X_T = X_C - X_L$ with the total impedance of the circuit being thought of as the voltage source required to drive a current through it.[14]

Other similar cases are V_{max} (maximum value for voltage), V_{out} (output voltage) or V_{AV} (average voltage). All these instances are employed on an *ad hoc* basis. On the contrary, standard abbreviations used in specialised domains should be taken into account as true terms, e.g. *Hz* (*hertz*) or *kWhr* (*kilowatt-hour*).

The logical step of addressing this issue is to consider what could have been done to make these false positive candidates have a lower SRC score, or even be removed from the candidates list. In this respect, the SRC values of the ngrams in the categories (i) and (ii) could certainly have been reduced if relative frequencies had been contrasted with those in other specialized corpora (cf. Sclano & Velardi, 2007), instead of applying the R metric to a general-language corpus. However, this alternative cannot become a workable solution in a multilingual setting such as DEXTER, where the user is free to create corpora of any specialized domain. A more realistic choice would have been to integrate the IATE data set into the ATE procedure, so that candidates pertaining to domains different from that of the corpus would be discarded. However, this approach is likely to increase the rate of silence in the system: *coil of wire* (Iron, steel and other metal industries), *photoelectric* (Earth sciences), *proximity sensor* (Information technology and data processing, Mechanical engineering, Technology and technical regulations), or *solar cell* (Energy, Environment), among others, wouldn't have been extracted because they are part of the domains given in brackets.

On the other hand, it could be assumed that C-value (Frantzi and Ananiadou 1996; Frantzi, Ananiadou, and Hideki 2000), which is described as "a method to improve the extraction of nested terms" (Frantzi, Ananiadou, and Hideki 2000, 122), would have yielded more

[14] This text has been extracted from http://www.electronics-tutorials.ws/accircuits/series-circuit.html.

appropriate scores for the false candidates in the category (iii), since this cohesion metric allows calculating the degree of independence of complex ngrams with regard to nestedness. However, Tables 11 and 12 demonstrate that C-value would not have been as efficient as expected.

Table 11. A sample of bigrams: C-value and position in the candidates list.

	C-value	pos
parallel RLC	0.81896	16
pulse width	0.76019	93
Schmitt trigger	0.73493	148
stator field	0.68044	281

Table 12. A sample of bigrams: C score and position in the candidates list.

	C	pos
Schmitt trigger	0.53126	11
pulse width	0.51953	16
parallel RLC	0.49061	21
stator field	0.23799	191

As shown in Table 12, the four bigrams are found in the top 200 C-ranked candidates, where *Schmitt trigger* and *stator field* are the only real terms, and *pulse width* and *parallel RLC* are false positive candidates. However, Table 11 clearly shows that C-value keeps the same noise but increases the silence by placing the two terms farther away from the top candidates. Indeed, the two false candidates would have been treated as more independent than the two real terms, which is not true at all. Most of the research with C-value has actually focused on long candidate terms in the medical domain, where 4-gram and even 5-gram candidates are relatively frequent, e.g. *adenoid cystic basal cell carcimona*. However, the metric (7) is the right choice to measure the cohesion of unigrams, bigrams and trigrams in small- and medium-sized corpora.

Finally, regarding the category (iv), all those ngrams with an embedded subscript tag (e.g. V_{max}) could have been easily discarded in the case of HTML documents. However, this strategy could not be implemented with plain text files.

5. Conclusions

Gold-standards, such as domain-specific dictionaries and thesauri, are not sufficient for the evaluation of the candidate terms discovered by ATE systems. A semi-automatic hybrid approach is accurately described in this chapter, where the IATE thesaurus together with a specialized dictionary is integrated with the human validation of the terminological outcome. We conducted the experiment with SRC, which is the composite measure for term extraction in DEXTER, a multilingual platform for data mining and

terminology management with small- and medium-sized specialized corpora. SRC is a user-adjustable metric based on the terminological notions of salience, relevance and cohesion.

6. Acknowledgement

Financial support for this research has been provided by the DGI, Spanish Ministry of Education and Science, grants FFI2011-29798-C02-01 and FFI2014-53788-C3-1-P.

7. References

Ahmad, Khurshid, Lee Gillam, and Lena Tostevin. "Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)." In *Proceedings of the 8th Text Retrieval Conference (TREC-8)*, ed. by E. M. Voorhees, and D. K. Harman, 717-724. Washington: National Institute of Standards and Technology. 2000.

Cabré, María Teresa. "Hacia la unidad terminológica del español". *IV Congreso Internacional de la Lengua Española*. Cartagena de Indias. 2007.

Claassen, Jurgen. "The gold standard: not a golden standard". *British Medical Journal*. 2005. 330 (7500): 1121.

Frantzi, Katerina, and Sophia Ananiadou. "Extracting Nested Collocations." In *Proceedings of the 16th International Conference on Computational Linguistics*, 41-46. 1996.

Frantzi, Katerina, Sophia Ananiadou, and Mima Hideki. "Automatic Recognition of Multi-Word Terms: the C-Value/NC-Value Method." *International Journal of Digital Libraries.* 2000. 3 (2): 115-130.

Graf, Rudolf F. *Modern Dictionary of Electronics*, 7th edition. Boston: Newnes. 1999.

Kageura, K. and Umino, B. "Methods of automatic term recognition: a review". *Terminology.* 1996. 3(2), 259-289.

Knoth, Petr, Marek Schmidt, Pavel Smrz, and Zdenek Zdráhal. "Towards a Framework for Comparing Automatic Term Recognition Methods." In *Proceedings of the 8th Annual Conference Znalosti*, 2009. 83-94. Bratislava: Informatics and Information Technology STU.

Park, Youngja, Roy J. Byrd, and Branimir K. Boguraev. "Automatic Glossary Extraction: Beyond Terminology Identification." In *Proceedings of the 19th International Conference on Computational Linguistics.* 2002. Vol. 1, 1-7. Stroudsburg, PA: Association for Computational Linguistics.

Pazienza, Maria Teresa, Marco Pennacchiotti, and Fabio Massimo Zanzotto. "Terminology Extraction: An Analysis of Linguistic and Statistical Approaches". In *Studies in Fuzziness and Soft Computing: Knowledge Mining*, ed. by Janusz Kacprzyk, and Spiros Sirmakessis, 2005. 255-279. Berlin-Heidelberg: Springer.

Salton, Gerard, and Christopher Buckley. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management.* 1988. 24 (5): 513-523.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A Vector Space Model for Automatic Indexing." *Communications of the ACM.* 1975.18 (11): 613-620.

Salton, Gerard, Chung-Shu Yang, and Clement T. Yu. "A Theory of Term Importance in Automatic Text Analysis." *Journal of the American Society for Information Science*. 1975. 26 (1): 33-44.

Sclano, Francesco, and Paola Velardi. "TermExtractor: A Web Application to Learn the Common Terminology of Interest Groups and Research Communities." In *Proceedings of the 9th Conference on Terminology and Artificial Intelligence*, Sophia Antinopolis. 2007.

Sinclair, Ian. Electronics Simplified. Oxford: Newnes-Elsewier. 2011

Vivaldi, J. & Rodríguez, H. "Evaluation of terms and term extraction systems: a practical approach". Terminology. 2007. 13 (2), pp. 225-248.

Wong, Wilson, Wei Liu, and Mohammed Bennamoun. "Determination of Unithood and Termhood for Term Recognition." In *Handbook of Research on Text and Web Mining Technologies*, ed. by Min Song, and Yi-Fang Wu. 2008. 500-529. Hershey-New York: IGI Global.

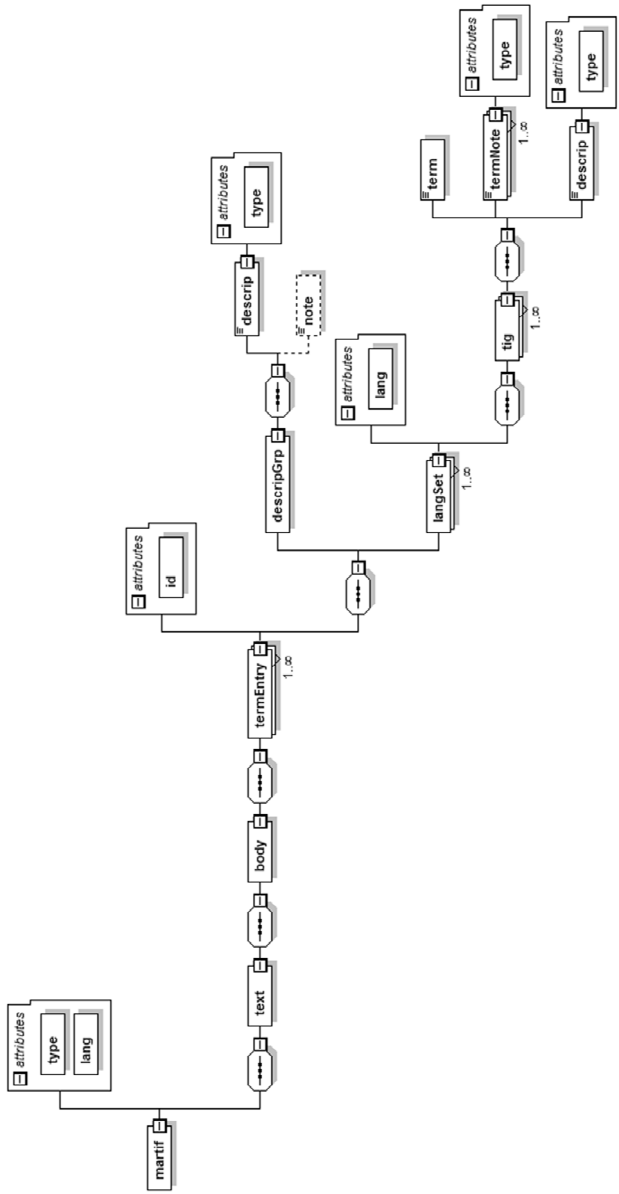
Zhang, Ziqi, José Iria, Christopher Brewster, and Fabio Ciravegna. "A Comparative Evaluation of Term Recognition Algorithms." In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. 2008. 2108-2113. Marrakech: ELRA.

Zorrilla-Agut, Paula "When IATE met LISE: LISE clean-up and consolidation tools take on the IATE challenge". In G. Budin & V. Lušicky (eds.) *Languages for Special Purposes in a Multilingual, Transcultural World*. *Proceedings of the 19th European Symposium on Languages for Special Purposes*. Vienna: University of Vienna, 2014. 536-545.

Appendix 1. Sample of XML entry in IATE.

```
<termEntry id="IATE-59612">
<descripGrp>
<descrip type="subjectField">4816001</descrip>
<note>NT1 road transport,NT2 vehicle fleet,NT3 motor car</note>
</descripGrp>
<langSet xml:lang="de">
<tig>
<term>Gaszug</term>
<termNote type="termType">fullForm</termNote>
<descrip type="reliabilityCode">3</descrip>
</tig>
</langSet>
<langSet xml:lang="en">
<tig>
<term>throttle cable</term>
<termNote type="termType">fullForm</termNote>
<descrip type="reliabilityCode">3</descrip>
</tig>
</langSet>
<langSet xml:lang="es">
<tig>
<term>acelerador</term>
<termNote type="termType">fullForm</termNote>
<descrip type="reliabilityCode">3</descrip>
</tig>
</langSet>
<langSet xml:lang="fr">
<tig>
<term>accélérateur</term>
<termNote type="termType">fullForm</termNote>
<descrip type="reliabilityCode">3</descrip>
</tig>
</langSet>
<langSet xml:lang="it">
<tig>
<term>filo del gas</term>
<termNote type="termType">fullForm</termNote>
<descrip type="reliabilityCode">3</descrip>
</tig>
 </langSet>
 </termEntry>
```

Appendix 2. IATE XML schema diagram.



CHAPTER TWELVE

DEVELOPING PARSING RULES WITHIN ARTEMIS: THE CASE OF DO AUXILIARY INSERTION

ANA DÍAZ GALÁN
INSTITUTO DE LINGÜÍSTICA ANDRÉS BELLO,
UNIVERSIDAD DE LA LAGUNA
AND MARÍA DEL CARMEN FUMERO PÉREZ
INSTITUTO DE LINGÜÍSTICA ANDRÉS BELLO,
UNIVERSIDAD DE LA LAGUNA[1]

1. Introduction

The research carried out in this chapter is situated within Functional Grammar Knowledge Base (FunGramKB) (Periñán-Pascual and Arcas Túnez 2010), a multipurpose lexico-conceptual knowledge base designed to be used in different Natural Language Processing (NLP) tasks, such as information retrieval and extraction, machine translation, dialogue-based systems, etc. In order to comply with any of these tasks, it is necessary to complement the knowledge base with a device that binds natural language fragments with their corresponding grammatical and semantic structures. This is the aim of ARTEMIS (Automatically Representing Text Meaning via an Interlingua–based System), a parsing device based on a sound linguistic model, Role and Reference Grammar (RRG). This lexical-functionalist description of language provided by Van Valin (2005) is suitable for NLP due to its typological and functional nature, as well as the

[1] This work has been developed within the framework of the research project "Desarrollo de plantillas léxicas y de construcciones gramaticales en inglés y español. Aplicación en los sistemas de recuperación de la información en entornos multilingües" (FF12011-29798-C02-02), funded by the Spanish Ministry of Science.

connection it presents between the syntactic and the semantic components provided by its linking algorithm (Periñán-Pascual and Arcas Túnez 2014).

The analysis presented in this chapter seeks to contribute to the development of ARTEMIS by studying a specific grammatical phenomenon: the insertion of the DO operator in simple sentences in English. The chapter is organized as follows: Section 2 offers a brief overview of ARTEMIS, FunGramKB and RRG, giving special emphasis to the aspects that facilitate the integration of the grammatical model into the NLP prototype and the role that semantic information stemming from FunGramKB plays in the process, namely the function of ontological concepts in building up the semantic representation of clauses. Section 3 is essentially descriptive, as it seeks to provide the information necessary to understand the processes that trigger the DO insertion. This is followed by the interpretation of such processes within the framework of Role and Reference Grammar (Section 4) and the subsequent necessary adjustments to be carried out for their integration in the parsing device within ARTEMIS (Section 5). Finally, some concluding and prospective remarks are presented in Section 6.

2. ARTEMIS, FunGramKB and RRG

As stated in Periñán-Pascual and Arcas Túnez (2014:181), ARTEMIS can be described as a bottom-up chart parser with top-down prediction, which transforms or transduces the natural language input it receives to its equivalent grammatical and semantic structures. To do so, it is provided with three components: the Conceptual Logical Structure Constructor (CLS constructor), the COREL-Scheme Builder and the Grammar Development Environment (GDE).

The Conceptual Logical Structure (CLS) is a text meaning representation that results from the extension of the RRG logical structure. One of the main differences between the two is that the CLS involves the use of FunGramKB ontological concepts instead of predicates, since the Ontology constitutes a pivotal component of the whole architecture of FunGramKB; another distinguishing feature is the fact these concepts are also assigned a thematic role. The representation is further enriched by an *Aktionsart* operator, which specifies the type of event, and one or more constructional operators that integrate argument-structure constructions, what in the Lexical Constructional Model (LCM) (Ruiz de Mendoza and Mairal Usón 2008; Mairal Usón and Ruiz de Mendoza 2009) are called L1-CONSTRUCTIONS (e.g. Kernel, Resultative, Caused Motion, etc.). This integration of constructional meaning into the CLS also has a direct

influence on the Layered Structure of the Clause (LSC), resulting in the addition of a new node, i.e. the L1-CONSTRUCTION node, which is situated between the CORE and the CLAUSE nodes.

As we can see in the enhanced LSC model represented in Figure 1, taken from Periñán-Pascual (2013: 221), L1-constructions can appear recursively and may add an argument to the structure.

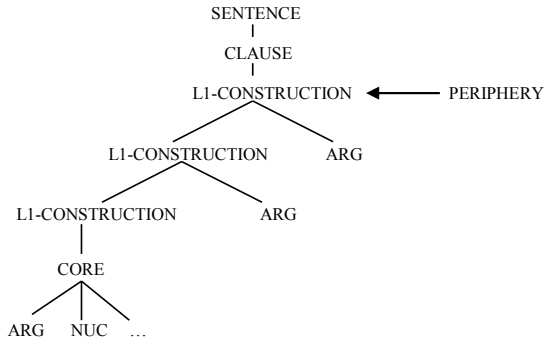


Figure 1. Enhanced model of LSC (unrefined tree)

A final process affecting the CLS is that it is transduced into a semantic conceptual representation in COREL (Conceptual Representation Language), i.e. the formal representation language in FunGramKB.

This evolution from the logical structure in RRG to the enhanced COREL scheme in FunGramKB can be illustrated with the example (1), taken from Periñán-Pascual and Arcas Túnez (2014: 175-177).

(1) Peter broke the glass.

RRG Logical structure:

$<_{IF}^{DEC}<_{TNS}^{PAST}<_{ASP}^{PERF}$ < [do` (peter, Ø)] CAUSE [BECOME broken`(glass)]>>>>

CLS:

$<_{IF}^{DEC}<_{TNS}^{PAST}<_{ASP}^{PERF}<_{CONSTR-L1}^{KER2}<[_{AKT}$ +BREAK_00(%PETER_00-Theme,$GLASS_00 Referent)]]>>>>

Extended COREL scheme:

+(e1: +DAMAGE_00 (x1: %PETER_00)$_{Theme}$ (x2: $GLASS_00)$_{Referent}$ (f1: (e2:+SPLIT_00 (x1)$_{Theme}$ x2$_{Referent}$))$_{Result}$)

The ultimate aim of ARTEMIS is, in fact, to build the CLS and the COREL scheme automatically. In this chapter, of the three components that make up ARTEMIS, it is the GDE that we are going to focus on, since our aim is to enrich the information which is necessary for an effective parsing of a common, yet very relevant, grammatical phenomenon in English: the insertion of the DO auxiliary in simple sentences.

The GDE comprises the set of rules necessary for the parsing of natural language expressions. They are feature-based production rules which are subject to the linear order of the constituents (Periñán-Pascual 2013: 222). Within the GDE, we can distinguish three different types of rules: syntactic, constructional and lexical. Syntactic rules aim at building the enhanced framework of the Layered Structure of the Clause (LSC) and are provided with the syntactic units we saw in Figure 1: nucleus (NUC), core (CORE), construction (CONSTR), periphery (PER) and clause (CL). As an example, we find the following syntactic rule in ARTEMIS version 1.0:

```
S -> CL
CL[Tense=?t, Template=?tpl] -> CONSTR-L1[Tense=?t, Template=?tpl] ll
CONSTR L1 [Tense=?t, Template=?tpl] PER
```

While constructional rules govern the insertion of constructions into the LSC, lexical rules offer morphosyntactic and semantic information about grammatical units. The lexical details contained in lexical rules are drawn from both the FunGramKB Ontology and the Lexicon, which is the language-dependent component of FunGramKB, that is, where we find morphosyntactic, collocational and pragmatic information related to lexical units. These lexical rules are formally represented as Attribute-Value Matrices (AVMs). The following AVM is an example of such a formal description. In this case, together with the actual concept which is being defined (and which must always be included), the features listed are those relevant to the noun category, i.e. countability and number:

```
</Category>
<Category Type= “NOUN”>
<Attribute ID= “Concept”/>
<Attribute ID= “Count”/>
<Attribute ID= “Num”/>
</Category>
```

3. Formal Description of the DO Operator

The transduction of the grammatical phenomenon under analysis in this study into its computational equivalent requires, in our opinion, a previous description of its most common behaviour in simple sentences in English. Therefore, an examination of the relevant features related to the appearance of the DO operator, as offered by well-known formal grammatical descriptions (cf. Quirk et al. 1985), is a previous necessary step. Accordingly, we dedicate this section to the revision of the uses of this operator, which will enable us to implement them computationally.

DO is one of the primary auxiliary verbs in English together with BE and HAVE. In its auxiliary use, its mission is to form the tense, mood and voice of other verbs and to operate on the sentence. According to Quirk et al. (1985: 133), in those cases in which there is no other operator present in the construction, the DO auxiliary is needed for negation, interrogation, emphasis and certain cases of inversion. In what follows, we briefly review and exemplify each of these uses of DO.

The DO support (also known as the DO periphrasis) refers to the use of DO when it acts as an empty or dummy operator in those sentences in which no other operator has been provided, that is, when there is no semantic reason for an operator other than DO to be present in the sentence. This is the case in the uses illustrated below:

a) To form the negative of simple indicative clauses in which the verb is in the simple present or in the simple past, we need to insert the DO auxiliary and the negator *not* before the verb. The lexical verb has to be in the bare-infinitive form, in such a way that it is the DO operator itself which indicates the tense and the person. The enclitic or contracted negative form (*n't*) is the most common, at least, in non-formal English, as in (2) and (3).

(2) My sister *doesn't* want any ice cream.

(3) They *didn't* leave early.

To form the negative of imperative clauses, we also need DO; in this case, however, the verb is not in the simple present or past tense, since the imperative is characterized by being tenseless. Nevertheless, the syntactic behaviour is exactly the same as the one described for simple indicative clauses, that is, DO + negator + bare infinitive; the only difference is that (in its most common use) this is a subjectless structure, as evidenced in (4).

(4) *Don't* be late

b) DO is also required to form the interrogative of constructions -either positive or negative- with verbs in the simple present or past tense. This includes both *yes-no questions* and *wh-questions*, with the exception of those in which the *wh-element* is the subject of the sentence, in which case the DO operator is not necessary. As the examples (5-8) show, the operator must be placed before the subject in both types of questions, i.e. DO + subject + bare infinitive / *wh-element* + DO + subject + bare infinitive.

(5) *Does* he want it?

(6) *Did* the man come yesterday?

(7) *Didn't* you hear the news?

(8) Who *do* they know?

c) DO is also required in the so-called "emphatic construction"; DO is necessary to indicate the simple present or past tense and the person, as in (9) and (10).

