ITHEA

International Journal INFORMATION TECHNOLOGIES & KNOWLEDGE

2010 Volume 4 Number 2

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE
Volume 4 / 2010, Number 2

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK as well as the **subscription fees** are given on ***www.foibg.com***.
Responsibility for papers published in IJ ITK belongs to authors.

General Sponsor of IJ ITK is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Vol.4, Number 2, 2010

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Institute of Mathematics and Informatics, BAS, Bulgaria,
Universidad Politécnica de Madrid, Spain.

Publisher **ITHEA**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, www.foibg.com, e-mail: info@foibg.com
Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

FREQUENCY EFFECTS ON THE EMERGENCE OF POLYSEMY AND HOMOPHONY

Gertraud Fenk-Oczlon, August Fenk

Abstract: *In this paper we try to answer the following questions: Why do frequently used words tend to polysemy and homophony? And what comes first - frequency or the higher number of meanings per word? We shall stress the key role of frequency in the emergence of polysemy and assume an interactive step-up initiated by frequency: High frequency not only favors reduction processes of words or the bleaching of meanings that can result in polysemy; it also plays a crucial role in the creation of metaphors or metonymies, i.e., the main sources of polysemy. Only familiar or frequent source words/concepts tend to be used in metaphorical or metonymical expressions. Through the conventionalization of the metaphors and metonymies, the source words get additional meanings. They now can be used in a higher number of contexts what in turn favors a more frequent use.*

A similar explanation might hold for the development of homophony: Shorter words are known for their tendency to homophony [Jespersen, 1933] and high token frequency. Our explanation: High frequency favors backgrounding processes, such as vowel reduction, lenition and deletion of consonants or even of syllables. This frequency-induced shortening of words often results in sound merger and in a relatively high proportion of homophonous words, i.e., words encoding unrelated meanings.

Keywords: *frequency, polysemy, homophony, metaphor, metonymy*

Introduction

Both polysemy and homophony refer to the phenomenon in which words are having the same phonological form but different meanings. Such words are either classified as polysemous or as homophonous. As polysemous, if they encode related meanings, as for instance *tongue* (part of the body, language, scales,…). And as homophonous, on the other hand, if they encode semantically unrelated meanings. For example *bat* ('flying mammal', 'wooden stick'); *inn, in; I, eye.*

This distinction between polysemy and homophony is, however, not as clear-cut as it might look at a first glance. Because originally related senses can "become so distant that they are perceived as unrelated, such as French *voler* 'to fly' and *voler* 'to steal'" [Nerlich and Clarke, 2003:11].

In this paper we investigate mutual dependencies between token frequency and polysemy and mechanisms involved in the development of both polysemy and homophony. Starting points of these considerations are classical findings such as Jespersen's [1933] observation of an association between shortness of words and homophony or Zipf's [1949] principles relating high token frequency of words to shortness as well as to semantic versatility. Both these principles seem to be motivated by economy principles in communication and cognition [cf. Fenk-Oczlon and Fenk, 2002].

Before studying the role of frequency in the development of polysemy and homophony we shall outline, in rather general terms, the role of frequency in language and cognition.

The Role of Frequency in Cognition and Communication

The realization of frequency as a determinant of our cognitive processes traces back at least as far as Aristotle. In the course of memorizing, he says, custom “takes the place of nature. Hence we remember quickly things which are often in our thoughts; for as in nature one thing follows another, so also in the actualization of these stimuli; and the frequency has the effect of nature…” Aristotle, quoted from [Suppes, 2009: 164]

In [Hume, 1777; 1993] “it is not reasoning which engages us to suppose the past resembling the future” (p. 25), but “Custom or Habit”. This principle explains “why we draw, from a thousand instances, an inference, which we are not able to draw from one instance…” (p. 28).

Needless to mention the frequency of the occurrence or co-occurrence of stimuli as a decisive factor in the best studied and most fundamental forms of learning, such as sensitization, habituation, and conditioning [Kandel and Schwartz, 1982].

A widely discussed advantage and effect of learning is that it prepares the organism for the (near) future; it enables the organism to anticipate (expect, predict, …), to some degree, the course of events, including the consequences of its own (re)actions. A second advantage and effect: The growing capability to anticipate what's going on has positive effects on the efficiency (speed and/or accuracy) of information processing in the respective domain. This second effect, i.e., the enhancement of the present perceptive-cognitive activities by previous perceptive-cognitive activities, is maybe too obvious to be noticed as extensively: Everybody has experienced that reading texts in a specific language improves through reading texts in that language.

In terms of information theory, *learning* means the extraction of a system's redundancy (patterns, invariants, periodicities). In Shannon's [1949] guessing game technique it is mainly the guessing person's (implicit) knowledge about the statistical structure of the respective language – frequency distributions of graphemes and words, transitional probabilities – what allows her to reduce the number of prognostic errors. In other words: She uses the redundancy of that language and she can use it the better the more familiar she is with that language.

No wonder that our “sensitivity to frequency” [Hasher and Chromiak, 1975] plays a crucial role in language acquisition [Saffran et al., 1996]. But here the point in question is the more or less indirect way in which the frequency of use changes the respective expressions. The probably best studied phenomenon of this kind is the shortening of words through frequent use [Zipf, 1929, Mandelbrot, 1954, Fenk-Oczlon, 1989], i.e., a frequency effect on the level of phonology and/or morphology.

But frequency also affects the level of semantics. Since Zipf [1949] we know that the number of different meanings of words increases with their frequency (Zipf's *Principle of the Economical Versatility of Words*) and that the length of words is inversely related to their relative frequency (Zipf's *Law of Abbreviation of Words*). In [Köhler, 1986] it is argued that frequency influences the word length that in turn influences polysemy. We assume slightly different mechanisms and an interactive step-up between frequent use and polysemy:

The Role of Frequency in the Development of Polysemy

We here suggest the idea of an interactive step-up between frequency and polysemy: Frequent use favours the tendency to shortness and polysemy, and shortness and polysemy favors frequent use for obvious reasons – the use of shorter expressions is economically motivated, and words encoding a higher number of meanings fit in a higher number of contexts. A chicken-and-egg problem? In view of the fundamental role of frequency in information processing (see the above section and [Fenk-Oczlon, 2001]) the strongest and initial impulses are assumed to come by frequency. The presumably relevant mechanisms:

The Emergence of Polysemy: Phonetic Reduction Processes

The influence of frequency on phonetic reduction processes has been documented in numerous works [e.g. Zipf, 1929; Manczak, 1980]. Backgrounding processes, such as vowel reduction, lenition and deletion of consonants or even of syllables are strongly associated with the token frequency of words. When analyzing data regarding rapid speech by American students [Kypriotaki, 1973] it can be shown that aphaeresis (deletion of the initial syllable) appears above all in words which belong to the 1000 most frequent words in English like *suppose*, *because*, *remember*, *almost, around,* etc. [Fenk-Oczlon, 1989]. And such reduction processes can result in polysemy:

around → *round* (verb, noun, adj, adv, prep)

remember → *member* (verb, noun)

because → *cause (*conjunction, noun*)*

The Emergence of Polysemy: Bleaching

Aitchison and Lewis [2003] argue that bleaching (fading of meaning) might be a further source for the development of polysemy. For instance "...words signifying catastrophic events like 'disaster' are subject to bleaching, and consequently, the development of polysemy" (p.263). "A prerequisite for the development of polysemy may be that the word must be widely used" (p.261). Frequent use favors a layering of the (still existing) original concept by more or less related meanings.

The Emergence of Polysemy: Metaphors and Metonymies

Metaphors and metonymies are considered to be the main sources of polysemous sense extensions. For instance: According to Blank [2003], metaphoric polysemy is “based on a more or less salient similarity between two concepts that belong to different or even distant conceptual domains” (p.268). As an example he mentions *mouse* (small rodent, computer device). Metonymic polysemy, he says, is “based on conceptual contiguity, i.e. the typical and salient co-occurrence or succession of elements in frames or scenarios or of these frames themselves” (p.269); e.g. *lingua* (tongue, language).

But metaphors and metonymies, i.e., the main sources for the development of polysemy, might in turn reflect frequency effects: When a new metaphor is created, only words being well-entrenched in the lexicon of the respective language community, or being familiar within this community, can be incorporated in this metaphor. The prerequisite for high familiarity is a rather high token frequency of these words. With the metaphorical/metonymical use of unfamiliar words one would risk the metaphor/metonymy not being understood or being misinterpreted, or that its processing and comprehension would at least require too much time.

Thus we see a strong tendency to metaphorical use of source words (concepts) showing a **high frequency** and/or **prototypicality**:

Prototypical and/or frequently mentioned animals such as *fox, bear, ox, dog, cow, lamb, frog, bird, elephant, camel etc.*, occur more often in metaphors than less prototypical and/or less frequently mentioned animals. (*You are a platypus (duckbill)*! would hardly be understood in Austria.)

Frequent verbs such as *see, hear, smell, touch, sit, stand, lie, eat, drink* etc. are more often used in metaphors than less frequent words.

Prototypical colors such as *black, white, red, green, yellow, blue* have more metaphorical meanings than e.g. *violet.*

In the metaphorical use of frequent verbs or color terms we tend to assume a lower cultural variation than in the use of animal names.

High frequency of predications in which the predicates express typical attributes, activities, or relations between two concepts, results in high conjoint frequency and favors metaphorical use:

The fox is cunning. → *He is a fox.*

German: *Der Fuchs hat ein rötliches Fell*. ‘The fox has a reddish coat’ → Bavarian: *Sie is fuchsat.* ‘she has red hair’

The cactus is prickly. → *He is a cactus*.

The parrot “parrots“. → *He is a parrot.*

The parrot is colorful. → *She dresses like a parrot*.

The same holds for metonymies

German: *Schiffe haben einen Kiel* ‘Ships have a keel’→ ”*Tausend Kiele näherten sich der Küste*“ ‘Thousand keels are approaching the coast’ [Keller, 1995:176]

He drinks a glass of wine/vodka/whisky → He drinks a glass.

An increasing **frequency** of the use of a certain metaphor/metonymy means a conventionalization of this metaphor with potentially two results:

1. The metaphorical character of the metaphor ‘bleaches’, and it may become a ‘frozen’ or a dead metaphor.

2. The relevant (vehicle) term may get an additional meaning; the respective word has become polysemous.

The Role of Frequency in the Development of Homophony

The main source of homophony is accidental sound merger. Linguistic history shows that it is often reduction processes that lead to a shortening of words (e.g. loss of final vowels, loss of parts of words). Words with different etymological origin become identical in sound and therefore homophonous. An example from [Blank, 2003]: Old English *earm* ‘upper limb of the body’, *arme* ‘weapon’ → Modern English *arm* ‘upper limb of the body’, ‘weapon’.

Already Jespersen [1933] stated an association between homophony and shortness of words: “The shorter the words, the more likely is it to find another word of accidentally the same sound”. In English he found about four times more monosyllabic than polysyllabic homophones. Ke [2006] also found high positive correlations between homophony and number of monosyllables in the 5000 most frequent words in English, Dutch and German. And when grouping the 5000 words into 14 frequency bands in decreasing order of frequencies she stated the highest degree of homophony in the highest frequency bands. Although languages vary widely in their number of monosyllables and their degree of homophony, there is a strong association between shortness of words, token frequency, and homophony.

The bridge to our frequency-based explanation concerning the development of homophony is again the association between high token frequency, phonetic reduction processes, and shortness. Frequently used words get shortened, and this can result in sound merger and therefore in homophony.

Conclusion

Regarding the ‘chicken and egg problem’ mentioned in the third section we may conclude: Frequency comes first! It seems to be the trigger in the emergence of polysemy: Frequent use plays a crucial role in phonetic reduction processes and in the bleaching of meanings that can lead to polysemy (i). High frequency/familiarity of words favors their use in metaphorical/metonymical expressions that are well-known as sources for polysemous sense extensions. Through frequent use metaphors and metonymies become conventionalized, the source words get additional meanings (ii). Polysemous words are apt to be used in a higher number of different (con-)texts. Thus, polysemy in turn increases the token frequency of the respective words (iii).

In our attempt to explain the development of homophony, token frequency again plays a central role. Phonetic reduction processes, such as lenition and deletion of consonants, vowel reduction etc., affect predominantly frequently used words. In linguistic diachrony words get shortened and this can lead to sound merger and thus also to homophony.

Bibliography

[Aitchison and Lewis, 2003] J.Aitchison and D.M.Lewis. Polysemy and bleaching. In: Polysemy: Flexible Patterns of Meaning in Mind and Language. Eds. B.Nerlich, Z.Todd, V. Herman, and D.D.Clarke. Mouton de Gruyter, Berlin, 2003.

[Blank, 2003] A.Blank. Polysemy in the lexicon and in discourse. In: Polysemy. Flexible Patterns of Meaning in Mind and Language. Eds. B.Nerlich, Z.Todd, V.Herman, and D.D.Clarke. Mouton de Gruyter, Berlin, 2003.

[Fenk-Oczlon, 1989] Geläufigkeit als Determinante von phonologischen Backgrounding-Prozessen. Papiere zur Linguistik 40, 91-103, 1989.

[Fenk-Oczlon, 2001] G.Fenk-Oczlon. Familiarity, information flow, and linguistic form. In: Frequency and the Emergence of Linguistic Structure. Eds. J.Bybee and P.Hopper. John Benjamins B.V., Amsterdam/Philadelphia, 2001.

[Fenk-Oczlon and Fenk, 2002] G.Fenk-Oczlon and A.Fenk. Zipf’s tool analogy and word order. Glottometrics, 5, 22-28, 2002.

[Hasher and Chromiak, 1975] L.Hasher and W.Chromiak. The processing of frequency information: an automatic mechanism. Journal of Verbal Learning and Verbal Behavior, 16, 173-184, 1975.

[Hume, 1777] D.Hume. An Enquiry Concerning Human Understanding. Ed. E.Steinberg. Hackett Publishing Company, Indianapolis, 1993.

[Jespersen, 1933] O.Jespersen. Monosyllabism in English. Linguistica. In: Selected Writings of Otto Jespersen, George Allen and Unwin LTD, London, (no year).

[Kandel and Schwartz, 1982] E.R.Kandel and J.H.Schwartz. Molecular biology of learning: Modulation of transmitter release. Science, 218, 433-443, 1982.

[Ke, 2006] J. Ke. A cross-linguistic quantitative study of homophony. Journal of Quantitative Linguistics, 13, 129-159, 2006.

[Keller, 1995]. R.Keller. Zeichentheorie: zu einer Theorie semiotischen Wissens. Francke, Tübingen, 1995.

[Köhler, 1986] R.Köhler. Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Brockmeyer, Bochum, 1986.

[Kypriotaki, 1973] L. Kypriotaki. Aphaeresis in rapid speech. American Speech, 45, 99-116, 1973.

[Manczak, 1980] W.Manczak. Frequenz und Sprachwandel. In: Kommunikationstheoretische Grundlagen des Sprachwandels. Ed. H.Lüdtke. Walter de Gruyter, Berlin/New York, 1980.

[Mandelbrot, 1954]. B.Mandelbrot. Structure formelle des textes et communication. Deux etudes. Word, 10, 1-27, 1954.

[Nerlich and Clarke, 2003] B.Nerlich and D.D.Clarke. Polysemy and flexibility: introduction and overview. In: Polysemy. Flexible Patterns of Meaning in Mind and Language. Eds. B.Nerlich, Z.Todd, V.Herman, and D.D.Clarke. Mouton de Gruyter, Berlin, 2003.

[Saffran et al., 1996] J.R.Saffran, E.I.Newport and R.N.Aslin. Word segmentation: The role of distributional cues. Journal of Memory and Language, 35, 606-621, 1996.

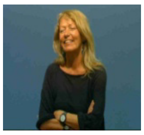
[Shannon, 1949] C.E.Shannon. The mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

[Suppes, 2009] P.Suppes. Neuropsychological foundations of philosophy. In: Reduction. Between the Mind and the Brain. Eds. A.Hieke and H.Leitgeb. ontos verlag, Heusenstamm, 2009.

[Zipf, 1929] G.K.Zipf. Relative frequence as a determinant of phonetic change. Harvard Studies in Classical Philology, 40, 1-95, 1929.

[Zipf, 1949] G.K.Zipf. Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology, Addison-Wesley. Cambridge, MA, 1949.

Authors' Information



Gertraud Fenk-Oczlon *– Department of Linguistics and Computational Linguistics, Alps-Adriatic University of Klagenfurt, Universitaetsstrasse 65-67, 9020 Klagenfurt, Austria; e-mail: gertraud.fenk@uni-klu.ac.at*

Major Fields of Scientific Research: Cognitive linguistics, Systemic typology, Language evolution

August Fenk *– Department of Media and Communication Sciences and Department of Psychology, Alps-Adriatic University of Klagenfurt, Universitaetsstrasse 65-67, 9020 Klagenfurt, Austria; e-mail: august.fenk@uni-klu.ac.at*

Major Fields of Scientific Research: Cognitive psychology, Quantitative linguistics, Theoretical semiotics

CONCEPTUAL MODELING IN SPECIALIZED KNOWLEDGE RESOURCES

Pamela Faber, Antonio San Martín

***Abstract**: Conceptual modeling is the activity of formally describing aspects of the physical and social world around us for purposes of understanding and communication. The conceptual modeler thus has to determine what aspects of the real world to include, and exclude, from the model, and at what level of detail to model each aspect [Kotiadis and Robinson, 2008]. The way that this is done depends on the needs of the potential users or stakeholders, the domain to be modeled, and the objectives to be achieved. A principled set of conceptual modeling techniques are thus a vital necessity in the elaboration of resources that facilitate knowledge acquisition and understanding.*

In this respect, the design and creation of terminological databases for a specialized knowledge domain is extremely complex since, ideally, the data should be interconnected in a semantic network by means of an explicit set of semantic relations. Nevertheless, despite the acknowledged importance of conceptual organization in terminological resources [Puuronen, 1995], [Meyer et al., 1997], [Pozzi, 1999], [Pilke, 2001], conceptual organization does not appear to have an important role in their design. It is a fact that astonishingly few specialized knowledge resources available on Internet contain information regarding the location of concepts in larger knowledge configurations [Faber et al., 2006].

Such knowledge resources do not take into account the dynamic nature of categorization, concept storage and retrieval, and cognitive processing [Louwerse and Jeuniaux, 2010], [Aziz-Zadeh and Damasio, 2008], [Patterson et al., 2007], [Gallese and Lakoff, 2005]. Recent theories of cognition reflect the assumption that cognition is typically grounded in multiple ways, e.g. simulations, situated action, and even bodily states. This means that a specialized knowledge resource that facilitates knowledge acquisition should thus provide conceptual contexts or situations in which a concept is conceived as part of a process or event. Since knowledge acquisition and understanding requires simulation, this signifies that horizontal relations defining goal, purpose, affordance, and result of the manipulation and use of an object are just as important, if not more so, than vertical generic-specific and part-whole relations.

Within the context of recent theories of cognition, this paper examines the frame-based conceptual modeling principles underlying EcoLexicon, a multilingual knowledge base of environmental concepts (http://ecolexicon.ugr.es/) [Faber et al., 2005, 2006, 2007].

***Keywords**: conceptual modeling, terminological knowledge base, cognition, specialized knowledge representation*

***ACM Classification Keywords**: J.5 Arts and Humanities – Linguistics*

1. Introduction

Conceptual modeling is the activity of formally describing aspects of the physical and social world around us for purposes of understanding and communication. The conceptual modeler thus has to determine what aspects of the real world to include, and exclude, from the model, and at what level of detail to model each aspect [Kotiadis and Robinson, 2008]. The way that this is done depends on the needs of the potential users or stakeholders, the domain to be modeled, and the objectives to be achieved. A principled set of conceptual modeling techniques are thus a vital necessity in the elaboration of resources that facilitate knowledge acquisition and understanding. Such resources would ideally allow non-experts to understand a given domain by focusing on and capturing essential knowledge.

2. Terminology, user needs and terminological knowledge bases

Terminology and specialized knowledge representation is basic to knowledge acquisition processes such as specialized translation and communication. Given that terms are the linguistic designations of specialized knowledge concepts, it goes without saying that they are inextricably linked to their representation, activation, transmission, and acquisition of specialized knowledge. According to [Sandrini, 2000: 1], concepts are at the center of all types of knowledge, and constitute the key elements of the knowledge space of a subject area. A knowledge space is made up of relations among concepts of a predefined domain, and is represented by statements. Concept systems are evidently a core element in conceptual knowledge representation and acquisition.

As is well-known, a major focus in both applied and theoretical Terminology and Specialized Communication has always been conceptual organization. In fact, a great deal has been written on the topic [Budin, 1994], [Puuronen, 1995], [Meyer and Mackintosh, 1996], [Meyer et al., 1997], [Pozzi, 1999], [Pilke, 2001], [Feliu, 2004], [Tebé, 2005], [Faber et al., 2007], [León 2009], *inter alia*. Given the fact that terms are specialized knowledge units that designate our conceptualization of objects, qualities, states, and processes in a specialized domain and are key to understanding, any theory of conceptual modeling and knowledge representation should aspire to psychological and neurological adequacy. Conceptualization processes as well as the organization of semantic information in the brain should underlie any theoretical assumptions concerning the access, retrieval, and acquisition of specialized knowledge as well as the design of specialized knowledge resources. However, quite often, this is not the case.

It is a fact that conceptual organization (of any sort), despite its acknowledged importance, does not appear to have an important role in the elaboration of specialized knowledge resources. Astonishingly few resources are conceptually organized, and even those that are based on meaning merely provide an overview of a specialized field, solely based on the IS_A or TYPE_OF conceptual relation. This overview usually consists of graphical representations of concepts in the form of tree or bracket diagrams. However, even this type of organization is a fairly rare occurrence since the great majority of terminological resources available on Internet contain little or no information regarding the location of specialized knowledge concepts in larger knowledge configurations [Faber et al., 2006].

Even when concept maps or representations are provided, they rarely respond to user needs or expectations. Our experience as thinkers tells us that the mainstream conceptual tree does not adequately reflect what is in our mind, and that our mental representations are much richer and more flexible than such representations of conceptual structure.

Since knowledge resources should reflect, to the extent possible, conceptual categories and the processes that actually occur in the brain, the question is how an awareness of the nature of mental processes can be applied to the representation of specialized knowledge concepts in order to enhance specialized knowledge acquisition.

3. Theories of cognition

As is well-known, standard theories of cognition are based on abstract, amodal representations of entities, events, and processes stored in semantic memory, which do not take into account the human and contextual factor of processors, their focus of attention, spatiotemporal situation, or context of perception [Barsalou, 2008: 618], [Mahon and Caramazza, 2008: 59]. As it happens, these conventional (though inadequate) theories of cognition are the same theories upon which mainstream conceptual representations (or conceptual trees) in specialized knowledge domains are currently based.

The question is what really happens in our mind when we think about something, and how we acquire permanent knowledge about it. Recently, a set of new theories of cognition have been proposed that provide new insights into conceptualization processes. These theories claim that cognition is situated, and that understanding is equated with sensory and motor simulation. In other words, when we encounter a physical object, we partially capture property information on sensory modalities so that this information can later be reactivated [Damasio and Damasio, 1994].

For example, to represent the concept, PEACH, neural systems for vision, action, touch, taste and emotion partially reenact the perceiver's experience of a peach. These reenactments or simulations are not the same thing as mental imagery, which is consciously evoked in working memory. Unlike mental imagery, these simulations seem to be relatively automatic processes that lie outside of our awareness [Simmons et al., 2005: 1602].

To date, brain-imaging experiments have largely involved the conceptualization of everyday objects such as cups, hammers, pencils, and food, which, when perceived, trigger simulations of potential actions. For example, the handle of a cup activates a grasping simulation [Tucker and Ellis, 1998, 2001]. Food activates brain areas related to gustatory processing as well as areas in the visual cortex representing object shape [Simmons et al., 2005]. When conceptual knowledge about objects is represented, brain areas represent the shape and color of objects, the motion they exhibit, and the actions that agents perform on them become active to represent these properties conceptually.

Such reenactments not only occur in the presence of the object itself, but also in response to words and other symbols. For precisely this reason, they should be taken into account in Terminology. Although few neuropsychological experiments of this type have ever been performed with specialized concepts, there is no reason to suppose that the brain would work any differently.

For example, when reading about hockey, experts were found to produce motor simulations absent in novices [Holt and Beilock, 2006]. In all likelihood, a similar result would be the obtained if the object were a tide gauge, pluviometer, or anemometer. The expert's brain would show motor simulations in brain areas that would not be activated in the case of non-experts to whom the object was unfamiliar. The information regarding simulated interaction is thus a vital part of conceptual meaning. The way that object concepts are represented in our brain seems to suggest that current methods and ways of elaborating specialized knowledge representations should be modified in order to take this information into account in order to facilitate knowledge acquisition.

4. Applying situated cognition to specialized knowledge representation

Yet, we may well ask ourselves if such research on cognition, however valuable, can be usefully applied to the creation of specialized knowledge resources. We believe that the answer is yes. First of all, situated conceptualizations reflect the fact that concepts are not processed in isolation, but are typically situated in background situations and events [Barsalou, 2003]. This signifies that context is crucial in knowledge representation. At any given moment in the perception of the entity, people also perceive the space surrounding it, including the agents, objects, and event present in it [Barsalou, 2009: 1283], and this can be applied to specialized knowledge modeling.

For example, EROSION is the wearing away of the earth's surface, but whether conceptualized as a process or the result of this process, erosion cannot be conceived in isolation. It is induced by an agent (wind, water, or ice) affects a geographic entity (the Earth's surface) by causing something (solids) to move away. Moreover, any process takes place over a period of time, and can be divided into smaller segments. In this sense, erosion can happen at a specific season of the year, and may take place in a certain direction. All of this information about erosion should be available for potential activation when we think about the concept, and wish to acquire knowledge about it. The meaning of a concept is constructed on-line, and is modulated by context.

4.1 Frame-based Terminology and dynamic knowledge representation

Accordingly, a knowledge resource that facilitates knowledge acquisition should not be in the form of a static term base with a list of unrelated data records. It should represent concepts as part of a larger context or situation in which the concept is related to others in a dynamic structure that can streamline the action-environment interface.

Frame-based terminology [Faber et al., 2005, 2006, 2007] uses a modified version of Fillmore’s Frames [Fillmore 1982, 1985], [Fillmore and Atkins, 1992] coupled with premises from Cognitive Linguistics to configure specialized domains on the basis of definitional templates and create situated representations for specialized knowledge concepts.

4.1.1 Event representation

In Frame-based Terminology, conceptual networks are based on an underlying domain event as well as a closed inventory of both hierarchical and non-hierarchical semantic relations. We have used these premises to construct an environmental knowledge base called EcoLexicon (http://ecolexicon.ugr.es/). The main focus is on conceptual relations as well as a concept’s combinatorial potential, extracted from corpus analysis. This prototypical domain event or action-environment interface [Barsalou, 2003] provides a template applicable to all levels of information structuring.

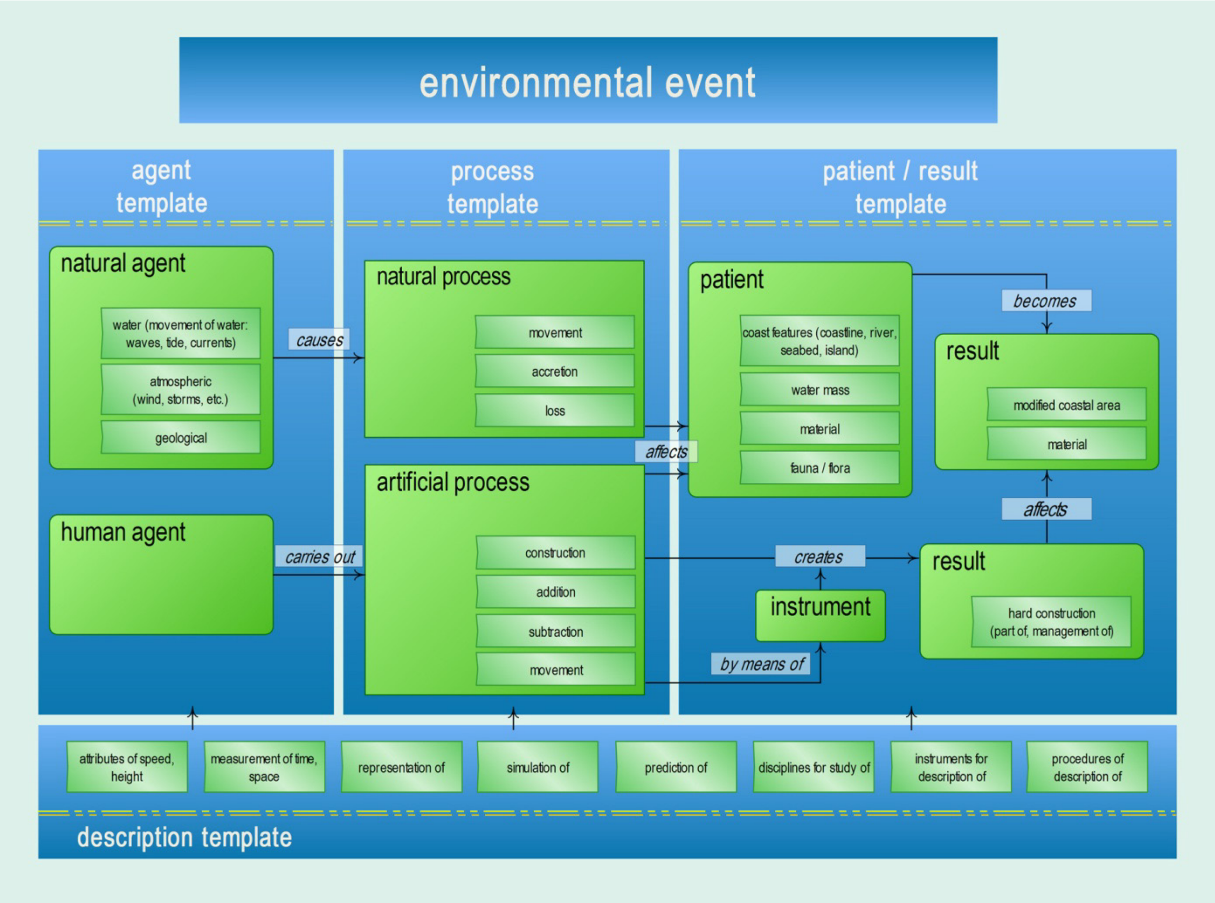
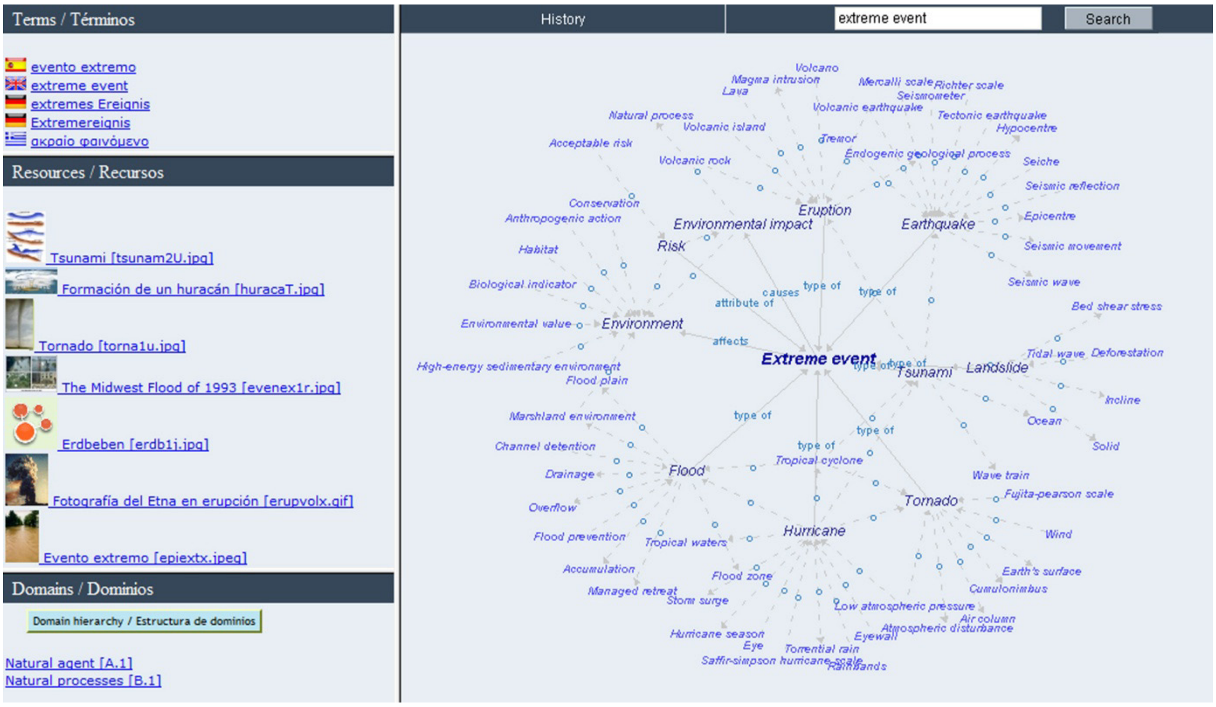


Figure 1. Environmental Event.

In EcoLexicon, knowledge can be accessed from top-level categories to more specific relational structures. The most generic level is the Environmental Event (EE), which provides a frame for the organization of all concepts in the knowledge base. As shown in Figure 1, the EE is conceptualized as a dynamic process that is initiated by an agent (either natural or human). This process affects a patient (an environmental entity), and produces a result. These categories (agent, process, patient, etc.) are the concept roles characteristic of this specialized domain. Additionally, there are peripheral categories which include instruments that are typically used during the EE, as well as a category where the concepts of measurement, analysis, and description of the processes in the main event are included. This event-based representation facilitates knowledge acquisition in text processing since conceptual categories are bound together by event knowledge.