(9) He *does* know the answer.

(10) He *did* ask her to marry him.

Quirk et al. (1985:124-125) distinguish two main uses of this emphatic operator, either to refute an implicit or explicit negative or to express purely emotive force with no contrastive meaning, as in (11) and (12).

(11) You *DID* speak to her? ["I thought you didn't"]

(12) I *DO* wish you would listen.

The so-called "persuasive imperative" introduced by DO would also be included in this category, as illustrated in (13) and (14).

(13) *DO* sit down.

(14) *DO* be quiet.

However, in our opinion, these uses of DO differ significantly from the others mentioned previously, since the presence of the operator seems to be pragmatically motivated. Quirk et al. themselves (1985: 134) spoke of the dubious nature of DO as an operator in this type of constructions.

Rather than carrying out a syntactic operation, the presence of DO in these sentences seeks to achieve an intended pragmatic effect. Therefore, we consider that this type of structures would be better described as what in FunGramKB are called "constructions", which Periñán-Pascual and Arcas Túnez defined as follows:

> A construction is a pairing of form and meaning serving as a building block in the compositionality of sentential semantics, whose meaning cannot be fully derived from the sum of the lexical meanings of the individual constructs taking part in the utterance. (Periñán-Pascual and Arcas Túnez 2014:172)

Within FunGramKB, constructions are stored in the Grammaticon. Following the LCM, four different modules are distinguished in this component: argumental (level 1), implicational (level 2), illocutionary (level 3) and discursive (level 4). We think that the fixed syntactic structure which characterizes this type of sentence, together with its inherent implicational meaning, allows us to classify DO emphatic structures as level-2 constructions. This implies that they should be described and stored in the Grammaticon and not in the GDE. Thus, DO emphatic constructions will not be included in our syntactic rules.

d) Finally, the DO operator appears in those cases in which there is inversion after an initial negative element. It can be either an element negative in meaning but not in form (e.g. *seldom*, *rarely*, *scarcely*, *barely* or *little*) or an element negative in form and meaning (e.g. *never* or *not*), as shown in (15-17).

(15) Very seldom *does* the programme have any documentation.

(16) Only rarely *does* a suitable piece of information present itself.

(17) Never *did* I dream before.

When the elements negated are non-subject constituents placed in the initial position, as in the example (18) provided by Quirk et al. (1985: 779), the sentences follow the same pattern displayed in the examples (15-17): *negative element* + DO + *subject* + bare *infinitive.*

(18) *Not one bottle* did we leave behind

However, the subjects affected by narrow-scope negation do not trigger the inversion and do not require the presence of DO, as in (19).

(19) No one listens to me.

Although its consequences lie beyond the scope of this chapter, we have observed an interesting parallelism between this type of local negation and *wh-questions* in their behaviour with respect to the insertion of DO. The disparate behaviour between negative subjects and objects in terms of the DO insertion, which we have just described, seems to correlate with the presence or absence of DO in *wh-questions* depending on whether the *wh-element* is a subject or an object argument, as shown in (20) and (21).

(20) Who came to the party?

(21) Who *did* you meet there?

The review of the most important uses of DO shows that this operator not only marks tense, number and person systematically but also implies a precondition on the structural features of the clauses in which the DO insertion takes place, namely, that the main verb is not marked for aspect. The rationale for this lies in the fact that aspectual modification triggers the introduction of an auxiliary verb other than DO (e.g. BE for progressive and HAVE for perfective), thus blocking the use of DO. Such a distributional variation is reflected in the pairs of the examples (22-24) for positive, negative and interrogative sentences.

(22) Marked for aspect:	John is/was eating
Unmarked for aspect:	John eats/ate
(23) Marked for aspect:	John isn't/wasn't eating
Unmarked for aspect:	John doesn't/didn't eat
(24) Marked for aspect:	Is/was John eating?
Unmarked for aspect:	Does/did John eat?

To conclude this section, in Table 13 we summarize the main syntactic patterns of the types of sentences in which the DO operator is required. Therefore, these patterns will have to be described in the GDE, with the exception of the emphatic DO, which, given its constructional nature, should be accounted for in the FunGramKB Grammaticon.

Table 1. Patterns of DO operator insertion in simple sentences according to formal grammars

Operation		aux		Lexical verb	example
Negation indicative sentences	subject	**DO**	negator	bare infinitive	*They didn´t leave early*
Negation imperative sentences	-	**DO**	negator	bare infinitive	*Don't be late*
Yes/no Questions		**DO**	subject	bare infinitive	*Does he want it?*
Wh-Questions (unless wh- el. is subject)	Wh-element	**DO**	subject	bare infinitive	*Who do they know?*
Emphatic construction	subject	**DO**		bare infinitive	*He DOES know the answer*
Inversion	Negative element	**DO**	subject	bare infinitive	*Never did I dream before*

4. DO Operator in RRG

From the point of view of the application of RRG to the parsing of DO in ARTEMIS, there are a number of issues to be addressed. One crucial feature is the fact that DO does not seem to participate in the constituent projection within the LSC in RRG. In general, the DO auxiliary is just a functional exponent of a number of operators: negation, tense or illocutionary force. This is a common phenomenon of all lexical items in RRG belonging to the AUX category. In this regard, it is worth mentioning that the status of AUX is not consistently addressed in RRG, as there are a few cases where such a node is inserted in the constituent structure. The example in Figure 2 is one of the few instances we have found in which Van Valin (2005:13-14) actually represented AUX in the constituent structure.

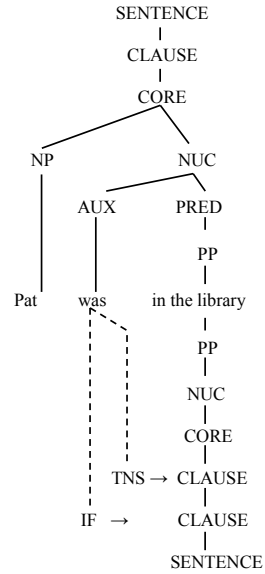


Figure 2. Instance of AUX in a syntactic RRG template

Note, however, that the author is referring to what is known as copular BE in formal grammars; other auxiliaries such as BE for progressive or HAVE for perfective are considered operators and thus do not appear explicitly in the constituent structure. There are also many other examples where this is not done, and auxiliaries such as DO are not bound to anything, as can be seen in Figure 3.

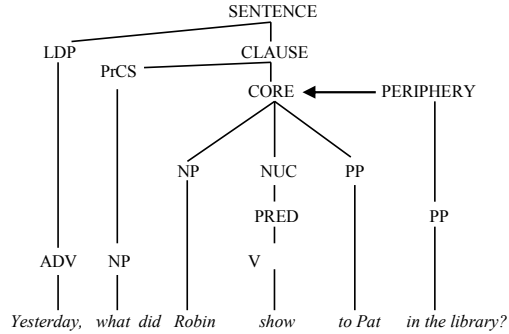


Figure 3. Example of DO auxiliary in an RRG syntactic template

With respect to the position of DO in the node sequence, Van Valin himself (2005: 8) indicated that "[…] the auxiliary *did* is not attached to anything, because it is not part of the nucleus, core or periphery. It is, rather, the morphological realization of a tense 'operator' which modifies the clause". Van Valin justified the non-insertion of elements such as DO in the constituent structure by saying that:

> Since operators are technically not part of the nucleus, core or periphery, but rather, are modifiers of these units and combinations thereof, it is reasonable that they should be represented separately from the predicates and arguments they modify. (Van Valin 2005: 11)

What he proposed, instead, is a formalization of the LSC where operators occupy a different projection (i.e. operator projection) from that of predicates and arguments (i.e. constituent projection). As we can see in Figure 4, taken from Van Valin (2005: 12), within the operator projection, we find grammatical categories such as aspect, tense, modality, negation, illocutionary force, etc. which modify the different layers of the clause.

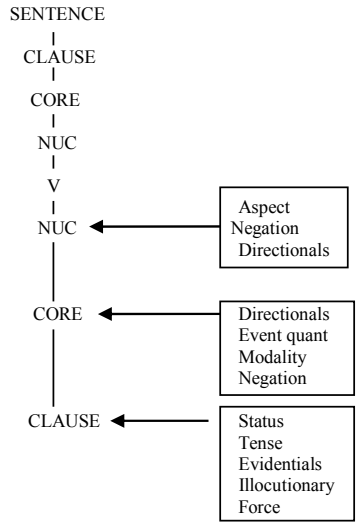


Figure 4. Layered structure of the clause with operator projection

Of the operators in Figure 4, the ones involved in the insertion of DO would be (i) aspect, which is located in the nucleus and is relevant because it is only when the main verb is not marked for aspect that the presence of DO is required, (ii) core negation to account for external or clausal

negation, (iii) tense, since one of the functions of the DO operator is that of tense carrier, and (iv) illocutionary force to mark the insertion of DO in interrogative sentences which require this operator.

To account for the insertion of the DO operator in ARTEMIS, we arrive at the conclusion that it is necessary to use the label AUX in a more extensive manner than RRG, as it is of paramount importance to describe a collection of functional items that must be explicitly included in the linearization rules within the GDE. Thus, our use of the AUX category will be more in line with unification-based approaches (cf. Sag et al. 2003: 59), which use this label to distinguish auxiliary verbs from lexical verbs and, therefore, allow us to classify the different types of auxiliaries which the parser may find.

Another factor we need to address is the computational representation of the AUX element. In other words, we need to solve the potential problem caused by the non-insertion of AUX within the constituent structure in RRG, since ARTEMIS would not be able to recognize it and parse it accordingly. In our opinion, this problem would be solved by spelling out a syntactic rule for each of the levels affected by the insertion of DO and formalizing in each of them the exact position the AUX element has to occupy. Van Valin himself (2005: 10) clearly states the relevance of the position of the operators in the syntactic sequence by saying that, in English, illocutionary force and tense are linked because "illocutionary force is indicated by the position of the tense marker in the main clause". Tense is core-internal in declaratives and core-initial in interrogatives, but imperatives are characterized by having no tense.

Following Van Valin (2005:12), for the three phenomena discussed in this study (i.e. negation, inversion and interrogation), we think that DO, which indicates both tense and illocutionary force, should be located in the core node. Negation will, therefore, affect the predicate and its arguments. In the case of inversion, DO will commonly appear after a negative element in the initial CORE periphery and immediately before the subject argument. The negative nature of the adverb that triggers the inversion will eventually have to be accounted for in the AVM of its category type, indicating its negative polarity. In those instances of inversion which are not triggered by the presence of a negative adverb but by a negative non-subject argument, DO still remains in the same pre-subject position. However, the inversion trigger remains in the core position in opposition to the negative adverb, which, as we have said, is in the initial CORE periphery. Finally, as stated above, illocutionary force in English is indicated by the position of the tense marker in the main clause: the core initial position for both *yes-no questions* and *wh-questions*. However, the

scope of the DO interrogative operator percolates through the clause because illocutionary force is a clausal operator modifying the clause as a whole. Figure 5 summarizes the location of the DO operator in relation to each of the three phenomena we are describing.

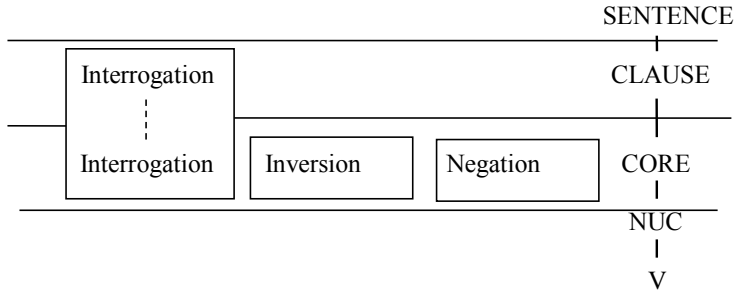


Figure 5. DO insertion and RRG levels

5. DO Operator in ARTEMIS

To achieve our purpose of designing the GDE rules that describe the insertion of the DO operator in English, the phenomena we have seen so far call for the redefinition of three syntactic categories in ARTEMIS**.** The first issue we have to address is the absence of the aspect that is implicit in the presence of the DO operator in the clause. Therefore, the nucleus category becomes relevant in the description of this phenomenon, because it is at this level of the LSC that we find aspect. Secondly, the core category in ARTEMIS needs to be refined to account for the actual insertion of the DO auxiliary that occurs at this level. In the third place, we need to enrich the clause category to deal with the illocutionary force which stems from the core but reaches the clause node.

Furthermore, since in ARTEMIS operators are substituted for attributes in AVMs, we have to elucidate what features should be described in relation to the grammatical phenomena triggered by the insertion of DO in the library of AVMs for grammatical units. Additionally, the insertion of DO in ARTEMIS will necessarily involve not only the revision of the AUX category, but also the creation of a new category for its negative counterpart: AUXN. The remainder of this chapter is devoted to these issues.

5.1. AVMs and Lexical Rules

The first AVM to be updated is the one encoding the features of the NUC category to add a new attribute for aspect, which will enable us to indicate that it is empty when the DO insertion takes place.

AVM for NUC (updated):
</Category>
<Category Type= “NUC”>
<Attribute ID= “Aspect”/>
<Attribute ID= “Num”/>
<Attribute ID= “Template”/>
<Attribute ID= “Tense”/>
</Category>

This also involves introducing the values for the two different types of aspect in the inventory of attributes; the description in XML format will be as follows:

Attribute for Aspect:
<Attribute ID=“Aspect” obl=“*”num=“s”>
<Value>?a</Value>
<Value>prog</Value>
<Value>perf</Value>
</Attribute>

This attribute for aspect reads in the following manner: it is not an obligatory feature since it is represented by obl=“*”; when present in a clause, it may conflate more than one value, as there can be progressive and perfective verb forms simultaneously (e.g. “*have been writing*”), so this is represented by num=“s” (where *s* stands for some); finally, the value ?a indicates that aspect is inherited from the corresponding auxiliary verb up to the NUC node.

The existing AVM for the core has already been extended by Cortés Rodríguez (2015), so that it now allows for polarity, which is necessary to account for the presence of DO in negative sentences.

AVM for CORE (updated):
</Category>
<Category Type= “CORE”>
<Attribute ID= “Mod”/>
<Attribute ID= “Neg”/>
<Attribute ID= “Num”/>
<Attribute ID= “Pol”/>

<Attribute ID= “Template”/>
<Attribute ID= “Tense”/>
</Category>

Accordingly, we have also created the corresponding attribute for polarity, which can be positive or negative:

Attribute for polarity:
<Attribute ID=“Pol” obl=“*”num=“s”>
<Value>?p</Value>
<Value>pos</Value>
<Value>neg</Value>
</Attribute>

Finally, the AVM that describes the clause, updated by Cortés Rodríguez (2015), now features the illocutionary-force attribute needed to describe the interrogative structures with DO.

AVM for CLAUSE (updated):
</Category>
<Category Type= “CL”>
<Attribute ID= “Illoc”/>
<Attribute ID= “Status”/>
<Attribute ID= “Template”/>
<Attribute ID= “Tense”/>
</Category>

As before, we have also designed the corresponding attribute for illocutionary force with its values: declarative, interrogative or imperative.

Attribute for illocutionary force:
<Attribute ID=“Illoc” obl=“+”num=“1”>
<Value>?i</Value>
<Value>declarative</Value>
<Value>interrogative</Value>
<Value>imperative</Value>
</Attribute>

Furthermore, we have to not only redefine the AUX category as it is currently in ARTEMIS, but also create a new one for its negative counterpart, which we have termed AUXN. This label is needed to account for enclitic or contracted negative forms with *n’t* (e.g. *don’t*, *didn’t*, etc). The list of simple part-of-speech tags in ARTEMIS already includes a tag for AUX verbs, so we only have to generate the

corresponding AVM. In the case of AUXN, we have designed both the tag and the AVM, as follows:

AVM for AUX:
</Category>
<Category Type= “AUX”>
<Attribute ID= “Aux”/>
</Category>

AVM for AUXN:
</Category>
<Category Type= “AUXN”>
<Attribute ID= “Auxn”/>
</Category>

It is also necessary to specify the different types of AUX the parser may find, which can be either primary or modal. Accordingly, we propose the following syntactic rule to distinguish the two types:

AUX -> AUXP II MOD

The equivalent rule for the negative counterpart to describe enclitic forms is as follows:

AUXN -> AUXPN II MODN

At the same time, the AUX or AUXN categories constitute a closed class and have to be included in the lexical rules in ARTEMIS. What we suggest is a reorganization of the existing tags and the addition of new ones which will eventually be defined. The following figure shows an example of this reorganization, which in this case affects negative enclitic auxiliaries. As the syntactic rule above reads, they can be either primary or modal; thus, we have to, firstly, create the corresponding tags (i.e. AUXPN and MODN) and AVMs, and, secondly, define all the elements listed under each category in Figure 6.

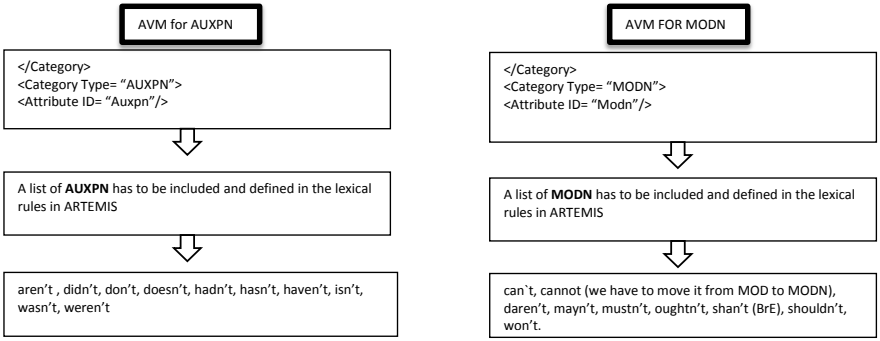


Figure 6. Example of reorganization of ARTEMIS AUX category

The following screen capture shows the ARTEMIS editor. By clicking on the tab "lexical rules", an interface such as the one shown in Figure 7 appears. Lexical rules have to be written in the blank textbox on the right.

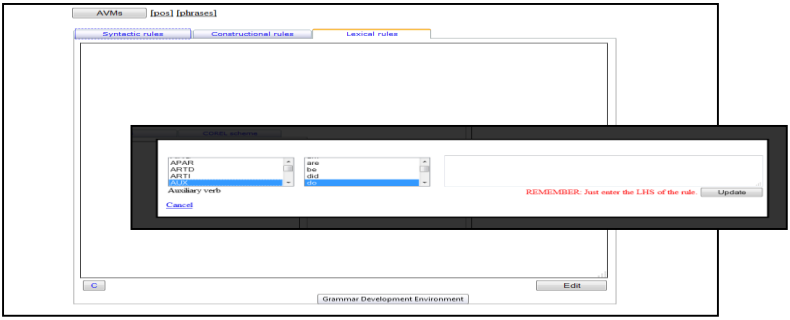


Figure 7. ARTEMIS editor

An example of a lexical rule could be the one we have created for *didn't*. This item is part of a set of negative primary auxiliaries in the past (e.g. *didn't, hadn't, wasn't, weren't*), for which we propose the label APASN, whose AVM would be the following:

</Category>
<Category Type= "APASN">
<Attribute ID= "Apasn"/>
<Attribute ID= "Aspect"/>
<Attribute ID= "Num"/>
<Attribute ID= "Per"/>
<Attribute ID= "Pol"/>
<Attribute ID= "Tense"/>
</Category>

The actual lexical rule for *didn't* would be [Pol:n,Tense:past], which only shows two values, i.e. negative polarity and past tense, since the other values present in the AVM for APASN are not saturated. This rule would now have to be inserted in the ARTEMIS editor.

5.2. Syntactic Rules

Once we have defined the categories AUX and AUXN, we have to consider the place these auxiliaries occupy in the syntactic structure, which, as we have said, is the CORE node. Therefore, the next step is to design what we think would be the initial rule for the CORE in ARTEMIS. Following Van Valin's (2005) syntactic templates, which in FunGramKB are known as kernel structures, we arrived at this basic rule for the CORE in simple sentences, which, for the sake of clarity, is presented in separate lines for each type of clause pattern:

Rule for the CORE

CORE -> RP-NUC II PER-RP-NUC-PER II PER-RP-NUC II RP-NUC-PERII
RP-NUC-PP II PER-RP-NUC-PP-PER II PER-RP-NUC-PP II RP-NUC-PP-PER II
RP-NUC-RP II PER-RP-NUC-RP-PER II PER-RP-NUC-RP II RP-NUC-RP-PER II
RP-NUC-RP-RP II PER- RP-NUC-RP-RP-PER II PER- RP-NUC-RP-RP II RP-NUC-RP-RP-PERII
RP-NUC-RP-PP II PER- RP-NUC-RP-PP-PER II PER- RP-NUC-RP-PP II RP-NUC-RP-PP-PER

Obviously, the rule is incomplete, since the attributes of each of the elements need to be added, but this will be the scope of future research. This basic rule for the CORE has to be modified according to each type of structure in which DO is inserted, in such a way that we now have to place AUX in its appropriate positions, resulting in the following syntactic rules:

Rule for the CORE in positive questions

CORE-> AUX-RP-NUC II PER- AUX-RP-NUC-PER II PER- AUX -RP-NUC II AUX- RP-NUC-PER II
AUX-RP-NUC-PP II PER AUX-RP-NUC-PP-PER II PER AUX-RP-NUC-PP II AUX-RP-NUC-PP-PER II
AUX-RP-NUC-RP II PER AUX-RP-NUC-RP-PER II PER AUX-RP-NUC-RP II AUX-RP-NUC-RP-PER II
AUX-RP-NUC-RP-RP II PER- AUX- RP-NUC-RP-RP-PER II PER AUX- RP-NUC-RP-RP II AUX-RP-NUC-RP-RP-PER II
AUX- RP-NUC-RP-PP II PER AUX- RP-NUC-RP-PP-PER II PER AUX- RP-NUC-RP-PP II AUX-RP-NUC-RP-PP-PER

To the previous rule for the CORE, in the case of positive sentences, we have added AUX in its appropriate initial position so that the parser can recognize both *yes-no questions* and *wh-questions*. In the case of the latter, the *wh-element* appears in the initial CORE periphery. Nevertheless, for *yes-no questions* we cannot have initial CORE periphery because the insertion of the AUX element displaces the periphery present in the declarative (e.g. *Yesterday they run in the park*) to the final position (e.g.