4.1.1.1 extreme event

For example, one of the concepts in EcoLexicon is EXTREME EVENT in its sense of natural disaster. Disasters in the environment include great earthquakes, floods, giant sea waves, hurricanes, tornadoes, etc., and their consequences. The concept of EXTREME EVENT is very complex since it is a natural agent that initiates a process (i.e. earthquakes or volcanic eruptions can produce tsunamis) but it can also be the process itself, which occurs in time and space. This information is represented in EcoLexicon as shown in Figure 2.



Terms / Términos

evento extremo
extreme event
extremes Ereignis
Extremereignis
ακραίο φαινόμενο

Resources / Recursos

Tsunami [tsunam2U.jpg]
Formación de un huracán [huracaT.jpg]
Tornado [torna1u.jpg]
The Midwest Flood of 1993 [evenex1r.jpg]
Erdbeben [erdb1j.jpg]
Fotografía del Etna en erupción [erupvolx.gif]
Evento extremo [epiextx.jpeg]

Domains / Dominios

Domain hierarchy / Estructura de dominios

Natural agent [A.1]
Natural processes [B.1]

Figure 2. Representation of EXTREME EVENT in EcoLexicon.

As shown in Figure 2, all of the concepts closest to the central concept are connected to it by a series of conceptual relations that are explicitly named (e.g. TYPE-OF, CAUSES, AFFECTS, etc.). Since EXTREME EVENT is a very general concept, the only visual information that can be associated with it is that of its subtypes (HURRICANE, TORNADO, EARTHQUAKE, FLOOD, etc.). The majority of relations at this level are thus TYPE_OF. However, EXTREME EVENT also activates non-hierarchical relations typical of the general event frame. As such, its principal attribute is RISK; it AFFECTS the environment; and CAUSES an environmental impact. As for the TYPE_OF relations, they can be regarded as access routes to more prototypical base-level concepts [Rosch, 1978], which do have a mental image, and can activate specific contexts. This set of subtypes (hurricane, tornado, flood, tsunami, etc.) take the form of constellations, each with their own set of subordinate concepts and conceptual relations, which encode more specific sub-event knowledge and representations.

4.1.1.2. Recontextualization: hurricane

According to [Barsalou, 2005], a given concept produces many different situated conceptualizations, each tailored to different instances in different settings. Thus, context can be said to be a dynamic construct that activates or restricts knowledge. This general event that codifies a natural disaster can thus be recontextualized at any moment to center on any of the more specific subevents. For example, when the EXTREME EVENT representation is recontextualized to focus on HURRICANE, it takes the following form.

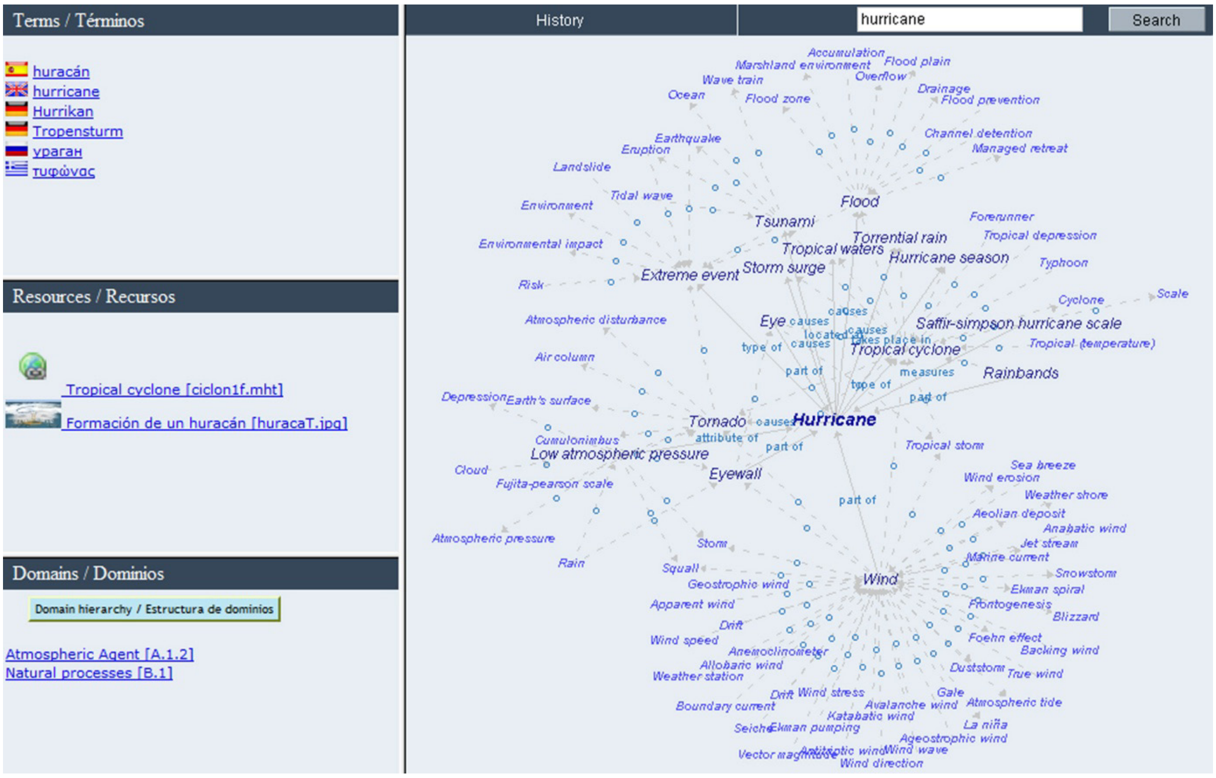


Figure 3. Representation of HURRICANE in Ecolexicon.

This type of recontextualization of EXTREME EVENT still contains a sector of the previous information, but varies the focus of attention so that hurricane is now the center of focus. Besides communicating the fact that hurricane is a type of extreme event, this new representation highlights the fact that wind and flooding are crucial participants in the event. Wind is part of a hurricane, and a hurricane causes floods. Not surprisingly, WIND and FLOOD are concepts that are susceptible to simulation since they can directly affect human life and health. It also mentions the attribute of low atmospheric pressure as well as the scale used for hurricane measurement (Saffir-Simpson hurricane scale), which codifies an important aspect of expert interaction with a hurricane.

4.1.2. Object representation

Object concepts can also be represented dynamically as parts of events. One of the basic characteristics of objects is knowledge of whether and how they can be manipulated. In the case of man-made objects, another important property is their function, or how they can be used. This would mean that an important part of the information in the representation of specialized engineering instruments would evidently involve how they are used by humans, for what purpose, and what is the result of the manipulation.

4.1.2.1. Recording instrument

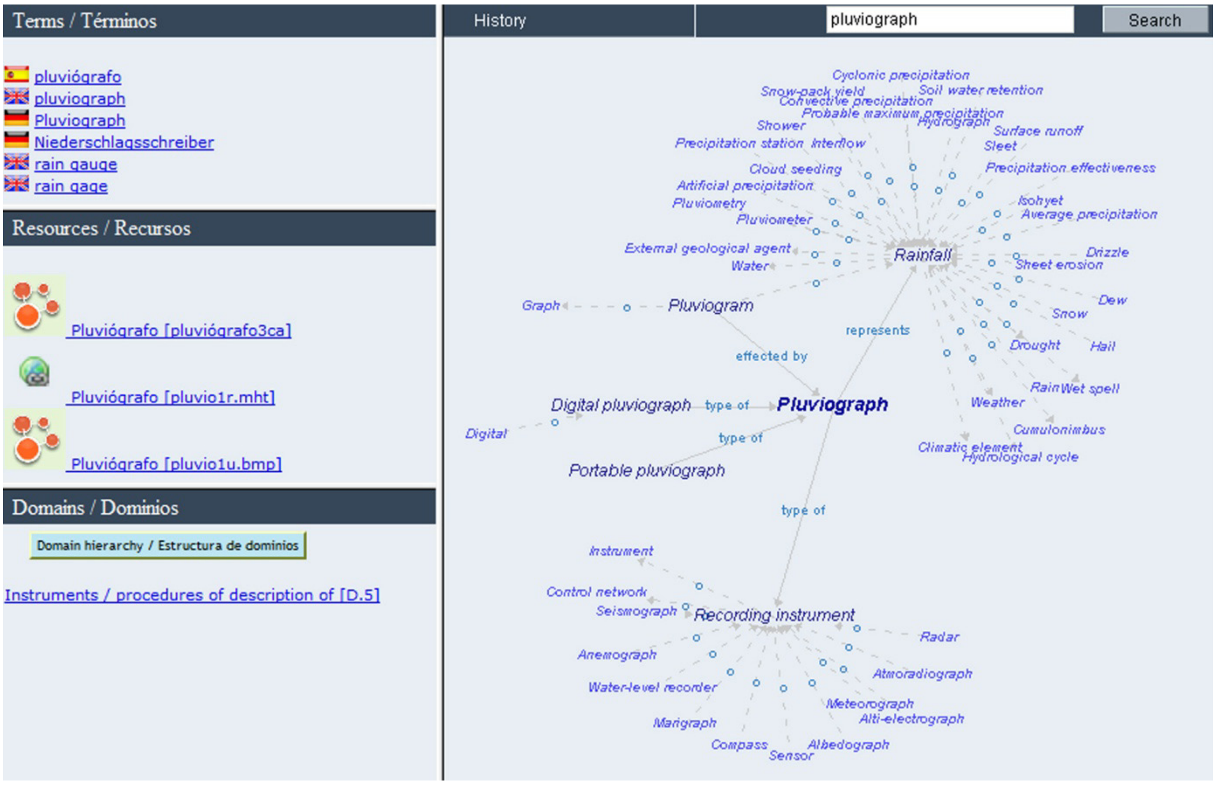


Figure 4. Representation of PLUVIOGRAPH in EcoLexicon.

For example, a RECORDING INSTRUMENT (e.g. marigraph, pluviograph, anemograph, etc.) is a subtype of INSTRUMENT. As a manmade manipulable artifact, a recording instrument has a function (i.e. recording) as well as an object that is recorded (tides, rain, wind, etc.). As a tool, it is strongly susceptible to human interaction, and activates a simulation frame in which much of the perceiver's knowledge of the artifact involves his/her ability to handle it and in some way to extract information from it. For example, Figure 4 shows the representation of PLUVIOGRAPH.

The representation of PLUVIOGRAPH, of course, includes TYPE_OF information. A pluviograph is a recording instrument, and has subtypes, such as digital pluviograph and portable pluviograph. However, it is also part of what might be called a RECORDING EVENT in which a human agent causes the machine to record and generate a representation of something (RAINFALL). The recording instrument used in this event is a pluviograph, which produces (or effects) a PLUVIOGRAM. As can be observed in Figure 4, this process is reflected in the non-hierarchical relations REPRESENTS and EFFECTED_BY.

5. Conclusions

In order to translate specialized texts, translators must acquire sufficient knowledge of conceptual content. Although it is not necessary to have the same depth of knowledge as an expert in the field, there is a minimum threshold that must be met. The knowledge acquisition process can be carried out in cost-effective time if translators have a set of search strategies developed and knowledge resources at their disposal.

One of the problems of knowledge acquisition is precisely the lack of translation-oriented terminological resources that reflect the complexity and dynamicity of conceptualization. Although in terminology theory, much emphasis is placed on conceptual representation, reality shows that very few specialized dictionaries or glossaries are concept-based, and those that are based on meaning, only offer static representations based on the IS_A or PART_OF relation.

A truly effective specialized knowledge resource should reflect recent advances in neurocognition which point to the following:

1. No specialized knowledge concept should be activated in isolation, but rather as part of a larger structure or event. A specialized knowledge resource that facilitates knowledge acquisition should thus provide conceptual contexts or situations in which a concept is related to others as part of a process or event.

2. Since knowledge acquisition and understanding requires simulation, this signifies that non-hierarchical relations defining goal, purpose, affordance, and result of the manipulation and use of an object are just as important as hierarchical generic-specific and part-whole relations.

3. Specialized domains are constrained by the nature of their members. This is reflected in clusters of conceptual relations that make up the general representational template, characterizing different categories.

All of these conclusions have been illustrated by examples from EcoLexicon, an environmental knowledge base (available at: http://ecolexicon.ugr.es/). EcoLexicon is a conceptually-organized, frame-based terminological resource that facilitates knowledge acquisition since it presents concepts as part of larger knowledge structures and permits dynamic processes such as the recontextualization of knowledge representations.

Bibliography

[Aziz-Zadeh and Damasio, 2008] L. Aziz-Zadeh and A. Damasio. Embodied semantics for actions: Findings from functional brain imaging. In: Journal of Physiology – Paris 102, 35–39. 2008.

[Barsalou, 2003] L. W. Barsalou. Situated simulation in the human conceptual system. In: Language and Cognitive Processes 18, 513–62. 2003.

[Barsalou, 2005] L. W. Barsalou. Situated conceptualization. In: Handbook of Categorization in Cognitive Science. Ed. H. Cohen and C. Lefebvre. Elsevier, St. Louis, 619–650. 2005.

[Barsalou, 2008] L. W. Barsalou. Grounded cognition. In: Annual Review of Psychology 59, 617–645. 2008.

[Barsalou, 2009] L. W. Barsalou. Simulation, situated conceptualization, and prediction. In: Philosophical Transactions of the Royal Society B, 1281–1289. 2009.

[Budin, 1994] G. Budin. Some hypotheses about concept representations. In: Proceedings of the 9th European Symposium on LSP, Bergen, Norway, 2-6 August. Fagbokforlaget , Bergen, 919-924. 1994.

[Cabré, 1999] M. T. Cabré. Terminology Theory, Methods and Applications. John Benjamins, Amsterdam/Philadelphia. 1999.

[Caramazza and Mahon, 2003] A. Caramazza and B. Z. Mahon. The organization of conceptual knowledge: the evidence of category-specific semantic deficits. In: Trends in Cognitive Sciences 7(8), 354–361. 2003.

[Damasio and Damasio, 1994] Damasio, A. and Damasio, H. Cortical systems for retrieval of concrete knowledge: the convergence zone framework. In:, Large-scale Neuronal Theories of the Brain. Ed. C. Koch and J. Davis. MIT Press, Cambridge, MA. 1994.

[Faber et al., 2007] P. Faber, P. León, J. A. Prieto, and A. Reimerink. Linking images and words: the description of specialized concepts. In: International Journal of Lexicography 20, 39–65. 2007.

[Faber et al., 2005] P. Faber, C. Márquez, and M. Vega. Framing Terminology: A Process-Oriented Approach. In: Meta 50 (4). 2005.

[Faber et al., 2006] P. Faber, S. Montero, M.R. Castro, J. Senso, J.A. Prieto, P. León, C. Márquez, M. Vega. Process-oriented terminology management in the domain of Coastal Engineering. In: Terminology 12(2), 189–213. 2006.

[Feliu, 2004] J. Feliu. Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica. PhD thesis. Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada (IULA), Barcelona. 2004.

[Fillmore, 1982] C. J. Fillmore. Frame semantics. In: Linguistics in the Morning Calm. Ed. Linguistics Society of Korea. Hanshin, Seoul, 111–137. 1982.

[Fillmore, 1985] C. J. Fillmore. Frames and the semantics of understanding. In: Quaderni di Semántica 6(2), 222–254. 1985

[Fillmore and Atkins, 1992] C. J. Fillmore and B. T. S. Atkins. Towards a frame-based lexicon: the semantics of risk and its neighbours. In: Frames, Fields and Contrasts. Ed. A. Lehrer and E. Kittay. Lawrence Erlbaum, Hillsdale, NJ, 75-102. 1992.

[Gallese and Lakoff, 2005] V. Gallese and G. Lakoff. The brain's concepts: the role of the sensory-motor system in conceptual knowledge. In: Cognitive Neuropsychology 22 (3/4), 455–479. 2005.

[Holt and Beilock, 2006] L. E. Holt and S. L. Beilock S. L. Expertise and its embodiment: examining the impact of sensorimotor skill expertise on the representation of action-related text. In: Psychonomic Bulletin and Review 13, 694–701. 2006.

[Humphreys and Forde, 2001] G. W. Humphreys and E. M. Forde. Hierarchies, similarity, and interactivity in object recognition: 'category specific' neuropsychological deficits. In: Behavioral and Brain Sciences 24, 453–509. 2001.

[Kotiadis and Robinson, 2008] K. Kotiadis and S. Robinson. Conceptual modeling: Knowledge acquisition and model abstraction. In: Proceedings of the 2008 Winter Simulation Conference, Miami Florida, 7-10 December 2008. Ed. S. J. Mason, R.R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler. IEEE Press, Austin. 2008.

[León, 2009] P. León. 2009. Representación multidimensional de conocimiento especializado. PhD thesis. University of Granada, Granada.

[Louwerse and Jeuniaux, 2010] M. M. Louwerse and P. Jeuniaux. The linguistic and embodied nature of conceptual processing. In: Cognition 114, 96–104. 2010.

[Mahon and Caramazza, 2008] M. Z. Mahon and A. Caramazza. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. In: Journal of Physiology–Paris 102, 59–70. 2008.

[Mahon and Caramazza, 2009] M. Z. Mahon and A. Caramazza. Concepts and categories: A cognitive neuropsychological perspective. In: Annual Review of Psychology 60, 27–51. 2009.

[Martin, 2007] A. Martin. The representation of object concepts in the brain. In: Annual Review of Psychology 58, 25–45. 2007.

[Meyer and Mackintosh, K. 1996] I. Meyer and K. Mackintosh. Refining the terminographer's concept-analysis methods: How can phraseology help? In: Terminology 3(1), 1–26.

[Meyer et al., 1997] I. Meyer, K. Eck, and D. Skuce. Systematic concept analysis within a knowledge-based approach to terminology. In: Handbook of Terminology Management. Ed. S. E. Wright and G. Budin. John Benjamins, Amsterdam/Philadelphia, 98–118. 1997

[Patterson et al., 2007] K. Patterson, P. J. Nestor, and T. T. Rogers. Where do you know what you know? The representation of semantic knowledge in the human brain. In: Nature Reviews Neuroscience 8, 976–988. 2007.

[Pavel and Nolet, 2001] S. Pavel and D. Nolet. Handbook of Terminology. Minister of Public Works and Government Services, Canada. 2001.

[Pilke, 2001] N. Pilke. Field-specific features of dynamic Concepts – What, when and why? In: Language for Special Purposes: Perspective for the New Millennium. Ed. F. Mayer. Gunter Narr, Tübingen. 2001

[Pozzi, 1999] M. Pozzi. The Concept of 'Concept' in Terminology: a Need for a New Approach. In: TKE'99 Terminology and Knowledge Engineering Proceedings, Fifth International Congress on Terminology and Knowledge Engineering, 23–27 August, 1999. Ed. P. Sandrini. TermNet, Vienna. 1999.

[Puuronen, 1995] N. Puuronen. On describing dynamic concepts – A philosophical and terminological approach. In: ITTF Proceedings of the 10th European LSP Symposium. Ed. G. Budin. TermNet, Vienna. 1995.

[Rosch, E., 1978] E. Rosch. Principles of categorization. In: Cognition and Categorization. Ed. E. Rosch and B. B. Lloyd. Erlbaum, Hillsdale, NJ, 27–8. 1978.

[Sandrini, 2000] P. Sandrini. Joint Activities at the Interface between Terminology and Knowledge Engineering. Paper presented at the Conference for a Terminology Infrastructure in Europe. Maison de l'Unesco, Paris. 13-15 March 2000. Available at: http://homepage.uibk.ac.at/~c61302//publik/paris.pdf

[Simmons et al., 2005] W. K. Simmons, A. Martin, and L. W. Barsalou. Pictures of appetizing foods activate gustatory cortices for taste and reward. In: Cerebral Cortex 15, 1602–1608. 2005.

[Smith and Mark, 1999] B. Smith and D. Mark. Ontology with human subjects testing: An empirical investigation of geographic categories. In: American Journal of Economics and Sociology 582, 245–272. 1999.

[Tebé, 2005] C. Tebé. 2005. La representació conceptual en terminologia: l'atribució temàtica en els bancs de dades terminològiques. PhD thesis. Universidat Pompeu Fabra, Institut Universitari de Lingüística Aplicada (IULA), Barcelona. 2005.

[Tucker and Ellis, 1998.] M. Tucker and R. Ellis. On the relations between seen objects and components of potential actions. In: Journal of Experimental Psychology: Human Perception and Performance 24, 830–46. 1998.

[Tucker and Ellis, 2001] M. Tucker and R. Ellis. The potentiation of grasp types during visual object categorization. In: Visual Cognition 8, 769–800. 2001.

[Warrington and McCarthy, 1983] E. K. Warrington and R. McCarthy. Category specific access dysphasia. In: Brain 106, 859–878. 1983.

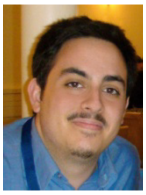
[Warrington and McCarthy, 1987] E. K. Warrington and R. McCarthy. Categories of knowledge: further fractionations and an attempted integration. In: Brain 110, 1273–1296. 1987.

[Warrington and Shallice, 1984] E. K. Warrington and R. McCarthy. Category-specific semantic impairment. In. Brain 107, 829–854. 1984.

Authors' Information

Pamela Faber *– Full Professor at the Department of Translation and Interpreting (University of Granada). Calle Buensuceso 11, 18079 Granada (Spain); e-mail: pfaber@ugr.es*

Major Fields of Scientific Research: Terminology, Specialized Translation, Cognitive Semantics, and Lexicography.



Antonio San Martín *– Research Fellow at the Department of Translation and Interpreting (University of Granada). Calle Buensuceso 11, 18079 Granada (Spain); e-mail: asanmartin@ugr.es*

Major Fields of Scientific Research: Terminology, Knowledge Representation, Cognitive Semantics, Lexicography, and Specialized Translation.

CONTEXT-BASED MODELLING OF SPECIALIZED KNOWLEDGE[1]

Pilar León Araúz, Arianne Reimerink, Alejandro G. Aragón

Abstract: *EcoLexicon is a terminological knowledge base (TKB) on the environment where different types of information converge in a multimodal interface: semantic networks, definitions, contexts and images. It seeks to meet both cognitive and communicative needs of different users, such as translators, technical writers or even environmental experts. According to Meyer et al. [1992], TKBs should reflect conceptual structures in a similar way to how concepts relate in the human mind. From a neurological perspective, Barsalou [2009: 1283] states that a concept produces a wide variety of situated conceptualizations in specific contexts, which clearly determines the type and number of concepts to be related to. The organization of semantic information in the brain should thus underlie any theoretical assumption concerning the retrieval and acquisition of specialized knowledge concepts as well as the design of specialized knowledge resources [Faber, 2010]. Furthermore, since categorization itself is a dynamic context-dependent process, the representation and acquisition of specialized knowledge should certainly focus on contextual variation. Context includes external factors (situational and cultural) as well as internal cognitive factors, all of which can influence one another [House, 2006: 342]. This view goes hand in hand with the perception of language as a kind of action, where the meaning of linguistic forms is understood as a function of their use [Reimerink et al., 2010]. In this paper we briefly describe each module of our resource and explain how EcoLexicon has been contextualized according to conceptual and terminological information. The conceptual contextualization of different entries in EcoLexicon has been performed according to role-based domains and contextual domains, whereas terminological contextualization is based on contextual domains and use situations. In this way, context is two-fold, since we account for the referential context of concepts in the real world and users’ own communicative and cognitive context.*

Keywords: *context, dynamism, reconceptualization, environmental knowledge, TKB.*

ACM Classification Keywords: *H.5.2 User interfaces – Natural language*

1. Introduction

EcoLexicon[2] is a terminological knowledge base (TKB) on the environment where different types of information converge in a multimodal interface: semantic networks, definitions, contexts and images. It seeks to meet both

[1] This research has been supported by project FFI2008-06080-C03-01/FILO, from the Spanish Ministry of Science and Innovation.

[2] http://ecolexicon.ugr.es/

cognitive and communicative needs of different users, such as translators, technical writers or even environmental experts. So far it has 3,146 concepts and 14,058 terms in Spanish, English and German, which converge in a multimodal interface with different types of information. Currently, four new languages are being added: Modern Greek, Russian, Dutch and French.

According to Meyer et al. [1992], TKBs should reflect conceptual structures in a similar way to how concepts relate in the human mind. From a neurological perspective, Barsalou [2009: 1283] states that a concept produces a wide variety of situated conceptualizations in specific contexts, which clearly determines the type and number of concepts to be related to. The organization of semantic information in the brain should thus underlie any theoretical assumption concerning the retrieval and acquisition of specialized knowledge concepts as well as the design of specialized knowledge resources [Faber, 2010].

Furthermore, since categorization itself is a dynamic context-dependent process, the representation and acquisition of specialized knowledge should certainly focus on contextual variation. Context includes external factors (situational and cultural) as well as internal cognitive factors, all of which can influence one another [House, 2006: 342]. This view goes hand in hand with the perception of language as a kind of action, where the meaning of linguistic forms is understood as a function of their use [Reimerink et al., 2010]. In other words, a given utterance does not have a meaning, but rather a meaning potential that will always be exploited in different ways dependent upon the discourse context [Evans, in press].

The notion of context has been widely discussed in the linguistic community [Austin, 1962; Gadamer, 1995; Grice, 1975; Sperber and Wilson, 1986, 1995], but all of the approaches seem to coincide in defining context as a dynamic construct. However, term bases are often restricted to generic-specific and part-whole relations, whereas conceptual dynamism can only be fully reflected through non-hierarchical ones. These are mostly related to the notions of movement, action and change, which are directly linked to human experience and perceptually salient conceptual features.

In the following sections we first explain how EcoLexicon has been designed and structured. Then we show how it has been contextualized according to conceptual and terminological information. The conceptual contextualization of different entries in EcoLexicon has been performed according to role-based domains and contextual domains, whereas terminological contextualization is based on contextual domains and use situations. In this way, context is two-fold, since we account for the referential context of concepts in the real world and users' own communicative and cognitive context (what users need and what they already know). First of all, in section 2, we briefly describe each module of our resource.

2. EcoLexicon: an environmental TKB

In EcoLexicon, all knowledge extracted from our specialized domain corpus has been organized at a macrostuctural level. This has resulted in a frame-like structure or prototypical domain event, namely, the Environmental Event (EE) [see Figure 1; Faber 2007, León Araúz *et al.* 2009, Reimerink and Faber 2009].

The EE provides a template applicable to all levels of information structuring from a process-oriented perspective. These macro-categories (AGENT → PROCESS → PATIENT/RESULT) are the semantic roles inherent to this specialized domain, and the EE provides a model to represent their interrelationships at different levels.

From a more fine-grained view, concepts appear in both dynamic networks and definitional statements linking them to all related concepts by means of a closed inventory of semantic relations especially conceived for the environmental domain. Figure 2 shows the network of GROYNE, associated to other concepts in a two-level hierarchy through both vertical (*type_of*, *part_of*, etc.) and horizontal relations (*has_function, located_at*, etc.).

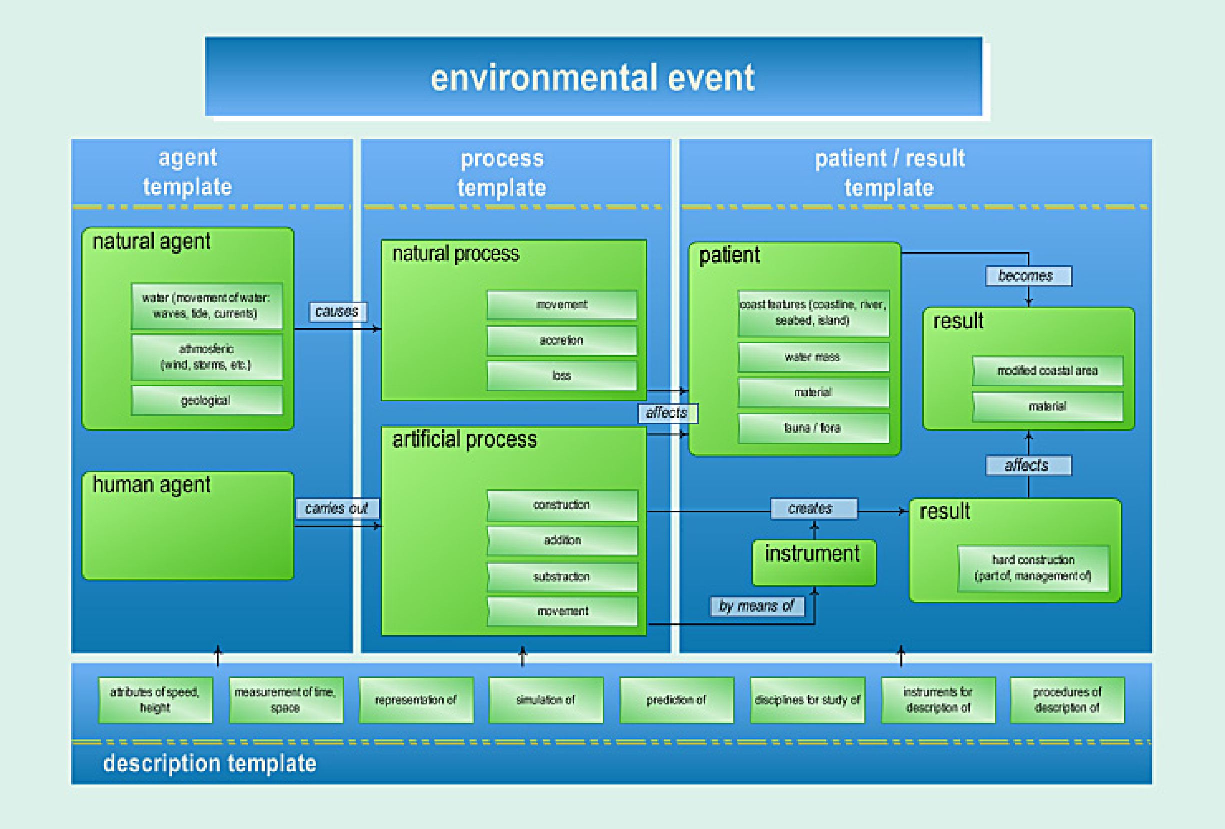


Figure 1. The Environmental Event

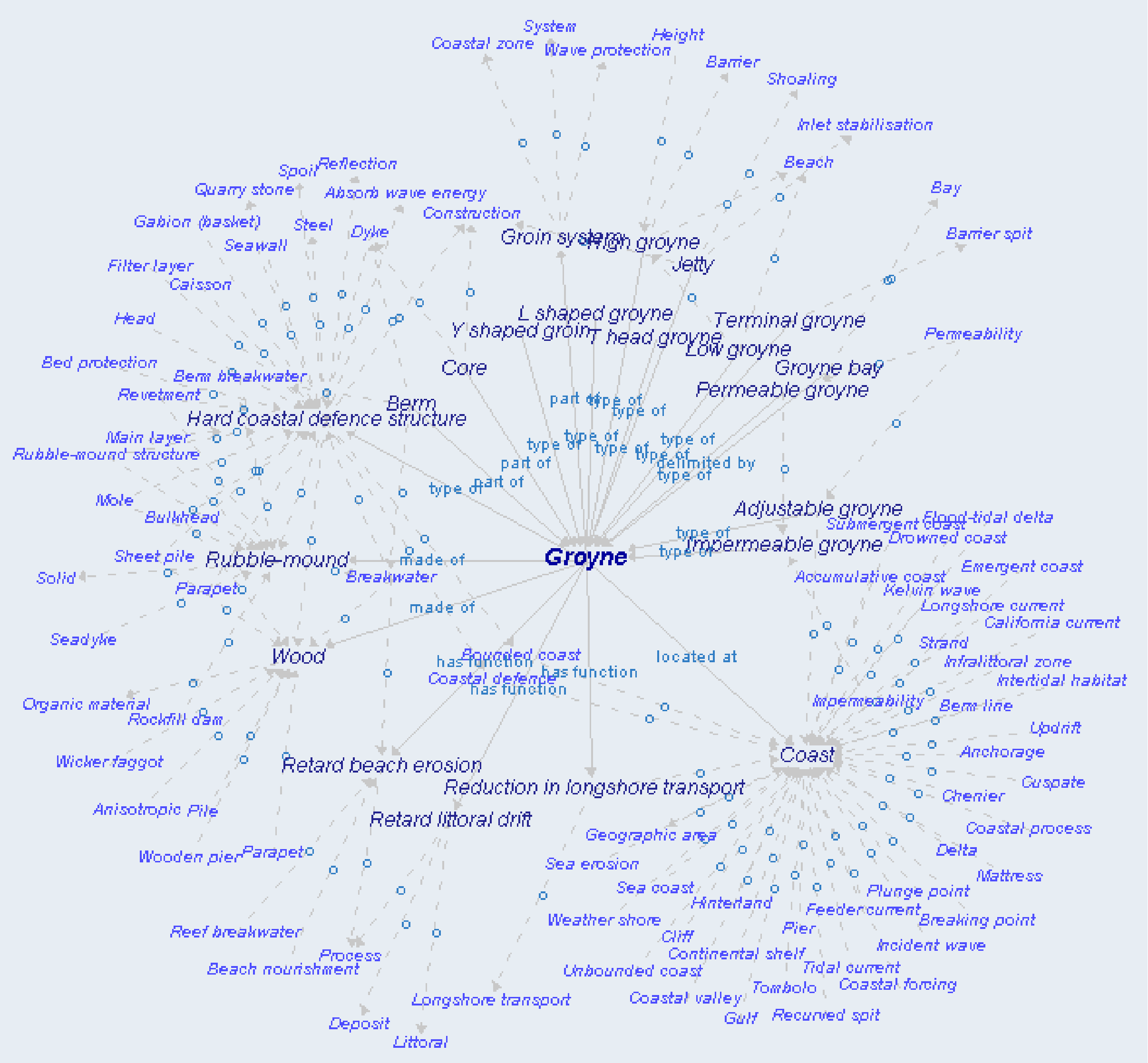


Figure 2. Conceptual network of GROYNE

In EcoLexicon definitions follow a category template [Faber et al., 2007] that constrains the definitional elements to be included (Figure 3). For example, the definitional statement of GROYNE is based on the number and type of conceptual relations defined for the category template HARD COASTAL DEFENCE STRUCTURE.