Did they run in the park yesterday?). The examples (25) and (26) show an instance of each type of interrogative.

(25) Why did Jane see him? PER-AUX-RP-NUC-RP

(26) Did they run in the park yesterday? AUX-RP-NUC -PER

Rule for the CORE in negative questions

```
CORE->  AUXN-RP-NUC II PER-AUXN-RP-NUC-PER II PER-AUXN-RP-NUC II AUXN-RP-NUC-PER II
        AUXN-RP-NUC-PP II PER AUXN -RP-NUC-PP-PER II PER-AUXN-RP-NUC-PP II AUXN-RP-NUC-
            PP-PER II
        AUXN-RP-NUC-RP II PER AUXN-RP-NUC-RP-PER II PER-AUXN-RP-NUC-RP II AUXN-RP-NUC-
            RP-PER II
        AUXN-RP-NUC-RP-RP II PER-AUXN- RP-NUC-RP-RP-PER II PER-AUXN-RP-NUC-RP-RP II AUXN-RP-NUC-RP-RP-
            PER II
        AUXN-RP-NUC-RP-PP II PER-AUXN- RP-NUC-RP-PP-PER II PER AUXN- RP-NUC-RP-PP II AUXN-RP-NUC-
            RP-PP-PER
```

In negative questions, the rule for the CORE would be the same but with AUXN instead of AUX. The same phenomenon concerning the displacement of the initial periphery takes place in *yes-no questions*, as in (28).

(27) Why didn't Jane see him? PER-AUXN-RP-NUC-RP

(28) Didn´t they run in the park yesterday? AUXN-RP-NUC -PER

Rule for the CORE in negative sentences

```
CORE ->  RP-AUXN-NUC II PER-RP-AUXN-NUC-PER II PER-RP-AUXN-NUC II RP-AUXN-NUC-PER II
         RP-AUXN-NUC-PP II PER-RP- AUXN-NUC-PP-PER II PER-RP-AUXN-NUC-PP II RP-AUXN-NUC-PP-
             PER II
         RP-AUXN-NUC-RP II PER-RP-AUXN-NUC-RP-PER II PER-AUXN-RP-NUC-RP II RP-AUXN-NUC-RP-
             PER II
         RP-AUXN-NUC-RP-RP II PER-RP-AUXN-NUC-RP-RP-PER II PER-AUXN-RP-NUC-RP-RP II RP-AUXN-NUC-RP-
             RP-PER II
         RP-AUXN-NUC-RP-PP II PER-RP-AUXN-NUC-RP-PP-PER II PER- RP-AUXN-NUC-RP-PP II RP-AUXN-NUC-
             RP-PP-PER
```

For negative sentences, we now have to add AUXN in its appropriate pre-nuclear position. With this rule we can describe clausal negation which involves enclitic negative constructions such as (29), whereas, to parse negative sentences with *not*, we would eventually have to create a different rule including this negative item.

(29) Yesterday John didn't play in the garden. PER-RP-AUXN -NUC -PER

Rule for the CORE for inversion with DO insertion

```
CORE ->  PER-AUX-RP-NUC II PER-AUX-R-NUC-PER II
         PER-AUX-RP-NUC-PP II PER-AUX-RP-NUC-PP-PER II
         PER-AUX-RP-NUC-RP II PER-AUX-RP-NUC-RP-PER II
         PER-AUX-RP-NUC-RP-RP II PER-AUX-RP-NUC-RP-RP-PER II
         PER-AUX-RP-NUC-RP-PP II PER-AUX-RP-NUC-RP-PP-PER II
```

This rule for inversion is shorter because it is restricted to those cases with an initial CORE periphery, since it is the insertion of a negative modifier phrase in this position which triggers the inversion, as in (30).

(30) Very seldom does the programme have any documentation.
PER-AUX -RP -NUC -RP

6. Conclusion

The analysis performed in this chapter is one of the first attempts at developing the elements that are part of the so-called GDE within ARTEMIS. In order to carry out this task, we have provided a description of the parsing rules and AVMs (together with their corresponding attributes) that would make an effective analysis of all the processes involving the DO insertion in English.

The aim of our study has been twofold. On the one hand, it has sought to enrich the NLP prototype providing a guide to the kind of analyses necessary for a functional description of the syntactic structures underlying a given piece of natural language. On the other hand, since the framework for grammatical description has been Role and Reference Grammar, we have also gone one step further in endowing this linguistic theory with computational adequacy.

Nevertheless, these goals will only be fully achieved after further extensive research is carried out. In particular, our analysis needs to be widened in the following aspects: (i) rules need to be completed with the attributes corresponding to each element, (ii) existing word classes should be reorganized and new ones have to be added, and (iii) new lexical rules are to be designed in order to define each lexical element.

7. References

Cortés Rodríguez, Francisco. "Towards the computational implementation of Role and Reference Grammar: Rules for the syntactic parsing of RRG phrasal constituents." VIAL (in consideration) (2015).

Mairal Usón, Ricardo and Francisco Ruiz de Mendoza "Levels of description and explanation in meaning construction." In Deconstructing Constructions, eds. Christopher Butler and Javier Martín Arista, 153–198. Amsterdam: John Benjamins, 2009.

Periñán-Pascual, Carlos and Francisco Arcas Túnez. "The Architecture of FungramKB." In Proceedings of the 7th International Conference on

Language Resources and Evaluation, 2667-2674. Malta: European Language Resources Association, 2010.

—. "The Implementation of the CLS Constructor in ARTEMIS." In Language Processing and Grammars, eds. Brian Nolan and Carlos Periñán- Pascual, 165-196. Amsterdam: John Benjamins, 2014.

Periñán-Pascual, Carlos. "Towards a Model of Constructional Meaning for natural Language Understanding." In Linking Constructions into Functional Linguistics: The role of constructions in grammar, eds. Brian Nolan and Elke Diedrichsen, 205–230. Amsterdam: John Benjamins, 2013.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. A Comprehensive Grammar of The English Language. London: Longman, 1985.

Ruiz de Mendoza Ibáñez, Francisco and Ricardo Mairal Usón "Levels of Description and Constraining Factors in Meaning Construction: an Introduction to the Lexical Constructional Model." Folia Linguistica 42-2 (2008): 355-400.

Sag, Ivan A., Thomas, Wasow and Emily M. Bender. Syntactic Theory: Formal Introduction. Stanford: Centre for the Study of Language and Information, 2003.

Van Valin, Robert D. Jr. Exploring the Syntax-Semantics Interface. Cambridge: Cambridge University Press, 2005.

CHAPTER THIRTEEN

COREFERENCE RESOLUTION WITH FUNGRAMKB

MARÍA JOSÉ RUIZ FRUTOS
UNIVERSIDAD DE BAYREUTH

Today, Natural Language Processing (NLP) systems and Semantic Web Technologies face the difficulty of properly connecting coreferences. For this reason, different projects have examined the necessity, possibilities and difficulties of including world knowledge, common-sense knowledge, and/or deep semantics in computer applications with the aim of simulating human reasoning in the process of coreference resolution. The way the lexical-conceptual knowledge base FunGramKB (Periñán Pascual and Arcas Túnez, 2005, 2007, 2010, 2011; Mairal Usón and Periñán Pascual, 2009; Periñán Pascual and Mairal Usón, 2010), the acronym of Functional Grammar Knowledge Base, could contribute to successfully resolving coreferences has already been examined (cf. Ruiz Frutos, 2012). Carrión Varela (2014a, 2014b) has also highlighted how the cultural knowledge stored in FunGramKB can help resolve coreferences. Furthermore, not only the ontological, procedural and encyclopedic knowledge stored within this knowledge base (cf. Periñán Pascual and Arcas Túnez, 2010) but also its reasoning mechanisms contribute to solving coreferences.

This chapter deals with automatic named-entity coreference resolution and is divided into the following parts: first, coreference (resolution) is briefly explained; second, the corpus and the methodology are shown; third, the integration of the linguistic, ontological, procedural and encyclopedic knowledge stored in FunGramKB is proposed for coreference resolution; finally, conclusions are drawn and some contributions of semantic web technologies for automatic coreference resolution are mentioned.

1. Coreference Resolution

Coreference consists of two or more elements that refer to the same extra-linguistic referent in the real world. Cross-document coreference occurs when the coreferential items appear in different texts. Indeed, if these coreferential items appear within the same text, they form coreferential chains, whose items establish links to one another: anaphora if the anaphor points back to the previously mentioned antecedent, and cataphora if one item points ahead to another item. Therefore, the process of anaphora resolution is the ability to determine the right antecedent. In order to do so, anaphors must be identified first, then the antecedent candidates must be found, and, lastly, the right antecedent must be determined. Anaphora resolution in NLP requires six knowledge levels: morphological, lexical, syntactical, semantic, discursive, and world knowledge.

The process of **coreference resolution** consists in identifying coreferential items (or chains) and finding the right referent in the real world by resolving different denominations for the same entity (cf. Hendrickx, Hoste and Daelemans, 2007). In order to do so, there are cases where world knowledge can become necessary, as in the case of the indirect anaphora in the bridging reference,[1] where the coreference is recognized with the help of the knowledge that is not explicitly mentioned in the text.

2. Corpus and Methodology

The analysis was carried out using a corpus of two online obituaries in English and Spanish on the death of the writer Carlos Fuentes, together with all of the readers' comments.[2] Both obituaries were published on 16 May 2012.

---

[1] Mitkov (2002: 15) explained that the indirect anaphora occurs when "a reference becomes part of the hearer's or reader's knowledge indirectly rather than by direct mention"; and Hendrickx, De Clercq and Hoste (2011: 2) added that in this case "resolution often requires some form of world knowledge – or at least information that is not explicitly represented in the textual context of the reference".

[2] Aguilar Sosa, Yanet. "El Aura de México, adiós a la última gran conciencia." El Universal, May 16, 2012. Accessed on April 11, 2015. http://www.eluniversal.com.mx/cultura/68749.html. DePalma, Anthony. "Carlos Fuentes, Mexican man of letters, dies at 83". The New York Times, May 15, 2012. Accessed on April 11, 2015. http://www.nytimes.com/2012/05/16/books/carlos-fuentes-mexican-novelist-dies-at-83.html?_r=3&pagewanted=all#comments.

First, we localized all coreferential items related to Carlos Fuentes in this corpus. Second, we focused on lexical Noun Phrase (NP) anaphors: either proper name phrases, e.g. "*Carlos Fuentes*", "*Mr. Fuentes*", "*Don Carlos*", "*Carlitos*", etc., or definite noun phrases, e.g. "*El escritor nacido el 11.11.1928*" (The writer born on 11 November 1928) or "*El autor de* Aura" (The author of *Aura*). Then, we also included indirect anaphoras, especially with *referential metonymies* (cf. Ruiz de Mendoza Ibáñez and Díez Velasco, 2004) such as "*A brave and creative voice*" or "*Mente brillante*" (Brilliant mind), whose resolution is a real challenge because of the semantic ties that must be taken into consideration.

Third, we conducted a theoretical analysis of how FunGramKB could contribute to automatic coreference resolution: not only because of its knowledge, but also because of its knowledge spreading. This was accomplished by consulting the Ontology and the English Lexicon by means of FunGramKB Suite. Due to the beginning stages of development in the Cognicon and the Onomasticon, we relied on the creation of scripts as well as on the information transferred from DBpedia (Auer et al., 2007; Bizer et al., 2009). In this way, we considered what kind of knowledge and reasoning in FunGramKB could be useful for automatic coreference resolution.

3. FunGramKB for Coreference Resolution

The architecture of FunGramKB has three models: the lexical model, the grammatical model and the conceptual model.

3.1. Linguistic Information

The lexical and morphological information of each language is stored within the lexical model. The information about the lexical units in the Lexicon, such as index, graphical variant, abbreviation, gender, dialect, style, domain, translation, etc. (cf. Mairal Usón and Periñán Pascual, 2009; Guerra García and Sacramento Lechado, 2011), facilitates coreference resolution in monolingual and multilingual texts. Moreover, the Morphicon helps with the inflectional morphology. The four levels identified in the Lexical Constructional Model (LCM, cf. Ruiz de Mendoza and Mairal Usón, 2008; Mairal Usón and Ruiz de Mendoza, 2009), which helps to construct the semantics-to-syntax linkage (cf. Van Valin and LaPolla, 1997; Van Valin, 2005), are stored within the

grammatical model (Grammaticon) of each language, currently in the process of development by Lexicom.[3]

3.2. Conceptual Knowledge

The conceptual knowledge shared by all languages is stored within the conceptual model: (a) semantic knowledge (i.e. lexical meaning) in the Ontology, (b) procedural knowledge (i.e. common sense) in the Cognicon, and (c) encyclopedic knowledge (i.e. world knowledge) in the Onomasticon. All of this knowledge is integrated because the system uses the same formal language, i.e. COREL (the acronym of COnceptual REpresentation Language, cf. Periñán Pascual and Mairal Usón, 2010). This is extremely advantageous for coreference resolution since the selection of the right coreferential item does not simply depend on semantic information, but on procedural or even encyclopedic information as well.

3.2.1. Ontological Knowledge

The word *Ontology* is usually understood as a formal representation of knowledge. In FunGramKB, the Ontology stores the concepts that a person has in mind. Each concept, which is lexicalized in each language with different lexical items, facilitates both monolingual and multilingual coreference resolution.

These concepts stored in the Ontology are organized in a subsumption hierarchy (IS-A) that allows for multiple (monotonic and non-monotonic) **inheritance**:[4] the concept expands its meaning by getting information from its superordinates. Let's look at this example from the corpus:

(1) *A great man* […].

In order to resolve this coreference, it would be very useful to have access to the semantic information of the concept to which the lexical item *man* is linked. The lexical units *man* [English], *hombre* [Spanish] and other similar words in different languages are linked to the concept +MAN_00. In the Ontology, the concept +MAN_00 is subsumed under

[3] More information about Lexicom can be found at <http://www.lexicom.es/drupal>.

[4] FunGramKB allows monotonic reasoning with strict predications and non-monotonic reasoning with defeasible predications (cf. Periñán Pascual and Arcas Túnez, 2007a).

+ADULT_00, which is again subsumed under +HUMAN_00, etc.; as we can see:

(2) #ENTITY >> #PHYSICAL >> #OBJECT >> #SELF_CONNECTED_OBJECT_00 >> + ARTIFICIAL_OBJECT_00 >> +SUBSTANCE_00 >> +SOLID_00 >> ORGANISM_00 >> **+HUMAN_00 >> +ADULT_00 >> +MAN_00**

On the other hand, **inferences** can also occur: the concept's meaning lies in the concepts used for its description (as if they were *imported concepts*). For example, the meaning of +MAN_00 lies in being a +ADULT_00 that is +MALE_00, i.e. a man is "an adult male person". At the same time, each concept can also expand its meaning by being used (as if it were an *exported concept*) to describe other concepts, so that new knowledge can be inferred. For example, +MAN_00 is used for describing +BEARD_00, +FATHER_00, +HUSBAND_00, etc. Its superordinate concept +HUMAN_00 is used for describing –among many others– the quality $AUDIBLE_00, as well as the entity +BODY_00 or the event $TOLL_00. From this network, it can be inferred that a man is an adult male person who can have beard, be a father and/or a husband, be audible, have a body and be able to toll a bell.

Regarding the indirect anaphora, let us comment on the following example:

(3) […] *A brave and creative voice* […].

The lexical units *voice* [English], *voz* [Spanish] and *Stimme* [German] are linked to the concept +VOICE_00. In the Ontology, the concept +VOICE_00 is not subsumed by +HUMAN_00, so both concepts are not related by inheritance. The metonymy *voice-human* in this indirect anaphora could be detected because +HUMAN_00 appears as a selectional preference in the second predication of its meaning postulate (i.e. *a human creates a voice with the mouth and throat*):

(4) Meaning Postulate of the entity +VOICE_00
+(e1: +BE_00 (x1: +VOICE_00)Theme (x2: +SOUND_00)Referent)
+(e2: +CREATE_00 (x3: **+HUMAN_00**)Theme (x1)Referent (f1: +MOUTH_00 & +THROAT_00)Instrument)

Let us look at another example of indirect anaphora:

(5) Mente brillante! [sic]
(Brilliant mind!)

The lexical units *mind, head, brain, intelligence* [English], *mente, cabeza, cerebro, inteligencia* [Spanish], *Verstand, Kopf, Gehirn, Intelligenz* [German] are linked to the concept +INTELLIGENCE_00; this helps to resolve coreferences when the definite NPs are synonyms. Again, in the Ontology the concept +INTELLIGENCE_00 is not subsumed by +HUMAN_00, so both concepts are not related by inheritance. The metonymy *mind-human* in this indirect anaphora could be detected because the nuclear concept +INTELLIGENCE_00 spreads its meaning on the basis of the concepts appearing in its meaning postulate: the superordinate concept +COGNITIVE_ATT_00 appears as a selectional preference in its first predication (i.e. *intelligence is a cognitive attitude*):

(6) Meaning postulate of the entity +INTELLIGENCE_00 :
+(e1: +BE_00 (x1: +INTELLIGENCE_00)Theme (x2: **+COGNITIVE_ATT_00**)Referent
*(e2: +THINK_00 (x3)Theme (x4)Referent (f1:x1)Means)

Again, regarding the concept +COGNITIVE_ATT_00, in the second predication of its meaning postulate (i.e. *a cognitive attitude exists in a human being*), the concept +HUMAN_00 also appears as a selectional preference:

(7) Meaning postulate of the entity +COGNITIVE_ATT_00:
+(e1: +BE_00 (x1: +COGNITIVE_ATT_00)Theme (x2: +ABILITY_00)Referent)
+(e2: +BE_01 (x3: **+HUMAN_00**)Theme (x1)Attribute)

Therefore, the inheritance and inference mechanisms in the MicroKnowing make the establishment of semantic ties between concepts possible, so they are extremely useful for indirect anaphoras when resolving bridging coreference.

3.2.2. Procedural knowledge

Let us look at this example of coreference:

(8) […] *el cuerpo será velado en la casa del escritor* […].
([…] a wake will be held at the writer's home […])

(9) Meaning postulate of the entity +BODY_00:
+(e1: +BE_00 (x1: +BODY_00)Theme (x2: +NATURAL_OBJECT_00 & +CORPUSCULAR_00 & +SOLID_00)Referent)
+(e2: +BE_02 (x1)Theme (x3: +HUMAN_00 ^ +ANIMAL_00)Location)

*(e3: +COMPRISE_00 (x1)Theme (x4: 1 +HEAD_00 & s +LIMB_00 & 1 +TRUNK_00)Referent)

As can be noted in Example 9, the selectional preference +HUMAN_00 appears in the meaning postulate of the concept +BODY_00. However, many other writers are also mentioned in the text. In order to identify *el cuerpo* with Carlos Fuentes, it would be very useful to know how certain societies behave after somebody's death: for example, what takes place when a wake is held. This type of procedural information is stored in the Cognicon in the form of scripts, i.e. "schemata in which a sequence of stereotypical actions is organized on the basis of temporal continuity" (Periñán Pascual, 2012: 186). If we relied on the creation of future scripts regarding some possible prototypical behaviour after somebody's death, we could access important information in order to resolve coreferences such as *el cuerpo* with Carlos Fuentes; in such a case, the script @HOLDING_A_WAKE_AT_HOME_00 would provide us with information about the typical actions that occur in a situation with a deceased body.

3.2.3. Encyclopedic knowledge

Let us look at a few other examples:

(10) El escritor mexicano nacido en Panamá […] el 11 de Noviembre de 1928 […].
(The Mexican writer born in Panama [...] on 11 November 1928)

(11) El autor de obras emblemáticas como La muerte de Artemio Cruz, Cristóbal Nonato, Aura y Las buenas conciencias […]
(The author of such emblematic works as The Death of Artemio Cruz, Christopher Unborn, Aura and The Good Conscience […])

World knowledge is extremely useful for identifying Carlos Fuentes with the coreference in Example 10. On the one hand, we need to know who *Carlos Fuentes* was, as well as what *Panamá* is. On the other hand, we also need to know Carlos Fuentes' biography, in particular, where and when he was born. This encyclopedic knowledge is stored in FunGramKB within the Onomasticon.

The role of named entities (i.e. the names of people, places, organizations, etc.) in coreference resolution has already been highlighted by Periñán Pascual and Carrión Varela (2011). For named-entity recognition, cultural knowledge is necessary. The cultural information stored in huge databases such as DBpedia can be imported into

FunGramKB as knowledge in the form of snapshots or stories, as long as the concepts required already exist in the Ontology. This task is carried out in FunGramKB by making use of COREL schemas created with DBpedia Mapper.

If we were to check what knowledge about the named entity Carlos Fuentes is available in the DBpedia, we could confirm that, after transferring the knowledge to the Onomasticon, we would be able to retrieve some information about Carlos Fuentes: a Mexican writer who was born in Panama on 11 November 1928. This would help us to resolve the coreference in Example 10. Furthermore, we would also know that he was the author of novels such as *La muerte de Artemio Cruz* or *Aura*, which would also facilitate the resolution of the other coreference in Example 11.