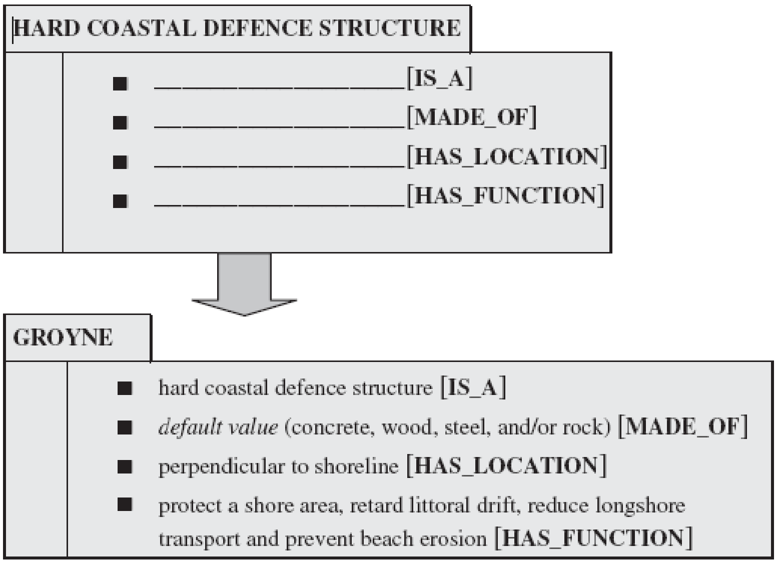

HARD COASTAL DEFENCE STRUCTURE
__________________[IS_A]
__________________[MADE_OF]
__________________[HAS_LOCATION]
__________________[HAS_FUNCTION]
GROYNE
hard coastal defence structure [IS_A]
default value (concrete, wood, steel, and/or rock) [MADE_OF]
perpendicular to shoreline [HAS_LOCATION]
protect a shore area, retard littoral drift, reduce longshore transport and prevent beach erosion [HAS_FUNCTION]

Figure 3. Activation of the HARD COASTAL DEFENCE template in the definition of GROYNE

All coordinate concepts of GROYNE make use of the same template. As functional entities, all HARD COASTAL DEFENCE STRUCTURES need the following information for an overall description: (1) the *is_a* relation marking category membership; (2) the material they are *made_of*, completed with the values of the construction material class; (3) their location, since a groyne is not a groyne if it is not *located_at* the sea; and (4) especially the purpose for which they are built.

Images are also an important part of EcoLexicon, which are chosen according to definitional structure. For example, in Table 1, five meaningful images are associated with the definition of GROYNE. *Is_a* and *made_of* relations are illustrated by a concrete iconic image, whereas *has_location* and *has_function* are described through more abstract and dynamic images.

Definitions are also complemented with other linguistic resources such as Textual Contexts (TCs) and concordances, which help users achieve a different level of understanding of the specialized domain. In Table 2, for example, GROYNE is not only defined as a coastal defence structure. Other relevant information is included as well: they are cost-effective and many coastal communities prefer other solutions.

Three types of concordances are included in each entry of EcoLexicon: conceptual, phraseological and verbal. These concordances allow the users to widen their knowledge from different perspectives. Conceptual concordances show the activation of conceptual relations in the real use of terms, which makes definitional templates empirical structures and adds new knowledge about the values of each sentence argument. Phraseological concordances help the user in acquiring specialized discourse. Thirdly, verbal concordances highlight the most frequent verbal collocations, which offer, again, both linguistic and conceptual information.

Figure 4 shows the conceptual concordances in the entry of GROYNE. Linguistic markers such as *designed to* and *provide* explicitly relate the concept to its function, SHORE PROTECTION and TRAP AND RETAIN SAND.

Table 1. Linguistic and graphic definitional information for groyne [Faber et al. 2007]

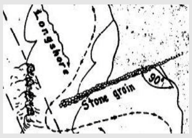
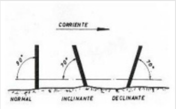
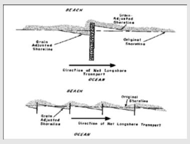
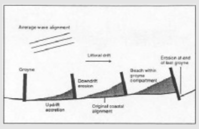
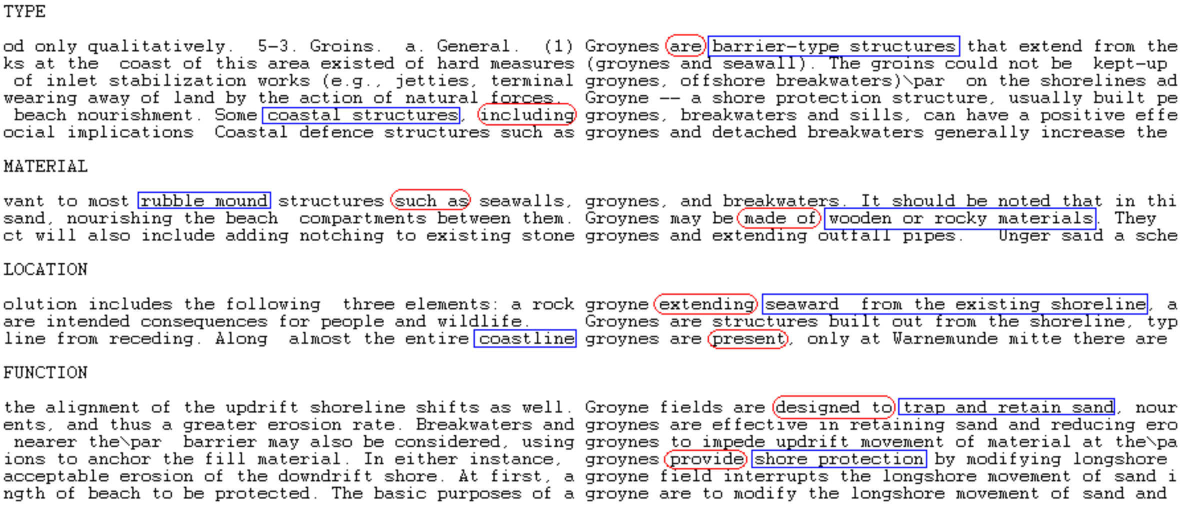
groyne		
▪ hard coastal defence structure [**IS_A**], ▪ *default value* (concrete, wood, steel, and/or rock) [**MADE_OF**]		
▪ perpendicular to shoreline [**HAS_LOCATION**]		
▪ protect a shore area, retard littoral drift, reduce longshore transport and prevent beach erosion [**HAS_FUNCTION**]		 Average wave alignment Littoral drift Groyne Downdrift erosion Beach within groyne compartment Erosion at end of last groyne Updrift accretion Original coastal alignment

Table 2. Textual context of *groyne*

Groynes are extremely cost-effective coastal defense measures, requiring little maintenance, and are one of the most common coastal defense structures. However, groynes are increasingly viewed as detrimental to the aesthetics of the coastline, and face strong opposition in many coastal communities.



```
TYPE

od only qualitatively.  5-3. Groins.  a. General.  (1) Groynes are barrier-type structures that extend from the
ks at the  coast of this area existed of hard measures (groynes and seawall). The groins could not be  kept-up
 of inlet stabilization works (e.g., jetties, terminal groynes, offshore breakwaters)\par  on the shorelines ad
wearing away of land by the action of natural forces.  Groyne -- a shore protection structure, usually built pe
 beach nourishment. Some coastal structures, including groynes, breakwaters and sills, can have a positive effe
ocial implications  Coastal defence structures such as groynes and detached breakwaters generally increase the

MATERIAL

vant to most rubble mound structures such as seawalls, groynes, and breakwaters. It should be noted that in thi
sand, nourishing the beach  compartments between them. Groynes may be made of wooden or rocky materials. They
ct will also include adding notching to existing stone groynes and extending outfall pipes.  Unger said a sche

LOCATION

olution includes the following  three elements: a rock groyne extending seaward  from the existing shoreline, a
are intended consequences for people and wildlife.     Groynes are structures built out from the shoreline, typ
line from receding. Along  almost the entire coastline groynes are present, only at Warnemunde mitte there are

FUNCTION

the alignment of the updrift shoreline shifts as well. Groyne fields are designed to trap and retain sand, nour
ents, and thus a greater erosion rate. Breakwaters and groynes are effective in retaining sand and reducing ero
 nearer the\par  barrier may also be considered, using groynes to impede updrift movement of material at the\pa
ions to anchor the fill material. In either instance,  groynes provide shore protection by modifying longshore
acceptable erosion of the downdrift shore. At first, a groyne field interrupts the longshore movement of sand i
ngth of beach to be protected. The basic purposes of a groyne are to modify the longshore movement of sand and
```

Figure 4. Conceptual concordances in the entry of GROYNE

Finally, different multilingual terminological choices are shown for every entry associated to the first hierarchical level of the searched concept. As can be seen in Figure 4, terminological variation is also reflected (two variants for Greek and English).

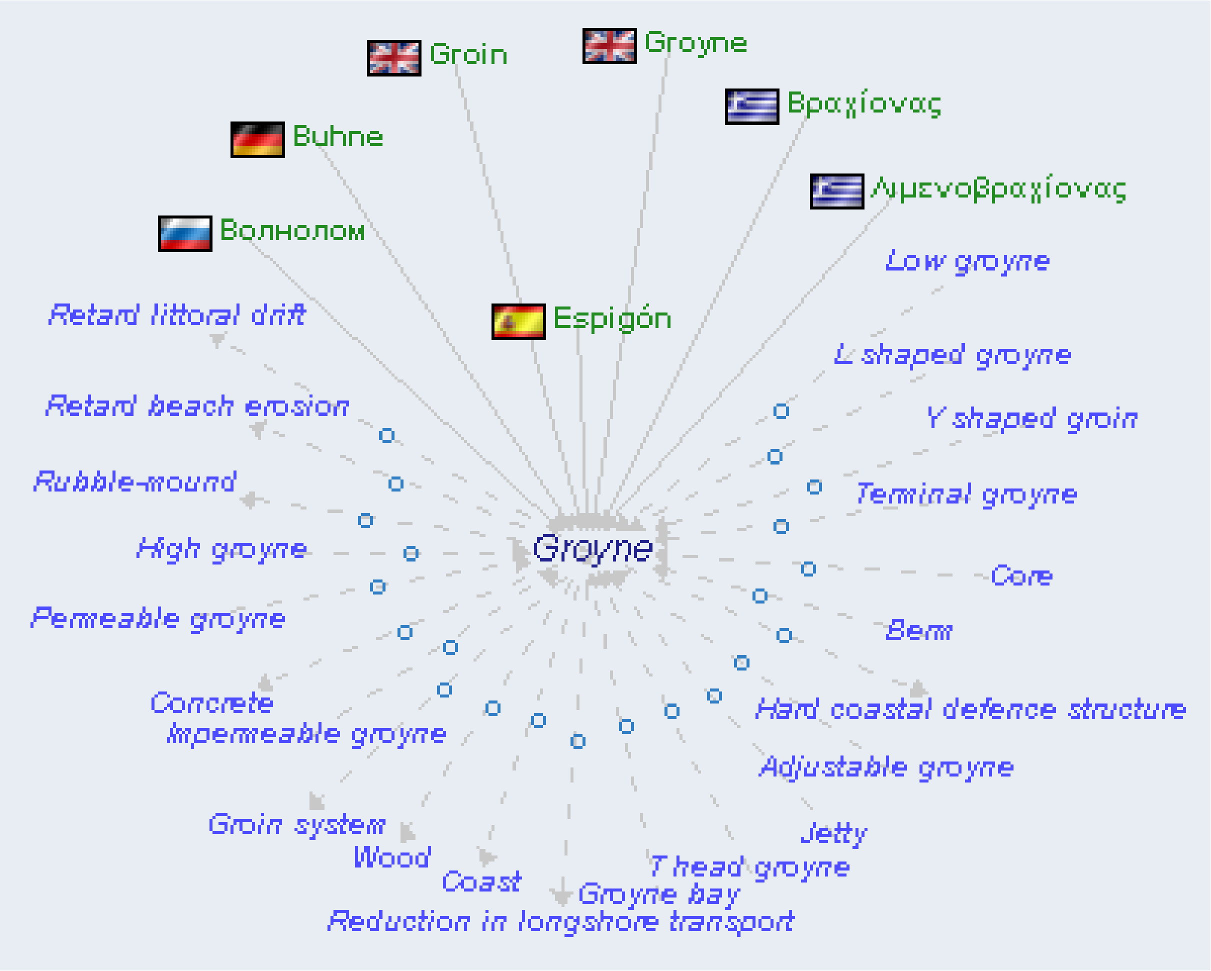
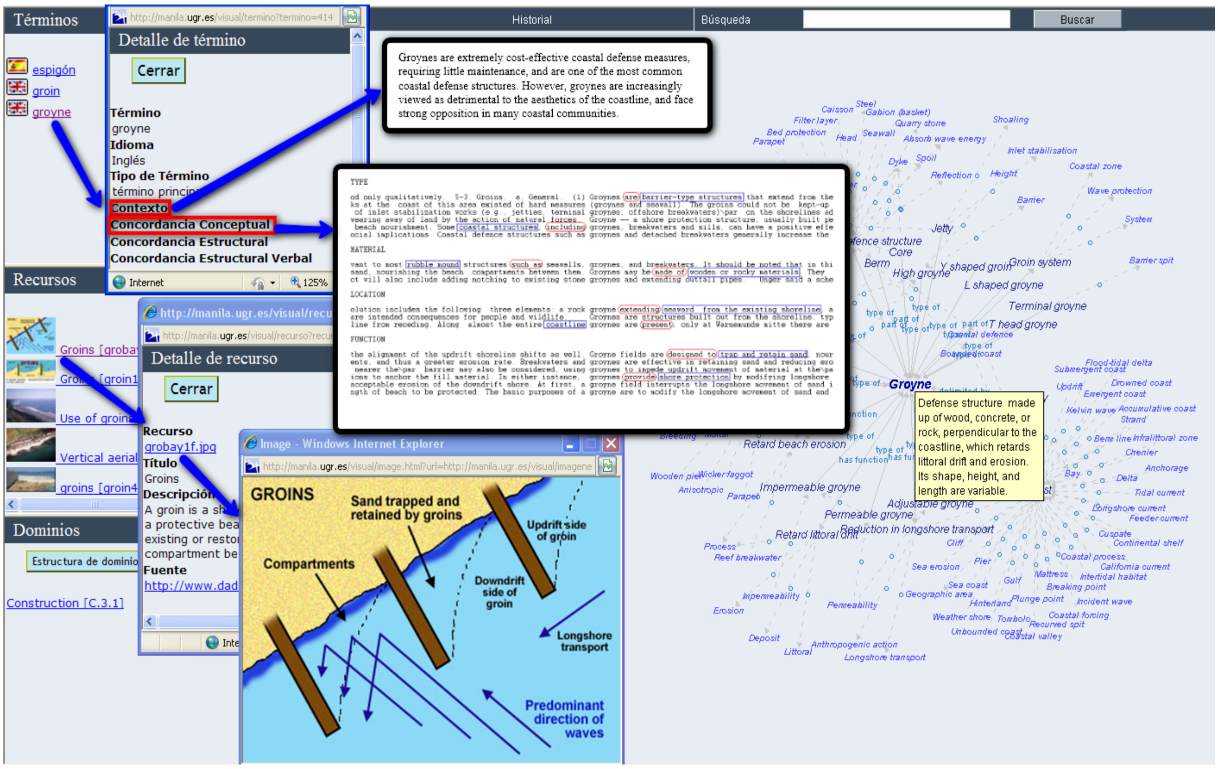


Figure 5. Multilingual terminological choices for GROYNE

In EcoLexicon, the visual representation of knowledge is thus two-fold. As shown in the groyne networks (Figures 2 and 5), our TKB provides users with an interface based on infographics. Infographics are widely used in statistics through charts, diagrams, tables, maps, and the like [*cf.* Newsom and Haynes 2004: 236] and are known to facilitate quick communication and create clear mental images of complex knowledge.

The user retrieves different configurations by surfing from one concept to another. Each configuration focuses on a two-level hierarchy of a central concept. Thus, these networks take a different shape every time a user clicks on a particular node. Therefore, each entry of EcoLexicon provides a great amount of interrelated information. In Figure 6, the entire GROYNE entry is shown. Users do not have to see all this information at the same time, but can browse through the different windows and resources according to their needs.

Under the tag 'Dominios' an ontological structure shows the exact position of the concept in the class hierarchy of the EE. GROYNE, for example, *is_a* CONSTRUCTION (bottom-left corner of the window). The concept definition is shown when the cursor is placed on the concept. Contexts (top window with black contour) and concordances (bottom window with black contour) appear when clicking on the terms and inform different users about both conceptual and linguistic aspects. Graphical resources are displayed when clicking on the links in the box 'Resources' (in the left-hand margin towards the middle), which are selected according to definitional information. At a more fine-grained level, conceptual relations are displayed in a dynamic network of related concepts (right-hand side of the window). The terminological units, under the tag 'Terms', designate the concept in English and Spanish: '*groyne*' and its variant '*groin*', and '*espigón*', respectively (top left-hand corner). In the next sections we focus on the contextualization of both dynamic networks and terminological units.



Término
groyne
Idioma
Inglés
Tipo de Término
término princip
Contexto
Concordancia Conceptual
Concordancia Estructural
Concordancia Estructural Verbal

Groynes are extremely cost-effective coastal defense measures, requiring little maintenance, and are one of the most common coastal defense structures. However, groynes are increasingly viewed as detrimental to the aesthetics of the coastline, and face strong opposition in many coastal communities.

```
TYPE

od only qualitatively.  5-3. Groins.  a. General.  (1) Groynes are barrier-type structures that extend from the
ks at the  coast of this area existed of hard measures (groynes and seawall). The groins could not be  kept-up
 of inlet stabilization works (e.g., jetties, terminal groynes, offshore breakwaters)\par  on the shorelines ad
wearing away of land by the action of natural forces.  Groyne -- a shore protection structure, usually built pe
 beach nourishment. Some coastal structures, including groynes, breakwaters and sills, can have a positive effe
ocial implications  Coastal defence structures such as groynes and detached breakwaters generally increase the

MATERIAL

vant to most rubble mound structures such as seawalls, groynes, and breakwaters. It should be noted that in thi
sand, nourishing the beach  compartments between them. Groynes may be made of wooden or rocky materials. They
ct will also include adding notching to existing stone groynes and extending outfall pipes.   Unger said a sche

LOCATION

olution includes the following  three elements: a rock groyne extending seaward  from the existing shoreline, a
are intended consequences for people and wildlife.     Groynes are structures built out from the shoreline. typ
line from receding. Along  almost the entire coastline groynes are present, only at Warnemunde mitte there are

FUNCTION

the alignment of the updrift shoreline shifts as well. Groyne fields are designed to trap and retain sand, nour
ents, and thus a greater erosion rate. Breakwaters and groynes are effective in retaining sand and reducing ero
 nearer the\par  barrier may also be considered, using groynes to impede updrift movement of material at the\pa
ions to anchor the fill material. In either instance,  groynes provide shore protection by modifying longshore
acceptable erosion of the downdrift shore. At first, a groyne field interrupts the longshore movement of sand i
ngth of beach to be protected. The basic purposes of a groyne are to modify the longshore movement of sand and
```

Recurso
grobay1f.jpg
Título
Groins
Descripción
A groin is a sh
a protective bea
existing or resto
compartment be
Fuente
http://www.dad

Defense structure made up of wood, concrete, or rock, perpendicular to the coastline, which retards littoral drift and erosion. Its shape, height, and length are variable.

Figure 6. The entry of GROYNE in the user interface of EcoLexicon

3. Conceptual contextualization

In knowledge modelling, concepts are very often classified according to very different dimensions (shape, function, colour, etc.). Multidimensionality [Kageura, 1997] is commonly regarded as a way of enriching traditional static representations, enhancing knowledge acquisition through different points of view in the same conceptual network [León Araúz and Faber, 2010]. As is well-known, the more relations that users are able to activate through a particular concept, the more knowledge they are likely to possess for the domain. In such a wide domain as the environment, multidimensionality increases the number of possible relations activated by specialized concepts, since it is also intimately linked to the semantic roles concepts may play. In a process-oriented domain [Faber et al., 2006] the same concept may act as an AGENT or a PATIENT, as an active PROCESS or a RESULT. For example, the concept WATER can be either an AGENT (in the process of EROSION) or a PATIENT (in WATER TREATMENT), which implies that WATER can be related to other concepts through the conceptual relation *causes* as well as *affected_by*. However, the environmental domain has caused a great deal of information overload, which ends up jeopardizing knowledge acquisition.

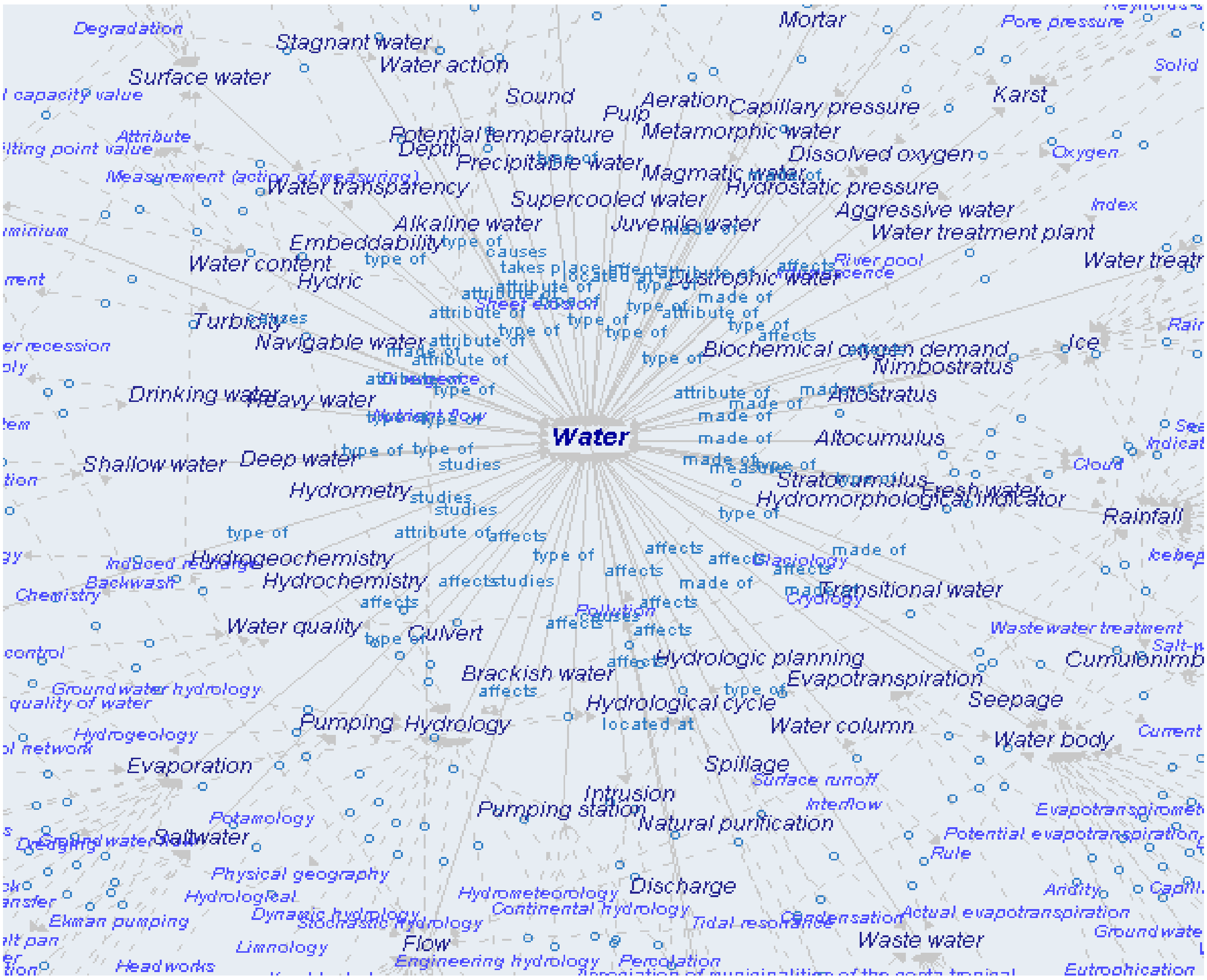


Figure 7. Information overload in the network of WATER

This is not only due to its wide scope, but especially to the fact that multiple dimensions are not always compatible but context-dependent. Although concepts are entrenched cognitive routines which are interrelated in various ways facilitating their co-activation, they actually retain enough autonomy that the execution of one does not necessarily entail the activation of all of the rest [Langacker, 1987: 162]. This is the case of certain concepts such as WATER (Figure 7). We call them versatile concepts because they have such a low degree of specificity that they can be involved in a myriad of events. For instance, even though WATER subtypes, such as PRECIPITABLE WATER, DRINKING WATER and NAVIGABLE WATER, all represent the same facet *function*, strictly speaking, they are not coordinate concepts, because they belong to different environmental paradigms that rarely coincide, if ever, in time or space. The same applies to WATER as an EROSION AGENT or as a PATIENT in a WATER TREATMENT PLANT.

Yeh and Barsalou [2006] state that when situations are not ignored, but incorporated into a cognitive task, processing becomes more tractable. In the same way, any specialized domain reflects different situations in which certain conceptual dimensions become more or less salient. As a result, a more believable representational system should account for reconceptualization according to the situated nature of concepts. Rather than being decontextualized and stable, conceptual representations should be dynamically contextualized to support diverse courses of goal pursuit [Barsalou, 2005: 628].

In EcoLexicon, overloaded concepts are reconceptualised according to two contextual factors: domain membership and semantic role. We have divided the environmental field in different contextual domains according to corpus information and expert collaboration: HYDROLOGY, GEOLOGY, METEOROLOGY, BIOLOGY, CHEMISTRY, ENGINEERING/CONSTRUCTION, WATER TREATMENT/SUPPLY, COASTAL PROCESSES and NAVIGATION. Domain membership restricts concepts’ relational behaviour according to how their referents interact in the real world.

On the other hand, semantic role reconceptualization is domain-independent and offers new networks in the form of upper-level conceptual classes. In this way, users can visualize how concepts like WATER behave either as an AGENT or a PATIENT in all kinds of events. This highlights certain relational constraints associated to the natural aspect of concepts and not to those of their referents. Thus, role-dependent networks will be characterized by a certain type of relations. Interestingly enough, hierarchical relations are invariable parameters [León Araúz and Faber, 2010]. Entities may *have parts* or be *part of* other wholes whether they are AGENTS or PATIENTS, but that is not the case for non-hierarchical relations. If an entity behaves like a PATIENT it cannot *affect* anything, as it would then become an AGENT. Prototypically, a PATIENT can only activate its inverse relation, *affected_by*.

3.1 Role-based constraints

Role-based relational constraints are applied to individual concepts according to their own perspective in a given proposition. For example, in WATER CYCLE *affects* WATER, WATER is a PATIENT. However, if a role-based domain was to be associated with WATER CYCLE, this would require the application of agent-based constraints. As mentioned before, role-based constraints apply for non-hierarchical relations. Hierarchical ones are always activated, whether concepts are AGENTS or PATIENTS. Moreover, this kind of constraints can only be applied to the first hierarchical level, since they are focused on a particular concept and not its whole conceptual proposition. In the next figures, the overloaded network of WATER (Figure 8) is restricted according to the AGENT role (Figure 9).

Actually, role-based domains by themselves are not sufficient to reconceptualize knowledge in a meaningful way. In the role-free network, WATER appears linked to 72 concepts, whereas in the role-based one, WATER is related to 50. Despite the difference, the concept still appears overloaded, especially once the second hierarchical level is displayed. However, contextual domains, although usually dominated by one role, restrict relational power of versatile concepts in a more quantitative way.

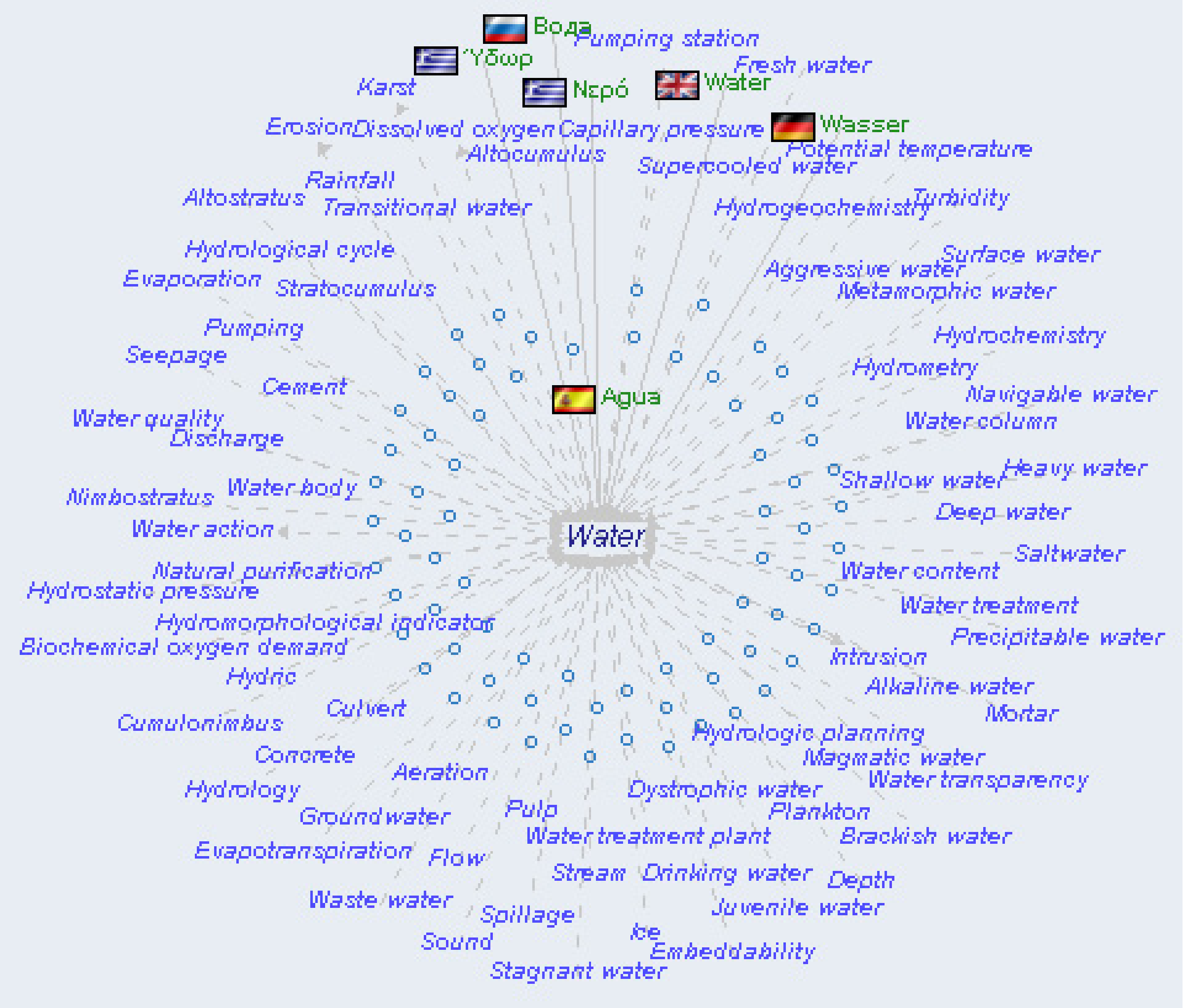


Figure 8. Role-free network of WATER

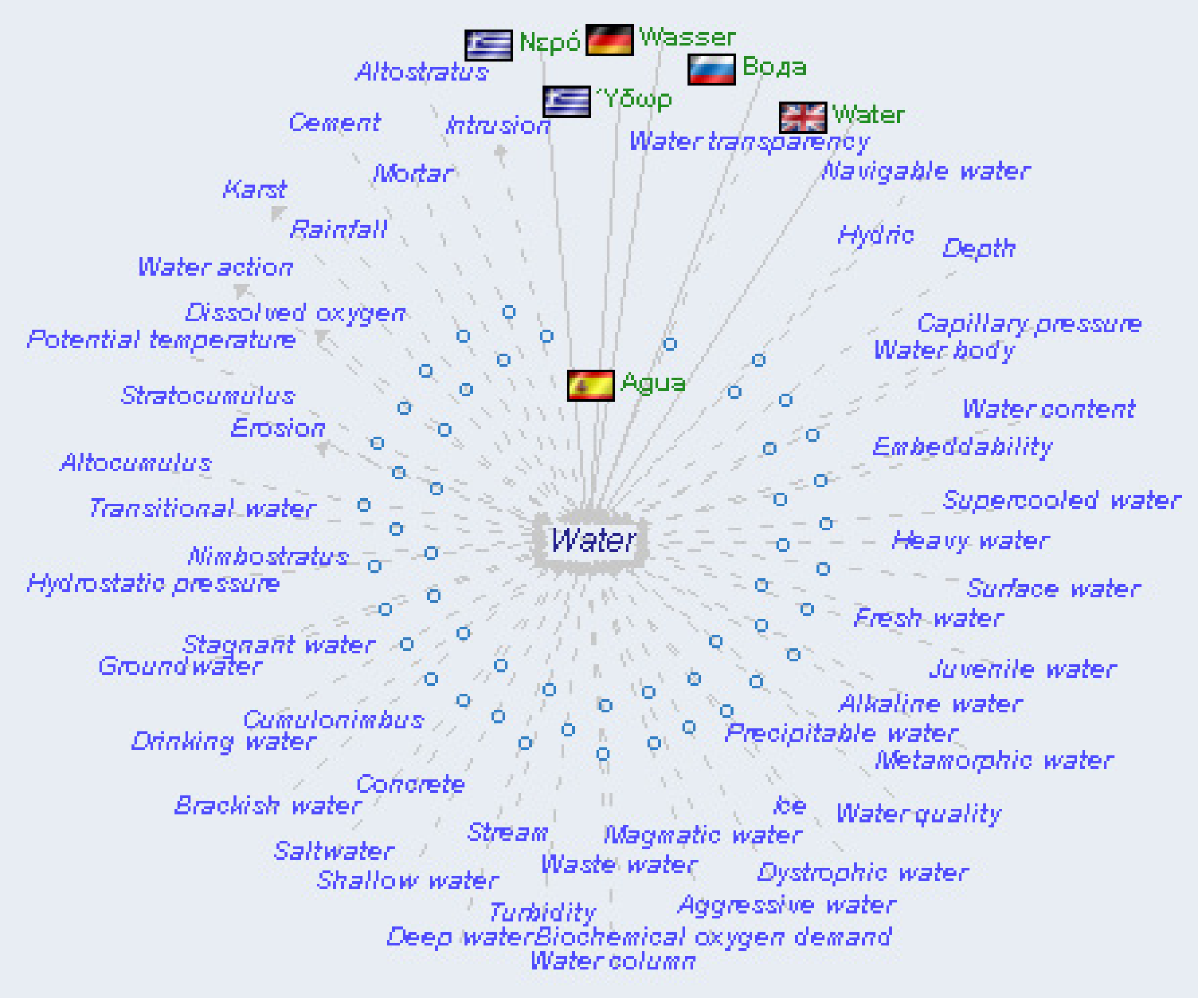


Figure 9. Agent-based network of WATER

3.2 Domain-based constraints

Our contextual domains have been allocated in a similar way as the General European Multilingual Environmental Thesaurus, whose structure is based on themes and descriptors, reflecting a systematic, category or discipline-oriented perspective [GEMET, 2004]. These domains can also be related to the notion of micro-theories, which are theories of some topic, e.g. a theory of mechanics, chemical elements, etc. Micro-theories might make different assumptions or simplifications about the world with contexts providing a mechanism for recording and reasoning using these assumptions [Guha 1991: 41].

In this way, our contextual domains provide the clues to simplify the background situations in which concepts can occur in reality. Domain-based constraints are neither applied to individual concepts nor to individual relations, since one concept can be activated in different contexts or use the same relations but with different values. Constraints are instead applied to conceptual propositions [León Araúz et al., 2009]. For instance, CONCRETE is linked to WATER through a *made_of* relation, but this proposition is irrelevant if users only want to know how

WATER naturally interacts with the landscape or how it is purified from contaminants. Consequently, the proposition CONCRETE *made_of* WATER will only appear in an ENGINEERING/CONSTRUCTION context. As a result, when constraints are applied, WATER only shows relevant dimensions for each contextual domain. In Figure 10 WATER is only linked to propositions belonging to the context of ENGINEERING/CONSTRUCTION:

However, in Figure 11 the GEOLOGY context shows WATER in a new structure with other concepts and relations:

The number of conceptual relations changes from one network to another, as WATER is not equally relevant in all contextual domains. Furthermore, relation types differ too, which also highlights the changing nature of WATER’s internal structure in each case. For example, in the ENGINEERING/CONSTRUCTION context domain, most relations are *made_of* and *affects*, whereas in the GEOLOGY domain, *causes* and *type_of* stand out. *Affects* is also shared by the GEOLOGY domain, but the arrow direction shows a different perspective: in geological contexts WATER is a much more active AGENT than in ENGINEERING/CONSTRUCTION, where the concept is more subject to changes (PATIENT). Finally, WATER is not always related to the same concept types. In ENGINEERING/CONSTRUCTION, WATER is only linked to artificial entities or processes (PUMPING, CONCRETE, CULVERT), while in GEOLOGY it is primarily related to natural ones (EROSION, GROUNDWATER, SEEPAGE).

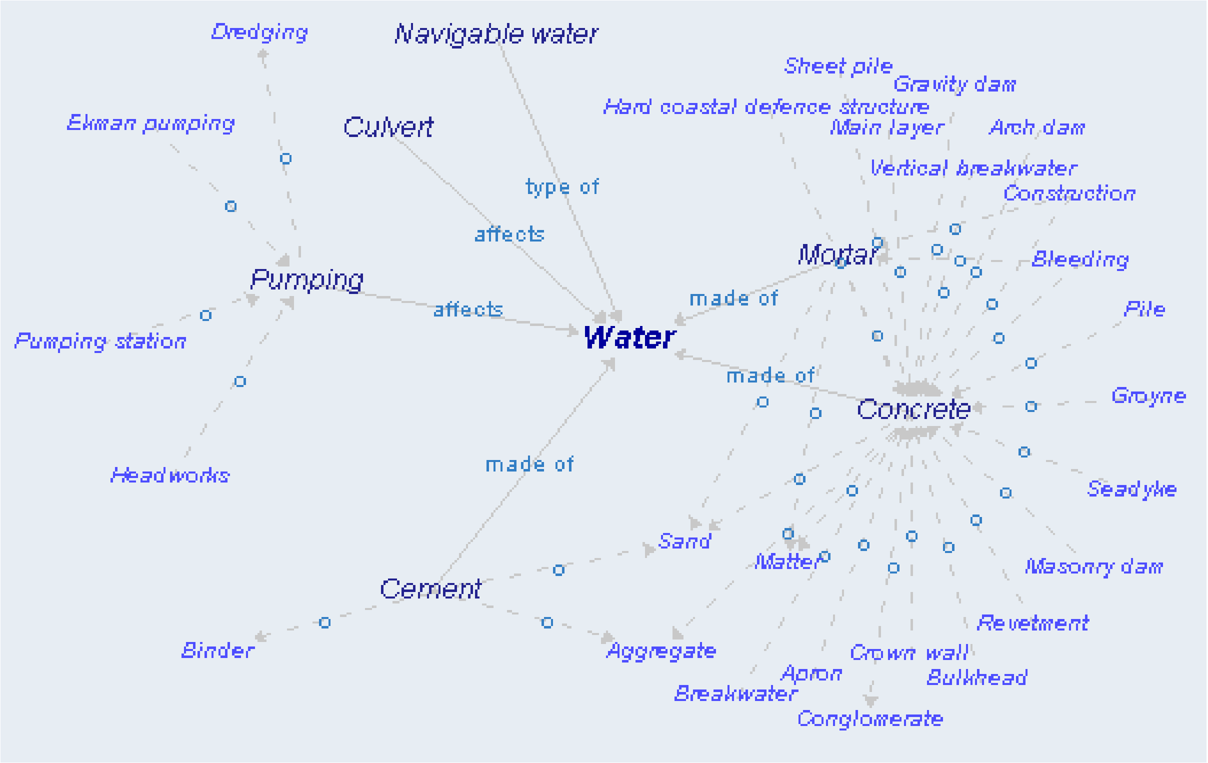


Figure 10. WATER in the ENGINEERING/CONSTRUCTION contextual domain

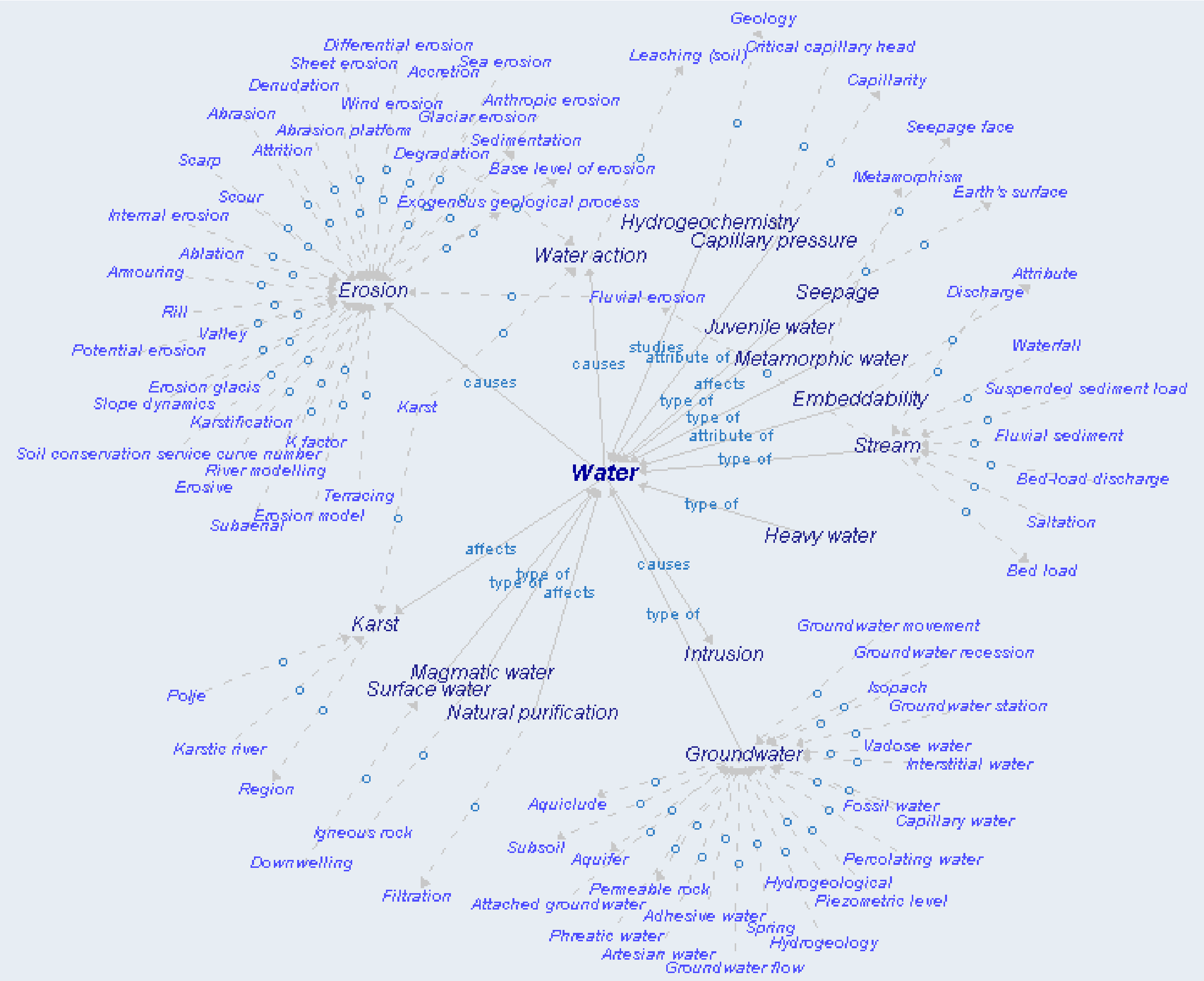


Figure 11. WATER in the GEOLOGY contextual domain

3.3 Intersection of role- and domain-based constraints

A new reconceptualization can take place with the intersection of role- and domain-based constraints. For example, WATER can be framed as an AGENT (Figure 12) or a PATIENT (Figure 13) or even both (Figure 14) within the HYDROLOGY context.

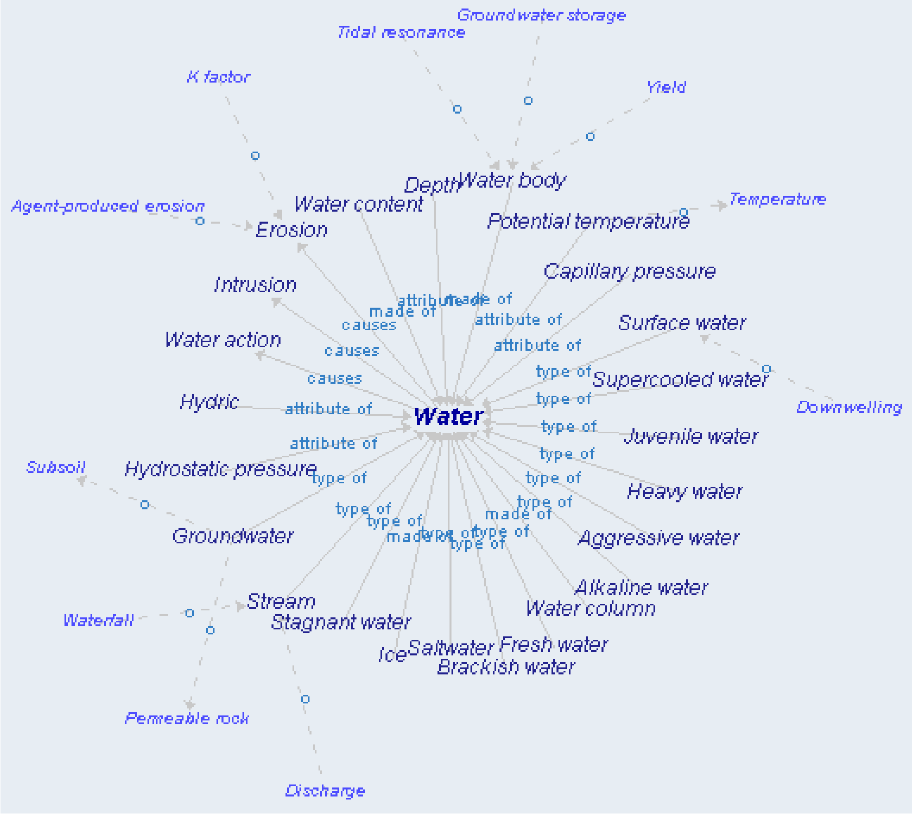


Figure 12. WATER as an AGENT in HYDROLOGY

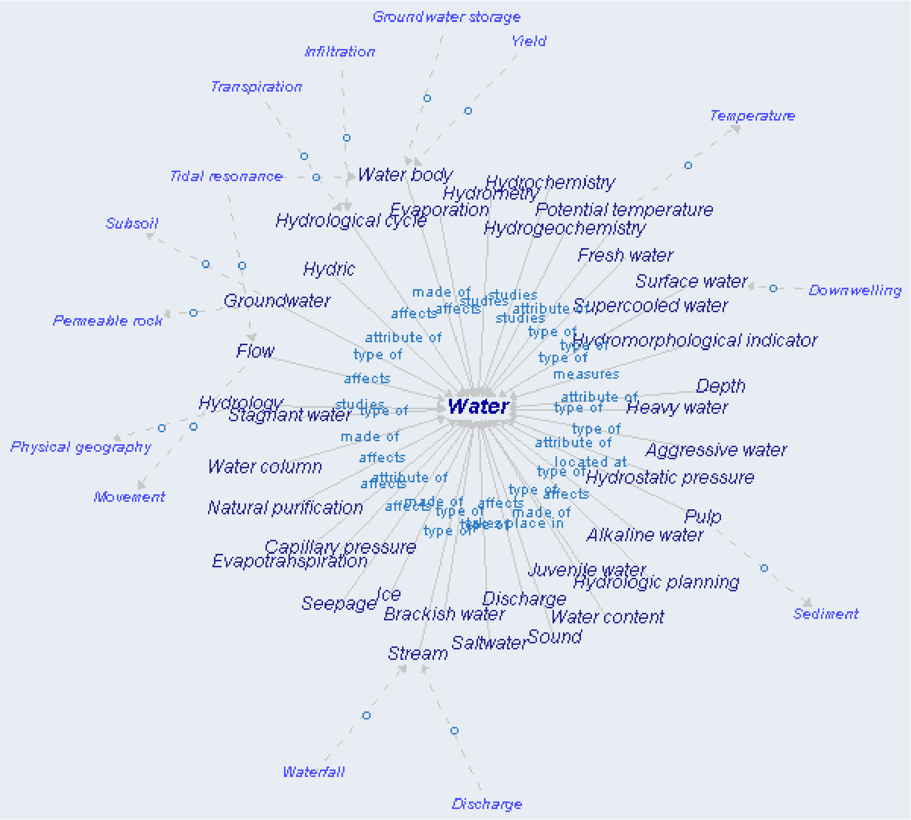


Figure 13. WATER as a PATIENT in HYDROLOGY

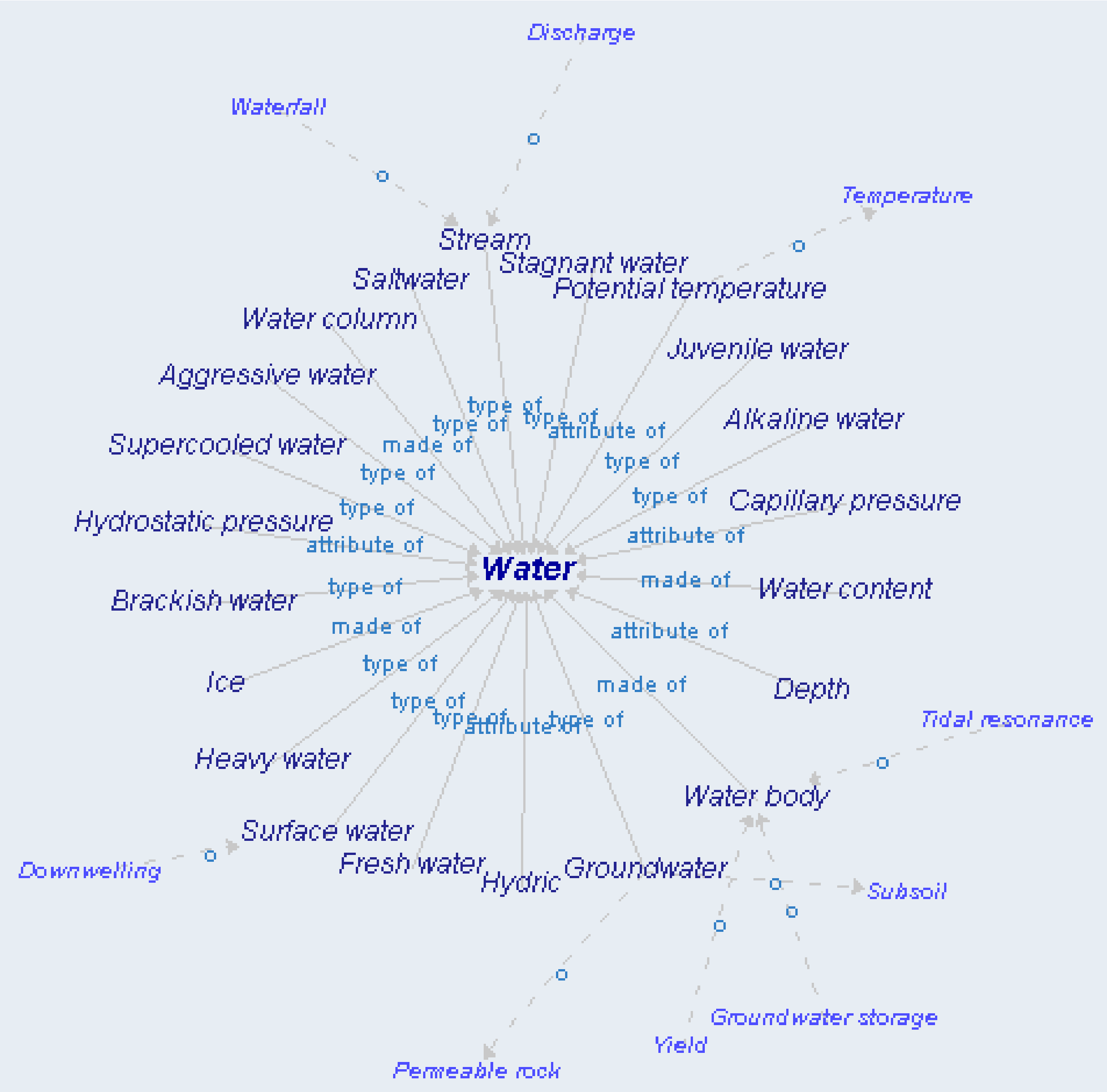


Figure 14. WATER as an AGENT and PATIENT in HYDROLOGY

Now, the first level appears constrained according to different roles in a particular contextual domain, which at the same time applies for the second level. It is worth noting that Figure 14 only shows hierarchical relations (*type_of, attribute_of*, *made_of*), because these are the only ones shared by concepts that can be AGENTS or PATIENTS. In Figure 12, however, the representation adds the relation *causes*, typical of AGENTS, and in Figure 13, it adds propositions where WATER is *affected_by*, *measured*, *studied* or *located_at*.

4. Terminological contextualization

The environmental domain is a recent and multidisciplinary science where conceptual overlapping can easily arise. Different disciplines deal with the same subject in different terms and, consequently, meaning and conceptual networks can vary. However, due to the lack of univocity in specialized languages, contextualization is also needed at the linguistic level. In this case, concepts may receive different designations and still be the same concept, which means that there will be no contextualization at the conceptual level but linguistic choices will be determined by different factors, such as contextual domains and user type or situation. For example, a term may be activated in several contextual domains – or even in the same domain – but it may refer to different concepts, thus creating the well-known phenomenon of polysemy. Also, different terms can express the same concept in the same or several contextual domains (synonymy). However, synonymy is only apparent, since terms should be selected according to domains and usage. Terminological contextualization is thus based on a more pragmatic perspective, whereas conceptual contextualization stems from a more cognitive point of view.

4.1 Domain-based constraints

The concept MUD may be referred to as *mud* or *sludge*. It is the same concept because it has exactly the same composition, colour, etc. At first sight they could be regarded as simple synonyms. In many dictionaries they are given as equivalents and are described in circular definitions. However, only in a WATER TREATMENT discourse *sludge* prevails over *mud*. As a result, with this other type of contextualization, users manage to cover both cognitive and communicative needs. Not only does the concept MUD show a different conceptualization in a WATER TREATMENT domain, but also a different designation.

Another example can be found in compound names, such as *beach erosion*. EROSION has different subtypes according to different dimensions, depending on its AGENT (wind erosion, water erosion, etc.), its PATIENT/LOCATION (beach erosion, soil erosion) or the RESULT it may cause (sheet erosion, rill erosion). Some of these dimensions can be overridden according to contextual domains. For example, if users searched for EROSION in the contextual domain of COASTAL PROCESSES, they would get the prototypical designation of *beach erosion*, which focalizes the potential meaning of EROSION.

4.2 Usage-based constraints

Other term preferences related to contextual domains are at the same time linked to use situations. For instance, *H_2O* and/or *water* are more or less frequently used depending on the domain. Both of the terms are used in ENGINEERING/CONSTRUCTION and WATER TREATMENT, whereas *H_2O* is never used in the GEOLOGY domain and is preferred over *water* in the CHEMISTRY domain.

Register thus plays an important role and triggers different variants e.g., geographical variants, such as *groyne* (British English) and *groin* (American English); and diaphasic variants, such as the continuum from formal to informal in *thermal low pressure system*, *thermal low*, *thermal trough* and *heat low*. These intralinguistic differentiations are extremely interesting for any kind of user group. Experts may be more interested in a more formal discourse where *thermal trough* is more likely to appear, but in the case of addressing lay receivers, they might prefer to use *heat low*. The same happens with translators, whose terminological choice will vary according to the text type they need to produce.

4.3 Intersection of domain- and usage-based constraints

The intersection of domain- and usage-based constraints in EcoLexicon is reflected through what we call Textual Contexts (TCs), which are carefully selected from real texts.

In Table 3, TCs have been assigned to different contextual domains according to the way in which *water* appears related to other domain-specific concepts (in bold). Users can appreciate the changing behaviour of WATER across three different domains (ENGINEERING/CONSTRUCTION, GEOLOGY, WATER TREATMENT) in two ways: (1) the specialized concepts surrounding water change from one domain to another and (2) the designation of a concept may also change accordingly. These relations become explicit through linguistic markers such as *consists of*, *causes*, etc. (in italics).

Table 3. *Water* TCs in ENGINEERING/CONSTRUCTION, GEOLOGY and WATER TREATMENT

TEXTUAL CONTEXT IN ENGINEERING/CONSTRUCTION	
1	Fine aggregate **concrete** *consists of* a mixture of Portland cement, fine aggregate (sand) and **water**, so proportioned and mixed as to provide a pumpable fine aggregate **concrete**. Fine aggregate **concrete** *has a* typical mix **water**/cement ratio of 0.65 to 0.75.
2	The heat evolution of **cement** A at w/c = 0.40 was measured *using* thermal calorimetry at 30ºC for hydration with both **H_2O** and D_2O.
TEXTUAL CONTEXT IN GEOLOGY	
3	Sometimes, as layers of **rock** are steeply uplifted, the bonding of one layer to another may be *weakened by the action of* **water** or *other agents of* **erosion**.
4	Running **water** *causes* **erosion** of **soil** and **rocks**. This is done by the friction between the constant movement of the **water** and the still **rock** or **soil**. Soil is *washed away* by running **water**.

TEXTUAL CONTEXT IN WATER TREATMENT	
5	Some of this clear **water** is decanted *from* the **tank** *into* an **effluent lagoon** where it *undergoes* **UV disinfection**, while some of the settled **sludge** is pumped to **sludge ponds**. Water exiting the treatment process at the downstream end of the **effluent lagoon** is pumped to storage, and *reused for* **irrigation**.
6	Under normal conditions the half-lifetime of the direct reaction of **chlorine** on **chlorite** has a value of 10 -^{5} s. When dissolved in **water**, clorine is **hydrolyzed** according to the reaction: Cl_2 + **H_2O** = HOCl + HCl

Use situations are reflected providing a different TC for each register within each domain. In the case of TC2 or TC6, the use of *H_2O* clearly implies a specialised register. In contrast, more didactic TCs, such as TC1, TC3 and TC5, show more explicit information thanks to a higher density of linguistic markers.

5. Conclusions

As the Environment is a recent and multidisciplinary field of study, building a coherent TKB implies a conscious and continuous effort to provide the changing and dynamic nature of the field to the TKB users, while at the same time helping them to extract exactly the kind of information they are looking for. EcoLexicon provides this to the end users by coherently organizing the information at all levels in a domain event and applying definitional templates, whereas at the same time it provides the necessary dynamicity through conceptual and terminological recontextualization.

Recontextualization provides a way of representing the dynamic and multidimensional nature of concepts and terms. On the one hand, conceptual contextualization offers a qualitative criterion for the representation of specialized concepts in line with the workings of the human conceptual system. Moreover, it is a quantitative solution to the problem of information overload, as it significantly reduces irrelevant context-free information. On the other hand, terminological contextualization guides users in selecting the most adequate term for each discourse, according to contextual domains and use situation.

Bibliography

[1962] Austin, John L. How to do things with words. Clarendon, Oxford, 1962.

[2005] Barsalou, L.W. Situated conceptualization. In H. Cohen. & C. Lefebvre. Eds. Handbook of Categorization in Cognitive Science p. 619-650. St. Louis, 2005.

[2009] Barsalou, L.W. Simulation, situated conceptualization and prediction. *Philosophical Transactions of the Royal Society of London*: Biological Sciences, 364: 1281-1289, 2009.

[in press] Evans, V. Cognitive linguistics. In Cummings, L. (ed.), *Encyclopedia of pragmatics*, in press. Available at: http://www.vyvevans.net/cognitiveLinguisticsPRAG-ENCYC.pdf

[2007] Faber, P., León Araúz, P., Prieto Velasco, J.A. and Reimerink, A. "Linking Images and Words: the description of specialized concepts (extended version)". International Journal of Lexicography 20:1, 39-65, 2007.

[2006] Faber, P., Montero Martínez, S., Castro Prieto, M.C., Senso Ruiz, J., Prieto Velasco, J.A., León Araúz, P., Márquez Linares, C.F. and Vega Expósito, M. Process-oriented terminology management in the domain of Coastal Engineering. *Terminology* 12: 2, 189-213, 2006.

[1995] Gadamer, Hans G. *Truth and Method*. Continuum, New York, 1995.

[2004] GEMET. "About GEMET. General Multilingual Environmental Thesaurus", 2004. Available at: http://www.eionet.europa.eu/gemet/about

[1991] Guha R. V. Contexts: A Formalization and Some Applications, Stanford PhD Thesis, 1991.

[2006] House, Juliane. Text and context in translation. *Journal of Pragmatics* 38: 338-358, 2006.

[1997] Kageura, K. Multifaceted/Multidimensional concept systems. In Wright, S.E. and Budin, G. (eds.), *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam/Philadelphia: John Benjamins. 119-32, 1997.

[1987] Langacker, R. W. *Foundations of cognitive grammar: theoretical prerequisites*, Vol 1. Stanford: Stanford University Press, 1987.

[2010] León Araúz, P. and Faber, P. Natural and contextual constraints for domain-specific relations. *Proceedings of Semantic relations. Theory and Applications*. Malta, 2010.

[2009a] León Araúz, P., A. Reimerink, P. Faber. "PuertoTerm and MarcoCosta: a frame-based knowledge base for the environmental domain". *Journal of Multicultural Research*. 1 (1), 47-70, 2009a.

[2009b] León Araúz, P., P. Magaña Redondo, and P. Faber. "Managing inner and outer overinformation in Ecolexicon: an environmental ontology". In *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*. Toulouse, France, 2009b.

[1992] Meyer, I., Bowker, L. and Eck, K. COGNITERM: An experiment in building a knowledge-based term bank, *Proceedings of Euralex* '92, 159–172, 1992.

[2004] Newsom, D. and J. Haynes. *Public Relations Writing: Form and Style* (7th ed.). Belmont, CA: Wadsworth, 236, 2004.

[2010] Reimerink, A., García de Quesada, M., Montero Martínez, S. Contextual information in terminological bases: a multimodal approach. *Journal of pragmatics*. 42-7, 2010.

[2009] Reimerink, A. and P. Faber. "A frame-based knowledge base for the environment". *Proceedings of Towards e-Environment*. Prague, 2009.

[1986] Sperber, Dan; Wilson, Deirdre. *Relevance: Communication and Cognition*. Mitt Press, Cambridge MA (2nd edition 1995. Blackwell, Oxford), 1986.

[1995] Sperber, Dan; Wilson, Deirdre. Postface to the second edition of *Relevance: Communication and Cognition*. Blackwell, Oxford, 1995.

[2006] Yeh, W. and Barsalou, L.W. The situated nature of concepts. *American Journal of Psychology* 119, 349-384, 2006.

Authors' information

Pilar León Araúz *– Junior Professor (Profesor Ayudante Doctor), Department of Translation and Interpreting, University of Granada, Calle Buensuceso, 11, 18002 Granada (Spain); e-mail:* pleon@ugr.es

Major Fields of Scientific Research: Terminology, Cognitive Linguistics, Knowledge Representation, and Knowledge Extraction.

Arianne Reimerink *– Junior Professor (Profesor Ayudante Doctor), Department of Translation and Interpreting, University of Granada, Calle Buensuceso, 11, 18002 Granada (Spain); e-mail:* arianne@ugr.es

Major Fields of Scientific Research: Terminology, Knowledge Representation, Cognitive Semantics, and Translation.

Alejandro G. Aragón *– PhD Student / Research Fellow (Becario FPU), Department of Translation and Interpreting, University of Granada, Calle Buensuceso, 11, 18002 Granada (Spain); e-mail:* aga@ugr.es

Major Fields of Scientific Research: Lexicography, Terminology, Specialized Translation, Knowledge Representation, and Modern Greek Studies.

COGNITIVE MODEL OF TIME AND ANALYSIS OF NATURAL LANGUAGE TEXTS

Xenia A. Naidenova, Marina I. Garina.

Abstract: *The extension to new languages is a well known bottleneck for any text analyzing system. In this paper, a cognitive model of time is proposed and the questions of extracting events and their time characteristics from texts are discussed. The cognitive model of time due to its independence of concrete natural language can be considered as a basis for constructing text mining systems intended for extracting temporary relations.*

Keywords: *Natural Language Processing, cognitive model, time model.*

ACM Classification Keywords: *I.2.7. Computing Methodologies - Artificial intelligence - Natural Language Processing.*

Introduction

Time representation and reasoning is an issue of many different disciplines. Studies in this area exploit several sources: cognition, language, perception as well as world knowledge and the difficulties of these studies are explained by the inherent complexity and multidimensionality of time [Elkin, 2008] as a human thinking category. Modern psycho-linguistic and neuro-linguistic investigations show that mechanisms of thinking and mechanisms of thinking verbalization are different from one another [Popova, & Sternin,2007]. But human minds have the ability to establish systematic relationships between linguistic forms and perceptually based knowledge. Kubrjakova E.S. [2004] has formulated a new cognitive – discursive investigation direction, insisting on the thesis that each language phenomenon can be adequately described and explained only if it has been studied in the framework of cognition and communication. A goal of cognitive linguistic is to find adequate cognitive construction for every language form. Our perspective aim is creating a cognitive model of time. This model being translated in different natural languages could serve as a basis for text mining and extracting information of temporary attributes of events. Presumably, the text processing system consists of:

- **Cognitive models of time** and events that oriented to a given domain application and the goals of text processing;
- **Translator** that is adjusted to a given NL;
- **Block of plausible (commonsense) reasoning** to infer consequences from established temporary relations between events in the text by means of meta-knowledge of cognitive models;
- **Dialogue Syntactical Analyzer** for a given NL;

- **Block of control** or operational subsystem of the translator.