3.3. Extending Semantic Knowledge

As mentioned before, FunGramKB allows two types of **reasoning mechanisms**: inheritance and inference. The inheritance mechanism in the Ontology consists of predications being transferred from superordinate concepts to subordinate concepts. Example 1 was a case of inheritance: […] >> +HUMAN_00 >> +ADULT_00 >> +MAN_00. The inference mechanism in the Ontology consists of structures shared between predications linked to conceptual units, which are not subsumed. We have already seen two cases of inference in Examples 3 and 5: we saw how the concept +VOICE_00 has the concept +HUMAN_00 in its meaning postulate and how the nuclear concept +INTELLIGENCE_00 can extend its meaning postulate to the concept +HUMAN_00 via the selectional preference +COGNITIVE_ATT_00.

In addition to these inheritance and inference mechanisms, there are two **reasoning processes** in FunGramKB (cf. Periñán Pascual and Arcas Túnez, 2005): the *Microconceptual-Knowledge Spreading* (MicroKnowing) between the concepts and their meaning postulates stored in the Ontology, and the *Macroconceptual-Knowledge Spreading* (MacroKnowing) between the semantic knowledge from the Ontology and the scripts stored in the Cognicon. Furthermore, the possibility of designing a reasoning engine for semantic, procedural and encyclopedic knowledge has already been pointed out by Periñán Pascual and Carrión Varela (2011: 99) as a facilitator to knowledge integration. This achievement could be extremely useful when it comes to resolving named-entity coreference. I propose calling this third reasoning process in FunGramKB the *Bioconceptual-Knowledge Spreading* (BioKnowing)

because the *bio-microstructures* and the *bio-macrostructures* from the Onomasticon are taken into account. This BioKnowing could help to resolve metaphors in indirect anaphora as in the following bridging coreference:

(12) Nuestro Virgilio.
(Our Virgil)

If we relied on the knowledge stored in the Onomasticon about the Roman poet Virgil as well as in the Cognicon with the complex script @PRAISING, consisting of the subscript @COMPARING WITH CELEBRITIES (where it would be noted that in some academic fields comparing someone to Greco-Roman authors is a form of praise), then Example 12 could rightly be identified not only with Virgil, but precisely with the text's topic: Carlos Fuentes.

4. Conclusions

Based on the results obtained from this analysis, we can conclude that, in order to resolve coreference by using FunGramKB, the following steps should be taken into account:

- In the case of proper names, other possible antecedents with the same denomination will be localized in the text, including variants and aliases.
- In the case of definite NPs, coreference candidates will be tested by checking compatibilities such as gender and number (especially in the Lexicon, the Ontology and the Onomasticon), semantic features such as +HUMAN_00 (especially in the Ontology and the Onomasticon) and semantic ties between concepts such as synonyms or hyperonyms (especially in the Lexicon and the Ontology with inheritance and inference mechanisms).
- The right candidate will be selected based on the information obtained from the BioKnowing, where knowledge is spread throughout the whole conceptual model, thus taking into account not only the meaning, but also common-sense and world knowledge. Nevertheless, it would also be very useful to take into account other features such as textual distance or textual frequency.

Finally, because of their great importance, some contributions of the **Semantic Web Technologies** to automatic coreference resolution will be

briefly mentioned. These technologies offer useful tools, data and knowledge such as (a) URIs for referring to concrete entities, (b) RDF for describing data (i.e. representing properties and relationships among resources), as well as RDF schemas for defining data (i.e. defining classes, class instantiations, properties and hierarchies) so that it is possible to express facts within the assertional knowledge (RDF) based on the terminological knowledge (RDFS), (c) SPARQL (the Query Language for RDF) for extracting data, (d) Description Logics for constructing complex descriptions (concepts/roles) from simple descriptions, as well as for deductions and reasoning, and (e) the Web Ontology Language OWL for relating classes, properties and individuals. All of this allows machines to get, ask for and produce information. Lastly, the linked data cloud has become a huge information resource (whereas DBpedia is just a small node) and offers a great amount of data for developing tasks such as named-entity recognition.

This chapter has analysed to what extent FunGramKB knowledge contributes to the resolution of coreferences with the linguistic information about the lexical items and especially with the ontological, procedural and encyclopedic knowledge about different conceptual schemata. All this information allows for monolingual and multilingual coreference resolution, as well as named-entity recognition, by using common-sense and world knowledge.

Attention has been drawn to reasoning with FunGramKB through the inheritance and inference mechanisms within the MicroKnowing (where synonyms and hyperonyms play an important role, especially for indirect anaphora), the MacroKnowing and what I have called the *BioKnowing*. Finally, conclusions were drawn and some contributions of the Semantic Web Technologies to automatic coreference resolution have been mentioned.

5. References

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary Ives. “DBpedia: a nucleus for a Web of open data”. In The Semantic Web. Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, edited by Karl Aberer (et al.), 722-735. Berlin and Heidelberg: Springer-Verlag, 2007. Accessed April 11, 2015. http://www.informatik.uni-leipzig.de/~auer/publication/dbpedia.pdf.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak and Sebastian Hellmann. “DBpedia: a

crystallization point for the Web of data". Journal of Web Semantics: Science, Services and Agents on the World Wide Web 7(3), (2009): 154-165.

Carrión Varela, María Ll. "Resolución de anáforas que requieren conocimiento cultural con la herramienta FunGramKB". Revista de lingüística y lenguas aplicadas, 9(1), (2014a): 01-13. Accessed January 24, 2015. http://polipapers.upv.es/index.php/rdlyla/article/view/2003/3181.

—. Implementación de reglas de proyección conceptuales en el marco de la semántica profunda para la reutilización de bases de conocimiento enciclopédico. PhD Dissertation. Madrid: Universidad Nacional de Educación a Distancia, 2014b. Accessed January 24, 2015. http://e-spacio.uned.es/fez/eserv.php?pid=tesisuned:Filologia-Mlcarrion&dsID=Documento.pdf.

Guerra García, Fátima and Elena Sacramento Lechado. „El módulo léxico de FunGramKB". ANGLOGERMANICA ONLINE, 8, (2011): 52-65. Accessed January 24, 2015. http://www.fungramkb.com/resources/papers/fgkb04.pdf.

Hendrickx, Iris, Orphée De Clercq and Véronique Hoste. „Analysis and Reference Resolution of Bridge Anaphora across Different Text Genres". In Anaphora Processing and Applications. DAARC 2011, edited by Iris Hendrickx, Sobha L. Devi, António Branco and Ruslan Mitkov, 1-11. Berlin: Springer-Verlag, 2011. Accessed January 24, 2015. https://biblio.ugent.be/input/download?func=downloadFile&recordOId=2000619&fileOId=2129088.

Hendrickx, Iris, Véronique Hoste and Walter Daelemans. Evaluating hybrid versus data-driven coreference resolution. In Anaphora: Analysis, Algorithms and Applications. DAARC 2007, edited by António Branco, 137-150. Berlin: Springer-Verlag, 2007. Accessed January 24, 2015. http://www.cnts.ua.ac.be/~iris/corea/publications/daarc2007.pdf.

Mairal Usón, Ricardo and Periñán Pascual, Carlos: "The anatomy of the lexicon component within the framework of a conceptual knowledge base". Revista Española de Lingüística Aplicada, 22, (2009): 217-244.

Mitkov, Ruslan. Anaphora resolution. London: Longman, 2002.

Periñán Pascual, Carlos. "The situated common-sense knowledge in FunGramKB". Review of Cognitive Linguistics, 10(1), (2012): 184-214. Accessed January 24, 2015. http://www.fungramkb.com/resources/papers/020.pdf.

Periñán Pascual, Carlos and Francisco Arcas Túnez. "Microconceptual-Knowledge Spreading in FunGramKB". In 9th IASTED International Conference on Artificial Intelligence and Soft Computing, 239-244. Anaheim-Calgary-Zurich: ACTA Press, 2005. Accessed January 24, 2015. http://www.fungramkb.com/resources/papers/002.pdf.

—. "Cognitive Modules of an NLP Knowledge Base for Language Understanding". Procesamiento del Lenguaje Natural, 39, (2007): 197-204. Accessed January 24, 2015. http://www.sepln.org/revistaSEPLN/revista/39/24.pdf.

—. "The Architecture of FunGramKB". In Proceedings of the Seventh International Conference on Language Resources and Evaluation, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner and Daniel Tapias, 2667-2674. Valletta, Malta: European Language Resources Association (ELRA), 2010. Accessed January 24, 2015. http://www.fungramkb.com/resources/papers/011.pdf.

—. "Introducción a FunGramKB". In ANGLOGERMANICA ONLINE, 8, (2011): 1-15. Accessed April 11, 2015. http://www.fungramkb.com/resources/papers/fgkb01.pdf.

Periñán Pascual, Carlos and María L. Carrión Varela. "FungramKB y el Conocimiento Cultural". ANGLOGERMANICA ONLINE, 8, (2011): 87-105. Accessed January 24, 2015. http://www.fungramkb.com/resources/papers/fgkb06.pdf.

Periñán Pascual, Carlos and Ricardo Mairal Usón. "La gramática de COREL: un lenguaje de representación conceptual". Onomázein, 21, (2010): 11-45. Accessed January 24, 2015. http://www.fungramkb.com/resources/papers/012.pdf.

Ruiz de Mendoza Ibáñez, Francisco J. and Olga I. Díez Velasco. "Metonymic Motivation in Anaphoric Reference". In Studies in Linguistic Motivation (Cognitive Linguistics Research), edited by Günter Radden and Klaus-Uwe Panther, 293-320. Berlin and New York: Mouton de Gruyter, 2004.

Ruiz de Mendoza Ibáñez, Francisco J. and Ricardo Mairal Usón. "Levels of description and constraining factors in meaning construction: an introduction to the Lexical Constructional Model". Folia Linguistica 42(2), (2008): 355-400. Accessed January 7, 2015. http://www.lexicom.es/drupal/files/RM_Mairal_2008_Folia_Linguistica.pdf.

Ruiz Frutos, María José. Resolución de la anáfora correferencial con FunGramKB. Master's thesis. Madrid: Universidad Nacional de Educación a Distancia, 2012. Accessed January 24, 2015.

http://62.204.194.43/fez/eserv/bibliuned:master-Filologia-TICETL-Mjfrutos/Frutos_mariajose_TFM.pdf.

Van Valin, Robert D. Exploring the Syntax-Semantics Interface. Cambridge: Cambridge University Press, 2005.

Van Valin, Robert D. and Randy J. LaPolla. Syntax: Structure, Meaning and Function. Cambridge: Cambridge University Press, 1997.

CHAPTER FOURTEEN

THE INTEGRATION OF THE CONCEPT +CRIME_00 IN FUNGRAMKB AND THE CONCEPTUALIZATION OR HIERARCHIZATION PROBLEMS INVOLVED

ÁNGELA ALAMEDA HERNÁNDEZ
UNIVERSIDAD DE GRANADA
AND ÁNGEL FELICES LAGO
UNIVERSIDAD DE GRANADA

1. Introduction

In the last few years, a few contributions have been made in an attempt to connect the Core Ontology of the knowledge base named FunGramKB (Periñán-Pascual and Arcas-Túnez 2010a, 2010b) to other domain-specific ontologies (Ureña Gómez-Moreno, Alameda-Hernández, Felices-Lago 2011; Felices-Lago and Ureña Gómez-Moreno 2012, 2014; Faber, Mairal-Usón and Magaña-Redondo 2011; Periñán-Pascual and Arcas-Túnez 2014). The main purpose of these studies has been to expand the lexical and conceptual repositories integrated in FunGramKB and facilitate the application of Natural Language Processing (henceforth, NLP) tasks to expert knowledge.

In the following lines, we will focus our attention on the research conducted so far in the *Globalcrimeterm* satellite ontology (integrated in FunGramKB) and, more precisely, on the methodological problems encountered in the conceptualization of the superordinate basic concept +CRIME_00 and its configuration as an umbrella concept involving a set

of subordinate concepts[1]. Section 2 will deal with the theoretical background of both FunGramKB and *Globalcrimeterm*. Section 3 will show how the selection of conceptual units, such as +CRIME_00, has been assisted by the compilation of a specialized corpus (i.e. semi-automatic process) and has also been followed by a thorough lexicographical analysis of general and specialized sources (i.e. manual process). This double-check method has helped the researcher determine whether the selected concepts belong to the general FunGramKB Core Ontology or whether they should be included in the domain-specific ontology (*Globalcrimeterm*). As it will be explained, +CRIME_00 has been included in the FunGramKB Subontology of Entities in the Core Ontology as a mirror concept[2] and, at the same time, it is itself a superordinate entity that comprises a variety of other terminal concepts which were properly defined and organized within the Core or the Satellite Ontology. In all the instances, the process has followed the COHERENT methodology described by Periñán and Mairal (2011), comprising the phases of conceptualization and hierarchization. Section 4 illustrates this process with one of the subordinate concepts of +CRIME_00, i.e. $CUCKOO_SMURFING_00, whose meaning postulates (henceforth, MPs) have also been created using the metalanguage known as COREL (Periñán-Pascual and Mairal-Usón 2010).[3] Finally, some conclusions will be presented in Section 5.

2. Theoretical Background of FunGramKB and the *Globalcrimeterm* Subontology

FunGramKB is an online environment for the semi-automatic construction of a multipurpose lexico-conceptual knowledge base for NLP systems, and more particularly for natural language understanding (Periñán-Pascual and Arcas-Túnez 2004, 2007, 2010a, 2010b; Periñán-Pascual and Mairal-Usón 2009, 2010). As observed in Figure 1,

[1] This satellite ontology has been developed with the financial support of the Spanish Ministry of Economy and Competitiveness, project FFI2010-15983, and has also been extended in the project FFI2014-53788-C3-1-P.

[2] The mirror concept is a concept which can be used both in the Core Ontology and in a satellite ontology. The Core Ontology stores the information shared by most speakers of a language and the satellite ontology would expand the basic MP included in the Core Ontology by the addition of the necessary predications to cover the expert knowledge.

[3] An MP is a group of logically connected predications which articulate the generic features of a concept.

FunGramKB comprises three major knowledge levels, consisting of several independent but interrelated modules:

1) Lexical level: The Lexicon stores morphosyntactic, pragmatic and collocational information about lexical units, and the Morphicon handles cases of inflectional morphology.
2) Grammatical level: The Grammaticon stores the constructional schemata which help construct the semantics-to-syntax linking algorithm (Van Valin and Lapolla 1997; Van Valin 2005).
3) Conceptual level: The Ontology is presented as a hierarchical catalogue of the concepts that a person has in mind, so here is where semantic knowledge[4] is stored in the form of MPs. The Ontology consists of a general-purpose module (i.e. Core Ontology) and several domain-specific terminological modules (i.e. satellite ontologies or subontologies). The Cognicon stores procedural knowledge by means of scripts, that is, conceptual schemata in which a sequence of stereotypical actions is organised on the basis of temporal continuity, and more particularly on Allen's temporal model (Allen 1983; Allen and Ferguson 1994), e.g. 'dine in a restaurant', 'arrest a suspect', 'launder money', etc. The Onomasticon stores information about instances of entities and events (i.e. episodic knowledge), such as The Pentagon, September 11, Bin Laden, etc. This module stores two different types of schemata (i.e. snapshots and stories), since instances can be portrayed synchronically or diachronically.

[4] These types of knowledge follow the distinctions established within the framework of cognitive psychology.

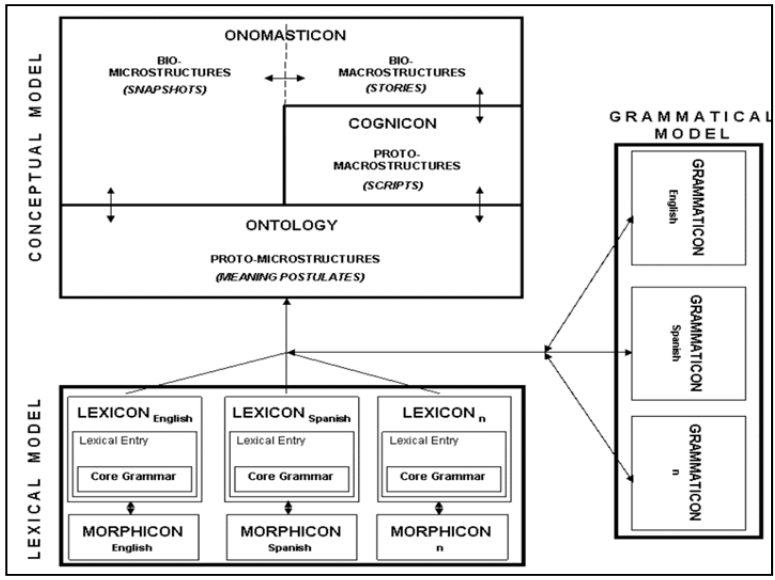


Figure 1. FunGramKB modules

Consequently, the FunGramKB Ontology is a conceptual taxonomy derived from linguistic concepts, in which interlinguistic differences in syntactic constructions do not involve conceptual differences. It is general-purpose but not domain-specific. However, since expert knowledge stems from general knowledge, it can be extended to include specialized knowledge by establishing links to domain-specific satellite ontologies, be it medicine, law, chemistry or accounting. The concepts of FunGramKB belong to three levels. The upper level is composed of 42 metaconcepts, marked with the symbol #. They constitute the upper level in the taxonomy as a result of the analysis of the most relevant linguistic ontologies, such as DOLCE (Gangemi et al. 2005), SIMPLE (Lenci 2008), SUMO (Niles and Pease 2001), etc. These metaconcepts are distributed in three subontologies: #ENTITY, #EVENT and #QUALITY. Second-level concepts, which are immediately under the metaconcepts, are marked by the sign + (e.g. +LAW_00). These concepts are used in the MPs that define basic and terminal concepts and also encode the selection restrictions in thematic frames. The third level is composed of terminal concepts, marked by $ (e.g. $WATERBOARDING_00). The difference between basic and terminal concepts is that basic concepts are used to define other concepts in MPs, whereas terminal concepts are not. Obviously, in the satellite ontologies for specialized knowledge, terminal

concepts in FunGramKB will have to be promoted to basic concepts under the adequate circumstances, or some basic concepts in the Core Ontology should become mirror concepts to cover the expert knowledge.

Globalcrimeterm is the subontology chosen for this study and lies within the framework of criminal law, even if many concepts are shared with the general world knowledge of non-expert speakers, as it is the case of mirror concepts such as +CRIME_00. The building of this subontology involved the compilation of the *Globalcrimeterm* Corpus (GCTC), which ultimately aimed at the purpose of structuring information for the retrieval of documents in professional contexts and for the solving of problem-oriented tasks in real situations. The chosen languages were English and Spanish, and the initial stages in the process of corpus compilation required a number of decisions and selections to collect and organize the input coherently and efficiently (Bowker and Pearson 2002; Koester 2010). To begin with, the legal subdomain of organized crime and terrorism was selected both for its international relevance and for the scarce up-to-date references, particularly in the area of ontology building.

A main concern in the development of a representative corpus relates to the selection of the sources, i.e. the institutional and relevant repositories whose texts serve to feed the corpus. For the building of the GCTC, two main sources were considered: international institutions and academic works. Hence, in order to collect a significant amount of texts, a few institutions were selected, particularly those related to the fight against terrorism and organized crime at the international level. In addition, these entities had to offer free-access websites where we could get the documents in digital format to facilitate their computer processing. Consequently, United Nations, The Criminal Court of Justice, Europol, Eurojust and OSCE, among others, were chosen. These organizations and their legal representatives are leaders in the field of the fight against terrorism and organised crime and offer a rich representation of the expert knowledge and specialised vocabulary officially used in a vast array of reports, acts and other documents. Other sources, such as academic reference works and journal articles, were also considered due to the probable concentration of specialised terms in their texts. Once the relevant documents were selected and downloaded, a series of manual and semi-automatic editing tasks were required to filter out typographical mistakes resulting from the reformatting of original formats (usually *pdf*) to plain text. These preparatory pre-processing of the texts was necessary because of the characteristics of the term extractor tool, part of the FunGramKB suite, which only works with raw texts.

The large amount of texts collected (621 documents) led the compilers to consider the need to create some sort of organized database in which all the relevant information related to the texts could be stored[5]. This database would serve as a storage system for easy text identification and access whenever necessary. As can be observed in Figure 2, the first field, "ID", assigns a unique numerical code to each text. The field "Language" contains information about the language in which the text is written. "Brief description" offers very succinct information about the contents of the text. "Title" provides a title that summarises the specific topic of the document. The field "Topic" records the subdomain the text belongs to; in the case of GCTC, a distinction is drawn between "Organised crime", "Terrorism" or "Both". Finally, the field "Type of document" contains information about the text type (e.g. joint action, agreement, green paper, proceedings, etc.), while "Source" adds a reference on the source from which the original document was extracted.

ID	Language	Brief description	Title	Topic	Type of document	Source
1	English	Fight against organised crime	EOAct (joint) law enforcement cooperation	Organised Crime	Joint Action	Eur-Lex
2	Spanish	Fight against organised crime	SOAct (joint) law enforcement cooperation	Organised Crime	Joint Action	Eur-Lex

Figure 2. Corpus database

The data gathered in the database has three main objectives. First, it serves as a guide to monitor criteria such as corpus balance and representativeness. Second, some of the data registered in the database will be used during the uploading of texts to the extractor, so they must be conveniently stored. Finally, the database also provides the documentary basis for the calculation of simple descriptive statistics about the corpus.