According to Figure 1, Text interacts with Translator and Syntactical Analyzer; Translator interacts with Syntactical Analyzer and Cognitive Models of Time and Event. As a result, events and their time moments or intervals are extracted from Text, and then the conclusions about temporary relations between events are done.

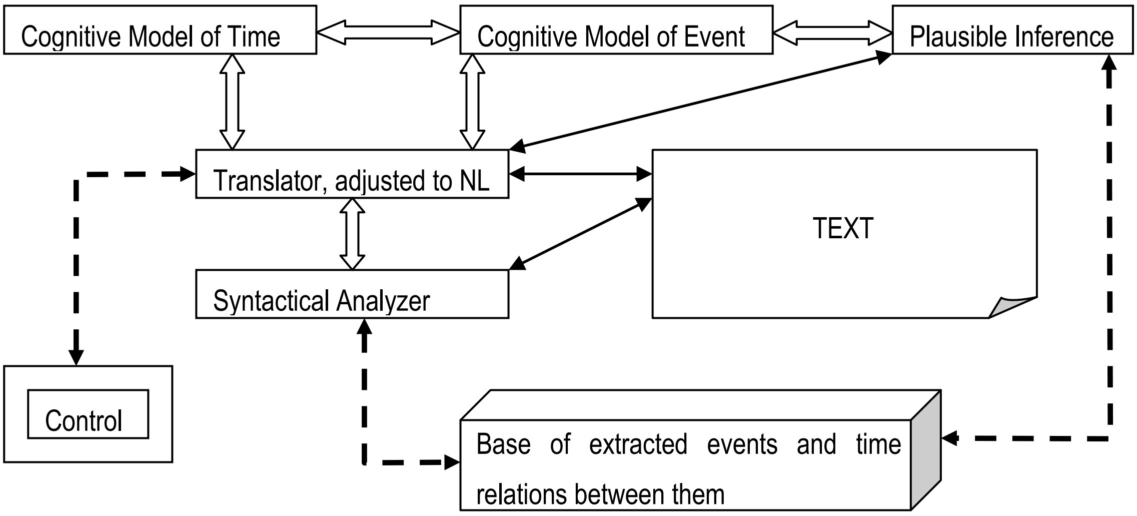


Figure – 1. Approximate structure of the text processing system.

State-of-the-art

The volume of the literary sources related to text mining temporary information about events is enormous. All sources can be divided into two groups: 1) the works developing the logical theory of time or logic of time [Moszkowsky, 2007]; and 2) the works, connected with extracting information about events and their time characteristics from the texts in different natural languages [Boguraev et al., 2006]. Both these directions have some limitations. The former is occupied by the problems of inferring consequences from the facts, already extracted from tests. The latter does not yet have a general platform for representing knowledge about the world and lingual structures, at least in limits of one of the lingual groups. Cognitive simulation is a "bridge", a connecting link, which is necessary for understanding the principles of interaction between knowledge, reasoning, and lingual abilities.

Model of time

The main purpose of time model and applying it to some text is to reveal **events** appearing in this text. Event as a concept can have the name and some other properties (may be empty) and, as a rule, it is associated with **time interval**.

In general, a time interval consists of two markers: (the beginning; the end) and has the duration. But sometimes we need to use intervals opened in the past (without beginning) or in the future (without end) or both.

Time interval can be expressed via some events, for example, "at dawn, to the first volleys of artillery", "long before the first sun rays". Events are used as time markers in this case. Also particular cases of time interval are: unit of time, a set of units of time, a moment of time. It is worth mentioning that the very moment of time can be an event, for example, "September began", "Days go".

The cognitive model of time includes the following elements:

- the **units** of time (year, month, spring, minute);
- the **time intervals** and their properties: the beginning, the end, duration, without beginning (opened in the past), without end (opened in the future), consisting of points;
- **environment**: nearest past, nearest future (about noon, soon after the beginning/the end, toward the evening);
- **various relations** between units and intervals: coincidence, contact, precedence, going after, intersection, inclusion, remoteness into the past/future time;
- **degree of relations**: the measures of remoteness, intersection and so on;
- **comparison relations of interval duration**: longer, less for long, shorter and so on;
- **uncertain (fuzzy) relations**: considerably later, once, early in the morning and so on.

Many references to dates or times in a text are not fully specified, with the result that some parts of extracted knowledge about time and events will have to be computed from the context during the interpretation stage.

In Figure 2, the basic cognitive model of time and the relation between its elements considered above is represented. Obviously, this model is recursive, i.e., an event is associated with a certain time interval, and a time interval can be expressed via some events (look above).

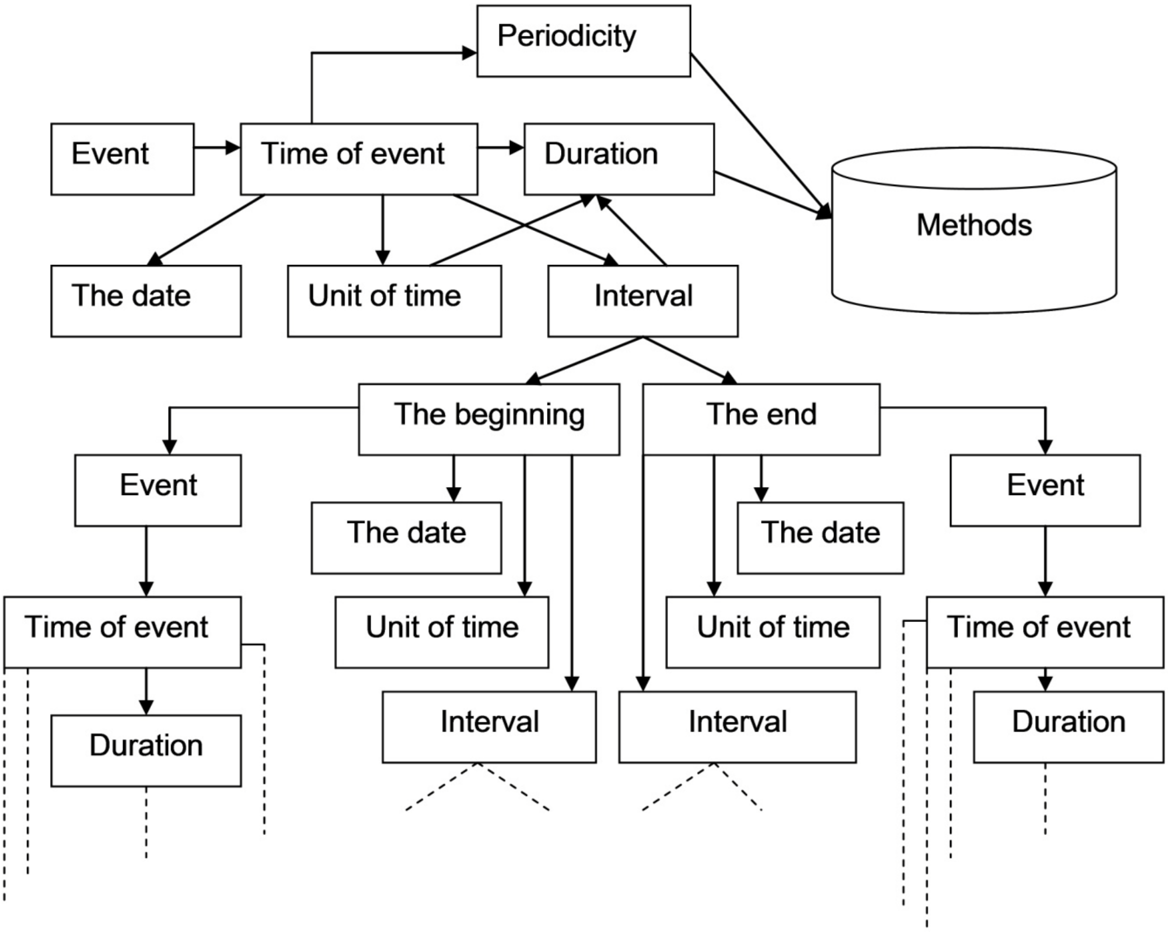


Figure - 2. The Basic Cognitive Model of Time

Methods

At first, it is important to be able to calculate the duration of time interval associated with some event. This duration can be calculated with the following **computation rules**:

- as the difference between the end and the beginning of time interval associated with the precise time markers (the dates);
- via the events being the markers of the beginning and the end of a time interval: <EVENT1> to <EVENT2>;
- as a set of time units (for example, 900 days).
-

Further, it is necessary to demonstrate the truthfulness of different enumerated above **relations between time intervals.** The most useful relations are listed below.

- **Sequence** (which event is earlier, which is later, which will be the next, which already happens). At best we concern a strict or weak order relation. Events in this relation can **contact** (in this case we have null duration between events).
- **Simultaneity** with it's particular cases:
  - the same time interval of events,
  - inclusion one interval into another,
  - intersection of intervals.

So the methods we discuss are natural rules for analyzing the relationships between time units/intervals. They include both **computations** and implicative assertions of the general kind:

$$E1.T.end < E2.T.begin \leftrightarrow E1\ R_{precedence}\ E2.$$

Here E1 and E2 are events, E.T is the interval associated with event E, T.end and T.begin are the end and the beginning of interval T, respectively, and $R_{precedence}$ is the precedence relation. In object-oriented design terms here we deal with access to fields of objects of classes Event and Time_Interval. Thus it's important to watch for the following conditions' performance:

- $T.begin \le T.end$ or $T.end \ge T.begin$ with the precise time markers;
- T.begin $R_{precedence}$ T.end if events are used as markers.

Next, it may be useful to detect some properties of events appearing in the text, such as:

- Fuzziness. Events can be indeterminate (fuzzy) in the time (considerably later, once upon a time).
- Speed. Events can flow in the time rapidly or slowly.
- Frequency. Periodic events can be frequent or rare.
- The temporary properties of events can be estimated in both objective and subjective manner.

A set of events can be associated with only one time interval, so those events can be indiscernible that will lead to a weak order relation. An event can be expressed both by only one word and by a set of proposals (maybe only one proposal). So there should be a method to parse complex expressions, especially in view of an event and the time interval associated with it can be in different proposals. It is possible that an active agent (including temporary moment) cannot be determined without the aid of referential relation. It is also necessary to take advantage that there are the events attached by default to the time intervals, such as dawn, sunset, school-leaving ball, dinner, supper, breakfast, the beginning of workday and so on.

Constructing a translator

The cognitive model of time does not depend on language, but is tuned into different natural languages. A translator of the cognitive model of time into language expressions for a given natural language and vice versa can be built as a trained system that learns by specially constructed phrases. For this goal, the following levels of natural language are considered: lexical, morphological, and syntactic ones.

Let’s discuss the **lexical level** now. A special type of time interval is the name of time unit, for example: *TI* = {century, year, month, twenty-four hours, the morning, day, evening, night, January, February, March, April, May, June, minute, second, winter, summer…}. There are some banal relations between these units, such as:

- Classification (“is-a”), for example: “seasons are winter, spring, summer, and autumn”;
- composition (“consist-of”), for example: “twenty-four hours consist of night and day”;
- part-whole, for example: “minute is a part of hour”;
- occurring in cycles, for example: “winter of one year follows after autumn of previous year”;
- inclusion;
- sequence, for example “spring comes after winter”.

In Figure 3, the relation of classification is shown with the aid of triangle connections while the relation of composition is shown by simple arrow. If the relation of composition is determined between the intervals of upper level, then it is determined between the interval-descendants, for example, June consists of twenty-four hours. Specific dimensionality can be determined only for the connections of the lower level. It cannot be said how many twenty-four hours year generally consists of, month generally consists of, but it can be said, how many twenty-four hours the leap year consists of, current year consists of, January consists of, etc.

A generalized model of event can be defined as follows: EVENT = $\langle E, R_e, Pat_e\rangle$, where E is a set of classes of events, R_e is the relationship: $R_e \subseteq E \times E$, $R_e = R_{class} \cup R_{comp}$, i.e., events are also organized into the hierarchy of classification “is- a” (R_{class}) and composition “consist-of” (R_{comp}). Pat_e is a set of the regular expressions (patterns, templates), which make it possible to take out the text candidates into the exemplars of events of each class. Each candidate can have some parameters (contextual properties), according to which the relevancy of candidate can be evaluated. After establishing a certain threshold value, it is possible to select only the candidates with high probability of being some events.

The search for the beginning and the end of an event can also be achieved with the aid of templates.

Knowing time intervals, associated with events, it is possible, being guided by the ordering relations, determined for the intervals, to establish the same relations for events too.

At lexical level, it is possible to take advantage of the special words, helping to reveal events and their environment. For example, in [Elkin et al., 2008], the groups of words reflecting so-called event-related time are given:

- multiple repetition of one and the same event can be revealed with such words as "daily, every week, quarterly, monthly, and yearly";
- single event can be revealed with the words "once only, one time, once";
- such word as "momentary, prolonged" can help to distinguish long and short duration of an event;
- the time before an event is accompanied by the following words: before, in advance, in good time, previously, before the appointed time, it is preliminary, it is premature, on the threshold of, it is earlier than, long before, thus far not, not in a long time, recently, as long as, the day before;
- the time after the event is accompanied by the following words: later, afterward, it is later, then, after, hence, hence-forward, forth, in future, from now on, after all, immediately afterward, further, when;
- these words establish event-related time or time attached to a concrete event: in one's life, from birth, originally, while;
- affirmation time is revealed with the words: sometimes, someday, in the course of time, then, once;
- "negative" time is revealed with the words : never.

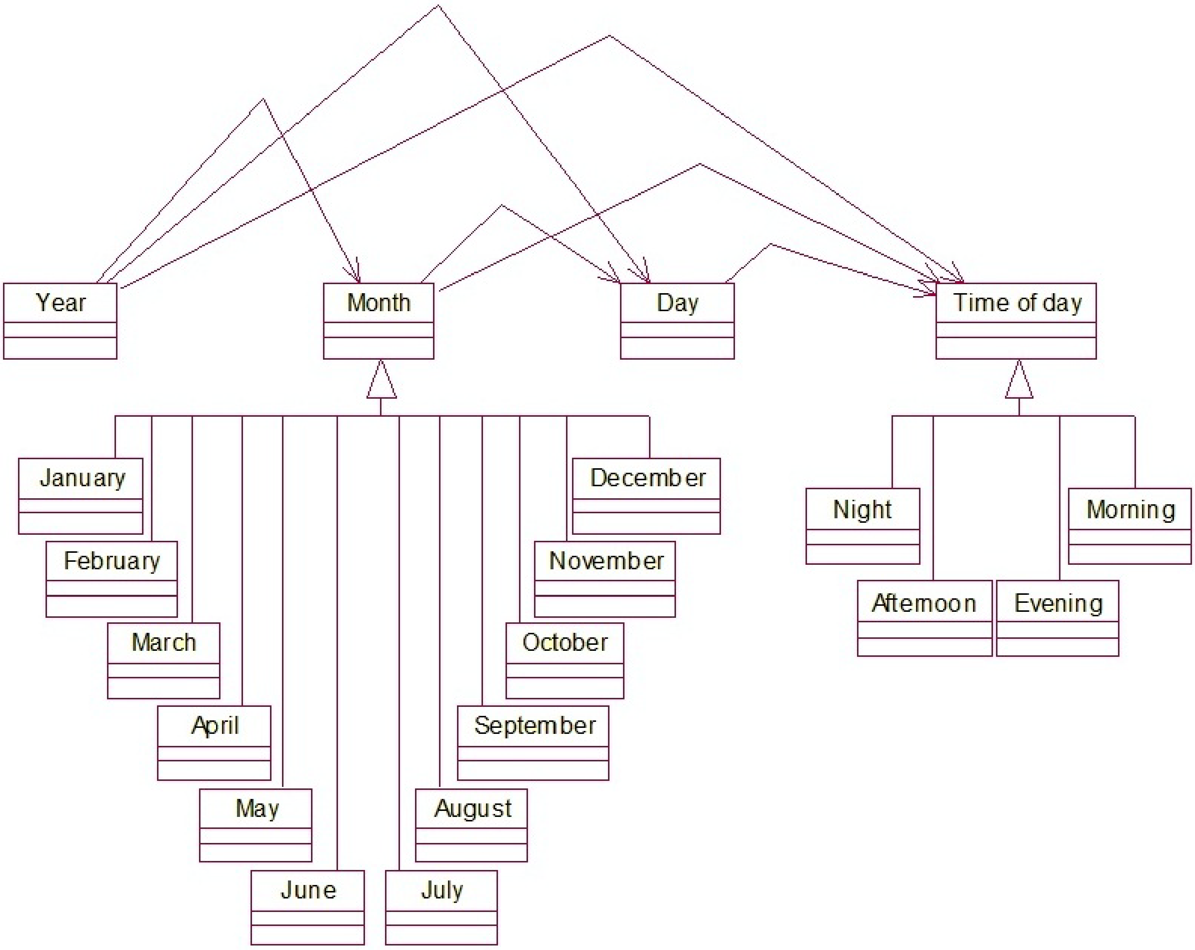


Figure - 3. Fragment of classification relation between time units

In [Kreydlin, 1997], the approximate classification of Russian temporary pretexts has been given (see, please, Table 1).

Table - 1. An example of temporary pretexts.

Relation	Pretext	Example	Temporary marker, the event
Simultaneity Extent Duration	For, During	For entire trip he said nothing During this year, he lives in town	The time interval is attached to event “trip” This year, the time interval is determined event “he lives”
Precedence	Approximately	We awaited approximately to midnight	Temporary marker: midnight; Time: the indeterminate half-interval. Event: we awaited

As to **syntactic level**, it is necessary to develop and use a set of syntactic patterns, such as Table 2 shows. Some principles of extracting syntactic patterns from texts are discussed in [Cimiano, 2006].

Table - 2. Some examples of syntactic patterns.

Relation	Structure	Syntactic pattern	Role in the sentence
Going after	Immediately afterward <action>/<event>, <event>	Adverb with the pretext «afterward» <action>/<event>, <event>	Adverbial modifier of time. Example: Immediately after the wedding and the parting words of parents, they left.
Inclusion	Including <the date>	Verbal Adverb <the date>	Adverbial modifier of time. Example: Including 2010

According to all above-stated translator work consists of following stages:

- At first translator searches for the supporting key words (time markers), which are associated with the expression of time in the text.
- Then translator, using lexical and syntactic models, attempts to determine the events, associated with the chosen time markers.
- If it is necessary, then the Syntactic Analyzer (Parser) is started.

Translator can repeatedly be turned first to the text or first to the cognitive model, then to the syntactic analyzer in order to search purposefully for the required (according to the rules of cognitive model) linguistic constructions.

While translator works assumptions about the events, extracted from the text, with their time intervals become. They can be represented as the list of the possible facts. Then the **Base of Events** will be filled up with the copies of events with their time characteristics. The **Block of Plausible Reasoning** also derives all consequences of the discovered facts (events, their properties, the relations between them).

Example

An example of text analysis is given on the narrative of V. Nekrasov “In the trenches of Stalingrad”. This example shows the result of the event-temporary text analysis with the use of cognitive models of time and events. There are some numbered proposals below, chosen from the source text. Some proposals without clearly time markers are passed.

1. For all my life, I can not recollect similar autumn. 2. September passed. 3. In the mornings, fish laps in Volga, and the big circles disperse on the mirror surface of the river. 4. At dawn, to the first volleys of the artillery ... it [the left shore of Volga river] is gentle …... 5. Some time it [the fog] still keeps over the river ……. 6. And long before the first sun rays, the first long-range gun shoots. 7. So the day begins. 8. Exactly at seven o’clock, at first sight imperceptible, the "frame" appears high in the sky. 9. It [The first ten of aircrafts] will determine the entire day.

The Table 3 illustrates the result of extracting time moments and associated events appearing in the text. At first the keywords definitely connected with the indication of time are revealed. The Syntactic Analyzer, Translator, and Cognitive Models are used for obtaining complete information about the events associated with the time indicators, calculating the duration of time intervals, and inferring all the consequences from the facts discovered.

Table - 3. The result of extracting events appearing in the text.

№	Event	Time interval	Inferred information
1	I do not recollect	Autumn	Autumn consists of «September, October, and November».
1	For all my life	The life of the author	
2	September passed	September	It precedes “October”; consequently, “October began”
3	Event 1: Fish laps in Volga; Event 2: The circles disperse	In the mornings	Each day in the morning; October;

	on the surface of the river		
4	X is gentle; X = the left shore of Volga river	At dawn; To the first volleys of the artillery;	At dawn = early in the morning; To (before) the first volleys of the artillery; **Event** = the first volleys of the artillery;
4	The first volleys of the artillery	At dawn;	Early in the morning; October; Autumn.
5	Event: X keeps; X = the fog	Some time	Some time, For a while; Early in the morning.
6	The long-range gun shoots	Long before the first sun rays	Before dawn
6	Event: First sun rays	At dawn	Early in the morning; October; Autumn
7	The day begins	The day	Day comes after morning; The beginning of the day;
8	The «frame» appears	At seven o’clock	At seven A.M.; the beginning of the day.
9	It will determine	The entire day	
9	It = the first ten of aircrafts	The entire day	The entire day = from the morning to the evening

In the second proposal, “September” is the subject. Predicate is expressed by the verb of passed time, whose semantics speaks that the time interval is finished, it left into the past. It is derived from the cognitive model of time that October goes after September, next month of autumn.

The subject in the fourth proposal is established with the aid of referential analysis of the previous proposal.

There is no explicit indication of time moment in the sixth proposal, but adverbial modifier of time «before the first sun rays» is associated with the dawn and the dawn – with the morning. That’s why we extract the event «the first sun rays» and associate it with «early in the morning». The proposal «It will determine the entire day» requires returning to the previous proposal in order to associate the word «it» with «the first ten of aircrafts». This action requires the complete syntactic analysis of previous proposal.

The analysis of the text results in obtaining the following sequence of the events: 1) the long-range gun shoots before dawn; 2) the first volleys of the artillery at dawn; 3) the day begins; 4) The «frame» appears at seven A.M.

Conclusion

The proposed model and methods are quite suitable for extracting events and their time characteristics from the text. For achievement of this purpose, the dialogue between the cognitive model of time, the translator and the syntactic analyzer is indispensable. It is necessary to note that the completeness and accuracy of the extracted knowledge depend on the cognitive model of time, its completeness and accuracy.

Subsequently it's necessary to work out the Cognitive Model of Time as completely as possible including all its elements, relations and methods. It is also necessary to take into account the uncertainty of time intervals. The Cognitive Model of Event depends greatly on the field of application. But this model contains also some universal cognitive elements: fact, process, action, result, subject, object, place of event, time of event, causal links between events and properties of object (subject). It's planned by us to refine the Cognitive Event Model through the knowledge of a concrete domain application (business, finances). It implies incorporating a mechanism of plausible inference over events into this model.

The next step of our project is to create the translator as a system of trained links between cognitive structures of time and events and correspondent patterns reflecting these structures in the natural language texts. The cognitive models are based on the knowledge about the world and therefore they can perform a semantic control of the Syntactic Analyzer's activity. However the translator's construction is the object of our further work.

Acknowledgements

The authors thank Doktor Vladimir Polutin, Director of HP Labs Russia, for the initiation and support of this work

Bibliography

[Boguraev, & Ando, 2005] B. Boguraev and R.K. Ando. Time ML-Compliant Text Analysis for Temporal Reasoning. Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'2005), pp. 997-1003. 2005.

[Boguraev, & Ando, 2006] B. Boguraev, R. Minoz, and J. Pustejovsky (Eds). ARTE: Annotating and Reasoning about Time and Event. Proceedings of the Workshop, Sydney, Australia: the Association for Computational Linguistics (ACL), 2006.

[Cimiano, 2006] P. Cimiano. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer Science+Business Media, New York, 2006, 347 pages.

[Elkin, et al., 2008] S.V. Elkin, V.V. Kulikov, E.C. Klishinsky, O.Y. Mansurova, V.Y. Maksimov, T.N.Musaeva, and S.N. Alineva, The Multi-model Character of Time and Temporal Relations in Semantic Language (SL). Science-Technical Information, Number 2, pp. 18-30, 2008.

[Kreydlin, 1997] G.E Kreydlin. Time through the Prism of Temporary Pretexts // Logical Analysis of Language. Language and Time, Moscow, 1997, 352 pages.

[Kubrjakova, 2004] E. S Kubrjakova. About the Purposes of Cognitive Science and Actual Problems of Cognitive Linguistic. The questions of cognitive linguistic – 2004, number 1, pp.6-17.

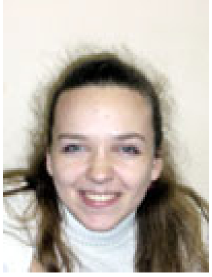
[Moszkowsky, 2007] Ben Moszkowski. Using Temporal Logic to Analyze Temporal Logic: A Hierarchical Approach Based on Intervals. Journal of Logic and Computation, 17(2):333–409, 2007. STRL Publication 2007-8.

[Popova, & Sternin,2007] Z.D. Popova and I.A. Sternin. Cognitive Linguistics. – M.: AST: «East – West». 2007.

Authors' Information

Xenia A. Naidenova – *Military medical academy, Saint-Petersburg, Stoikosty street, 26-1-248, e-mail: ksennaidd@gmail.com*

Major Fields of Scientific Research: Natural Language Processing, Machine Learning, Common-sense Reasoning



Marina I. Garina – *State university of transport ways, 190031, Saint-Petersburg, Moscow Avenue, 9;*

e-mail: MIGarina@gmail.com

Major Fields of Scientific Research: Multi-criteria Decision Aid, Natural Language Processing, Graph Theory

A FORMAL REPRESENTATION OF CONCEPT COMPOSITION

Daniel Schulzek, Christian Horn, Tanja Osswald

***Abstract**: This paper centers on argument saturation in relational-noun compounds. We argue that these compounds can be analyzed in terms of conceptual types, as introduced by [Löbner 1985, to appear]. He distinguishes between sortal, individual, functional, and proper relational concepts. To describe argument saturation in compounding, we use frames in the sense of [Barsalou 1992] since frames give a decompositional account of concepts and in particular reflect the conceptual types in their structure. Subsequently, we investigate relational-noun compounds in German as derived from their conceptual types. That is, we analyze in how far the conceptual types of the compound constituents determine the concept type of the compound as a whole. For possessive constructions, [Löbner, to appear] argues that a construction with a functional head inherits the type of the modifier. We demonstrate that for constructions with a relational head the case is less straightforward: the construction inherits the relational dimension of the modifier and the non-uniqueness from the head noun. However, we show that the combinations for compounds can follow complex compositional rules.*

***Keywords**: word formation, frames, compounds, lexical semantics*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings Languages, Theory*

1 Compounding in German

In German, compounding is a very frequent and productive type of word formation. On the linguistic surface, it consists in juxtaposing two or more lexemes, called the compound constituents. German compounds are right-headed; i.e. the last constituent determines the grammatical category of the compound. Morphological classifications of compounds are approached from a paradigmatic as well as from a syntagmatic perspective: the paradigmatic perspective is concerned with the grammatical categories of the constituents, while the syntagmatic perspective deals with the number of constituents being combined. Hypothetically, in German nearly all grammatical categories occur in compounding and there is no limit in combining constituents (as to formation rules in compounding see [Neef 2009]). In practice, however, the most frequent German compounds are binary noun-noun combinations, as shown in the empirical work of [Ortner et al. 1991]. In the following, we refer to the first constituent as the modifier and to the second constituent as the head of the compound.

From the perspective of cognitive semantics, the meaning of compounds is based on an implicitly given relationship between the concepts that are activated by the constituents. Noun-noun compounds seem to be the most flexible type in that they present a broader range of relationships than other types of compounds [cf. Ortner et al.]. With respect to the interpretation of compounds, [Olsen 1986] distinguishes between occasional and opaque compounds: the meaning of occasional compounds can be deduced from the compound constituents whereas the meaning of opaque compounds is non-transparent. Note that opaqueness and occasionality are not disjunctive categories but rather opponent features of a continuum. Apart from absolute opaque compounds, whose meanings must be stored in the lexicon, the interpretation of occasional compounds is a special type of concept combination in that the compound’s meaning results from creating a relationship between the concepts of its constituents; e.g. *Holztisch*, lit: ‘wood table’, can be interpreted as “table made of wood”, where the relative clause paraphrases the relationship between the concepts *Holz* ‘wood’ and *Tisch* ‘table’. Although compounds are potentially ambiguous in most cases [cf. Heringer 1984], [Kanngießer 1987] argues that the realm of possible relationships is restrained by the concepts of the compound constituents. Accordingly, the interpretation of compounds has to be considered as a matter of patterns rather than rules. This view is confirmed by the empirical work of [Maguire et al. 2010] for English who state that the semantic properties of the modifier and the head noun statistically correlate to the interpretation patterns they are preferably used in.

The semantics of compounding can be investigated on a descriptive and a computational level. The descriptive level focuses on documenting the relationships underlying the meanings of compounds while on the computational level, the mechanisms of deducing these relationships from the compound constituents are analyzed and explained. For German, the descriptive aspects of nominal compounding are well documented by [Ortner et al. 1991] who described more than 30 relationships in noun-noun compounds. [Fleischer & Barz 1995] point to similar relationships, but provide additional subtypes of each category. Furthermore, there have also been attempts to develop more abstract categorizations, e.g. the categorization in determinative, copulative, and possessive compounds that can be found in most traditional word-formation grammars. As it is, most of these categorizations are inconsistent [cf. Scalise & Bisetto 2009].

The computational level has been and still is widely debated in linguistics. Early approaches arise in Generative Grammar, but they have mainly been rejected in contemporary linguistic work because of their problematic basic assumptions [cf. ten Hacken 2009]. With respect to recent approaches, [Wisniewski 1997] differentiates between two explanatory accounts, the *thematic-relation view* and the *schema approach*. Both accounts assume that interpreting a (new) compound consists in creating a relationship between the concepts of the compound constituents, but they differ in explaining the way these relationships are created. According to the thematic-relation view, the interpretation is guided by a set of abstract thematic relations that have been deduced from already existing compounds; e.g. the thematic relation underlying the compound *Holztisch* in the above-mentioned interpretation would be “Y is made of X”. These thematic relations offer open variables that the concepts of the constituents can instantiate, depending on whether they fulfill selectional restrictions required by the particular thematic relations. Thus, interpreting a compound consists in selecting the appropriate thematic relation. Although the number of assumed thematic relations varies between the different proponents (see [Coolen et al. 1991], [Gleitman & Gleitman 1970], [Levi 1978], [Gagne 2001]), it seems widely accepted that the mentioned set is restricted. In contrast, the schema approach does not assume a fixed set of thematic relations. Instead, the interpretation of compounds is explained as a matching process of schemata understood as mental representations that are activated by the compound constituents. The proponents of the schema approach either draw on existing theories of mental representation or embed their explanation in an own theory. [Cohen & Murphy 1984] as well as [Wisniewski 1997], for instance, use schemata in the sense of [Minsky 1979], while [Lieber 2009] postulates so-called skeletons as concept-representation format to capture the semantics of compounding.

[Wisniewski, 1997] argues that the schema approach is cognitively more plausible than the thematic-relation view: in several experiments he demonstrates that subjects are able to create new interpretations of compounds spontaneously and that some of these interpretations cannot be captured by current thematic-relation sets postulated by proponents of the thematic-relation view. However, the schema approach lacks explanatory value since the different approaches are either too restricted in their range of application or they use vague notations: [Lieber 2009] is merely able to explain synthetic or copulative compounds. On the other hand, [Cohen & Murphy 1984] and [Wisniewski 1997] only propose possibilities of capturing compounds, but they do not offer a way to implement them within a consistent framework.

In this paper we will explain compounding as operations on frames as they have been introduced by [Barsalou 1992] and modeled as directed graphs by [Petersen 2007]. In contrast to the above-mentioned approaches of [Cohen & Murphy 1984] and [Wisniewski 1997], the frame model relies on a consistent formal basis and is flexible enough to capture a broader range of interpretation patterns than [Lieber, 2009] does. We will demonstrate its explanatory power by applying it to a class of compounds we refer to as relational-noun compounds (e.g. *Parteivorsitzender* lit: 'party chairman', *Whiskeyliebhaber* lit: 'Whiskey fancier"). They correspond to what [Fanselow 1981] calls "relationale Rektionskomposita": Fanselow coins the term in contrast to "Rektionskomposita" (english: synthetic compounds), where the modifier saturates an argument of the deverbal head. In opposition to synthetic compounds, nominal relational compounds are not formed with deverbal nouns but with non-derived relational nouns as heads. Thus, relational-noun compounds are understood as noun-noun compounds, where the head noun is a relational noun whose argument is saturated by the modifier noun.

2 Conceptual noun types

Relational nouns have long been distinguished from sortal nouns. The distinction is generally taken as a distinction between one-place predicates and two- (or more-) place predicates (cf. [Asudeh 2005], [Behaghel 1923], [Partee 1983/1997]). [Vikner & Jensen 2002] argue that relational nouns also exhibit a certain kind of semantic relation inherently determined in their primary interpretation. In contrast, sortal nouns do not exhibit an inherent relation. Their interpretations in possessive constructions depend on the linguistic specification, or on the context of utterance. [Löbner 1985, to appear] amends the distinction between sortal nouns [–R] and relational nouns [+R] (and their concepts, respectively) by introducing a uniqueness property [±U]. As a consequence, four basic noun concepts are distinguished: functional nouns ('FN'; *roof, chancellor, end, wife, trunk*) share the properties [+R] and [+U]. Functional nouns are construed in a way that there is only one possible referent once the possessor argument is saturated. For example, a house has only one roof and a roof is always the roof of a house. Proper relational nouns ('RN', which Löbner refers to as 'relational nouns') such as *chapter, piece, advisor, user,* or *member* are [+R] but in contrast to functional nouns [–U]; hence, the number of their potential referents is not restricted (an association may have many members, a book generally has several chapters). Individual nouns ('IN'; *Kreml, pope, bible*) are [–R], [+U] and construed as referring uniquely to one entity (without further contextual disambiguation, we may refer to the bible, to the Kreml). Sortal nouns ('SN'; *tree, cake*) are [–R], [–U]. Support for the conceptual noun type distinction is provided by typological (cf. [Gerland & Horn 2010], [Löbner, to appear]) and empirical investigations [cf. Horn & Kimm, to appear].

[Löbner, to appear] claims that the lexical referential properties of nouns influence the way they are used grammatically. In accordance with their referential properties, functional and relational nouns are predisposed for possessive use. Due to their inherent uniqueness, individual and functional nouns have a predisposition for definite use. Consider the examples in (1) for relational and in (2) for functional nouns:

1. a. *A member of the Academy of Science has died.*

   b. *He only read one chapter of the book.*

2. a. *The end of the movie was very sad.*

   b. *The chancellor of Germany is Angela Merkel.*

In use, however, all nouns can be shifted to a different type. Sortal nouns for example are also frequently used in definite NPs when referring to unique entities (*the book, the tree*). In other cases, the possessor argument of a relational or functional noun may be omitted when the possessor can be retrieved from the context of the utterance. For the purposes of this paper, however, only the semantic properties of the conceptual noun types are focused; shifts are consequently not considered here. The question addressed here is how the conceptual types combine in compounds (as we will see in section 4) and how this composition can be formally modelled.

3 The representation of nominal concepts as frames

As a representation of conceptual knowledge, frames as introduced in [Petersen 2007] are based on [Barsalou 1992] and [Carpenter 1992]. Frames give a decompositional account of concepts. In this, they are in the tradition of Carpenter's feature structures. Those are labeled directed graphs which have a root. As argued by Petersen, not all concepts are adequately analyzed by a rooted graph. Thus, frames are more general than classical feature structures. Formally, a frame is represented by a connected directed graph with one central node (marked by a double border). The nodes of the frame are labeled with types which are given by a type signature, and the arcs of the frame are labeled with attributes. On the latter, we have the constraint that attributes are functional; i.e. there cannot be two arcs labeled with the same attribute going out from one node. Note that this does not exclude incoming arcs at the central node; hence, frames are more general than feature structures. Concept frames feature a marker for open arguments. On the frame graph, we indicate an open argument by a rectangular node. Apart from that, referential uniqueness is marked by a definiteness marker, pictured by an incoming arrow without a source node. The type signature includes a hierarchy of types; that is, it is based on a partially ordered set which is a join semilattice. In addition, the type hierarchy conveys information about the possible attributes for nodes; i.e. it gives types and values that can be in the range and the domain of an attribute.

Conceptual types are reflected in the concept's frame representation. Relationality is indicated by an argument node that is not the central node. Uniqueness is indicated by a path from a definite node to the central node (that path can have length zero). Definite nodes are those that have a marker for unique reference. Therefore, they have a definiteness marker or they are non-central argument nodes. The arguments count as definite in this context since once they are filled, they are definite. Thus, frames representing sortal nouns have one argument node which is the central node [-R], and no path from a definite node to the central node [-U].

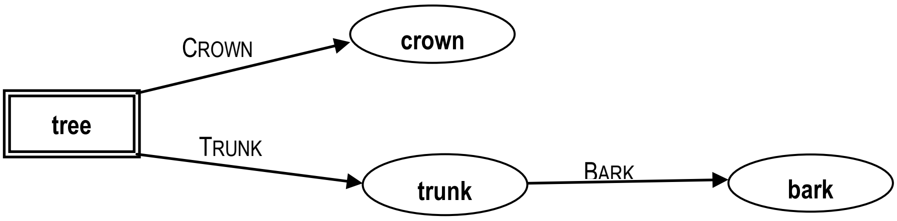


Figure 1. **Frame of the SC *tree***

For example, see Figure 1. Here, we have a frame for the sortal concept *tree*. The central node is the only argument node and in this particular case it is the root node of the graph; thus there is no incoming arc at the central node, in particular not from a definite node. In Figure 2, we have such a determining arc [+U]. *Kremlin* is not relational [-R]. Thus, the frame represents an individual concept.

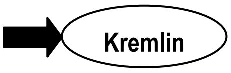


Figure 2. **Frame for the IC *Kremlin***

Relational concepts are those that have an argument node that is not the central node. As an example for a proper relational concept, regard the frame for *brother* in Figure 3. A brother is analyzed as something that is male and shares a mother with someone else. As a brother is always the brother of someone, this someone else is an argument for brother [+R]. Note that there is no directed path from the argument node to the central node [-U].

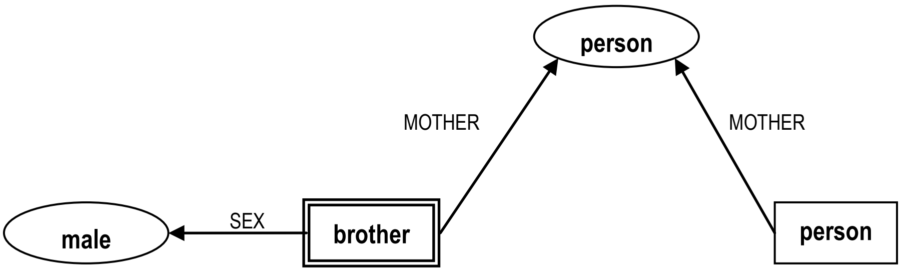


Figure 3. frame for the RC *brother*

Functional concepts have both, an argument apart from the central node [+R] and a path from a definite node to the central node [-U]. The example in Figure 4 shows the frame for the concept *mother*. A mother is something that is female and has something it is mother of. As soon as the argument is filled, the mother's identity is determined. Thus, the concept's referent depends functionally on the value of the argument node.

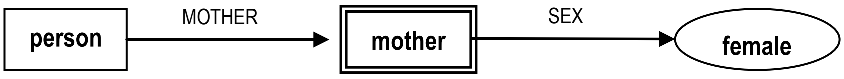


Figure 4. **Frame of the FC *mother***

4 Frame analysis of nominal relational compounds

[Petersen & Osswald 2009] demonstrate that argument saturation in general can be captured in terms of frames by inserting the possessor frame into the open argument of the frame whose argument is satisfied. In the following, we analyze the argument saturation in relational-noun compounds with RNs and FNs as heads. As we distinguish four conceptual types, in each case the argument of the relational or functional head can be saturated in four different ways. Thus we analyze eight combinatorial types of nominal relational compounds. Argument saturation of FN and RN in possessive constructions has already been investigated by [Löbner, to appear]. His findings are summarized in Table 1. In the following, we argue that argument saturation in compounding reflects a similar but not identical pattern: most of the combinatorial types behave correspondingly to argument saturation in possessive construction, except for combinations of SN and FN.

Table 1: Type composition for head plus possessor combinations [cf. Löbner, to appear: 35]

possessor		**head**		**head with possessor**	
SN	*car*	RN	*door*	SN	*door of a car*
RN	*sister*	RN	*aunt*	RN	*sister of an aunt*
IN	*pope*	RN	*brother*	SN	*brother of the pope*
FN	*mother*	RN	*uncle*	RN	*uncle of a mother*
SN	*boy*	FN	*father*	SN	*father of a boy*
RN	*aunt*	FN	*mother*	RN	*mother of an aunt*
IN	*Croatia*	FN	*capital*	IN	*capital of Croatia*
FN	*mother*	FN	*father*	FN	*father of a mother*

Figure 5 shows the frame of the **SN-RN compound** *Kuchenstück* (lit: *Kuchen* 'cake' *Stück* 'piece'). Since a cake is a special kind of an object, the frames can be unified: the *cake* frame saturates the open argument in the *piece* frame so that the possessor node transforms into a round node. Since there is no further open argument the result of the unification is an SN.

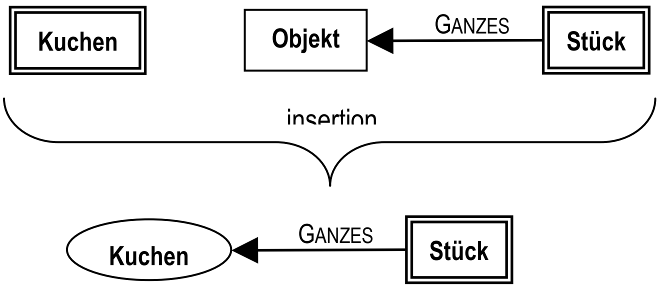


Figure 5. **Kuchen** '**cake**',**Objekt** '**object**', **Ganzes** '**whole**', **Stück** '**piece**'

|

Kapitel 'chapter'): the open argument in the *chapter* frame is saturated by the *bible* frame.

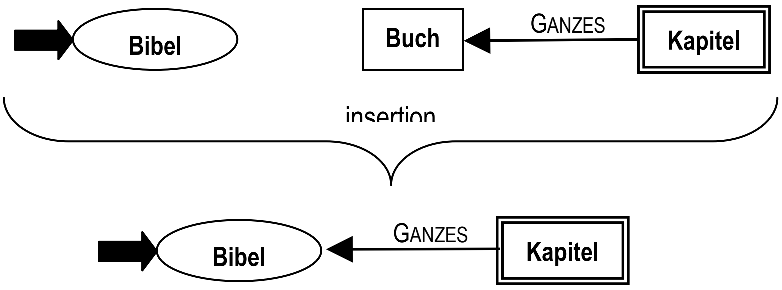


Figure 6. **Bibel** '**bible**', **Buch** '**book**', **Ganzes** '**whole**', **Kapitel** '**chapter**'

The determination of both compounds confirms the validity of our frame-based analysis: Both compounds can be used indefinitely:

3. Ein Kuchenstück ist schon angebissen. "A piece of cake is already been bitten off."

4. Wir mussten ein ganzes Bibelkapitel lesen. "We had to read a whole chapter of the holy bible."

RN-RN and **FN-RN compounds** seem to behave in a similar way: in both cases the result is a RN. Figure 7 shows the frame of the compound *Mitgliederberater* (*Mitglied* 'member' *Berater* 'adviser'), and Figure 8 the frame of the compound *Vorstandsmitglied* (*Vorstand* 'management' *Mitglied* 'member'). The relationality of these compounds results from the fact that the *institution* nodes in both cases are linked to the bridging frame by outgoing and therefore non-determining arcs. Furthermore, the relationality is reflected in the possibility to use the compounds possessively, but indefinitely:

5. ein Mitgliederberater des Tennisclubs "an adviser of members of the tennis club"

6. ein Vorstandsmitglied der Deutschen Bank "a board member of the German Bank"

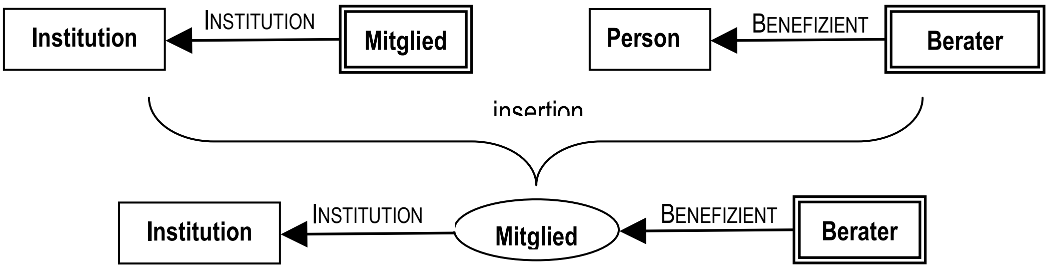


Figure 7. **Institution** '**institution**', **Mitglied** '**member**', **Benefizient** '**beneficient**', **Person** '**person**', **Berater** '**adviser**'

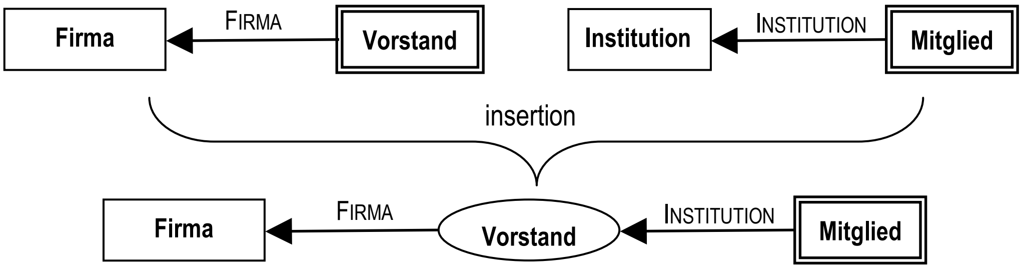


Figure 8. **Firma** ‘**company**’, **Vorstand** ‘**management**’, **Institution** ‘**institution**’, **Mitglied** ‘**member**’

IN-FN compounds result in INs. Figure 9 demonstrates the unification underlying the compound *Kremldach* (lit: *Kreml* ‘*Kremlin’* Dach ‘roof’). In contrast to IN-RN compounds, the functional head inherits the uniqueness of the IN since the arc labeled ROOF is a determining one, and thus it is uniquely determined to which roof the compound refers.

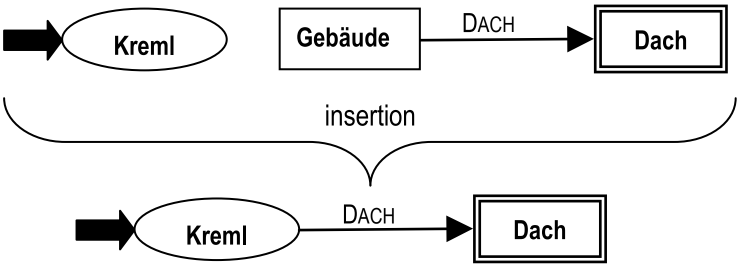


Figure 9. **Kreml** ‘**Kremlin**’, **Gebäude** ‘**building**’, **Dach** ‘**roof**’

RN-FN compounds are relational but not functional. Figure 10 shows the frame of the compound *Benutzername* (lit: *Benutzer* ‘user’ *Name* ‘name’), since the *name* frame inherits the relationality of the *Benutzer* frame in which the INSTITUTION attribute is unspecified.

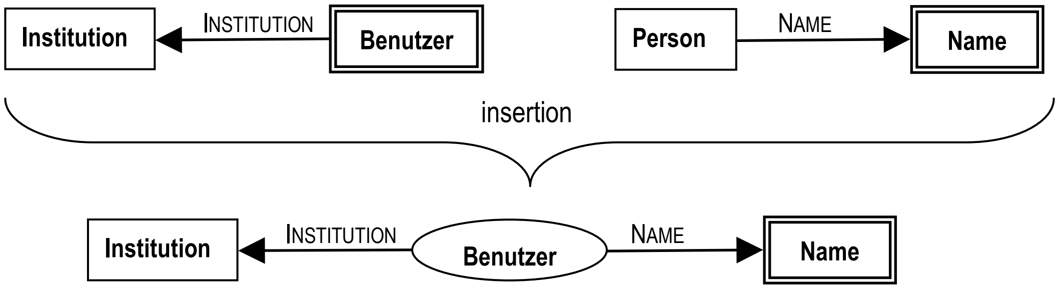


Figure 10. **Institution** '**institution**', **Benutzer** '**user**', **Person** '**person**', **Name** '**name**'

FN-FN compounds are functional. Figure 11 shows the frame structure of the compound *Kanzlergattin* (lit: *Kanzler* 'chancellor' *Gattin* 'wife') in which both constituents are functional. The result is a FN, as it can be deduced from the frame resulting from the unification of the frames of the compound constituents. The attribute CHANCELLOR and the attribute WIFE are series-connected determining arcs; thus, the instantiation of the *nation* node determines the value of the chancellor node that, in turn, determines the value of the *wife* node. In other words: the compound has one open argument and thus it is a functional noun.

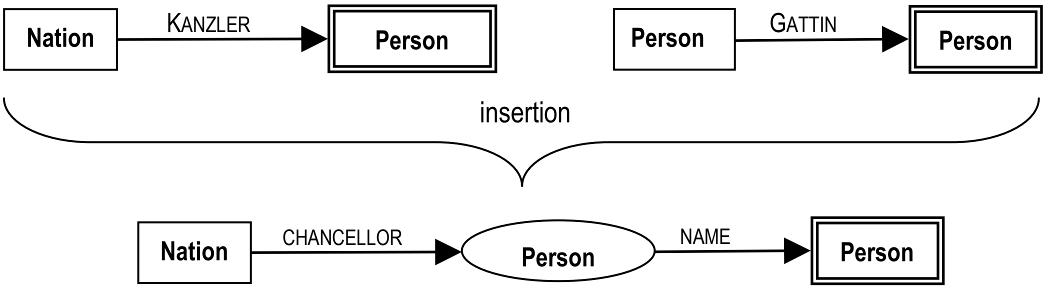


Figure 11. **Nation** '**nation**', **Kanzler** '**chancellor**', **Person** '**person**', **Gattin** '**wife**'

As mentioned above, the analysis of **SN-FN compounds** is not as unambiguous as that of possessive constructions. Instead, compounds of this type can be a SN as well as a FN. In the compound *Filmende* (lit: *Film* 'movie' *Ende* 'end'), the SN *Film* seems to saturate the possessor of the functional head *Ende*. However, the possessor can still be saturated on the linguistic surface (see 7), and the indefinite use of the compound is most heavily marked (see 8). Thus, the modifier does not saturate the argument, but rather constrains the possessor specification: the argument can merely be saturated by concepts that are sub-concepts of the modifier. For this reason, sentence 9 is inconsistent. On the other hand, the compound Baumstamm (Baum 'tree' Stamm 'trunk') can be used indefinitely (see 10) which is an indication that the compound is a SN. Thus, SN-FN compounds require two frame representations: In Figure 12, the argument is still open, while it is saturated in Figure 13.

7. Das Filmende von “Vom Winde verweht” ist traurig. ”The movie end of Gone with the Wind” is sorrowful.”

8. ??ein Filmende ist manchmal traurig. ??”A movie end is sometimes sorrowful.”

9. §§das Filmende des Buches §§”the movie end of the book”

10. Dort drüben liegt ein Baumstamm. ”A trunk of a tree is lying over there”

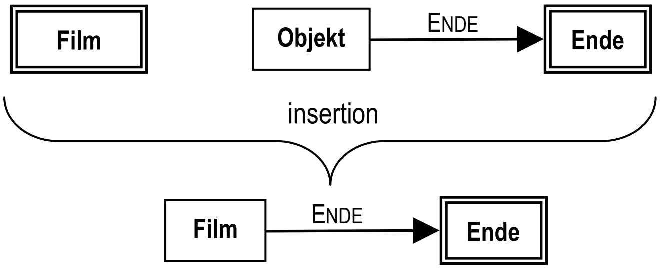


Figure 12. **Film** ‘**movie**’, **Objekt** ‘**object**’, **Ende** ‘**end**’

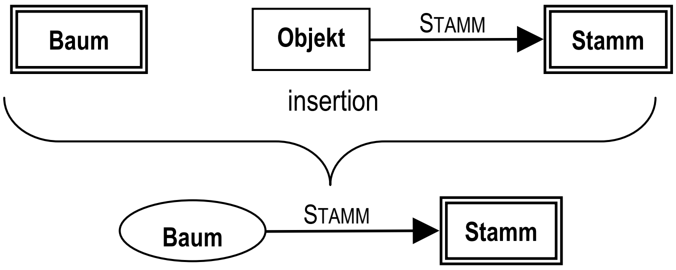


Figure 13. **Baum ‘tree**’, **Objekt ‘object**’, **Stamm ‘trunk**’

5 Conclusion

We presented a formal model of concept composition in compounding with respect to conceptual types in the sense of [Löbner 1985, to appear] to analyze relational-noun compounds. It has been shown how the modifier determines the conceptual type of the whole compound in most cases (see Table 2). The exception is the resulting type of SN-FN compounds that can either be functional or sortal. Thus, SN-FN compounds that are still functional are no relational-noun compounds in the narrow sense because the argument is not saturated by the modifier. Instead the modifier merely constrains the range of possible concepts which can function as arguments.

Table 2. Combinatorics of conceptual types in compounding

Type of the modifier	Type of the head noun	Resulting type
SN	RN	SN
RN	RN	RN
IN	RN	SN
FN	RN	RN
SN	FN	SN
RN	FN	RN
IN	FN	IN
FN	FN	FN or SN

Bibliography

[Asudeh 2005] Ash. Asudeh. Relational nouns, pronouns, and resumptions. Linguistics and Philosophy. 28 (4). 375-446, 2005.

[Barsalou 1992] L. Barsalou. Frames, Concepts, and Conceptual Fields. In: Frames, Fields, and Contrasts. New Essays in Semantic and Lexical Organisation. Ed. A. Lehrer and E.F. Kittay. Erlbaum, Hillsdale NJ, 21-74, 1992.

[Behagel 1923] O. Behagel. Deutsche Syntax. Eine geschichtliche Darstellung. Bd. I: Die Wortklassen und Wortformen. A. Nomen. Pronomen. Heidelberg: Carl Winter's Universitaetsbuchhandlung, 1923.

[Cohen & Murphy 1984] B. Cohen and G.L. Murphy. Models of concepts. In: Cognitive Science, 8, 27-58, 1984.

[Coolen et al. 1991] R. Coolen, H.J. van Jaarsveld and R. Schreuder. The interpretation of isolated novel nominal compounds. Memory & Cognition, 19, 341-352, 1991.

[Fanselow 1981] G. Fanselow. Zur Syntax und Semantik der Nominalkomposition. Niemeyer, Tübingen, 1981.

[Fleischer & Barz 1995] W. Fleischer and I. Barz. Wortbildung der deutschen Gegenwartssprache. Max Niemeyer, Tübingen, 1995.

[Gagne 2001] C.L. Gagne. Relation and lexical priming during the interpretation of noun-noun combinations. In: Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 236-254, 2001.

[Gerland & Horn 2010] D. Gerland and C. Horn. Referential properties of nouns across languages. In: Universal grammar and individual languages. Proceedings of SiCoL 2010. Ed. D. Choi et al. Seoul, 2010.

[Gleitman & Gleitman 1970] L.R. Gleitman and H. Gleitman. Phrase and paraphrase. Norton, New York, 1970.

[Heringer 1984] H.-J. Heringer: Wortbildung: Sinn aus dem Chaos. Deutsche Sprache 12, 1-13, 1984.

[Horn & Kimm To appear] C. Horn and N. Kimm. Concept Types and Frames. Applications in Language, Cognition, and Philosophy. In: Studies in Linguistics and Philosophy. Ed. Gamerschlag, T.; Gerland, D.; Osswald, R.; Petersen, W. Springer, Dordrecht, to appear.

[Kanngießer 1987] S. Kanngießer. Kontingenzen der Komposition. In: Neuere Forschungen zur Wortbildung und Historiographie der Linguistik. Festgabe für Herbert E. Brekle zum 50. Geburtstag. Ed. B. Asbach-Schnitker and J. Roggenhofer. Narr, Tübingen, 3-30, 1987

[Levi 1978] J.N. Levi. The syntax and semantics of complex nominals. Academic Press, New York, 1978.

[Lieber 2009] R. Lieber. A Lexical Approach to Compounding. In: The Oxford Handbook of Compounding. Ed. R. Lieber Rochelle and P. Štekauer. Oxford University Press, Oxford, 78-104, 2009.

[Löbner 1985] S. Löbner. Definites. In: Journal of Semantics 4. 279-326, 1985.

[Löbner, to appear] S. Löbner. Conceptual types and determination. Journal of Semantics, to appear.

[Maguire et al 2010] P. Maguire, E.J. Wisniewski and G. Storms. A corpus study of semantic patterns in compounding. Corpus Linguistics and Linguistic Theory 6: 49-73, 2010.

[Neef 2009] M. Neef. IE, Germanic: German. In: The Oxford Handbook of Compounding. Ed. R. Lieber Rochelle and P. Štekauer. Oxford University Press, Oxford, 386–399, 2009.

[Olsen 1986] S. Olsen. Wortbildung im Deutschen. Eine Einführung in die Theorie der Wortstruktur. Stuttgart, Kröner, 1986.

[Ortner et a. 1991] L. Ortner, E. Müller-Bollhagen, M. Pümpel-Mader and H. Gärtner. Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Vierter Hauptteil. Substantivkomposita. Berlin: De Gruyter. 145-657, 1991.

[Partee 1983/1997] B. Partee. Uniformity vs. versatility: the genitive, a case study. Appendix to Theo Janssen (1997): Compositionality. In: van Benthem, J., and A. ter Meulen (eds.): The Handbook of Logic and Language. Elsevier, 1983/1997.

[Petersen 2007] W. Petersen: Representation of Concepts as Frames. In: The Baltic International Yearbook of Cognition, Logic and Communication (3). Ed. Latvijas Universitate. New Prairie Press, Manhattan, 151-170, 2007.

[Petersen & Osswald 2009] W. Petersen and T. Osswald. A Formal Interpretation of Frame Composition. Presentation given at ctf09, Duesseldorf, 2009.

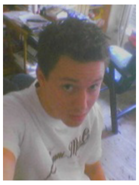
[Scalise & Bisetto 2009] S. Scalise and A. Bisetto. The classification of compounds. In: The Oxford Handbook of Compounding. Ed. R. Lieber and P. Štekauer. Oxford University Press, Oxford, 34-53, 2009.

[ten Hacken 2009] P. ten Hacken. Early Generative Approaches. In: The Oxford Handbook of Compounding. Ed. R. Lieber Rochelle and P. Štekauer. Oxford University Press, Oxford, 54-77, 2009.

[Vikner & Jensen 2002] S. Vikner and P. A. Jensen. A semantic analysis of the English genitive. Interaction of lexical and formal semantics. Studia Linguistica 56.191-226, 2002.

[Wisniewski 1997] E.J. Wisniewski. When Concepts Combine. Psychonomic Bulletin & Review 4: 167-183, 1997.

Authors' Information



Daniel Schulzek *– General linguistics at Heinrich Heine University Düsseldorf, Germany; e-mail: schulzek@phil.uni-duesseldorf.de*

Major Fields of Scientific Research: formal and cognitive semantics, word formation, metaphors and metonymies, frame theory

Christian Horn *– General linguistics at Heinrich Heine University Düsseldorf, Germany; e-mail: chorn@phil.uni-duesseldorf.de*

Major Fields of Scientific Research: lexical and cognitive semantics, definiteness, reference, frame theory

Tanja Osswald *– General linguistics at Heinrich Heine University Düsseldorf, Germany; e-mail: chorn@phil.uni-duesseldorf.de*

Major Fields of Scientific Research: Theories of Concepts, Modal Logic, Formal Frame Theory

THE ARGUMENT BASED COMPUTATION: SOLVING THE BINDING PROBLEM

Alona Soschen, Velina Slavova

***Abstract**: In this paper, we further developed the argument-based model of syntactic operations that is argued to represent the key to basic mental representations. This work concentrates on formal descriptions of the observed syntax-semantics dependencies. We briefly review our up do date experimental work designed to test this hypothesis, and offer the results of our most recent experiment. The results of our experiments confirmed that semantic relations between the images in conceptual nets influence syntactic computation. The binding problem that arises when the same noun can be represented either as Subject (ex. The cat chases the mouse) or Object (ex. The mouse chases the cat), was successfully resolved.*

Keywords: Cognitive Models of Language Phenomena, Formal Models in Language and Cognition, Psycholinguistics and Psycho semantics

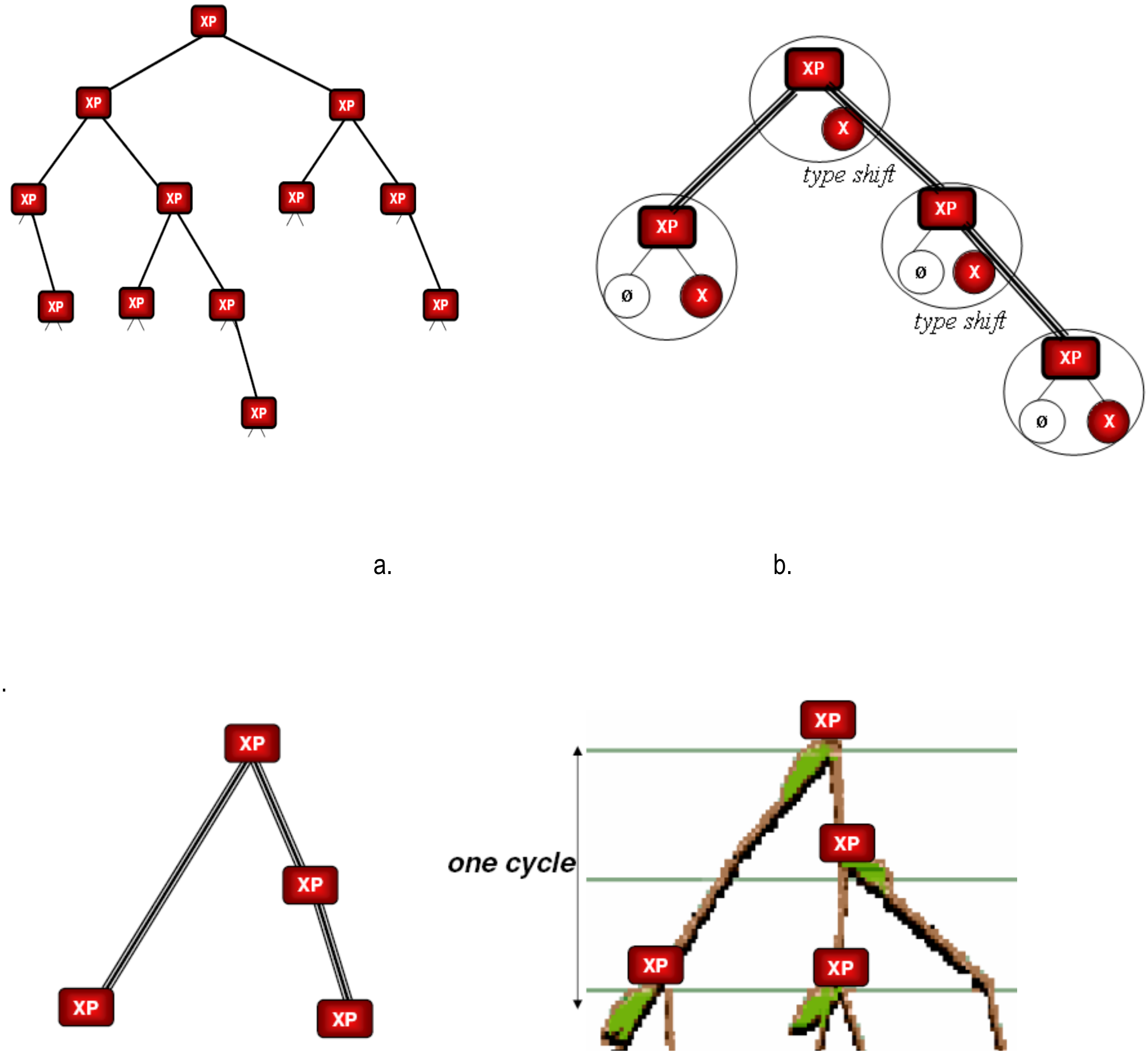
***ACM Classification Keywords**: **ACM Classification Keywords**: I.2 Artificial Intelligence, 1.2.0. Cognitive simulation*

Introduction

Following one of the widely accepted linguistic theories, the key component of Faculty of Language (FL) is a computational system (narrow syntax) that generates internal representations and maps them into the conceptual-intentional interface by the (formal) semantic system (Hauser et al., 2002). There is a consensus that the core property of FL is *recursion*, which is attributed to narrow syntax. In other words, the process of mental generation of syntactic structures relies on the capacity of the human brain to perform specific operations in compliance with the principles of efficient computation. The claim in the recent theories is that this computation is based on a primitive operation that takes already constructed objects to create a new object. This basic operation (Merge) provides 'a language of thought', an internal system to allow preexistent conceptual resources to construct expressions (Chomsky, 2006). Although these questions receive a lot of attention, there are no convincing proposals yet concerning the precise type of resources on which such computation is performed in a recursive manner to build syntactic structures.

In Slavova and Soschen (2007), syntactic structures, presented in the traditional sense of Chomskyan theory (Bare Phrase Structures, XP-structures), were re-defined in terms of finite recursive binary trees. The structure obtained in this way is a tree of Fibonacci (figure 1. a) that complies with the principles of optimization, namely with the principle of efficient growth (Soschen 2006, 2008).

This tree can be seen as an operator – it "performs" a bottom-up Merge (fig.1.a.); its nodes are the results of Merge. In the model under development (fig.1.b), XPs are *sets*, Xs are 'unbreakable' *entities*, and Merge can be applied to two non-equivalent substances (the tree has ordered nodes). We called the tree in (fig.1.b) "Argument-Based Syntactic Tree". According to the hypothesis put forward in Soschen (2005, 2006, 2008), a general rule governing efficient growth applies in syntax in such a way that minimal syntactic constituents incorporate arguments (*agent, recipient, theme*) which are related to each other. In the Fibonacci-tree model, the type of merge configuration determines the type of relation between arguments.



a.

b.

.

c.