FunGramKB Suite includes a term extractor tool for the assisted retrieval of sets of potentially relevant terms for the fields under study. The extractor applies a series of filters to an input corpus, mainly the removal of non-textual characters, numbers and punctuation marks. It is upon this cleaned-up text that the statistical extraction process operates. FunGramKB Extractor calculates a *tf-idf* score for each lexical unit in the corpus. As a result, the terminologist can work on a list of candidate terms

[5] The English corpus accounted for 5,698,754 tokens and 31,860,476 characters. The robustness of the GCTC resides mainly in the rich variety of texts that it contains and the fairly large amount of words it holds contributing to capturing legal language. The vast array of text types includes reports, agreements, declarations, regulations, acts, treaties, resolutions and journal articles, among others, adding up to a total of 45 different text types.

ranked according to their semantic weight, so that candidates that appear higher in the list are statistically relevant, while elements that show a *tf-idf* index below 3 are not statistically relevant. It is important to notice that the extraction process in FunGramKB is semi-automatic and that the ultimate decision of what counts as a specialised term relies on the criterion of the terminologist. Figure 3 shows the main menu of the extractor, containing the principal functions of the tool:

Figure 3. Main menu of FunGramKB Term Extractor

From the top leftmost button, the menu is described as follows. The "Pre-processing" button allows displaying an area for testing new features for the extractor. The "Processing (indexing)" button is used for uploading corpus texts to the extractor. "Processing (statistics)" is a key function allowing the terminologist to automatically obtain the list of candidate terms from the corpus. "View" allows the expert to filter false terms by means of a series of removal options. The "Search" button is a secondary tool for searching strings of text in a corpus. Finally, "Corpus" shows basic descriptive statistics concerning the number of indexed texts making up a given corpus as well as the number of tokens included; this button also allows showing a terminological box containing a list of false candidates that were discarded during the filtering process performed in the "View" function.

Once the whole process of automatic extraction and filtering was taken, the total number of relevant n-grams was dramatically reduced to 57,502 tokens. Identifying a close list of winning terms was not a straightforward task; rather, it raised a number of theoretical problems and complex decisions. To facilitate term identification, a four-criterion methodology was proposed to be considered by terminologists during the manual filtering process. These four criteria are presented in a sequential and logical order: (i) statistical significance, (ii) ontological grounding, (iii) lexicological features, and (iv) consultation of specialised dictionaries (Felices-Lago and Ureña Gómez-Moreno 2014: 264-266). The final number of winning conceptual candidates was 406. Among these winning units, one of the most relevant and representative concepts in the domain-specific ontology under scrutiny was +CRIME_00.

3. The Concept +CRIME_00: Conceptualization and Hierarchization

This section focuses on one of the basic and central concepts in the legal field, and hence a key one in the *Globalcrimeterm* satellite ontology presented above, namely the concept *crime*. The aim of this section is to describe the methodological procedure for the inclusion of this concept in the subontology. It is worth mentioning that its importance derives not only from its pertinence in the legal field but also from the intricacy and complexity encountered throughout the whole process that led to its conceptualization and hierarchization in the domain-specific subontology on law.

To begin with, the first impression that we get when we think of the concept *crime* is that no one needs to be an expert to know, in general terms, what it means. Thus, even if perfectly obvious, it seems relevant to point out that the concept *crime* is part of a person's general knowledge. As such, *crime* is commonly understood to be "an evil act, punishable by law". It is equally clear that the concept *crime* will have a more detailed meaning for legal experts, such as a lawyer or a judge, since they need to be more precise in their professional practice in order to delimit with accuracy what actually constitutes a crime together with its typology, legal consequences and the like. Therefore, the definition of *crime* in a specialized legal dictionary will be lengthy and more detailed than in a general dictionary. For example, in the *Oxford Dictionary of Law* the definition of *crime* is as follows:

> **Crime** n. An act (or sometimes a failure to act) that is deemed by statute or by the common law to be a public wrong and is therefore punishable by the state in criminal proceedings. Every crime consists of an actus reus accompanied by a specified mens rea (unless it is a crime of strict liability), and the prosecution must prove these elements of the crime beyond reasonable doubt. Some crimes are serious wrongs of a moral nature (e.g. murder or rape); others interfere with the smooth running of society (e.g .parking offences). Most prosecutions for crime are brought by the police (although they can also be initiated by private people); some require the consent of the Attorney General. Crimes are customarily divided into indictable offences (for trial by judge and jury) and summary offences (for trial by magistrates); some are hybrid. Crimes are also divided into arrestable offences and non arrestable offences. The punishments for a crime include death (for treason), life imprisonment (e.g. for murder), imprisonment for a specified period, suspended sentences of imprisonment, conditional discharges, probation, binding over, and fines; in most cases judges have discretion in deciding on the punishment. Some crimes may

> also be civil wrongs; for example, theft and criminal damage are crimes punishable by imprisonment as well as torts for which the victim may claim damages.

Similarly, in *Black's Law Dictionary* we find that *crime* is defined as:

> An act that the law makes punishable; the breach of a legal duty treated as the subject-matter of a criminal proceeding. - Also termed criminal wrong.

In addition, together with this definition, the dictionary entry for *crime* includes more than fifty subentries to thoroughly analyse and present what this concept entails for the expert user of the dictionary and, hence, in the specialized field. Thus, the concept *crime* is not exclusive to the legal field. Indeed, it is perceived differently depending on whether it is being used by a lay person or a legal expert. It follows, then, that general knowledge and specialized knowledge are not two independent realms, but they are deeply connected, since "specialized knowledge is based on and derived from everyday knowledge" (Van Dijk 2003: 27).

As a consequence, and translating the above discussion to the domain of ontologies, it stems from this that the concept *crime* has to be present both in the Core Ontology —since it stores the concepts that a person has in mind for its use in general purpose situations— and in the *Globalcrimeterm* satellite ontology —since it is one of the domain-specific modules in FunGramKB which stores the set of concepts specific to a particular field of specialty (Periñán-Pascual and Arcas-Túnez 2011: 3). Therefore, the task of the terminologists and knowledge engineers was not only to properly populate both ontologies but also to establish the conceptual connection between them. The linking of the two ontologies has already been addressed (Faber et al. 2011) and cases similar to the concept *crime* have been solved with the creation of "mirror concepts" (Felices-Lago and Ureña Gómez-Moreno 2012; Carrión-Delgado 2012). As previously explained, mirror concepts were proposed as a strategy for reusing the general conceptual information of the Core Ontology in the satellite ontologies, where definitions had to be enriched with greater density of content or granularity. In order to make it possible for the FunGramKB reasoner to establish the connection between the concept in the Core Ontology and its mirror concept in the satellite ontology, it is necessary that both concepts share the same conceptual path. The conceptual path expresses the set of concepts that a given concept is subordinated to in the general hierarchical organization of the ontology. Consequently, both concepts would inherit the same conceptual properties from their superordinates but would differ in the definitions and further

specifications, such as the number and length of predications or the number of subordinated concepts. In addition, both concepts would have an identical notation in COREL, that is, +CRIME_00.

In order to gain more insight into this complex process, it is necessary to scrutinize the representation of the concept *crime* and its conceptual path in the Core Ontology. Following the FunGramKB notation system, the concept *crime* is represented as +CRIME_00. As a basic concept, it is preceded by the symbol +. For +CRIME_00, the conceptual path is as follows: #ENTITY > #PHYSICAL > #PROCESS > +OCCURRENCE_00 > +CRIME_00. Hence, it belongs to the #ENTITY subontology and to the metacognitive dimensions #PHYSICAL and #PROCESS. Its immediate superordinate concept is +OCCURRENCE_00. It means that +CRIME_00 inherits all the conceptual properties of +OCCURRENCE_00. The MP of +CRIME_00, as notated in COREL, is as follows:

+(e1: +BE_00 (x1: +CRIME_00)Theme (x2: +OCCURRENCE_00)Referent)
+(e2: n +BE_00 (x1)Theme (x3: +LEGAL_00)Attribute)

This MP is made up of two predications (headed by the symbol "e") whose translation into natural language can be (1) "Crime is an occurrence" and (2) "It is something which is not legal". The operator "n" in the second predication negates the concept that follows, namely, +BE_00. Moreover, this concept has no subordinates, either basic or terminal, since it has no further specificity in the repertoire of concepts of a lay person. By contrast, integrating the concept +CRIME_00 in *Globalcrimeterm* involved the creation of a conceptually dense MP to account for the complexity or granularity of the concept in the domain of criminal law (see Figure 4).

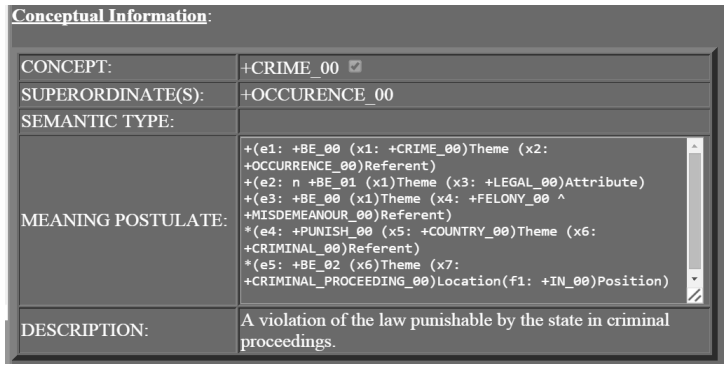


Figure 4. Representation of the concept +CRIME_00 in the FunGramKB editor

As mentioned above, when populating the satellite ontology and deciding on the concept position in the hierarchical organization of FunGramKB, the conceptual path of the new mirror concept +CRIME_00 in the domain-specific ontology had to be identical to the one in the Core Ontology. In other words, the immediate superordinate concept of +CRIME_00 has also to be +OCCURRENCE_00. However, the new MP has to include the granularity involved by the degree of specialization of the subontology. As can be observed in Figure 4, the MP of +CRIME_00 in *Globalcrimeterm* is represented as follows:

+(e1: +BE_00 (x1: +CRIME_00)Theme (x2: +OCCURRENCE_00)Referent)
+(e2: n +BE_01 (x1)Theme (x3: +LEGAL_00)Attribute)
+(e3: +BE_00 (x1)Theme (x4: +FELONY_00 ^ +MISDEMEANOUR_00)Referent)
*(e4: +PUNISH_00 (x5: +COUNTRY_00)Theme (x6: +CRIMINAL_00)Referent)
*(e5: +BE_02 (x6)Theme (x7: +CRIMINAL_PROCEEDING_00)Location(f1: +IN_00)Position)

This MP has five predications. The first two ones are similar to the MP of this concept in the Core Ontology. The remaining three expand the expert information, for example, about possible punishment of the criminal by the state or the subsequent starting of a criminal proceeding. The predications 4 and 5 are preceded by the symbol *, which marks defeasible predications (Periñán-Pascual and Mairal-Usón 2010: 32), i.e. those predications that include features which, although being characteristic or frequent in most cases, are not strict, so they can be refuted when contradictory data are provided. For instance, as regards the predication e4, the criminal is typically punished by the state, though it may not always be the case. These five predications cover the higher granular content and description in the specialized field of law.

However, experts possess special knowledge not only about what a crime is but also about its typology and other legal categories: organized crime, terrorism, cybercrime, crimes derived from corruption, crimes involving lack of freedom, etc. In other words, this specialized knowledge needs to be organized in a specific set of divisions. In this line, the hierarchical organization of the concept +CRIME_00 in the *Globalcrimeterm* ontology has taken into account previous terminological and corpus-based work (Felices-Lago et al. 2011; Ureña Gómez-Moreno et al. 2011), retrieving a list of terms linked to crimes prototypically associated with organized crime and terrorism, such as *assault, mayhem, fencing, smuggling, embezzlement, phishing, trafficking, slavery, chattel*

slavery, cuckoo smurfing, tax evasion, waterboarding, bribery, collusion or *abduction*, among others. On the whole, 66 terms were listed. With the aim of understanding and visualizing the conceptual organization of this criminal area, the first step was a thorough analysis of the lexicographical sources available in order to deeply understand the meaning specifications and internal connections of each of these terms. We extensively consulted various dictionaries, including both general and specialized dictionaries, as well as monolingual and bilingual dictionaries (English-Spanish), such as *Cambridge Advanced Learner's Dictionary, Longman Dictionary of Contemporary English, Oxford Advanced Learner's Dictionary, DRAE: Diccionario de la Lengua Española* (Real Academia), *CLAVE*, Alcaraz-Varó bilingual (English-Spanish) dictionaries of legal terms, *Black's Law Dictionary, Routledge Dictionary of Terrorism,* and *Oxford Law Dictionary*. Similarly, we consulted various academic sources (particularly, specialized journals and academic reference works) and the official websites of some institutions (e.g. OSCE, UN, INTERPOL and Eurojust), which deal not only with general legal aspects but also with questions of terrorism and organized crime as international issues. All this work helped us determine and clarify the conceptual organization of the concept *crime* on this ground. As a result, two broad umbrella concepts were identified: ***personal crimes*** and ***property crimes***, i.e. those offenses which affect persons and those which deal with property. In turn, each one of these two groups is itself divided into subgroups. On the one hand, the concept ***personal crimes*** is divided into *crimes which affect physical integrity, crimes against freedom, trafficking of human beings* and *sexual content offenses*. On the other hand, the concept ***property crimes*** is divided into *trafficking, theft, financial crimes and corruption* and *cybercrimes*. Each type of crime is included in one of these subgroups. For example, *physical integrity*, which is a subgroup of ***personal crimes***, includes *assault, mayhem, torture* and *waterboarding*. Another example is *financial crimes*, which is a subgroup of ***property crime*** and includes *money laundering, cuckoo smurfing, tax evasion* and *layering*. Thus, *crime* is itself an umbrella concept that encompasses a set of other concepts that are semantically associated.

However, once the conceptual organization of the concept *crime* was identified, it became apparent again that some types of crime in the list above belong to a person's general knowledge. Hence, further lexicographical work and research in the field helped us discriminate and isolate the types of crime that belong to the expert's knowledge, that is, the specialized concepts. All this work helped us refine and rearrange the original list of terms by deciding (i) which of the potentially relevant terms

should be included in the domain-specific ontology or should be part of the Core Ontology (possibly as terminal concepts), (ii) which concepts should be integrated as basic or terminal, and (iii) which terms should be defined and considered as lexical units associated with other concepts. In this process of the refinement and filtering of the original list of candidate terms, the following result was obtained: only nineteen out of sixty-six concepts associated with types of crime should be considered genuinely as specialized concepts and populate the *Globalcrimeterm* satellite ontology. That is the case of *alternative remittance, bulk cash smuggling, bootlegging, carousel fraud, chattel slavery, copyright infringement, cuckoo smurfing, curtain slashing, fencing, gambling, layering, market manipulation, mayhem, mixed larceny, pharming, point shaving, self-laundering, vishing* or *waterboarding*. Hence, crimes such as *abduction, smuggling* or *phishing* are not considered expert knowledge, as they can be easily found in learners' dictionary definitions. As Van Dijk (2003: 27) pointed out, "at first this knowledge may still be specialized, but at least part of it is assumed to be shared by others in society, such as journalists, teachers and others who distribute and popularize specialized knowledge". As a consequence, although some concepts were specialized concepts in origin, they have come to be part of a lay person's general knowledge in a process known as "banalization". This is particularly true of the terms we are dealing with in this study since recent terrorist events and their subsequent spread in the news have popularized types of crime which were originally used by experts only.

The nineteen specialized concepts come to populate the *Globalcrimeterm* satellite ontology as terminal concepts subordinated to the basic concept +CRIME_00. They are considered terminal concepts because they constitute the final link in the conceptual hierarchy of the ontology and, as such, they lack definitional potential to take part in the MP of other concepts (Jiménez-Briones and Luzondo-Oyón 2011:16). Following the notation system, these concepts are preceded by the symbol $. In FunGramKB Suite, these concepts are alphabetically listed since the conceptual organization of the field is based on deep semantics and IS-A relations (Periñán-Pascual and Arcas-Túnez 2007). In other words, the mechanisms of inference and inheritance that characterize the MicroKnowing (Micro-conceptual Knowledge Spreading) in FunGramKB (Periñán-Pascual and Arcas-Túnez 2005: 241) connect concepts and predications among the MPs of concepts, so that the reasoner can automatically work out the deep conceptual organization of the specialized field. Furthermore, the concepts representing the types of crime which were rejected for their inclusion in the specialized satellite ontology on the

grounds of being part of general knowledge can eventually be included in the Core Ontology as terminal concepts subordinated to +CRIME_00. Therefore, the work in the specialized domain not only has been productive for the *Globalcrimeterm* itself but also has enriched the Core Ontology by providing more terminal concepts.

Let us now illustrate the process of building ontological meaning in the domain-specific legal ontology *Globalcrimeterm* of FunGramKB.

4. The Case of +CUCKOO_SMURFING_00

In *FunGramKB*, conceptualisation is defined as the process by virtue of which the MPs of conceptual units are defined in COREL. MPs hold essential semantic information about the specific properties of a concept (i.e. event, entity or quality). As has been explained, in satellite ontologies, the construction of MPs is carried out on the basis of the semantic knowledge found in specialised dictionaries. The role of the knowledge engineer at this stage is to gather the semantic content of a term from a selected number of dictionaries and to produce a general description in natural language which encompasses all the different lexicographical definitions.

In order to illustrate how to define terms other than the mirror concept +CRIME_00, let us consider the example of $CUCKOO_SMURFING _00, a very specific term in the domain of organized crime which refers to an organized, transnational and highly coordinated technique of money laundering. It is the crime of putting illegal money from one country into the bank accounts of unwitting third parties in another country so that the money can be integrated into the legitimate economy. This new method started to be identified in the early 2000s by international authorities concerned with fighting financial illegal activities at international level. As a result, the term *cuckoo smurfing* has been recently coined and is still restricted to specialized contexts. The term was coined after the nesting behaviour of the cuckoo bird and the Smurfs, the tiny blue figures of the popular Belgian cartoon. On the one hand, the cuckoo bird is known for the practice of laying its eggs in the nests of other birds, which then hatch the chicks as their own. Just like these other birds, the innocent bank customer accepts the deposit of the criminal's illicit funds and uses them assuming they are legitimate. On the other hand, "smurfing" refers to the size of the cartoon figures and, hence, to the division of large sums of criminal money into smaller amounts.

Bearing in mind the above explanation and the fact that the *Globalcrimeterm* satellite ontology is primarily concerned with the field of

organized crime and terrorism, it seems clear that the concept *cuckoo smurfing* is prototypical in this satellite ontology since it involves the participation and coordination of criminals at an international level. Likewise, its inclusion in the domain-specific ontology is justified on the basis that this term is restricted to expert use, as can be derived from the fact that the term *cuckoo smurfing* cannot be found in general dictionaries, such as *Cambridge Dictionary, Longman Dictionary of Contemporary English* or *Oxford Learner's dictionary*. Nevertheless, information about this technique of money laundering can be found —and indeed explained in detail— in the documents of specialized institutions such as Austrac (by the Australian Government), USA Patriot Act and the international body FATF (Financial Action Task Force). In addition, even though cuckoo smurfing is a technique of money laundering, its definition contains sufficient features and specificities to consider it a different and independent concept, and not a lexical unit associated with an already existing concept such as the terminal concept $MONEY_LAUNDERING_00, as could have initially been suggested.

The need to create a new concept whenever a lexical unit in one language has a meaning that does not match any of the concepts already stored in the knowledge base has been highlighted by Mairal-Usón and Periñán-Pascual (2009: 222-223). According to the various definitions found, the crime of *cuckoo smurfing* specifically implies the involvement of innocent people whose bank accounts are used by criminals to launder illicit money from another country; in other words, in cuckoo smurfing, the bank account's holder is not aware of the fact that illicit money is being transferred to their accounts. Thus, the creation of the new terminal concept $CUCKOO_SMURFING_00 is pertinent and necessary for the satellite ontology to capture expert knowledge in this field. Any other label could have been chosen to refer to this concept, but FunGramKB has always used English as a convenient metalanguage and, for the sake of facilitating reference, the concept is usually named after the lexical unit that prototypically refers to this concept in English. It is for this reason that the newly created concept is a bigram. In other words, it is made up of two concepts, but the meaning of *cuckoo smurfing* cannot directly be inferred from the addition of the meanings of "cuckoo" and "smurf". It is not a transparent bigram (as opposed, for example, to terms like "hostage taking"). And hence, a concept made up of two elements had to be created and considered as a unit.

After analysing the definition and explanation of this concept in various specialized sources, the formalization in COREL produced the following MP:

+(e1: +BE_00 (x1: $CUCKOO_SMURFING_00)Theme (x2: +CRIME_00)Referent)
+(e2: +TRANSFER_00 (x3: +CRIMINAL_00)Agent (x4: p +MONEY_00)Theme (x5: +BANK_00)Origin (x6: +BANK_00)Goal (f1: (e3: n +BE_01 (x4)Theme (x7: +LEGAL_00)Attribute)Attribute) (f2: (e4: n +KNOW_00 (x6)Theme (x8: f1)Referent)Condition) (f3: (e5: +BECOME_00 (x4)Theme (x9: +LEGAL_00)Attribute)Purpose)

This MP is made up of two main predications. The first one includes the immediate superordinate concept of $CUCKOO_SMURFING_00 so as to identify it as conceptually "hanging" from the basic concept +CRIME_00 to inherit all its features. The second predication is a longer and rather complex one which includes three more predications (i.e. e3, e4 and e5) as part of the information regarding attribute, condition and purpose. To begin with, the event +TRANSFER_00 with its corresponding Thematic Frame has been chosen to explain that a criminal (Agent) transfers money (Theme) from a bank account (Origin) to another bank account (Goal). It should also be noted that the participant "x4" is preceded by the operator "p", which means "few or little of something". In this case, it means "small amounts of money", being characteristic of this method of money laundering. Following the definition of *cuckoo smurfing*, it is also essential to express that the money which is transferred is not legal. This has been achieved with the inclusion of the satellite "f1", which, followed by the predication "e3", expresses an attribute. In the predication "e3", the event +BE_00 is preceded by the operator "n" to negate it, and thus we read that the participant "x4" (the money) is not legal. A further distinctive feature that differentiates this concept from other money laundering techniques is the fact that the recipient of the money does not know that it is illegal. This meaning has been formalized in the satellite "f2", which expresses the condition that the participant "x6" (the goal) does not know what "f1" expresses, as formalized in *(x8: f1)Referent*. Finally, a third satellite (i.e. f3) has been included in order to express the purpose of the transfer of illegal money. Again, this satellite is filled with a predication (i.e. e5) that refers to the money (x4) eventually becoming legal.