Figure.1.

The question of the height of this obtained tree is deeply related to the question of the limits of human cognitive resources. This tree expresses a label-free structure that does not have lexical items; what it has are the paths of connecting smaller units in order to produce a larger meaningful unit. It could be suggested that the limits of this structure are determined in the same way as the number of nodes and relations that can be treated by the human brain within a semantically meaningful argument structure. The analysis has shown that the paths of merging terminal Xs that result in their final configuration, obtained at the root, are finite and well determined. The final configuration is a precise scheme that incorporates terminal Xs, which corresponds to one given path of merge.

Going back to the syntactical sense of XP, we may interpret the properties of the XP Fib trees as follows: the merge-tree defines the maximal number of XPs that can be merged into a configuration, to the root where a meaningful relation between these arguments is established. The type of configuration of merging arguments determines the type of relation between arguments. We showed what the kind of relation is set at the root of merge configuration (fig. 1.a.). The maximal configuration (fig 1.b.) corresponds to the determined in linguistics as: subject, recipient and theme and the semantic-merge pattern corresponds to the principle of efficient growth.

The argument-based model assigns a primary syntactic role to *entities*, usually expressed as nouns. This viewpoint is in contrast with the *verb*-centered model of syntax. Our efforts were further focused on experimental work in support of the argument-based model providing support for the argument-based model. In our recent paper (Slavova & Soschen, 2009) we provided experimental evidence that identifies the semantic role of entities (nouns) as primary in syntax. Our experiment was based on the fact that Bulgarian lost its case markers; the Genitive and Dative cases are both expressed by means of a preposition *‘на’* (na). In Bulgarian, all the grammatically correct sentences of type:

Subject Verb Object *на* **Y**.

are ambiguous: they may assign two different meanings to Y - either that of the *Recipient* or the *Possessor*. We used 12 such double-meaning sentences and asked about 100 native Bulgarians and fluent French speakers to translate them in French, where preposition *на* is translated as “à” (to) for the Recipient-meaning and as “de” (of) for the Possessor-meaning. Only two subjects noticed the double meaning. The experimental data supported the idea that mental computation of syntax is influenced by the inter-conceptual relations between the images of entities in a semantic space. The assumption was made that the syntactic treatment includes Merge that operates on the images of the concepts. In the present paper, we call this operation “Semantic Merge” (SeM).

In the course of the experiment, the result of one of the two treatments is rejected. The assumption is that the intermediate images μ of either the Recipient scheme (fig.2.a.) or the Possessor scheme (fig.2.b.) are rejected by activating semantic relations between concepts. The experiment showed that the treatment follows the Recipient scheme (fig. 2a) even when there are no reasons to reject μZ* as Possessor of Y*. The explanation of this result is that the argument-based syntactic structure is “calculated” as primary. If there are no reasons to reject the result of one of the treatments, it is accepted as final.

Our conclusion is that the argument-based syntax has a fundamental character. The role of *entities* (nouns in this case) is confirmed as primary; the relations between the images of the concepts in the conceptual nets influence the final result of the syntactic computation.

We showed that the reiterative operation assigns a primary role to entities as the key components of syntactic structure. The schemes on figures 2 represent the stages of syntactic treatment with SeM for the Recipient assignment and the Possessor assignment. The images μ obtained at each step are provided in order to develop the mechanism of the treatment and to analyze it in accordance with the results of the experiments. Our assumptions concerning the ways the argument structure is computed have led to the development of the argument-based model of basic syntax.

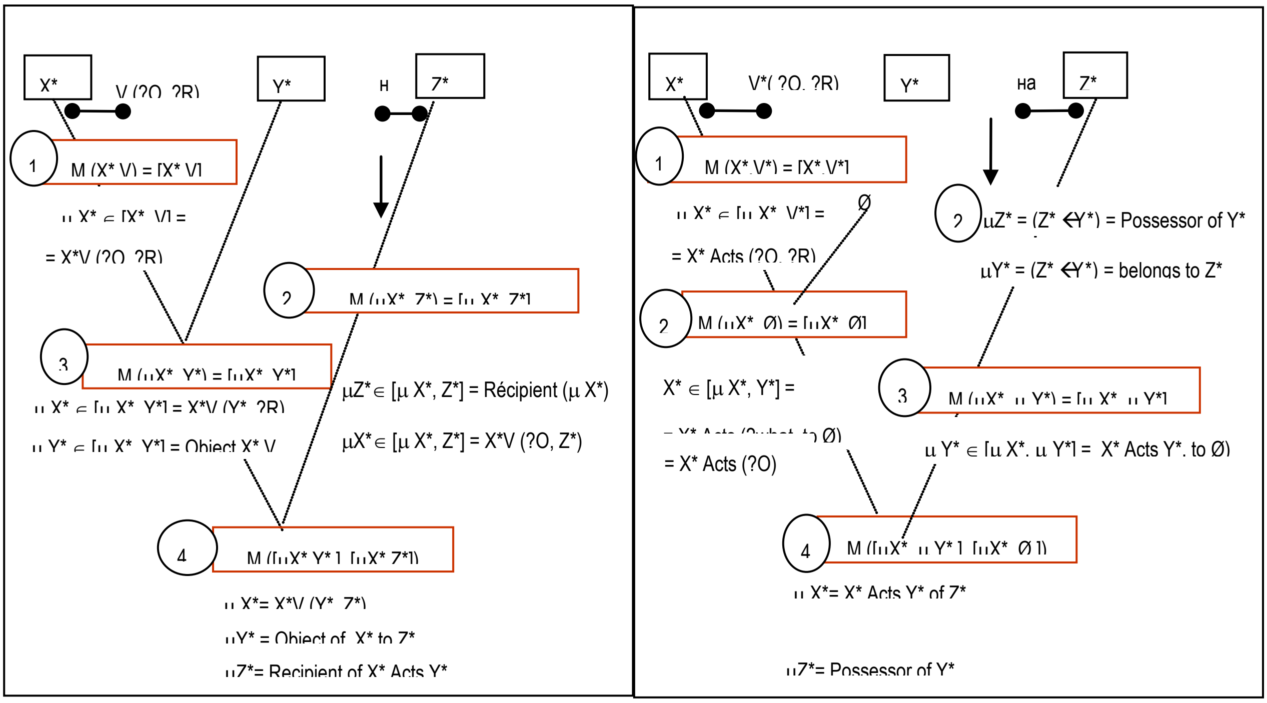


Fig 2 a) Recipient scheme b) Possessor scheme

SEMANTIC MERGE

Semantic Merge (SeM) was modeled as a binary operation, performed in sequential progression on the concepts (X*, Y* and Z*) expressed as nouns in a sentence that also includes verb V*. The general idea is that SeM complies with the principles of the argument-based syntax. The result of SeM consists of temporal semantic images μ retained in working memory up till the final stage of the syntactic treatment. The formal description of the stages obtained in the course of syntactic treatment corresponds to the experimental results. The hypothesis that syntactic rules comply with operations on semantic primitives is thus supported.

Our attention in this experiment was focused on SeM of Z* as either the Recipient or the Possessor; SeM of X* and Y* defines X* and Y* as Subject and Object of the verb V*, accordingly. The edge: *{X, Verb}* entails SeM between X* and V. The result of SeM is a pair, in which each element obtains image μ, which represents the concept in the semantic context that includes the other member of the pair. The syntactic structure begins to be assembled on the basis of the semantic information.

In the argument-based syntactic model, Subject is primary in the treatment.

$$M\ (X^*V^*) = [X^*, V]\ (\underline{?O}) \qquad (1)$$

For example, image μ of the concept X* within the couple [X*V] corresponds to image of X* as Subject, performing V:

$$\mu\ X^* \in [X^*V] = X^*\ \text{Acts}\ (\underline{?O}) \qquad (2)$$

In our test sentences Subject and Verb were grammatically marked. However, in the experiment presented in this paper, we analyze a language in which the grammatical rules do not always mark Subject and Object. In this experiment, the calculation of Subject vs. Object is dependent solely on the interaction of mental images of the concepts X* (Subject) and Y* (Object).

Recall that in the model offered for your consideration, the argument-centered representations are based on the primary function of the theme in respect to the agent; objects are grouped according to their primary function with respect to the participant (Soschen, 2008). This approach resolves certain problems for a neural instantiation (van der Velde & de Kamps, 2006). One of them, the binding puzzle, concerns the way in which neural instantiations of parts (constituents) can be related (bound) temporarily in a manner that preserves the structural relations between the constituents. Assume that words like cat, chases, and mouse each activate specific neural structures. The problem then would be that Noun cat and Noun mouse are bound to the role of agent and theme, respectively, of Verb chases in the sentence The cat chases the mouse and to the role of theme and agent of chases in the sentence The mouse chases the cat.” In the present theory, however, no binding by V is necessary; the semantic roles (Subject vs. Object) are determined on the basis of the interaction of the concepts X and Y*.*

A mathematical theory for semantic analysis is feasible when at some level a finite set of principles is available to determine the basic rules that underlie this interpretive part of language. The structure of a sentence is given by a recursive rule, as this provides the means to derive an infinite number of sentences using finite means. For the same reason, semantics employs recursive procedures that assign a certain meaning to a sentence based on the relations that exist between its elements.

SEMANTIC MERGE: EXPERIMENTAL PROOF

Bulgarian is an Indo-European language, a member of the Slavic branch. Bulgarian exhibits certain peculiarities that set it apart from other Slavic languages, such as elimination of Case marking and the development of a suffixed definite article. Although the Bulgarian nouns are rarely marked for Case, the word order is rather free. Thus, in a Bulgarian the sentence ‘The children have read the letter’ can be expressed as the following:

(SVO): децата прочетоха писмото. (3)

‘The children have read the letter’

(SOV): децата писмото прочетоха. (4)

‘The children the letter have read’

(OSV): писмото децата прочетоха. (5)

‘The letter the children have read’

(OVS): писмото прочетоха децата. (6)

‘The letter have read the children’

(VOS): прочетоха писмото децата. (7)

‘Have read the letter the children’

(VSO): прочетоха децата писмото. (8)

‘Have read the children the letter’.

Although SVO is the basic one, all permutations are possible; they are grammatically correct, even thought some are used mostly in poetry. According the Institute of Bulgarian at the Bulgarian Academy of Sciences, this grammatical particularity of permutation is possible because of the agreement between Subject and Verb which clarifies the role of Object.

As shown in (Fig.3), the grammatical relation between the first noun (concept X*) and verb V* vs. the relation between the second noun (concept Y*) and verb V* provide two distinct Merge-patterns. The first mental operator in this treatment is to merge nouns with V* in stages:

Stage I : Merge Subject-Verb

$$M\ (X^*V^*) = [X^*, V]\ (\underline{?O}) \qquad \mu\ X^* \in [X^*V] = X^* \text{ Acts } (\underline{?O}) \tag{9}$$

or

$$M\ (Y^*V^*) = [Y^*, V]\ (\underline{?O}) \qquad \mu\ Y^* \in [Y^*V] = Y^* \text{ Acts } (\underline{?O}) \tag{10}$$

Stage II : Merge Object with acting subject

$$M\ (\mu X^*, Y^*) = [\mu X^*, Y^*] \qquad \mu\ Y^* \in [\mu\ X^*, Y^*] = \text{Object } X^*\ V \tag{11}$$

or

$$M\ (\mu Y^*, X^*) = [\mu Y^*, X^*]\ \ \mu\ X^* \in [\mu\ Y^*, X^*] = \text{Object } Y^*\ V \tag{12}$$

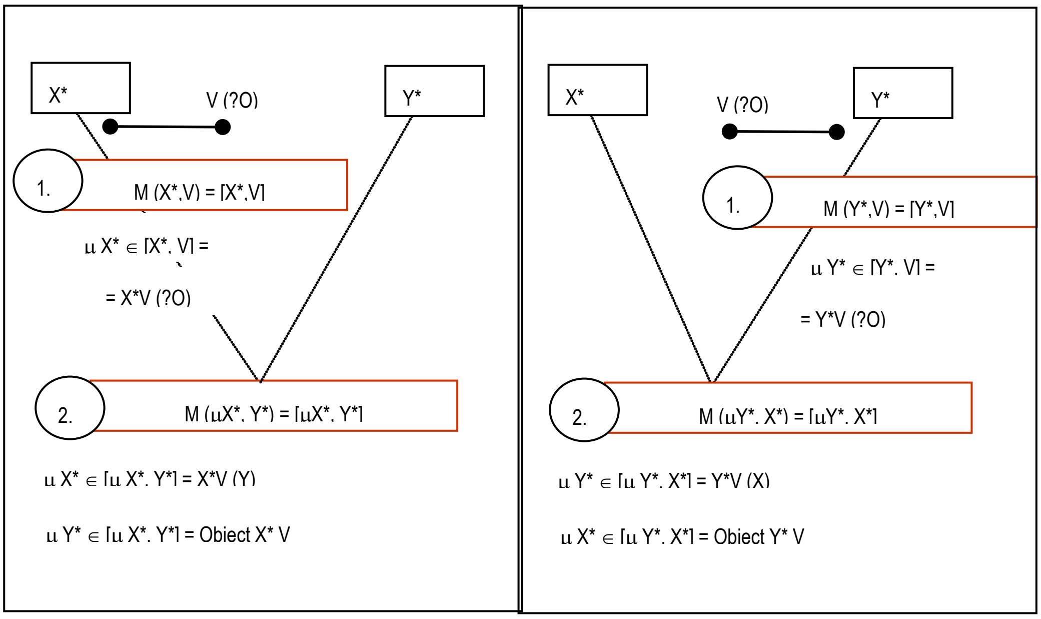


Fig 3. Stages of Semantic Merge.

The resulting mental representation μ is an image of the concept X* performing the action V. The second merge assigns to the second ('unbound) noun the role of Object of the already obtained image μ. However, there is a problem with this approach. What will happen if the verb is in agreement with *both* nouns that represent the concepts X* and Y*?

There exists in Bulgarian a particular grammatical operation (clitic doubling) that marks Object in cases of a reverse word order, as in:

(OVS): Иван ***го*** поздрави Мария. (means : Maria greeted Ivan) (13)

*'Ivan **him** greeted Maria'*

The clitic (го/ги) is obligatory only when the subject and the object are in the third person, and they are both either singular or plural. When the meaning is clear from the context, the clitic can be omitted, for example:

(OVS): Ролите озвучиха артистите: (followed by a list of the artists) (14)

The roles sound-screened the artists: (followed by a list of the artists)

'The artists (from the list) sound-screened the roles'.

There are no grammatical markers that allow one to distinguish Subject from Object in the above sentence. As shown in (Fig.4), both patterns of merge are grammatically possible. However, native Bulgarian speakers do not perceive the sentence as ambiguous; it is interpreted as having OVS word order.

In this paper, we assume that both operations are semantically supported by the meaning of the concepts X* and Y*. The absence of the Accusative Case marker which canonically marks Object in Slavic languages is compensated by means of other mechanisms. The cognitive basis of these mechanisms was discussed in (Slavova & Koujumdjieff, 2009). Cognitive analytism was defined there as the phenomenon of deciding upon meaning of ambiguous phrases, where there is no marking distinction whatsoever. The decision about the function (role) of the referent is made exclusively on the basis of its place in the cognitive (mental) space.

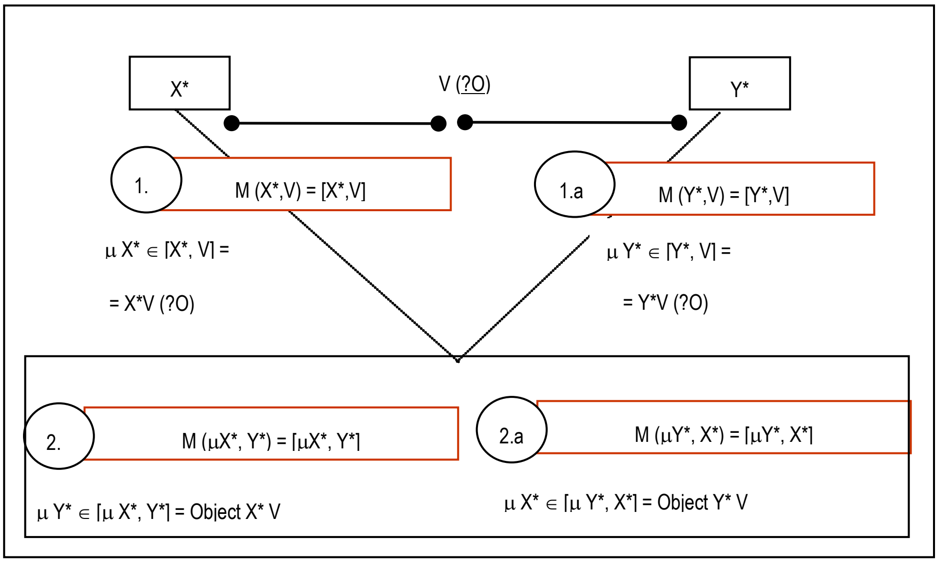


Fig.4. Double pattern of Semantic Merge.

EXPERIMENT: DESCRIPTION

The difficulty of designing an appropriate experiment is that mental computation runs on a deep (pre-linguistic) level and cannot be captured on the lexical level by standard experimental means. One possible way to extract information about the basic mental mechanisms is to induce ambiguity resolution on the lexical level, and then analyze the system's response.

The following experiment is based on the semantic ambiguity involved of certain Bulgarian sentences of the kind X(noun) Y(noun) V(verb). X and Y are nouns; V is in agreement with both X and Y. In this case, either X or Y can appear as Object:

$$M\ (X^*V^*) = [X^*, V]\ (\underline{?O}) \qquad \mu\, X^* \in [X^*V] = X^*\ \text{Acts}\ (\underline{?O}) \tag{15}$$

$$M\ (Y^*V^*) = [Y^*, V]\ (\underline{?O}) \qquad \mu\, Y^* \in [Y^*V] = Y^*\ \text{Acts}\ (\underline{?O}) \tag{16}$$

The following folk verses that contain sentences of the above kind were used in our experiment:

Живееше мишка, сива и красива	Живееше куче, сиво и красиво
Once upon a time there lived a mouse, grey and beautiful	*Once upon a time there lived a dog, grey and beautiful*
нейде на тавана, дето бе дивана.	нейде на тавана, дето бе дивана.
Somewhere in the attic, where the sofa was	*Somewhere in the attic, where the sofa was*
Появи се тук Котанчо, котаракът на Стоянчо.	Появи се тук Котанчо, котаракът на Стоянчо.
There appeared Kotantcho, the cat of Soyantcho	*There appeared Kotantcho, the cat of Soyantcho*
Мишката Котанчо хвана и изчезна под дивана.	Кучето Котанчо хвана и изчезна под дивана.
The mouse Kotantcho caught and disappeared under the sofa	*The dog Kotantcho caught and disappeared under the sofa*

The two sentences in the above verses (15) and (16) have identical structure X* Y* V* and ***there are no grammatical markers indicating which noun is Subject and which is Object.***

301.ex Мишката Котанчо хвана и изчезна под дивана. (17)

The mouse Kotantcho caught and disappeared under the sofa

304.ex Кучето Котанчо хвана и изчезна под дивана. (18)

The dog Kotantcho caught and disappeared under the sofa

We designed our experiment as a translation task (from Bulgarian to French). In contrast with Bulgarian, French the word order is fixed. Thus, our subjects had to assign a fixed word order to their French translations of our sentences, thus bringing out their interpretation of the same noun as either Subject or Object.

The experimenter asked the subjects to perform two tasks: 1) translate the verse (rhyming optional), 2) retell the story in two sentences. As in our previous experiment, the conditions were created where no attention was called to the ambiguity of the sentence(s).

The subjects of our experiment were 36 students with a variety of backgrounds (economists, sociologists, biologists, linguists, engineers, etc.). They were the students in the Masters program at the Francophone Institute for Management in Sofia, all of them fluent speakers of French and native speakers of Bulgarian. The Bulgarian verses were presented to them in a written form, on small separate sheets of paper.

EXPERIMENT: RESULTS AND ANALYSES

The results of the experiment confirmed once again that the relations between the images of the concepts in the conceptual nets influence the final result of the syntactic computation.

In sentences where ‘the dog’ appeared as the first noun and ‘Kotantcho’ (the cat) as the second noun, in both tasks all subjects assigned the Subject role to the first noun. All translations were structured as S V O (le chien a attrapé le chat); the word order was changed.

The dog Kotantcho (the cat) caught. (19)

the dog --> Subject	caught	the cat --> Object	17 (100%)

In sentences where ‘the mouse’ appeared as the first noun and ‘Kotantcho’ (the cat) as the second noun, in all cases of ‘*retell the story in two sentences’* condition the first noun was considered Object.

The mouse Kotantcho (the cat) caught. (20)

The cat --> Subject	caught	the mouse --> Object	12 (100%)

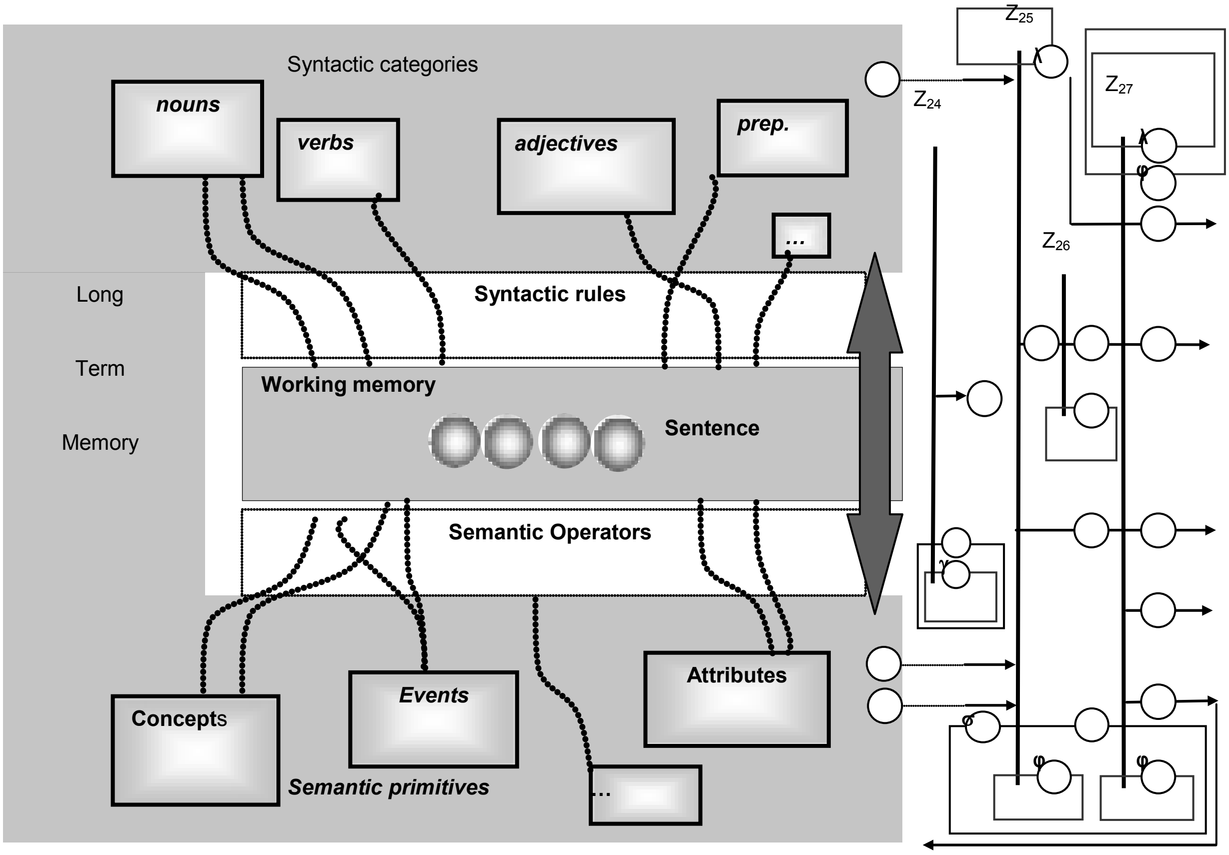
All translations were structured as S V O (le chat a attrapé la sourris); the word order was changed. The results of the translation task are as follows:

The mouse Kotantcho (the cat) caught. (21)

the cat --> Subject	caught	the mouse --> Object	16 (72,7%)
the mouse --> Subject	caught	the cat --> Object	4 (18,2%)
not clear --> Subject	caught	not clear --> Object	2 (9,1%)

It is clear that in the translation task the subjects were more confused as they attempted to respect the original word order. The cases which are not clear represent mot-à-mot translations, so the result in French does not make sense because of the word order. An attempt to respect the original word order is seen in a curious way in a couple of cases where the word order "subject first" is respected, but the sentences are translated in a passive form "*La sourris* ***a été*** *par Kotantcho* ***attrapée***" (The mouse has been caught by the cat).

Going back to the basic model (see Slavova, Soschen 2009) the language units (word-forms) have images as semantic primitives such as "concepts", "attributes", "events" etc, and the grammatical rules comply with semantic operations on these primitives. The detailed examination of the information flow using formal model, developed for simulating the cognitive process of natural language comprehension (Slavova 2004) has led to the suggestion that the procedures on the net must use semantic and syntactic knowledge in parallel (figure 5). Following this model, the cognitive system first assembles a fractional representation of the sentence-meaning structure (coupled words for example) and uses working memory loops for checking the semantic consistency.



.

Figure 5. Information treatment of a sentence, based on language and semantics

The grammatical features of the verb and the nouns in our sentences create two images:

Stage I Merge Subject-Verb (first noun merged with preference)

$M\ (Y^*V^*) = [Y^*, V]\ (\underline{?O}) \qquad \mu\, Y^* \in [Y^*V] = Y^*\ \text{Acts}\ (\underline{?O})$

$\boxed{\text{TheMouseCaught}}$ (22)

AND

$M\ (X^*V^*) = [X^*, V]\ (\underline{?O}) \qquad \mu\, X^* \in [X^*V] = X^*\ \text{Acts}\ (\underline{?O})$

$\boxed{\text{TheCatCaught}}$ (23)

Both images are semantically correct (as it is not “the cat flies” for example) and stored in working memory. So, both steps 1 and 1a from the merge-tree on figure 4 are performed.

Following the experiment, the result of one of the treatments is **rejected**. We suppose that this is done on the second step of the merge where object is assigned.

Stage II Merge Object with acting subject

$M\ (\mu Y^*, X^*) = [\mu Y^*, X^*] \qquad \mu\, X^* \in [\mu\, Y^*, X^*] = \text{Object}\ Y^*\ V$

The Cat is the Object of $\boxed{\text{TheMouseCaught}}$ (24)

AND

$M\ (\mu X^*, Y^*) = [\mu X^*, Y^*] \qquad \mu\, Y^* \in [\mu\, X^*, Y^*] = \text{Object}\ X^*\ V$

The Mouse is the Object of $\boxed{\text{TheCatCaught}}$ (25)

The rejection of (24) is done and (25) is accepted.

CONCLUSIONS

Once again, it was confirmed is that whether a particular noun is interpreted as either Subject or Object does not depend on the verb. For one and the same verb, the interpretation of a noun “switches” from one meaning to another. The analysis shows that the syntactic decision is influenced by the semantic images of the nouns themselves. The mental representation μ is an image of either the concept X* or the concept Y*, both involved in some action expressed as V. The second merge in (Fig. 4) assigns to the second (unbound) noun the role of Object of the already obtained image μ. We conclude that the operations in question are supported by the content of the concepts X* and Y* in their inter-relational semantic space.

In this work, the argument-centered approach shifts the focus from verb to the noun, from propositional to the non-propositional logic of grammar. The minimal building block that enters into linguistic computation is identified as symmetrical conjunct, or *a relation between individuals*. As a result, the true structure of language is characterized within a remarkably weak formal system, which is expected to develop into a more complex one to handle a broader range of data.

Bibliography

Chomsky, Noam (2006), Biolinguistic Explorations: Design, Development, Evolution, West Hall, Bathish Auditorium, AUB.

Hauser, M., N. Chomsky and W.T. Fitch (2002), The Faculty of Language: What is it, who has it, and how did it evolve? In: Science Vol. 298.

Slavova, Velina (2004) A generalized net for natural language comprehension. In: Advanced Studies in Contemporary Mathematics, vol 8, Ku-Duk Press, 131-153.

Slavova, V., and A. Soschen (2007), A Fibonacci-tree model of cognitive processes underlying language faculty, in: Proceedings of the 3-rd International Conference on Computer Science, NBU, University of Fulda, Boston University, pp. 196-205.

Slavova V. and Kujumdjieff T. (2009), The cognitive analytism of Bulgarian language, in: Text processing and cognitive technologies, No 14, proc. of the 11-rd international conference in Cognitive Modelling in Linguistics, Constanza, Romania, p.-p. in print, Conslanza, Romania, in print, 9 ;

Slavova V., and A. Soschen (2009), Experimental support of syntactic computation based on semantic merge of concepts,, in: International Journal Information theories & applications, International Journal "Information Technologies & Knowledge" Vol.3 / 2009

Soschen, Alona (2005). Derivation by phase: Russian applicatives. Canadian Linguistic Association Conference proceedings.

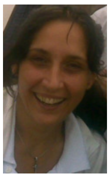
Soschen, Alona (2006). Natural Law and the Dynamics of Syntax (MP). Linguistics in Potsdam 25. Optimality Theory and Minimalism: a Possible Convergence? Hans Broekhius and Ralf Vogel (eds.): ZAS, Berlin.

Soschen, Alona (2008). On the Nature of Syntax. Bio-Linguistics Journal, vol. 2/2.

Soschen A., and V. Slavova (2007), Cognitive modeling of recursive mechanisms in syntactic processing. In: Proceedings of the IX international conference Cognitive Modeling n Linguistics, Text processing and cognitive linguistics, pp. 334-343.

Van der Velde, Frank and Marc de Kamps (2006), Neural Blackboard Architectures of Combinatorial Structures in Cognition, Behavioral and Brain Sciences, 29, pp. 37-70.

Authors' Information



Velina Slavova *- New Bulgarian University, Department of Computer Science, 21 Montevideo str., 1618 Sofia, Bulgaria, e-mail: vslavova@nbu.bg*

Major Fields of Scientific Research: Cognitive modeling in linguistics, formal models of language phenomena

Alona Soschen *- Massachusetts Institute of Technology, MIT Department of Linguistics and Philosophy, 77 Massachusetts Ave. 32-D808Cambridge, MA 02139-4307, USA, e-mail:* soschen@mit.edu

Major Fields of Scientific Research general features of biological systems present in linguistic structures

MULTILINGUAL REDUCED *N*-GRAM MODELS

Tran Thi Thu Van and Le Quan Ha

Abstract: *Statistical language models should improve as the size of the n-grams increases from 3 to 5 or higher. However, the number of parameters and calculations, and the storage requirement increase very rapidly if we attempt to store all possible combinations of n-grams. To avoid these problems, the reduced n-grams' approach previously developed by O'Boyle [1993] can be applied. A reduced n-gram language model can store an entire corpus's phrase-history length within feasible storage limits. Another theoretical advantage of reduced n-grams is that they are closer to being semantically complete than traditional models, which include all n-grams. In our experiments, the reduced n-gram Zipf curves are first presented, and compared with conventional n-grams for all Irish, Chinese and English. The reduced n-gram model is then applied for large Irish, Chinese and English corpora. For Irish, we can reduce the model size, compared to the 7-gram traditional model size, with a factor of 15.1 for a 7-million-word Irish corpus while obtaining 41.63% improvement in perplexities; for English, we reduce the model sizes with factors of 14.6 for a 40-million-word corpus and 11.0 for a 500-million-word corpus while obtaining 5.8% and 4.2% perplexity improvements; and for Chinese, we gain a 16.9% perplexity reductions and we reduce the model size by a factor larger than 11.2. This paper is a step towards the modeling of Irish, Chinese and English using semantically complete phrases in an n-gram model.*

Keywords: *Reduced n-grams, Overlapping n-grams, Weighted average (WA) model, Katz back-off, Zipf's law.*

ACM Classification Keywords: *I. Computing Methodologies - I.2 ARTIFICIAL INTELLIGENCE - I.2.7 Natural Language Processing - Speech recognition and synthesis*

Introduction

Shortly after this laboratory first published a variable *n*-gram algorithm by [Smith and O'Boyle, 1992], [O'Boyle, 1993] proposed a statistical method to improve language models based on the removal of overlapping phrases.

A distortion in the use of phrase frequencies had been observed in the small railway timetable Vodis Corpus when the bigram “RAIL ENQUIRIES” and its super-phrase “BRITISH RAIL ENQUIRIES” were examined. Both occur 73 times, which is a large number for such a small corpus. “ENQUIRIES” follows “RAIL” with a very high probability when it is preceded by “BRITISH.” However, when “RAIL” is preceded by words other than “BRITISH,” "ENQUIRIES" does not occur, but words like "TICKET" or "JOURNEY" may. Thus, the bigram "RAIL ENQUIRIES" gives a misleading probability that "RAIL" is followed by "ENQUIRIES" irrespective of what precedes it. At the time of their research, O'Boyle reduced the frequencies of "RAIL ENQUIRIES" by subtracting the frequency of the larger trigram, which gave a probability of zero for "ENQUIRIES" following "RAIL" if it was not preceded by "BRITISH." The phrase with a new reduced frequency is called a reduced phrase.