Corpus-based work motivated the creation of the concept $CUCKOO_SMURFING_00 from the statistically relevant presence of the term *cuckoo smurfing* in the analysed body of texts. Therefore, *cuckoo smurfing* is a lexical unit associated with this new terminal concept. Knowledge engineers and terminologists could not find any other lexical units linked to this concept. This is not a surprising fact since technical language is traditionally characterized by the principle of univocity. This

property of clarity or uniformity means that a concept is generally associated with only one term. Indeed, scientific and technical concepts ideally require the lack of ambiguity to serve their purpose. Polysemy undermines comprehension and causes problems when the precision of meaning is paramount.[6] Legal terms which belong to a specialized domain should not have more than one meaning. This is not the case in common language; for example, as presented by Jiménez-Briones and Luzondo-Oyón (2011: 26-27), the basic concept +TRANSLATE_00 agglutinates lexical units such as *translate, interpret, render, transcribe* and *transliterate*. Likewise, those legal terms which are no longer restricted to the use by expert speakers but have been extended to the general public are not typically univocal. For example, if we take some of the concepts subordinated to +CRIME_00 in the Core Ontology, we can find that the terminal concept $SMUGGLING_00 can be lexicalized as *smuggling* and *contraband*; $TAX-EVASION_00 is associated with the lexical units *tax evasion* and *tax fraud*; and the concept $MONEY_LAUNDERING_00 agglutinates in Spanish as many lexical units as *blanqueo de dinero, lavado de capitales, blanqueo de capitales, legitimación de capitales* and *lavado de activos*. In the case of *cuckoo smurfing*, the new money laundering technique created a new concept which motivated and required the creation of a new lexical unit, which could differentiate the new concept from other methods such as general money laundering and smurfing, but since it is still a term restricted to the specialized domain, it is univocal.

Finally, when attempting to find lexical units associated to the concept $CUCKOO_SMURFING_00 in some of the languages implemented in FunGramKB, it is relevant to point out that it was not possible, for example, to find a lexical unit that referred to this concept in Spanish. This is not a problem since, as already introduced in Section 2, the FunGramKB Ontology is lexically motivated as well as language independent. Thus, the fact that a language lacks a lexical unit to refer to a particular concept does not impede the creation of that concept. In other words, as long as there is a language that lexicalizes a certain concept, that concept has to be created. In Spanish, there are lexical units that, though similar, refer to different concepts. The lexical units *pitufeo*, *trabajo de hormiga* o *estructuración* were discarded because they refer to the division of vast sums of money into smaller amounts so as to defeat suspicion of money laundering. However, unlike *cuckoo smurfing*, in this crime the recipients

[6] Terminology research over the last 30 years has proved that this claim is not true and the problems of polysemy, ambiguity, etc. are countless. However, it is also a well-known fact that the vast majority of scientific terms tend to be precise.

of the money are aware of its illegitimate nature. Hence, in Spanish, there does not seem to be a term that lexicalizes this concept. In consequence, when a typology of money-laundering crimes is listed, this concept is either paraphrased, as in *el uso de cuentas bancarias para depósitos de dinero ilícito o transferencias* (Jiménez-Sanz 2011), or translated next to the English term, as in "el llamado "cuckoo smurfing" ("pitufeo cuco")" [The so called "cuckoo smurfing" ("pitufeo cuco")] (Prieto-del-Pino 2010: 15).

5. Conclusions

Previous research on the Core Ontology of FunGramKB had highlighted the advantages that deep semantics can bring to the world of Artificial Intelligence. Going a step further, the inclusion of the domain-specific *Globalcrimeterm* satellite ontology in FunGramKB will make it possible for experts in the legal field to develop tasks that require natural language processing in their specialized field. The present chapter has focused on the process followed to integrate the key concept *crime*, notated as +CRIME_00, in *Globalcrimeterm*, particularly with the purpose of building conceptual meaning within a domain-specific ontology linked to this knowledge base. The deep semantics that characterizes the construction of FunGramKB will facilitate the necessary conceptual connections for the reasoner to interpret the complex conceptual organization of the legal domain and to establish the link with the conceptual information stored in the Core Ontology. As presented in this chapter, the work connecting the two ontologies (core and satellite) was rather a challenging and as well as a productive process.

Since *crime* is a concept that is part of both general and specialized knowledge, it had to be included in both the Core Ontology and the satellite ontology. More precisely, in the latter one, +CRIME_00 was incorporated as a mirror concept, which shared the same conceptual path as +CRIME_00 in the Core Ontology, but with a higher granularity in the MP so as to represent the specificity of the concept for experts in the legal field. In addition, further lexicographical and corpus-based work provided a list of terms semantically connected to the concept *crime* that were analysed and conceptually organized. The refinement and distillation process that followed reduced the initial list of potentially specialized concepts to nineteen types of prototypical crimes associated with organized crime and terrorism. The case study of the specialized concept $CUCKOO_SMURFING_00 has shed light on the process to build ontological meaning in a satellite ontology and on the challenging and

productive approach to establish the connection with concepts in the Core Ontology.

6. References

Allen, James F. "Maintaining knowledge about temporal intervals". Communications of the ACM 26/11 (1983): 832-843.

Allen, James F and George Ferguson, G. "Actions and events in interval temporal logic". Journal of Logic and Computation 4/5 (1994): 531-579.

Bowker, Lynne and Jennifer Pearson. Working with Specialized Language. A Practical Guide to Using Corpora. London, New York: Routledge, 2002.

Carrión-Delgado, M. de Gracia. "Extracción y análisis de unidades léxico-conceptuales del dominio jurídico: un acercamiento metodológico desde FunGramKB." Revista Electrónica de Lingüística Aplicada 11 (2012): 25-39.

Faber, Pamela, Ricardo Mairal-Usón, and Pedro Magaña-Redondo. "Linking a domain-specific ontology to a general ontology." Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, edited by R. Charles Murray and Philip M. McCarthy, 564-569. Menlo Park, California: The AAAI Press, 2011.

Felices-Lago, Ángel and Pedro Ureña Gómez-Moreno. "Fundamentos metodológicos de la creación subontológica en FunGramKB." Onomázein 26 (2012/2): 49:67.

—. "FunGramKB Term Extractor: a key instrument for building a satellite ontology based on a specialized corpus". In Language processing and grammars: The role of functionally oriented computational models (Studies in Language Series), edited by Brian Nolan and Carlos Periñán-Pascual, 251-269. Amsterdam, New York: John Benjamins, 2014.

Felices-Lago, Ángel, Pedro Ureña Gómez-Moreno and Ángela Alameda Hernández. "FunGramKB y la adquisición terminológica." Anglogermánica Online 8 (2011): 66-86.

Gangemi, Aldo, Maria-Teresa Sagri, and Daniela Tiscornia, D. "A Constructive Framework for Legal Ontologies". In Law and the Semantic Web, edited by V.Richard Benjamins et al, 97-124. Berlin: Springer Verlag, 2005

Jiménez-.Briones, Rocio and Alba Luzondo-Oyón. "Building ontological meaning in a léxico-conceptual knowledge base". Onomázein 23 (2011/1): 11-40.

Jiménez-Sanz, César. El blanqueo de capitales. Madrid: Editorial Académica Española, 2011.

Koester, Almut. "Building small specialised corpora". In The Routledge Handbook of Corpus Linguistics, edited by Anne O' Keeffe Michael McCarthy, 66-79. London: Routledge, 2010.

Lenci, Alessandro. From context to meaning: distributional models of the lexicon in linguistics and cognitive science. A foreword. Italian Journal of Linguistics 20/1 (2008):1-31.

Mairal-Usón, Ricardo and Carlos Periñán-Pascual. "The anatomy of the lexicon component within the framework of a conceptual knowledge base". Revista Española de Lingüística Aplicada 22 (2009): 217-244.

Niles, Ian and Adam Pease. "Towards a Standard Upper Ontology". In Formal Ontology in Information Systems. Collected Papers from the 2nd International Conference (FOIS-01), edited by Christopher Welty and Barry Smith, 2-9. Ogunquit, Maine, 2001.

Periñán-Pascual, Carlos and Francisco Arcas-Túnez. "Meaning postulates in a lexico-conceptual knowledge base". Proceedings of the 15th International Workshop on Databases and Expert Systems Applications, 38-42. Los Alamitos (California): IEEE, 2004

—. "Microconceptual-Knowledge Spreading in FunGramKB". Proceedings on the 9th IASTED International Conference on Artificial Intelligence and Soft Computing, 239- 244. Anaheim-Calgary-Zurich: ACTA Press, 2005.

—. "Deep semantics in an NLP knowledge base". In Proceedings of the 12th Conference of the Spanish Association for Artificial Intelligence, edited by Daniel Borrajo, Luis Castillo and Juan Manuel Corchado, 279-288. Salamanca: Universidad de Salamanca, 2007.

—. "Ontological commitments in FunGramKB". Procesamiento del Lenguaje Natural 44 (2010a): 27-34.

—. "The architecture of FunGramKB". In Proceedings of the 7th International Conference on Language Resources and Evaluation, 2667-2674. Edited by European Language Resources Association. Valletta, Malta: European Language Resources Association, 2010b.

—. "Introducción a FunGramKB". Anglogermánica Online (2011): 1-15.

—. "La ingeniería del conocimiento en el dominio legal: la construcción de una ontología satélite en FunGramKB." Signos 47/84 (2014): 113-139.

Periñán-Pascual, Carlos and Ricardo Mairal-Usón. "La gramática de COREL: un lenguaje de representación conceptual." Onomázein 21 (2010): 11-45.

—. "Bringing Role and Reference Grammar to natural language understanding". In Procesamiento del Lenguaje Natural 43 (2009): 265-273.

Prieto del Pino, Ana María, Deborah Isabel García Magna, and Antonio Martín Pardo. "La deconstrucción del concepto de blanqueo de capitales." InDret, Revista para el Análisis del Derecho 3 (2010): 1-36.

Ureña Gómez-Moreno, Pedro, Ángela Alameda-Hernández and Ángel Felices-Lago. "Towards a specialized corpus of organized crime and terrorism." In La investigación y la enseñanza aplicadas a las lenguas de especialidad y a la tecnología, edited by Carrión, Mª Luisa Carrió et al., 301-306. Valencia: Universitat Politecnica de Valencia, 2011.

Van Dijk, Teun A. "Specialized discourse and knowledge. A case study of the discourse of modern genetics." Cadernos de Estudos Lingüísticos 44 (2003): 21-55.

Van Valin, Robert. The Syntax-Semantics-Pragmatics Interface: An Introduction to Role and Reference Grammar. Cambridge: Cambridge University Press, 2005.

Van Valin, Robert D, and Randy J. LaPolla. Syntax, Structure, Meaning and Function, Cambridge: Cambridge University Press, 1997.

Dictionaries

Black's Law Dictionary. Saint Paul: Thomson Reuters, 2009.

Oxford Dictionary of Law. Oxford: Oxford University Press. 2003.

CHAPTER FIFTEEN

ASSISTING THE PROCESS OF BUILDING A SATELLITE ONTOLOGY OF MENTAL DISORDERS IN FUNGRAMKB USING A LATENT SEMANTIC ANALYSIS-BASED TOOL*

ISMAEL IVÁN TEOMIRO GARCÍA
UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA (UNED)
AND MARÍA BEATRIZ PÉREZ CABELLO DE ALBA
UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA (UNED)

1. Introduction

The goal of this work is the creation of a satellite ontology of mental disorders. To this effect we will follow the protocol for the creation of subontologies in FunGramKB set up by Felices Lago and Ureña Gómez-Moreno (2012) and Periñán Pascual and Arcas Túnez (2014). The main new element regarding our approach is that we propose the use of a tool based on Latent Semantic Analysis (LSA hereafter) to assist the process of corpus compilation and term extraction and identification (Jorge-Botana et al., 2011). This tool has the advantage for us that it gives access to a corpus made up of two international classifications: the ICD-10 (International Classification of Diseases) and the DSM-IV (Diagnostical Statistical Manual of Mental Disorders). With this goal in mind, the

* This work has been partially funded by research projects FFI2011-29798-C02-01 (Ministerio de Economía y Competitividad) and 2013-014-UNED-PROY (Universidad Nacional de Educación a Distancia - UNED). We would like to thank Alba Luzondo Oyón and Guillermo de Jorge Botana for their useful comments and suggestions.

organization of this work goes as follows. In Section 2 we present the protocol for the creation of subontologies in FunGramKB. In Section 3 we propose the creation of a subontology of mental disorders in FunGramKB with the help of LSA-based tools. We give a short account of LSA in Section 4. Section 5 presents the process of creation of the subontology of mental disorders with the tools mentioned above. Finally, in Section 6 we put forward the conclusions of this work and some future lines of research.

2. Creating a Subontology in FunGramKB

Before diving into the process of the creation of a satellite ontology within FunGramKB, a brief account of that knowledge base is in order. FunGramKB is a multipurpose and multilingual lexico-conceptual knowledge base designed for Natural Language Processing (NLP) tasks (Periñán Pascual & Arcas Túnez, 2010). As shown in Figure 1, this knowledge base includes three levels of information: a lexical level, a grammatical level and a conceptual level. Each one of these three levels consists of several independent but interrelated modules. Within the lexical level, we find the lexicon, which contains morphosyntactic, pragmatic and collocational information about lexical units, and the morphicon, which provides information about inflectional morphology. In the grammatical level, the grammaticon stores constructional schemata of four different types. The cognitive level consists of an ontology, an onomasticon and a cognicon, each corresponding to the three types of knowledge distinguished in cognitive psychology, namely, semantic, episodic and procedural knowledge (cf. Mairal Usón & Periñán Pascual, 2010; Periñán Pacual & Arcas Túnez, 2007, 2010). The ontology has a general-purpose module (i.e. core ontology) and several domain-specific terminological modules (i.e. satellite ontologies). Within the core ontology we find a multi-level model of metaconceptual, basic and terminal levels.

While both the lexical and the grammatical levels are language-dependent, and therefore we will find a different lexicon, morphicon and grammaticon for each language contained in the knowledge base, the conceptual level is claimed to have a universal status. What is really noteworthy is that the linguistic level, which includes both the lexical and the grammatical models, is connected up to the cognitive level, and, what is more, the cognitive level constitutes the pivot of the whole machinery (Pérez Cabello de Alba, 2011).

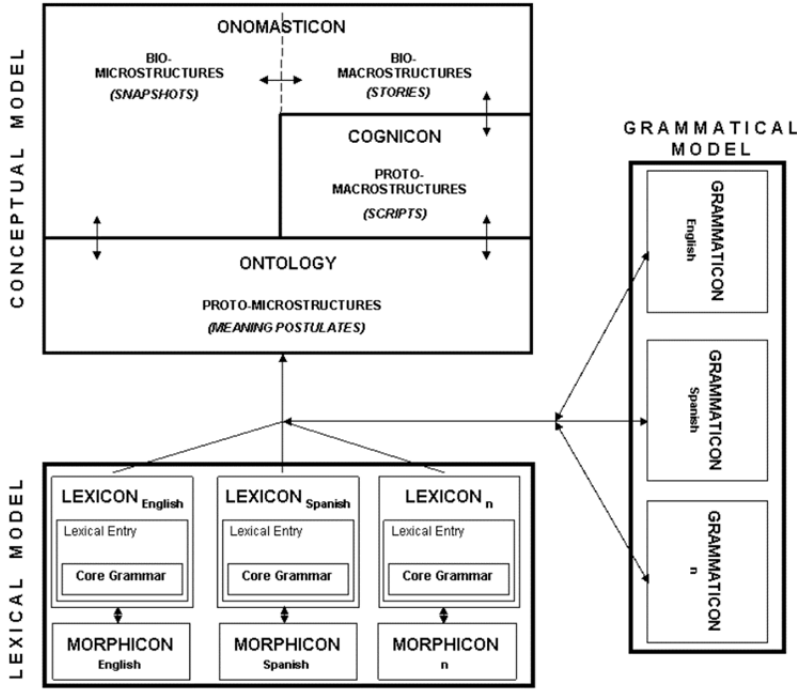


Figure 1. FunGramKB modules (Periñán-Pascual and Arcas-Túnez, 2011: 3).

Felices Lago and Ureña Gómez-Moreno (2012) worked on the assumption that the multi-level model of the core ontology (i.e. metaconceptual, basic and terminal levels) can be connected to a subontology in order to minimize redundancy and maximize knowledge. Within FunGramKB, subontologies contain concepts which belong to specialized-knowledge domains such as crime and terrorism (Felices Lago and Ureña Gómez-Moreno, 2012), mental disorders (this work), etc. The purpose of creating subontologies is thus to enable FunGramKB to perform natural language processing and reasoning tasks in different specialized-knowledge domains.

Following Felices Lago and Ureña Gómez-Moreno (2012), the methodology to build a satellite ontology consists of three phases: first the compilation of a specialized corpus, second the automatic extraction of specialized lexical units from the corpus, and third the modelling of specialized knowledge through the creation and hierarchization of concepts. We are going to give a brief account of each of the phases

proposed by the authors and then we will explain what has been useful for the elaboration of our satellite ontology of mental disorders.

Specialized corpus compilation

For the compilation of a representative corpus of a specialized domain, a significant number of relevant texts have to be identified. The corpus has to contain as many terms as possible, and it must show an ample variety of textual genres so that it reflects the topics and the lexicon of that domain. Then an electronic repository of texts will be created.

Terminology automatic extraction

In FunGramKB terminology extraction can be done with the aid of the *FunGramKB Term Extraction* application. It automatically obtains a list of representative terms from a corpus. It uses a statistical method based on the tf-idf (*term frequency - inverse document frequency*) of all the lexical units that compose the corpus. The higher the tf-idf of a term is, the more statistical relevance that term has and the more probable it is for it to be a terminological unit of that domain. The extractor can identify three types of potential units: *unigrams*, which are units formed by one word, *bigrams*, which are formed by two words, and *trigrams*, which are formed by three words. After the automatic processing, a manual edition is necessary in order to decide which candidates are finally relevant for knowledge modelling and which ones are "false candidates".

Knowledge modelling: Conceptualization and hierarchization

Once the terminology of a specialized domain has been identified, the next step consists in populating the subontology by projecting such terminological units onto conceptual units. As a result, a set of basic concepts, terminal concepts and subconcepts will be created and linked to the lexical units that lexicalize them in each supported language (cf. Periñán Pascual & Arcas Túnez, 2010; Jiménez Briones & Luzondo Oyón, 2011; Mairal Usón, Periñán Pacual & Samaniego Fernández, 2011).

The conceptualization in FunGramKB involves building thematic frames of the different concepts as well as their meaning postulates through COREL (Conceptual Representation Language) (Periñán Pascual & Mairal Usón, 2010).

Thematic frames are the prototypical cognitive scenarios of events, which indicate the number of participants involved. Meaning postulates

contain semantic information that is essential to define the properties of a specific concept, whether it be an event, an entity or a quality. In the subontologies, meaning postulates are created with the help of specialized dictionaries. Felices Lago and Ureña Gómez-Moreno (2012) gave the definition of the concept $WATERBOARDING_00, as shown in (1), as an example of a characteristic concept in the domain of terrorism and crime. In (1a) the definition in natural language is given, whereas in (1b) this definition is translated into meaning postulates through COREL. In (1c) the lexical terms in different languages are provided.

(1) $WATERBOARDING_00

a. Lexicographical description in natural language:
A form of torture in which water is poured over the face of a supine, immobilized victim whose head is pulled back so that the victim cannot avoid inhaling water, and thus experiences the sensation of drowning. In some variations, fabric or plastic may be draped over the victim's face or the victim may be gagged before the water is poured. [*Black's Law Dictionary* (BLD)].

b. Meaning postulates:
+(e1: +BE_00 (x1: $WATERBOARDING_00)Theme (x2: +TORTURE_00) Referent)
*(e2: +CHANGE_00 (x3: $TERRORIST_00 ^ +HUMAN_00)Theme (x4: $VICTIM_00)Referent (f1: (e3: +FEEL_00 (x3)Agent (x4)Theme (x5: +DAMAGE_00 | +PAIN_00 | +FEAR_00)Attribute))Result (f2: (e4: +SUFFER_00 (x4)Theme))Result) (f3: (e5: n +MOVE_00 (x4)Referent))Condition (f4: (e6: +DESCEND_00 (x3)Agent (x5: +LIQUID_00)Theme (x6)Location (x7)Origin (x8: +HEAD_00) Goal))means (f5: (e7: n +BREATH_00 (x4)Theme))Result) (f6: (e8: +EXPERIENCE_00 (x3)Agent (x4)Theme (x2)Referent (f7: (e9: ing +DROWN_00 (x1)Theme (x2)Referent))Result)
*(e10: +PUT_00 (x3)Agent (x9: +CLOTH_00 ^ +BAG_00)Theme (x10)Origin (x8)Goal)

c. Terms:
"waterboarding" (English), "submarino" (Spanish), "waterboarding" (Italian).

Once the basic concepts of the subontology have been identified and defined (i.e. both their thematic frames and their meaning postulates), the next step consists in their hierarchization, that is, for each concept we must determine its corresponding superordinate, subordinate or sibling concepts. For example, the concept $WATERBOARDING_00 in (1) has the conceptual route in (2b).

(2) $WATERBOARDING_00
a. Superordinate concept:
$TORTURE_00
b. Conceptual route:
#ENTITY > #PHYSICAL > #PROCESS > +OCCURRENCE_00 > +CRIME_00 > $TORTURE_00

Our goal in this work is to lay the foundations for the creation of a satellite ontology of mental disorders reusing existing resources in order to complete Felices Lago and Ureña Gómez-Moreno's phases 1 and 2. More specifically, we use LSA-based tools that are described in Section 4.

3. Creating a Subontology of Mental Disorders in FunGramKB

We propose the creation of a subontology of mental disorders in FunGramKB in order to enable it to perform natural language processing tasks in a specialized domain, e.g. machine translation or automatic diagnosis in the disciplines of mental health and psychiatry.

The mental health discipline has a terminology that is standardized in two international classifications: the ICD-10 in Europe and the DSM-IV in the USA. Therefore, we think it is crucial that the subontology of mental disorders reflects the concepts and terms of these two classifications since they constitute the basis of the knowledge of the discipline.

We pursue Felices Lago and Ureña Gómez-Moreno's (2012) protocol for the creation of subontologies. However, we propose using LSA-based tools in order to assist the phase 1 (corpus compilation) and the phase 2 (term extraction). More specifically, we propose using the corpus found in the research project "El Semántico" (http://www.elsemantico.com) by Jorge-Botana et al. (2011) for several reasons:

- "El Semántico" consists of a corpus made up of the two international classifications of mental disorders, namely the ICD-10 and the DSM-IV. The corpus is publicly available.
- It also provides the researcher with the semantic space of the corpus expressed by analytical tools based on LSA techniques.

Not only the corpus (phase 1) but also the LSA-based tools (phase 2) are immediately available. These mathematical tools serve to cast important information on which terms are the most informative in the corpus and also the most frequent.

In the next section we explain what LSA is and why it is an important technique for the creation of a subontology of specialized knowledge. We subsequently provide the basis for the creation of the subontology and propose the conceptualization and hierarchization of several basic concepts of this specialized domain that may or may not be present in the core ontology of FunGramKB.

4. Latent Semantic Analysis

LSA is a theory and a method for extracting and representing the contextual usage of the meaning of words by statistical computations applied to a large corpus of text (Landauer et al., 1998, p. 259).

LSA can be construed in two ways. First, LSA as a model of the computational processes and utilization of knowledge (Landauer & Dumais, 1997) constitutes a fundamental computational theory of the acquisition and representation of knowledge, able to account for Plato's problem or the problem of the "poverty of the input". Second, LSA as a method for characterizing word meaning represents the words used in a text and any set of these words as points in a very high-dimensional (e.g. 50-1,500) "semantic space". By doing so, LSA produces measures of word-word, word-passage and passage-passage relations. These measures are well correlated with several human cognitive phenomena involving association or semantic similarity despite the limitations derived from not making use of word order (i.e. syntactic relations) or morphology. The statistical methods that are used in LSA are closely akin to factor reduction. Consequently, the resulting dimensions of description are analogous to the semantic features often postulated as the basis of word meaning. This is very important when creating an ontology in FunGramKB since it will shed light on what predications should be present in each concept. LSA is a practical method for the characterization of word meaning. This is the way LSA is presented and used throughout this chapter.

LSA is a fully-automatic mathematical and statistical technique for extracting and inferring relations of the expected contextual usage of words in passages of discourse. LSA takes as input only raw text parsed into words as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs. Once the text is parsed, it is represented as a matrix where each row stands for a unique word, and each column stands for a text passage or other context. Therefore, each cell contains the frequency with which the word of its row appears in the passage denoted by its column. In the next step, the cell

entries undergo a preliminary transformation by which each cell frequency is weighted by a function that expresses both the word's importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general. Finally, LSA applies Singular Value Decomposition (SVD) to the matrix. Through this process, the rectangular matrix is decomposed into the product of three other matrices: one component matrix that describes the original row entities as vectors of derived orthogonal factor values, another matrix that describes the original column entities in the same way, and a third matrix that is a diagonal matrix containing scaling values such that, when the three components are matrix-multiplied, the original matrix is reconstructed. SVD reduces the number of dimensions to the most informative ones. However, the number of dimensions retained in LSA is an empirical issue.

Once we have a reduced dimensional space, we can obtain two measures:

- The cosine between vectors (each representing a word) measures the similarity computed in the reduced dimensional space.
- The length or magnitude (module) of LSA vectors reflects how much was said about a topic rather than how central the discourse was to the topic. It is a measure of the frequency of the words within the corpus.

The importance of LSA is that the word and passage meaning representations derived from it have been found capable of simulating a variety of human cognitive phenomena such as developmental acquisition of vocabulary recognition, word categorization, sentence-word semantic priming, discourse comprehension and judgment of essay quality. The adequacy of LSA for the reflection of human knowledge is shown by the following facts:

- Its scores overlap those of humans on standard-vocabulary and subject-matter tests.
- It mimics human word sorting and category judgments.
- It simulates word-word and passage-word lexical priming data.
- It accurately estimates:
  - passage coherence (Foltz, Kintsch, & Landauer, 1998),
  - learnability of passages by individual students (Wolfe et al., 1998), and
  - the quality and quantity of knowledge contained in an essay (Rehder et al., 1998).

To conclude, LSA has been used with good results to mimic synonym, antonym, singular-plural and compound-component word relations, to simulate aspects of imputed human representation of single digits and to replicate semantic categorical clusterings of words found in certain neuropsychological deficits (Laham, 1997). It has also been used to formally represent polysemy in texts (Jorge-Botana et al., 2011). Therefore, we defend that LSA can provide with important insights as to what words are important in order to define the concepts of an ontology as well as the semantic features that constitute their foundations, which will be defined by means of predications in our case.

5. A Proposal for the Creation of a Subontology of Mental Disorders

In this section we proceed to describe the process of the creation of the subontology on mental disorders. Although we follow Felices Lago and Ureña Gómez-Moreno (2012), we depart from them for the phase 1 (corpus extraction) and the phase 2 (terminology extraction). Instead, we make use of the corpus from "El Semántico" project and of the LSA-based mathematical indices described above. Subsequently, we model four basic concepts of the subontology.

5.1. Procedure

The corpus of mental disorders can be found in the web page of the research project "El Semántico". This is composed of two international classifications of mental disorders in Spanish: the ICD-10 and the DSM-IV. The web page lets users type in terms, and it provides the semantic neighbours with their corresponding cosine and magnitude. The cosine of each semantic neighbour tells us how similar that term is to the one in the search field. The magnitude gives information on the frequency of that term in the whole corpus and hence on the contribution of that term to the whole corpus.

Since this work is a preliminary study in order to establish the bases for a subontology on mental disorders, we proceeded to search the semantic neighbours of the bigram "mental disorder", which we consider the central concept of the subontology, as well as the semantic neighbours of the words "illness", "symptom" and "disorder", since these terms are crucial for the definition of the central concept "mental disorder". Besides the corpus from "El Semántico", we looked for the definitions of each term in general and specialized dictionaries in both English and Spanish,

since dictionaries can be used for the conceptualization of terms (cf. Jiménez Briones & Luzondo Oyón, 2011).

"El Semántico" builds a semantic vector space of the corpus so that semantic neighbours can be obtained for each term along with their cosine (i.e. similarity to the entry term) and magnitude or module (i.e. frequency within the whole corpus).

For the bigram "trastornos mentales" (mental disorders), we obtained the list of semantic neighbours, summarized in Table 1, which was ordered by module in order to see which terms are the most central in the ontology.

Table 1. List of semantic neighbours of "trastornos mentales"

Wordform	Cosine	Module
p	0.34055048	9.065974236
ej	0.364646522	8.73374939
síntomas	0.343713395	8.693949699
trastornos	0.894607284	8.684278488
ansiedad	0.310347492	8.377811432
enfermedad	0.352147218	7.765708923
mayor	0.288035372	6.304538727
ánimo	0.301402985	6.132365704
problemas	0.28603392	6.078482628
diagnóstico	0.385165628	5.990134239
criterios	0.3027593	5.762640953
mental	0.50504667	5.70413065
psicológicos	0.30093786	5.175914288
individuos	0.312437064	5.148649693
médica	0.328753096	5.081880093

Note that this corpus is not lemmatized, so, for example, plural and singular forms of the same term appear as different terms. There is also noise produced by abbreviations ("p" and "ej"), functional words ("si", *if*), and misspelt words ("ntomas" for "síntomas", *symptoms*). Therefore, we had to group the first together and get rid of the second.

As said before, we made use of two LSA-based measures:

- The **module** (i.e. frequency of occurrence in the corpus) was used to detect the most frequent terms; these will define the

conceptualization of the ontology (e.g. basic concepts, terminal concepts and subconcepts).

- The **cosine** (i.e. neighbourhood) was used to detect the terms related to a given term; these will be used to guide the definition of the meaning postulates of the concepts.

From the data obtained and summarized in Table 1, we can see that the most central terms of the corpus are "síntomas" (symptoms), "trastornos" (disorders), "ansiedad" (anxiety) and "enfermedad" (illness). Therefore, we proceed to revise the concepts +SYMPTOM_00 and +ILLNESS_00, which are present in the core ontology. We also create the new concepts +DISORDER_00 and +MENTAL_DISORDER_00 since we consider them central to the subontology given their magnitude. Although the concept +ANXIETY_00 is present in the core ontology, it requires a mirror concept because its definition is not extensive enough for the purposes of a specialized subontology. However, we leave the modelling of this concept for further research due to the diagnostic complexity of such a term and for space reasons.

5.2. Knowledge Modelling

In what follows, we provide the information we obtained for each of the aforementioned terms, i.e. "síntomas" (symptoms), "trastornos" (disorders) and "enfermedad" (illness), and we model their corresponding concepts expressed through their meaning postulates. The phase of knowledge modelling has also been assisted with definitions in natural language from general and specialized dictionaries as well as from definitions found in the DSM-IV.

5.2.1. +ILLNESS_00

This concept is present in the core ontology of FunGramKB with the following properties:

(3) +ILLNESS_00
  a. Description:
  Impairment of normal physiological function affecting part or all of an organism.
  b. Meaning postulates:
  +(e1: +BE_00 (x1: +ILLNESS_00)Theme (x2: +STATE_00)Referent)
  +(e2: +BE_01 (x3: +HUMAN_00 ^ +ANIMAL_00)Theme (x4: +SICK_00)Attribute (f1: x1)Scene)

c. Lexical units (English and Spanish lexica):
Enfermedad (Spanish lexicon)
Illness / disease (English lexicon)
d. Conceptual path:
#ENTITY >> #ABSTRACT >> +STATE_00 >> +ILLNESS_00

Despite the fact that this concept is already present in the core ontology, a mirror concept is needed in the satellite ontology that will serve the purpose of bringing the two ontologies together. For this end, several definitions in both specialized and non-specialized natural language have been used as the base for the creation of the mirror concept:

- **[Oxford Dictionary of English - illness]**: a disease or period of sickness affecting the body or mind.
- **[Oxford Dictionary of English - disease]**: a disorder of structure or function in a human, animal, or plant, especially one that produces specific symptoms or that affects a specific location and is not simply a direct result of physical injury.
- **[Diccionario General de la Lengua española VOX - enfermedad]**: Alteración leve o grave del funcionamiento normal de un organismo o de alguna de sus partes debida a una causa interna o externa.
- **Diccionario Médico de Bolsillo Dorland (23rd edition) (McGraw-Hill):** Alteración o desviación del estado fisiológico en toda la economía, o en alguna de sus partes, órganos o sistemas (o combinación de ellos), que se manifiesta por un conjunto característico de síntomas y signos cuyas etilología, patología y pronóstico pueden conocerse o ser conocidos.
- **DSM-IV**: there is no definition of illness or general medical condition, but there is a paragraph to distinguish illness or general medical condition ("enfermedad médica") from mental disorder ("trastorno mental"):

> The terms *mental disorder* and *general medical condition* are used throughout this manual. The term *mental disorder* is explained above. The term *general medical condition* is used merely as a convenient shorthand to refer to conditions and disorders that are listed outside the "Mental and Behavioural Disorders" chapter of ICD. It should be recognized that these are merely terms of convenience and should not be taken to imply that there is any fundamental distinction between mental disorders and general medical conditions, that mental disorders are unrelated to physical or biological factors or processes, or that general medical conditions are

> unrelated to behavioural or psychosocial factors or processes. (American Psychiatric Association, 1994: XXV)

Moreover, we have obtained the semantic neighbours of the terms "enfermedad" (illness) and "enfermedades" (illnesses) from "El Semántico", along with their cosines and modules, as shown in Tables 2 and 3 respectively.

Table 2. List of semantic neighbours of "enfermedad"

Wordform	Cos	Mod
médica	0.468002516	5.081880093
síntomas	0.383593184	8.693949699
diagnóstico	0.321560117	5.990134239
depresión	0.274200793	9.102970123
fisiológicos	0.249346807	3.139629602
cáncer	0.240816127	5.046553612
laboral	0.240425989	3.970389843
significativo	0.239854253	3.526536226
individuo	0.239266604	6.36037302
vida	0.235140333	6.111838818
criterio	0.234991773	8.425487518
factores	0.234510559	6.118657589
trastornos	0.234297567	8.684278488
actividad	0.233300163	5.797355652
demencia	0.232761223	5.632662296

Table 3. List of semantic neighbours of "enfermedades"

Wordform	Cos	Mod
trastornos	0.272051101	8.684278488
síntomas	0.233730705	8.693949699
salud	0.229239291	7.986397743
problemas	0.21423656	6.078482628
físicas	0.210022859	2.196476698
cáncer	0.208386009	5.046553612
estrés	0.199231285	7.762517929
crónicas	0.194370819	0.985747457
síndrome	0.189993928	5.149578094

depresión	0.189538646	9.102970123
médico	0.181908093	5.487432957
riesgo	0.181758538	4.90538168
diagnóstico	0.178157716	5.990134239
diabetes	0.173226757	1.17820251
médica	0.168954839	5.081880093

As a result, we have come up with the specialized definition of +ILLNESS_00 in (4) and with the meaning postulate in (5).

(4) Semantic features of +ILLNESS_00 in natural language:
- Illness is a state (+STATE_00).
- It affects the body (+BODY_00) and/or the mind (+MIND_00).
- It is a disorder (+DISORDER_00) of structure or function in a human, animal or plant.
- It produces symptoms (+SYMPTOM_00) and signs (+SIGN_00).
- There exists a pathology [+PATHOLOGY_00][1] that may be known or not.
- There exists an aetiology [+AETIEOLOGY_00][2] that may be known or not.
- There exists a prognostic [+PROGNOSTIC_00][3] that may be known or not.

(5) Meaning postulate of +ILLNESS_00:
+(e1: +BE_00 (x1: +ILLNESS_00)Theme (x2: +STATE_00)Referent)
+(e2: +CHANGE_00 (x1)Theme (x3: +BODY_00 | +MIND_00)Referent (f1: (e3: +CREATE_00 (x1)Theme (x4: +DISORDER_00)Referent))Result)
+(e4: +DO_00 (x1)Theme (x5: +SYMPTOM_00 & +SIGN_00)Referent)
+(e5: +EXIST_00 (x6: +AETIEOLOGY_00)Theme (f2: x1)Scene)
*(e6: pos +KNOW_00 (x7: +HUMAN_00)Theme (x6)Referent (f2)Scene)
+(e7: +EXIST_00 (x8: +PATHOLOGY_00)Theme (f2)Scene)
*(e8: pos +KNOW_00 (x7)Theme (x8)Referent (f2)Scene)
+(e9: +EXIST_00 (x9: +PROGNOSTIC_00)Theme (f2)Scene)
*(e10: pos +KNOW_00 (x7)Theme (x9)Referent (f2)Scene)

[1] In this work we do not define this concept due to space reasons. Moreover, we believe this concept is part of our common knowledge, so it should also be defined in the core ontology. This would require defining it in the satellite ontology as a mirror concept.
[2] In this work we do not define this concept due to space reasons. It will be included in the satellite ontology in future works.
[3] See previous footnote.

5.2.2. +SYMPTOM_00

The concept +SYMPTOM_00 is present in the core ontology with the properties listed in (6). We take it as a mirror concept and extend the information this concept encodes in the specialized subontology.

(6) +SYMPTOM_00
  a. Description:
     Any sensation or change in bodily function that is experienced by a patient and is associated with a particular disease.
  b. Meaning postulates:
     +(e1: +BE_00 (x1: +SYMPTOM_00)Theme (x2: +PHYSICAL_ATT_00 ^ +PSYCHOLOGICAL_ATT_00)Referent)
     *(e2: +SHOW_00 (x3: +HUMAN_00)Theme (x1)Referent (f1: (e3: +BE_01 (x3)Theme (x4: +SICK_00)Attribute))Scene)
  c. Lexical units (English and Spanish lexica):
     Síntoma (Spanish lexicon)
     Symptom (English lexicon)
  e. Conceptual path:
     #ENTITY >> #ABSTRACT >> +ATTRIBUTE_00 >> +PHYSICAL_ATT_00 >> +SYMPTOM_00

Several definitions in both specialized and non-specialized natural language have been used as the base for the creation of the mirror concept:

- **[Oxford Dictionary of English - symptom]:** a physical or mental feature which is regarded as indicating a condition of disease, particularly such a feature that is apparent to the patient
- **[Diccionario General de la Lengua española VOX – síntoma]:** Alteración del organismo que pone de manifiesto la existencia de una enfermedad y sirve para determinar su naturaleza.
- **[Diccionario Médico de Bolsillo Dorland, 23rd edition, McGraw-Hill**]: (1) Cualquier prueba subjetiva de enfermedad o del estado de un paciente; p. ej., dicha prueba como la percibe el paciente. (2) Cambio en la evolución del paciente que indica cierto estado corporal o mental. Cf. signo.
  - *SIGNO:* (1) indicación de la existencia de algo. (2) Cualquier prueba objetiva de una enfermedad, p.ej., las pruebas perceptibles para el médico que examina al paciente, a diferencia de las sensaciones subjetivas (síntomas) que percibe este último.
- **DSM-IV:** no definition given.

Moreover, we have obtained the semantic neighbours of the terms "síntoma" (*symptom*) and "síntomas" (*symptoms*) from "El Semántico", along with their cosines and modules, as shown in Tables 4 and 5 respectively.

Table 4. List of semantic neighbours of "síntoma"

Wordform	Cos	Mod
enfermedad	0.366595444	7.765708923
depresión	0.350373087	9.102970123
criterio	0.349441444	8.425487518
trastornos	0.330639732	8.684278488
ansiedad	0.321430377	8.377811432
abstinencia	0.303221373	7.115295887
estado	0.273597437	6.577159405
esquizofrenia	0.254817486	6.360538006
mayor	0.23991253	6.304538727
social	0.236819527	7.795681953
individuo	0.236573038	6.36037302
tratamiento	0.235690083	7.520925999
pacientes	0.226083045	7.050404549
paciente	0.186592545	7.864774704
sujetos	0.178459998	6.558637142

Table 5. List of semantic neighbours of "síntomas"

Wordform	Cos	Mod
conversión	0.469321023	4.004915237
somatización	0.354914669	3.624280453
neurológico	0.32567008	1.91571939
facticio	0.281596602	1.631866932
médica	0.27411133	5.081880093
exacerbación	0.265731765	2.21228075
simulación	0.259407282	2.292881727
afonía	0.255815083	1.399656653
criterio	0.252012939	8.425487518
diplopía	0.248050242	1.410321474
nudo	0.245026244	1.37112534
urinaria	0.243850974	1.368526936
dolor	0.243690015	8.173914909
sordera	0.24335685	1.359883189
déficit	0.237759536	4.950461388

With the data presented above, we have come up with the specialized definition of +SYMPTOM_00 in (7) and with the meaning postulate in (8).

(7) Semantic features of +ILLNESS_00 in natural language:
- Symptom is a physical or psychological attribute.
- It shows that there exists an illness.
- It is subjective (perceived by the doctor or patient but not measured by means of objective measures like body temperature, x-ray, etc.)

(8) Meaning postulate of +ILLNESS_00:
+(e1: +BE_00 (x1: +SYMPTOM_00)Theme (x2: +PHYSICAL_ATT_00 ^ +PSYCHOLOGICAL_ATT_00)Referent)
+(e2: +SHOW_00 (x1)Theme (x3: (e3: +EXIST_00 (x4: +ILLNESS_00)Theme))Referent)
+(e3: +BE_01 (x1)Theme (x5: +SUBJECTIVE_00)Attribute)

5.2.3. +DISORDER_00

The concept +DISORDER_00 does not exist in the core ontology and is needed in order to define the concept +ILLNESS_00, the mirror concept that links the subontology to the core ontology, as well as the concept +MENTAL_DISORDER_00, which is central in the subontology. We have looked for definitions in natural language in general and specialized dictionaries, as in (9) and (10).

(9) Definitions in natural language of +DISORDER_00:
- **[Oxford Dictionary of English –disorder]**: an illness that disrupts normal physical or mental functions.
- **[Diccionario General de la Lengua española VOX –mental]**: [2] Alteración en el funcionamiento de un organismo o de una parte de él o en el equilibrio psíquico o mental de una persona.

(10) Definitions in natural language of +DISORDER_00:
- **Diccionario Médico de Bolsillo Dorland (23rd edition) (McGraw-Hill)**: (1) Desviación de lo que se considera normal. (2) Alteración o perturbación de una función física o psíquica.

Moreover, we have obtained the semantic neighbours of the term "trastorno" (*disorder*) from "El Semántico", along with its cosines and modules, as shown in Table 6.

Table 6. List of semantic neighbours of "trastorno"

Wordform	Cos	Mod
síntomas	0.330785087	8.693949699
ansiedad	0.316390619	8.377811432
personalidad	0.264325161	7.021609306
estado	0.23077349	6.577159405
tratamiento	0.230354263	7.520925999
enfermedad	0.224013836	7.765708923
social	0.18724046	7.795681953
conducta	0.178355927	6.790324688
pacientes	0.166578208	7.050404549
sujetos	0.161435298	6.558637142
criterio	0.157824747	8.425487518
salud	0.145187373	7.986397743
fobia	0.136756312	7.266702652
abstinencia	0.13659769	7.115295887
estrés	0.135069305	7.762517929

With these data, we have come up with the specialized definition of +DISORDER_00 in (11) and with the meaning postulate in (12).

(11) Semantic features of +DISORDER_00 in natural language:
- Disorder is a state (+STATE_00).
- It is a change (+CHANGE_00) in the function or structure of an organism.
- It produces an unbalance in the function or structure of an organism.
- This unbalance is considered non-normal.

(12) Meaning postulate of +DISORDER_00:
+(e1: +BE_00 (x1: +DISORDER_00)Theme (x2: +STATE_00)Referent)
+(e2: +CHANGE_00 (x1)Theme (x3: +ORGANISM_00)Referent)
+((e3: +DO_00 (x1)Theme (x4: +UNBALANCE_00)Referent (f1: x3)Location)(e4: n +BE_01 (x4)Theme (x5: +NORMAL_00)Attribute))

5.2.4. +MENTAL_DISORDER_00

The concept +MENTAL_DISORDER_00 does not exist in the core ontology. We consider it the central concept of the subontology since basically every type of mental disorder will be defined as its hyponym. Moreover, this concept has ontological status in the DSM-IV, which points towards its crucial status in an ontology of mental disorders. The

DSM-IV does not define it but has an article that specifies some of its properties as used throughout the text. In (13) an excerpt is given, presenting properties of this concept that make it different from the concept +DISORDER_00.

(13)

> In DSM-IV, each of the mental disorders is conceptualized as a clinically significant behavioral or psychological syndrome or pattern that occurs in an individual and that is associated with present distress (e.g., a painful symptom) or disability (i.e., impairment in one or more important areas of functioning) or with a significantly increased risk of suffering death, pain, disability, or an important loss of freedom. In addition, this syndrome or pattern must not be merely an expectable and culturally sanctioned response to a particular event, for example, the death of a loved one. Whatever its original cause, it must currently be considered a manifestation of a behavioral, psychological, or biological dysfunction in the individual. Neither deviant behavior (e.g., political, religious, or sexual) nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the individual, as described above. (American Psychiatric Association, 1994: XXI-XXII)

Moreover, we have obtained the semantic neighbours of the term "trastorno mental" (mental disorder)" from "El Semántico", along with its cosines and modules, as shown in Table 7.

Table 7. List of semantic neighbours of "trastorno mental"

Wordform	Cos	Mod
asociados	0.566176874	3.13164854
enfermedades	0.515892626	4.862985611
diagnostican	0.500971865	0.524608731
médicas	0.450169646	2.518219471
descriptivas	0.433689487	1.327895164
crónicas	0.431753134	0.985747457
adaptativos	0.426617009	0.610964179
etiología	0.412887355	1.90322125
reumatismo	0.401086037	0.203672752
amnésicos	0.391235904	0.744638562
diagnóstico	0.385165628	5.990134239
endógenas	0.384382096	0.304860711
envenenamiento	0.382458007	0.552573919
psiquiátricos	0.380421865	0.901178718
clasificación	0.373685702	1.379248023

We propose the specialized definition of +MENTAL_DISORDER_00 in (14) and the meaning postulate in (15).

(14) Semantic features of +MENTAL_DISORDER_00 in natural language:
- Mental disorder is a kind of disorder (+DISORDER_00).
- It consists of [1] a set of syndromes (+SYNDROME_00) and/or signs (+SIGN_00) or [2] a set of behaviours and/or mental states that is significant [that produces a change in the individual or in his/her life or his/her body or his/her mind].
- It is associated with a feeling of uneasiness/being sick.

(15) Meaning postulate of +MENTAL_DISORDER_00:
+(e1: +BE_00 (x1: +MENTAL_DISORDER_00)Theme (x2: +DISORDER_00)Referent)
+(e2: +COMPRISE_00 (x1)Theme (x3: i +SYNDROME_00 & i +SIGN_00)Referent)
+(e3: +FEEL_00 (x1)Agent (x2:+HUMAN_00)Theme (x3:+UNEASY_00)Attribute)

6. Conclusion and Future Lines of Research

As a conclusion, we want to draw attention to the reusability of an existing tool since it may prove to be a time-saving and inexpensive option. Moreover, since it is based on two real international classifications of diseases, it may be a reality check for the design of specialized ontologies with no room for introspection.

Furthermore, we would like to point out some lines for future research for the development of the satellite ontology of mental disorders. First, new basic and terminal concepts are needed in the core ontology, such as +DISORDER_00, $PATHOLOGY_00 and $MENTAL_DISORDER_00. Other mirror concepts such as +ANXIETY_00 are also needed in the satellite ontology in order to extend the specialized knowledge represented in the ontology. In this work, the principles for the creation of a specialized subontology have been set out. We believe this may open the way to the development of the entire satellite ontology of mental disorders. Finally, such an ontology may be the base for the development of an NLP application for the automatic diagnosis of mental disorders by means of descriptions of patients' symptoms. This can be achieved by combining the specialized knowledge encoded in the satellite ontology of mental disorders proposed in this work with the general-purpose knowledge encoded in the core ontology, together with the computational power of a reasoning engine in FunGramKB (cf. Periñán Pascual & Arcas Túnez, 2005).

7. References

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (fourth ed.). Washington DC: American Psychiatric Association.

Felices Lago, Á., & Ureña Gómez-Moreno, P. (2012). Fundamentos metodológicos de la creación subontológica en FunGramKB. *Onomázein*, *26*(2), 49-67.

Foltz, P. W., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2&3), 285-307.

Jiménez Briones, R., & Luzondo Oyón, A. (2011). Building ontological meaning in a lexico-conceptual knowledge base. *Onomázein*, *23*, 11-40.

Jorge-Botana, G., León, J. A., Olmos, R., & Escudero, I. (2011). The representation of polysemy through vectors: Some building blocks for constructing models and applications with LSA. *International Journal of Continuing Engineering Education and Long Learning*, *21*(4).

Laham, D. (1997). *Latent semantic analysis approaches to categorization*. Proceedings from 19th Annual Conference of the Cognitive Science Society, Mahwah, NJ.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2&3), 259-284.

Mairal Usón, R., Periñán Pacual, C., & Samaniego Fernández, E. (2011). Using ontologies for terminological knowledge representation. A preliminary discussion. In N. Talaván, E. Martín, & F. Palazón (Eds.), *Technological innovation in the teaching and processing of LSPs: Proceedings of TISLID'10* (pp. 267-280). Madrid: UNED.

Mairal Usón, R., & Periñán Pascual, C. (2010). Role and Reference Grammar and Ontological Engineering. In J. L. Cifuentes, A. Gómez, A. Lillo, J. Mateo, & F. Yus (Eds.), *Los caminos de la lengua. Estudios en homenaje a Enrique Alcaraz Varó* (pp. 649-665). Alicante: Universidad de Alicante.

Pérez Cabello de Alba, M. B. (2011). Ontological Semantics in the Lexical Constructional Model. *RAEL*, *07/11*, 187-202.

Periñán Pacual, C., & Arcas Túnez, F. (2005). *Microconceptual-Knowledge Spreading in FunGramKB*. Proceedings from 9th IASTED

International Conference on Artificial Intelligence and Soft Computing, Anaheim-Calgary-Zurich.
Periñán Pacual, C., & Arcas Túnez, F. (2007). *Deep semantics in an NLP knowledge base*. Proceedings from 12th Conference of the Spanish Association for Artificial Intelligence.
Periñán Pascual, C., & Arcas Túnez, F. (2010). *The architecture of FunGramKB*. Proceedings from 7th International Conference on Language REsources and Evaluation, Valeta (Malta).
Periñán Pascual, C., & Arcas Túnez, F. (2011). Introduction to FunGramKB. *Anglogermánica Online* 8, 1-15.
Periñán Pascual, C., & Arcas Túnez, F. (2014). La ingeniería del conocimiento en el dominio legal: La construcción de una Ontología Satélite en FunGramKB. *Revista Signos. Estudios de Lingüística*, *47*(84), 113-139.
Periñán Pascual, C., & Mairal Usón, R. (2010). La gramática de COREL: Un lenguaje de representación conceptual. *Onomázein*, *21*, 11-45.
Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*(2&3), 337-354.
Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, *25*(2&3), 308-336.
World Health Organisation. (1992). *ICD-10 Classification of Mental and Behavioural Disorder: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organisation.

Dictionaries

BLD: Garner, B. (2010) *Black's Law Dictionary.* 9th edition. Minnesota: West.
Larousse Editorial (2012) *Diccionario General de la Lengua Española Vox.* Larousse Editorial, S.L.
Oxford University Press (2010) *Oxford Dictionary of English.* Oxford: Oxford University Press.
Saunders Company (1989) *Diccionario Médico de Bolsillo Dorland.* 23rd edition. Madrid: McGraw-Hill / Interamericana de España.

BIODATA

Angela Alameda-Hernández is an Assistant Lecturer at the Department of English Studies, University of Granada (Spain). Her main research lines are discourse analysis and terminology. She got her Ph.D. on English Language and Linguistics (2006) with a thesis on the discursive construction of national identity, analysed within the tradition of critical discourse analysis. She has published articles not only on this topic but also on media discourse, terminology and foreign language didactics. Her main contributions in this line have been published in a number of journals, such as *National Identities, The Linguistics Journal, Text Perspectives in Media and Academic Discourse, Anglogermánica Online* and *Odisea.* She currently lectures on English for Specific Purposes at the University of Granada.

Aysun Balkan is a Faculty Member at Bogazici University School of Foreign Languages in Istanbul, Turkey. She received her M.Sc. (2009) in Applied Linguistics from Georgetown University, M.A. (2006) in Applied Linguistics/TESOL from Ohio University and B.A. (2003) in Foreign Language Education from Bogazici University. During her graduate studies, she specialized in Second Language Acquisition (SLA). She has a wide range of university-level teaching experience in the USA and Turkey. She taught linguistics courses at Georgetown University and Ohio University. She has been teaching EFL in the English Preparatory Unit at Bogazici University. Her research interests include cross-linguistic influence in SLA (e.g. language transfer), cognitive semantics (e.g. the polysemy network of spatial nouns), and task-based instruction (e.g. the effects of explicit/implicit corrective feedback). As relevant to her research interests, she presented individual and joint research projects in various international conferences in the USA and Europe. She speaks Turkish, English and German.

Ana Díaz Galán studied English Philology at the University of La Laguna, where she also completed her Ph.D. degree (2001) on the study of lexical cohesion within the framework of functional grammars. She is currently a lecturer in the English Department at the University of La Laguna. Her research interests have been focused on discourse analysis

and functional grammar and lexis. She is a member of the Lexicom Research Project and a researcher of the research project entitled "Construction of a Core-Grammar Spanish-English database within the Lexical Constructional Model", funded by the Spanish Ministry of Science.

Elke Diedrichsen is a German linguist based in Dublin. She works as an International Project Manager at Microsoft European Headquarters in Dublin. In her career to date, she has worked as a researcher and a deputy professor in several universities across Germany. She has also worked as a Project Manager in an international team of linguists for the development and quality assurance with Google speech products (NLP, ASR, TTS) in Google's European Headquarters in Dublin. She is a member of the Computational and Functional Linguistics Research Group at the Institute of Technology Blanchardstown (ITB), Dublin. She has widely published about functional linguistics, Role and Reference Grammar, constructions as grammatical objects, NLP, pragmatics and the semiotic, cultural and interactional potential of memes. She has recently published a co-edited volume with Brian Nolan and Gudrun Rawoens entitled *Causation, Permission and transfer – Argument realisation in GET, GIVE, PUT and TAKE verbs*. Another recent publication is a volume co-edited with Brian Nolan (2013), entitled *Linking constructions into functional linguistics – The role of constructions in grammars.* Both book publications appeared in John Benjamins' *Studies in Language Companion Series.* She is also the author of the paper "A Role and Reference Grammar parser for German", which appeared in *Language processing and grammars: The role of functionally oriented computational models,* edited by Brian Nolan and Carlos Periñán-Pascual (2014).

Elena Even-Simkin is a Post-Doctoral Fellow at the Ben-Gurion University of the Negev, a researcher, a lecturer and a reviewer for academic journals, e.g. David Publishing Company, USA. She received her Ph.D. *Summa cum Laude* in Linguistics at the Ben-Gurion University of the Negev. She is a member of International Society for the Linguistics of English and of Columbia School Linguistics Society. Her scholarly interests range from historical linguistics, applied linguistics, semiotics, discourse and text analysis, linguistic theory to language acquisition, ESL teaching and language disorders. The results of her findings have been published in books, anthologies and different international academic journals.

Ángel Felices Lago is a Full Professor and has worked as a member of the Department of English and German Philology (University of Granada, Spain) since 1987. He teaches English and Spanish for business and tourism. His main areas of research interest go from lexicology, discourse analysis and axiological linguistics to NLP applied to LSP. He has co-authored or co-edited 10 scholar and pedagogical books or textbooks and has also published over 80 scholarly articles and reviews in specialized national and international journals and volumes. He has served as invited reviewer or member on the editorial and scientific boards of a dozen journals and has also taken part in various international academic projects funded by the Spanish Ministry of Education or the European Union (Tempus, Leonardo, Erasmus, Erasmus-Mundus, etc.) He has recently led an international and interdisciplinary R&D project which aims at creating a legal ontology based on deep semantics, with the cooperation of 9 universities and other institutions.

Aoife Finn is a Ph.D. student at Trinity College Dublin. The working title of her dissertation is "A characterization of the case-system and grammatical relations in Māori in a Role and Reference Grammar account". Her research interests include Māori, syntax, reflexivity, case-system, unaccusativity and ergativity. After initially graduating with a degree in Electronic Engineering from Dundalk IT, she received a B.A. in Italian and French from University College Dublin. Upon gaining a Certificate in English Language Teaching, she spent some years teaching general and aviation English. She also has experience in teaching children with special needs and in educational administration. Having attended Trinity College Dublin to complete an M.Phil in linguistics, she received a distinction in her dissertation "Towards a characterization of reflexivity and reciprocity in Māori in a Role and Reference Grammar account". In addition, she enjoys cooking, vegetarianism, yoga and gardening.

María del Carmen Fumero Pérez studied at the University of La Laguna, where she also received her Ph.D. degree in English Philology in 2001. She is a lecturer in the English Department at the University of La Laguna. Since her doctoral dissertation on pragmatic functions, her main research interests have been focused on academic discourse analysis and, more recently, as a member of the Lexicom Research Project, on the interaction between lexis and grammar within functional and cognitive models. She is currently a researcher of the research project entitled "Construction of a Core-Grammar Spanish-English database within the

Lexical Constructional Model", funded by the Spanish Ministry of Science.

Svetlana Kiseleva is the Head of Meaning and Sense Group, which focuses on the understanding of the cognitive mechanisms of meaning construction. The group is housed in the National Research University "Higher School of Economics" of St. Petersburg in Russia. She has co-authored over 100 publications, including three monographs about word meaning. She is a leading and respected research scientist with twelve years of research experience in semantics and cognitive linguistics. Her current research interests in the group include (1) the creation of the substantive core of the polysemantic verb and (2) the creation of the meaning of the sentence. The group was formed in 2013 and established an extremely fruitful collaboration in 2014-2015 with Professor Nella Trofimova in order to better pursue different aspects of meaning.

Ciro Antunes de Medeiros is an undergraduate student of Biological Sciences at State University of Campinas (UNICAMP). Since the second semester of 2013, his research has focused on categorization and language evolution and development, resulting in four international-conference presentations until May 2015.

Eva M. Mestre-Mestre is an associate professor at Universidad Politécnica de Valencia. Since her Ph.D. thesis on the pragmatic implications of errors in English as a second language, her research has focused around three main axes, i.e. error analysis, English learning in higher education, and corpus management, resulting in publications indexed in nationally and internationally prestigious catalogues. Apart from several book chapters, she also wrote a monograph about teaching English as a second language for Higher Education in the Common European Framework of Reference for Languages. She was a visitor researcher in Aston University (UK), Université d'Angers (France) and College of William & Mary (USA). She is currently the manager of the panel on pragmatics in the Spanish Society for Applied Linguistics Conference.

Brian Nolan is the Head of School of Informatics and Engineering at the Institute of Technology Blanchardstown, Dublin (Ireland). His research interests include computational approaches to speech and language processing, computational linguistics, linguistic theory at the morphosyntactic-semantic interface, argument structure and valence,

constructions in grammar, event structure in language and the architecture of the lexicon. His linguistic work has been in the functional linguistic model of Role and Reference Grammar. In 2012 he published his book with Equinox UK on the linguistic structure of Irish in a Role and Reference Grammar account entitled *The structure of Modern Irish: A functional account*. In 2013 John Benjamins published his co-edited volume *Linking constructions into functional linguistics: The role of constructions in grammar* in their Studies in Language Companion series. His co-edited volume on computational linguistics and linguistic theory, *Language processing and grammars: The role of functionally oriented computational models*, was published in 2014, also in John Benjamins' Studies in Language Companion series. He also co-edited a Benjamins book on *Causation, transfer and permission* in linguistic theory, which appeared in early 2015. At present, he is co-editing a Benjamins book on complex predication entitled *Argument realisation in complex predicates and complex events: Verb-verb constructions at the syntax-semantic interface* to appear in 2016. He has over 40 years of national and international experience within the computer industry, with almost two decades in academia in a variety of senior roles. He is also a widely published professional linguist. He is a Fellow of the Irish Computer Society.

Beatriz Pérez Cabello de Alba is an Associate Professor of English Language and Linguistics at the UNED in Madrid (Spanish National University for Distance Education), where she teaches Linguistics, English for Specific Purposes (ESP) and Translation (legal, scientific-technical and economic-commercial English). She also teaches several courses in the UNED European Masters of English Applied Linguistics. Her research interests cover lexicology, lexicography, ontological semantics and natural language processing. She has collaborated in several competitive research projects funded by the Spanish Science and Research Ministry. She is currently implementing a subontology within FunGramKB. She has been a visiting scholar at the Universities of Amsterdam and Verona. She has also been a visiting professor at Chulalongkorn University in Bangkok, an associate professor at Kingston University and an assistant professor at the London School of Economics and Political Science.

Carlos Periñán-Pascual received his Ph.D. degree in English Philology at UNED in 1999 (Spain). Since his doctoral dissertation on the resolution of word-sense disambiguation in machine translation, his main research interests have included knowledge engineering, natural language

understanding and computational linguistics. More particularly, his research has been focused on the cognitive and computational treatment of lexical information, constructional meaning, conceptual representation, and reasoning, among many other tasks. Since 2004, he has been the director of FunGramKB, a lexico-conceptual knowledge base, together with a suite of tools, for the automatic processing of language. The regularity of his scientific production leads to more than 40 publications, including journal articles, book chapters and conference papers. Most of his 20 journal articles are indexed in prestigious international catalogues. He has been invited not only to deliver lectures on natural language processing in several universities but also to take part in international conferences as a plenary speaker. Finally, he has been the principal investigator in four funded research projects as well as the chair of the organizing committee in many scientific events, including international workshops and conferences. He is currently an associate professor at Universitat Politècnica de València in Spain.

Alena Poncarová graduated in 2013 at Charles University in Prague, Faculty of Arts. Since then, she is a Ph.D. student at Institute of Czech Language and Theory of Communication at the same university. She studies syntax and text linguistics and focuses on principles of text development and relations within the text. Her work is focused on coreferential chaining from the perspective of both information and constituent structures. In her dissertation project, she applies Centering Theory on authentic Czech data and examines the possibilities this application brings. In addition to her doctoral studies, she cooperates with Institute of the Czech National Corpus, Institute of Theoretical and Computational Linguistics and Institute of Formal and Applied Linguistics.

María José Ruiz Frutos studied Spanish language and literature as well as German philology in Valladolid (Spain) and Bayreuth (Germany). She got her M.A. degree from *Universidad Nacional de Educación a Distancia* (Spain) with her Master Thesis entitled *Coreferential anaphora resolution with FunGramKB*. In this work she theoretically examined how the knowledge stored in FunGramKB and the reasoning performed in it could contribute to successfully resolving coreferences. Since 2005, she has taught Spanish as a foreign language at the Bayreuth University, where she is currently the team coordinator for the Spanish courses at the Language Centre. She is interested in language education and technology, coreference resolution and semantic web technologies.

Ismael Iván Teomiro García is an Assistant Professor of English Language and Linguistics at the Spanish National Distance Education University (*Universidad Nacional de Educación a Distancia, UNED*). He holds a Ph.D. from the Universidad Autónoma de Madrid (2011) and an M.A. (2005) from the Universiteit Utrecht in Linguistics, as well as an M.Sc. (2005) from the Universiteit Utrecht and a B.A. (2001) from the Universidad Pontificia Comillas in Psychology. He teaches general linguistics, theoretical and applied syntax, and English for Specific Purposes in different bachelor degrees at the UNED (English Studies, Spanish Language and Literature, and Psychology). He also teaches English grammar and lexicology in two master degrees at the UNED (English Literature and English Applied Linguistics, respectively). His research interests cover the syntax of Romance, Germanic and Celtic languages, language acquisition (both L1 and L2), and lexicon-syntax interface phenomena, as well as the codification of time in the lexical-conceptual interface. He has collaborated in several national research projects (FFI2011-29798-C02-01, FFI2011-23829/FILO, FFI2008-01584/FILO and HUM2005-01728). He has written several articles in national and international journals (e.g. Onomázein, Topics in Linguistics, RESLA, Revista de Lingüística y Lenguas Aplicadas, and Brathair, among others) as well as one book on Binding Theory (Axac, 2011). Currently, he is leading a research project (ICP-FUNGRAMKB) at the UNED that is focused on Construction Grammar and its computational implementations. The goal is to implement the pronominal constructions of Spanish into FunGramKB, a knowledge base for natural language processing.

Nelly Trofimova is a professor of linguistics at the National Research University "Higher School of Economics" of St. Petersburg in Russia, where she works on the problem of meaning with Professor Svetlana Kiseleva. Her current research is focused on the dynamic construction of sentence meaning. She is a leading and respected research scientist with ten years of research experience in the semantic and pragmatic area. She has co-authored over 100 publications, including two monographs about the multidimensional sense of the linguistic sign.