Therefore, a phrase can occur in a corpus as a reduced *n*-gram in some places and as part of a larger reduced *n*-gram in other places. In a reduced model, the occurrence of an *n*-gram is not counted when it is a part of a larger reduced *n*-gram. One algorithm to detect/identify/extract reduced *n*-grams from a corpus is the so-called reduced *n*-gram algorithm. In [O'Boyle, 1992], O'Boyle was able to use it to analyse the Brown corpus of American English [Francis and Kucera, 1964] (of one million word tokens, whose longest phrase-length is 30), which was a considerable improvement at the time. The results were used in an *n*-gram language model by O'Boyle, but with poor results, due to lack of statistics from such a small corpus. We have developed here a modification of his method, and we discuss its usefulness for reducing *n*-gram perplexity.

Similar Approaches and Capability

Recent progress in variable *n*-gram language modeling has provided an efficient representation of *n*-gram models and made the training of higher order *n*-grams possible. Compared to variable *n*-grams, class-based language models are more often used to reduce the size of a language model, but this typically leads to recognition performance degradation. Classes can alternatively be used to smooth a language model or provide back-off estimates, which have led to small performance gains. For the LOB corpus, the varigram model obtained 11.3% higher perplexity than the word-trigram model [Niesler and Woodland, 1996.]

[Kneser, 1996] built up variable-context length language models based on the North American Business News (NAB-240 million words) and the German Verbmobil (300,000 words with a vocabulary of 5,000 types.) His results show that the variable-length model outperforms conventional models of the same size, and if a moderate loss in performance is acceptable, that the size of a language model can be reduced drastically by using his pruning algorithm. Kneser's results improve with longer contexts and a same number of parameters. For example, reducing the size of the standard NAB trigram model by a factor of 3 results in a loss of only 7% in perplexity and 3% in the word error rate. The improvement obtained by Kneser's method depended on the length of the fixed context and on the amount of available training data. In the case of the NAB corpus, the improvement was 10% in perplexity.

Table 1. Comparison of combinations of variable *n*-grams and other Language Models.

COMBINATION OF LANGUAGE MODEL TYPES								
Basic *n*-gram	**Variable *n*-grams**	**Category**	**Skipping distance**	**Classes**	**#params**	**Perplexity**	**Size**	**Source**
Trigram√					987k	474	1M	LOB
		Bigram√			-	603.2		
		Trigram√			-	544.1		
	√	√			-	534.1		
Trigram√					743k	81.5	2M	Switch-board Corpus
	Trigram√				379k	78.1		
	Trigram√		√		363k	78.0		
	Trigram√		√	√	338k	77.7		
	4-gram√				580k	108		
	4-gram√		√		577k	108		
	4-gram√		√	√	536k	107		
	5-gram√				383k	77.5		
	5-gram√		√		381k	77.4		
	5-gram√		√	√	359k	77.2		

[Siu and Ostendorf, 2000] developed Kneser’s basic ideas further and applied the variable 4-gram, thus improving the perplexity and word error rate results compared to a fixed trigram model. They obtained word error reductions of 0.1 and 0.5% (absolute) in development and evaluation test sets, respectively. However, the number of parameters was reduced by 60%. By using the variable 4-gram, they were able to model a longer history while reducing the model size by more than 50% compared to a regular trigram model, and improved both the test-set perplexity and recognition performance. They also reduced the model size by an additional 8%.

Other related work are those of [Seymore and Rosenfeld, 1996]; [Hu, Turin and Brown, 1997]; [Blasig, 1999]; and [Goodman and Gao, 2000.]

In order to obtain an overview of variable *n*-grams, Table 1 combines all of their results.

Reduced N-Gram Algorithm

The main goal of this algorithm [Ha, Seymour, Hanna and Smith, 2005] is to produce three main files from the training text

- The file that contains all the complete *n*-grams appearing at least *m* times is called the *PHR* file ($m \geq 2$.)
- The file that contains all the *n*-grams appearing as sub-phrases, following the removal of the first word from any other complete *n*-gram in the *PHR* file, is called the *SUB* file.
- The file that contains any overlapping *n*-grams that occur at least *m* times in the *SUB* file is called the *LOS* file.

The final list of reduced phrases is called the *FIN* file, where

$$FIN := PHR + LOS - SUB \quad (1)$$

Before O'Boyle's work, a student Craig [O'Boyle, 1993] in an unpublished project used a loop algorithm that was equivalent to *FIN*:=*PHR*–*SUB*. This yields negative frequencies for some resulting *n*-grams with overlapping, hence the need for the *LOS* file.

There are 2 additional files

- To create the *PHR* file, a *SOR* file is needed that contains all the complete *n*-grams regardless of *m* (the *SOR* file is the *PHR* file in the special case where *m*=1.) To create the *PHR* file, words are removed from the right-hand side of each *SOR* phrase in the *SOR* file until the resultant phrase appears at least *m* times (if the phrase already occurs more than *m* times, no words will be removed.)
- To create the *LOS* file, O'Boyle applied a *POS* file: for any *SUB* phrase, if one word can be added back on the right-hand side (previously removed when the *PHR* file was created from the *SOR* file), then one *POS* phrase will exist as the added phrase. Thus, if any *POS* phrase appears at least *m* times, its original *SUB* phrase will be an overlapping *n*-gram in the *LOS* file.
-

The application scope of O'Boyle's reduced *n*-gram algorithm is limited to small corpora, such as the Brown corpus (American English) of 1 million words [Smith and O'Boyle, 1992], in which the longest phrase has 30 words. Now their algorithm, re-checked by us, still works for medium size and large corpora. In order to work well for very large corpora, it has been implemented by file distribution and sort processes.

By re-applying O’Boyle and Smith’s algorithm, Ha et al. [2005] investigated a reduced *n*-gram model for the Chinese TREC corpus of the Linguistic Data Consortium (LDC) (www.ldc.upenn.edu), catalog no. LDC2000T52. Later on [Ha, Hanna, Stewart and Smith, 2006] obtained reduced *n*-grams from two English large corpora and a Chinese large corpus. The two English corpora used in their experiments were the full text of articles appearing in the Wall Street Journal (WSJ) [Paul and Baker, 1992] of 40 million tokens respectively; and the North American News Text (NANT) corpus from the LDC (catalog no. LDC95T21 and LDC98T30) sizing 500 million tokens. Their employed Chinese corpus was the compound word version in [Ha, Sicilia-Garcia, Ming and Smith, 2003] of the Mandarin News corpus with 50,000 word types, originally from the LDC, catalog no. LDC95T13 of over 250 million syllables.

Reduced N-Grams and Zipf’s Law

By re-applying O’Boyle and Smith’s algorithm, we obtained the Zipf curves [Zipf, 1949] for the English, Chinese and Irish reduced *n*-grams.

The Irish is a highly-inflected Indo-European Celtic language. Both the beginning and end of words are regularly inflected. The Irish corpus in our experiments is taken from a corpus of 17th and 18th century Irish from the Royal Irish Academy (www.ria.ie) with sizes 7,122,537 tokens with 449,968 types [Harvey, Devine and Smith, 1994.]

We next present the Zipf curves for the English, Chinese and Irish reduced *n*-grams: All of our reduced *n*-grams were created on a Pentium II 586 of 512MByte RAM.

Wall Street Journal corpus (English)

The WSJ reduced *n*-grams can be created by the original O’Boyle-Smith algorithm for over 40 hours, the disk storage requirement being only 5GBytes.

The Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 1 showing all the curves have slopes within [-0.6, -0.5]. The WSJ reduced bigram, trigram, 4-gram and 5-gram curves become almost parallel and straight, with a small observed noise between the reduced 4-gram and 5-gram curves when they cut each other at the beginning. Note that information theory tells us that an ideal information channel would be made of symbols with the same probability. So having a slope of –0.5 is closer than –1 to this ideal.

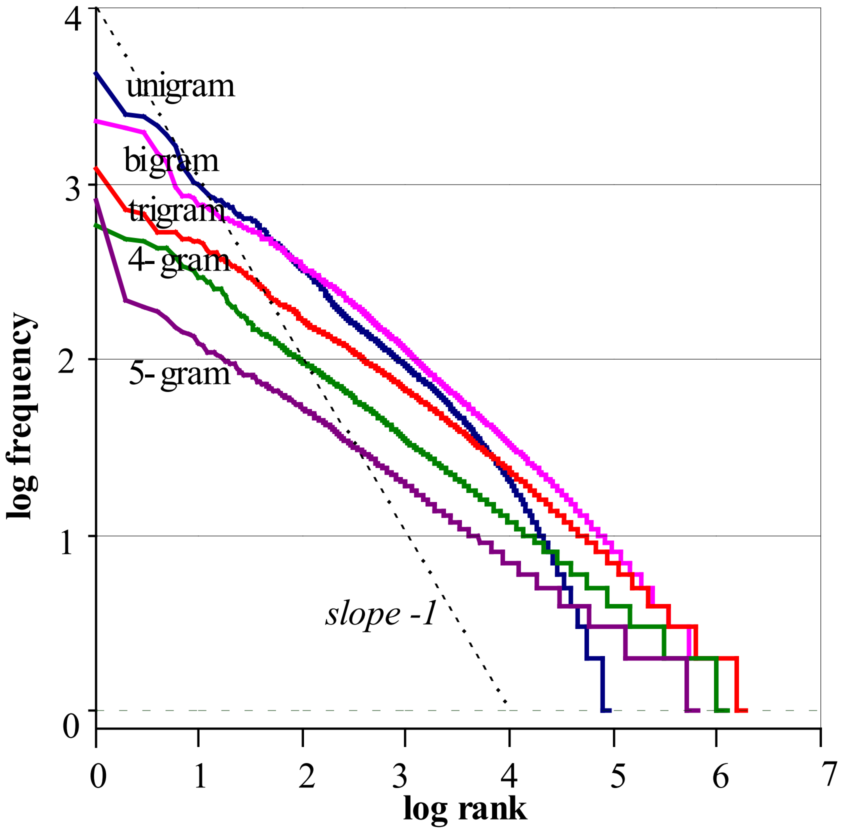


Figure 1. The WSJ reduced *n*-gram Zipf curves.

Table 2. Most common WSJ reduced *n*-grams.

Rank	Unigrams		Bigrams		Trigrams	
	Freq	Token	Freq	Token	Freq	Token
1	4,273	Mr.	2,268	he said	1,231	terms weren’t disclosed
2	2,469	but	2,052	he says	709	the company said
3	2,422	and	1,945	but the	664	as previously reported
4	2,144	the	1,503	but Mr.	538	he said the
5	1,918	says	1,332	and the	524	a spokesman for
6	1,660	or	950	says Mr.	523	the spokesman said
7	1,249	said	856	in addition	488	as a result
8	1,101	however	855	and Mr.	484	earlier this year
9	1,007	while	832	last year	469	in addition to
10	997	meanwhile	754	for example	466	according to Mr.

The conventional 10-highest frequency WSJ words have been published by [Ha, Sicilia-Garcia, Ming and Smith, 2002] and the most common WSJ reduced unigrams, bigrams and trigrams are shown in Table 2. It illustrates that the most common reduced word is not THE; even OF is not in the top ten. These words are now mainly part of longer *n*-grams with large *n*.

North American News Text corpus (English)

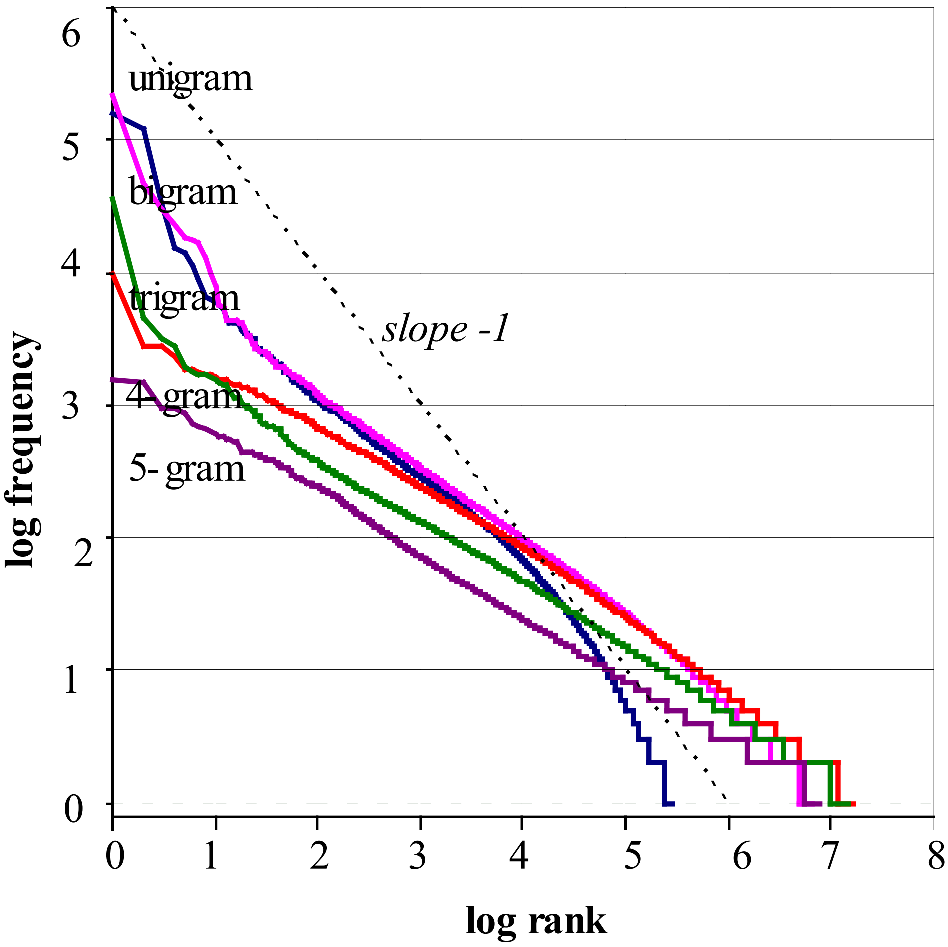


Figure 2. The NANT reduced *n*-gram Zipf curves.

The NANT reduced *n*-grams are created by the improved algorithm after over 300 hours processing, needing a storage requirement of 100GBytes.

Their Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 2 showing all the curves are just sloped around [-0.54, -0.5]. The reduced unigrams of NANT still show the 2-slope behavior when it starts with slope

–0.54 and then drop with slope nearly –2 at the end of the curve. We have found that the traditional *n*-grams also show this behaviour, with an initial slope of –1 changing to –2 for large ranks [Ha and Smith, 2004; Ferrer and Solé, 2002.]

Mandarin News compound words

The Mandarin News reduced word *n*-grams were created in 120 hours, using 20GB of disk space. The Zipf curves are plotted in Figure 3 showing that the unigram curve now has a larger slope than –1, it is around –1.2. All the *n*-gram curves are now straighter and more parallel than the traditional *n*-gram curves, have slopes within [-0.67, -0.5]. Usually, Zipf's rank-frequency law with a slope –1 is confirmed by empirical data, but the reduced *n*-grams for English and Chinese shown in Figures 1, 2 and 3 do not confirm it. In fact, various more sophisticated models for frequency distributions have been proposed by [Baayen, 2001] and [Evert, 2004.]

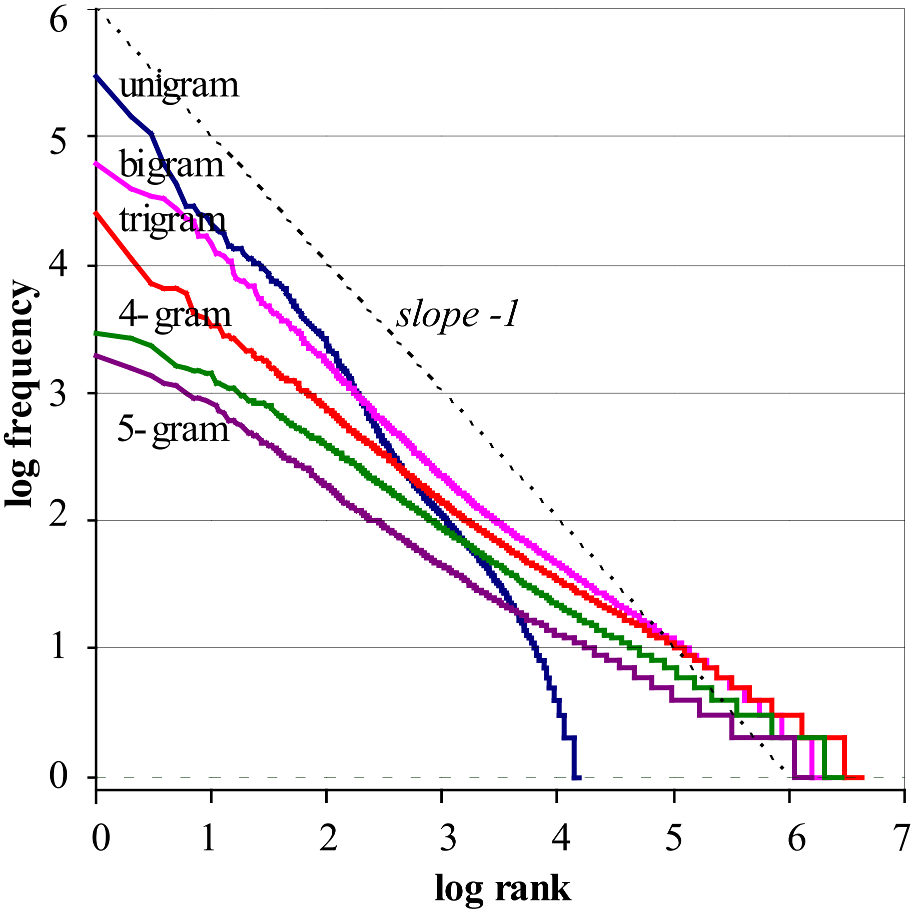


Figure 3. The Mandarin News reduced *n*-grams Zipf curves.

The Irish RIA corpus

The Irish reduced *n*-grams can be created by the original O'Boyle-Smith algorithm in 15 hours, the disk storage requirement being only 600 MBytes.

The most common Irish reduced and traditional unigrams are shown in Table 3 but nine words in the top ten are the same ones. This is very different from English and Chinese that only one or two most common words re-appeared within the top ten reduced unigrams. This fact is coming from the numerous word inflections in Irish language so that there are less overlapping Irish words and phrases.

The Irish Zipf curves are plotted for reduced unigrams and *n*-grams in Figure 4. Because of the word inflections, the Irish reduced unigram curve has a slope of -0.92, closer to the original Zipf's law than the previous English and Chinese reduced unigrams while the Irish reduced bigram, trigram, 4-gram and 5-gram Zipf curves have slopes within [-0.67, -0.45].

Table 3. The ten most common Irish traditional and reduced unigrams.

Rank	Traditional unigrams			Reduced unigrams		
	Freq	**Token**	**Meaning**	**Freq**	**Token**	**Meaning**
1	265,288	a	[relative particle]	74,505	a	[relative particle]
2	260,938	agus	and	71,145	an	the
3	259,093	do	to your	68,675	do	to your
4	253,247	an	the	45,063	na	the (pl.)
5	148,703	na	the (pl.)	37,252	agus	and
6	125,486	ar	on	34,006	go	that
7	111,170	go	that	31,398	ar	on
8	92,142	is	is	19,252	i	in
9	66,809	i	in	16,753	is	is
10	57,160	sin	that	15,991	#NO	[number]

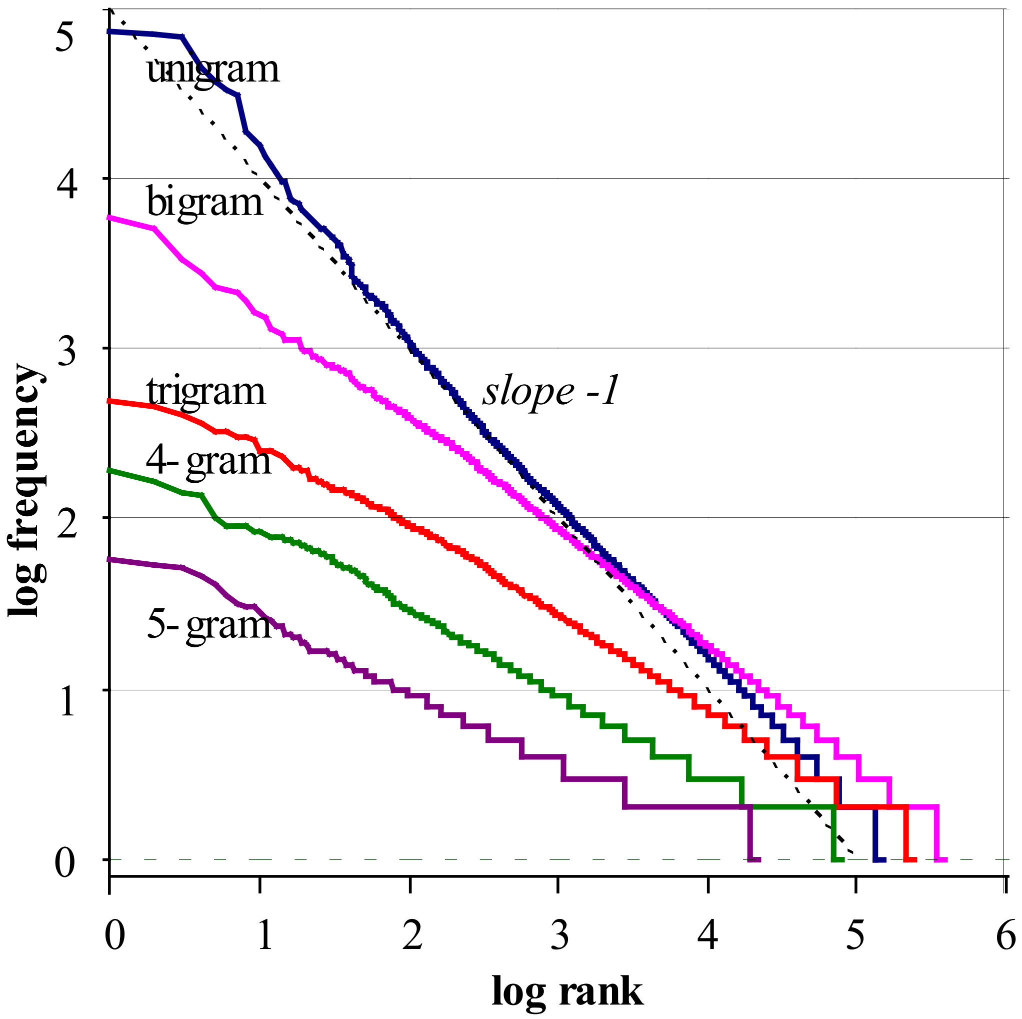


Figure 4. The Irish reduced *n*-gram Zipf curves.

Methods of Testing

The reduced *n*-gram approach was used to build a statistical language model based on the weighted average model of [O'Boyle, Owens and Smith, 1994.] We rewrite this model in formulae (2) and (3)

$$wgt\left(w_j^i\right) = \log\left(f\left(w_j^{i-1}\right)\right) \times 2^{i-j+1} \tag{2}$$

$$P_{WA}\left(w_i \middle| w_{i-N+1}^{i-1}\right) = \frac{wgt(w_i) \times P(w_i) + \sum_{l=1}^{N-1} wgt\left(w_{i-l}^i\right) \times P\left(w_i \middle| w_{i-l}^{i-1}\right)}{\sum_{l=0}^{N-1} wgt\left(w_{i-l}^i\right)} \tag{3}$$

This averages the probabilities of a word w_i following the previous one word, two words, three words, etc. (i.e. making the last word of an *n*-gram.) The averaging uses weights that increase slowly with their frequency and rapidly with the length of *n*-gram.

The probabilities of all of the sentences w_1^m in a test text are then calculated by the weighted average (WA) model

$$P\left(w_1^m\right) = P_{WA}\left(w_1\right)P_{WA}\left(w_2 \middle| w_1\right)...P_{WA}\left(w_m \middle| w_1^{m-1}\right) \tag{4}$$

and an average perplexity of each sentence is evaluated using Equation (5)

$$PP\left(w_1^m\right) = \exp\left(-\frac{1}{L}\sum_{i=1}^{L} Ln\left(P_{WA}\left(w_i \middle| w_1 w_2 ... w_{i-1}\right)\right)\right) \tag{5}$$

This weighted average model is a variable length model that gives results comparable to the Katz back-off method [Katz, 1987], but is much quicker to use.

Perplexity for Reduced N-Grams

[Ha et al., 2005, 2006] already investigated and analysed the main difficulties arising from perplexity calculations for our reduced model: a statistical model problem, an unseen word problem and an unknown word problem. Their solutions are applied in this paper also. Similar problems have been found by other authors, e.g. [Martin, Liermann and Ney, 1997]; [Kneser and Ney, 1995.]

The perplexity calculations of reduced *n*-grams includes statistics on phrase lengths starting with unigrams, bigrams, trigrams, etc. and on up to the longest phrase which occur in the reduced model.

The nature of the reduced model makes the reporting of results for limited sizes of *n*-grams to be inappropriate, although these are valid for a traditional *n*-gram model. Therefore we show results for several *n*-gram sizes for the traditional model, but only one perplexity for the reduced model. The perplexities of the WSJ reduced model by the weighted average model and the Katz Back-off method are shown in Table 4, North American News Text corpus in Table 5, Mandarin News words in Table 6 and the Irish RIA corpus in Table 7.

Table 4. Reduced perplexities for English WSJ.

Unknowns	**Tokens**	0	
	Types	0	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	762.69	762.69
	Bigrams	144.33	108.04
	Trigrams	75.36	59.71
	4-grams	60.73	53.16
	5-grams	56.85	52.84
	6-grams	55.66	51.61
	7-grams	55.29	51.10
Reduced Model by WA model		70.98	
%Improvement of Reduced Model on baseline WA trigrams		5.81%	
Model size reduction		14.56	

Table 5. Reduced perplexities for English NANT.

Unknowns	**Tokens**	24	
	Types	23	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	1,442.99	1,442.99
	Bigrams	399.61	339.26
	Trigrams	240.52	217.22
	4-grams	202.59	189.24
	5-grams	194.06	181.55
	6-grams	191.91	179.09
	7-grams	191.23	178.97
Reduced Model by WA model		230.46	
%Improvement of Reduced Model on baseline WA trigrams		4.18%	
Model size reduction		11.01	

Table 6. Reduced perplexities for Mandarin News words.

Unknowns	**Tokens**	84	
	Types	26	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	1,620.56	1,620.56
	Bigrams	377.43	328.32
	Trigrams	179.07	158.24
	4-grams	135.69	116.27
	5-grams	121.53	105.61
	6-grams	114.96	102.69
	7-grams	111.69	102.17
Reduced Model by WA model		148.71	
%Improvement of Reduced Model on baseline WA trigrams		16.95%	
Model size reduction		11.28	

Table 7. Reduced perplexities for Irish RIA corpus.

Unknowns	**Tokens**	36	
	Types	36	
Traditional Model	**Phrase length**	**WA model**	**Katz back-off**
	Unigrams	412.80	412.80
	Bigrams	162.99	144.81
	Trigrams	134.93	134.56
	4-grams	133.47	130.81
	5-grams	133.25	126.97
	6-grams	133.19	126.11
	7-grams	133.18	125.89
Reduced Model by WA model		78.75	
%Improvement of Reduced Model on baseline WA trigrams		41.63%	
Model size reduction		15.1	

In all cases of Irish, Chinese and English, their reduced models produced various perplexity improvements over the traditional 3-gram model (41.63% for Irish, 16.95% for Chinese and 4.18% for English). However, [Ha et al., 2006] obtained a significant reduction in model size, from a factor of 11.2 to almost 15.1 compared to the traditional Irish, Chinese and English model sizes. The Irish reduced model produces a perplexity improvement much better than previous results in English and Chinese languages and the reason is that the Irish language has numerous word inflections and inflected words' meanings are much related together.

For further work, we also need missing word tests.

Conclusion

The conventional *n*-gram language model is limited in terms of its ability to represent extended phrase histories because of the exponential growth in the number of parameters. To overcome this limitation, we have re-investigated the approach of [O'Boyle, 1993] and created a reduced *n*-gram model for Irish language. Our aim was to try to create an *n*-gram model that used semantically more complete *n*-grams than traditional *n*-grams in the expectation that this might lead to an improvement in language modeling. The good improvements in perplexity and model size reduction are better for Irish than for similar work by [Ha et al., 2005, 2006] in English and Chinese because Irish is a highly inflected language. It has numerous Irish word inflections while the inflected words' meanings are related together. So this represents an encouraging step forward, although still very far from the final step in language modelling.

Acknowledgements

The authors would like to thank the Royal Irish Academy for their Irish corpus. Our special acknowledgement is heart-fully sent to Prof Vladimir Polyakov and Prof Valery Solovyev.

Bibliography

[Baayen, 2001] H.R. Baayen. Word Frequency Distributions. Kluwer Academic Publishers, 2001.

[Blasig, 1999] R. Blasig. Combination of Words and Word Categories in Varigram Histories. In: ICASSP'99, Vol. 1, 529-532. 1999.

[Evert, 2004] S. Evert. A Simple LNRE Model for Random Character Sequences. In: Proc. of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles, 411-422. 2004.

[Ferrer and Solé, 2002] R. Ferrer I. Cancho and R.V. Solé. Two Regimes in the Frequency of Words and the Origin of Complex Lexicons. In: Journal of Quantitative Linguistics, 8(3):165-173. 2002.

[Francis and Kucera, 1964] N. Francis and H. Kucera. Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Department of Linguistics, Brown University, Providence, Rhode Island, 1964.

[Goodman and Gao, 2000] J. Goodman and J. Gao. Language Model Size Reduction by Pruning and Clustering. In: ICSLP'00. Beijing, China, 2000.

[Ha and Smith, 2004] L.Q. Ha and F.J. Smith. Zipf and Type-Token rules for the English and Irish languages. In: MIDL workshop. Paris, 2004.

[Ha, Hanna, Stewart and Smith, 2006] L.Q. Ha, P. Hanna, D.W. Stewart and F.J. Smith. Reduced n-gram Models for English and Chinese Corpora. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, 309-315. Sydney, Australia, 2006.

[Ha, Seymour, Hanna and Smith, 2005] L.Q. Ha, R. Seymour, P. Hanna and F.J. Smith. Reduced N-Grams for Chinese Evaluation. In: CLCLP, 10(1):19-34. 2005.

[Ha, Sicilia-Garcia, Ming and Smith, 2002] L.Q. Ha, E.I. Sicilia-Garcia, J. Ming and F.J. Smith. Extension of Zipf's Law to Words and Phrases. In: COLING'02, Vol. 1, 315-320. 2002.

[Ha, Sicilia-Garcia, Ming and Smith, 2003] L.Q. Ha, E.I. Sicilia-Garcia, J. Ming and F.J. Smith. Extension of Zipf's Law to Word and Character N-Grams for English and Chinese. In: CLCLP, 8(1):77-102. 2003.

[Harvey, Devine and Smith, 1994] A. Harvey, K. Devine and F.J. Smith. Archive of Celtic-Latin Literature ACLL-1 Royal Irish Academy. Dictionary of Medieval Latin from Celtic sources. Brespols, 1994.

[Hu, Turin and Brown, 1997] J. Hu, W. Turin and M.K. Brown. Language Modeling using Stochastic Automata with Variable Length Contexts. In: Computer Speech and Language, Vol. 11, 1-16. 1997.

[Katz, 1987] S.M. Katz. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. In: IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-35, 400-401. 1987.

[Kneser and Ney, 1995] R. Kneser and H. Ney. Improved Backing-off for M-Gram Language Modeling. In: ICASSP'95, Vol. 1, 181-184. Detroit, 1995.

[Kneser, 1996] R. Kneser. Statistical Language Modeling Using a Variable Context Length. In: ICSLP'96, Vol. 1, 494-497. 1996.

[Martin, Liermann and Ney, 1997] S.C. Martin, J. Liermann and H. Ney. Adaptive Topic-Dependent Language Modelling Using Word-Based Varigrams. In: EuroSpeech'97, Vol. 3, 1447-1450. Rhodes, 1997.

[Niesler and Woodland, 1996] T.R. Niesler and P.C. Woodland. A Variable-Length Category-Based N-Gram Language Model. In: ICASSP'96, Vol. 1, 164-167. 1996.

[Niesler, 1997] T.R. Niesler. Category-based statistical language models. St. John's College, University of Cambridge, 1997.

[O'Boyle, Owens and Smith, 1994] P. O'Boyle, M. Owens and F.J. Smith. A weighted average N-Gram model of natural language. In: Computer Speech and Language, Vol. 8, 337-349. 1994.

[O'Boyle, 1993] P.L. O'Boyle. A study of an N-Gram Language Model for Speech Recognition. PhD thesis. Queen's University Belfast, 1993.

[O'Boyle, McMahon and Smith, 1995] P. O'Boyle, J. McMahon and F.J. Smith. Combining a Multi-Level Class Hierarchy with Weighted-Average Function-Based Smoothing. In: IEEE Automatic Speech Recognition Workshop. Snowbird, Utah, 1995.

[Paul and Baker, 1992] D.B. Paul and J.B. Baker. The Design for the Wall Street Journal based CSR Corpus. In: Proc. of the DARPA SLS Workshop, 357-361. 1992.

[Seymore and Rosenfeld, 1996] K. Seymore and R. Rosenfeld. Scalable Backoff Language Models. In: ICSLP'96, 232-235. 1996.

[Siu and Ostendorf, 2000] M. Siu and M. Ostendorf. Integrating a Context-Dependent Phrase Grammar in the Variable N-Gram framework. In: ICASSP'00, Vol. 3, 1643-1646. 2000.

[Siu and Ostendorf, 2000] M. Siu and M. Ostendorf. Variable N-Grams and Extensions for Conversational Speech Language Modelling. In: IEEE Transactions on Speech and Audio Processing, 8(1):63-75. 2000.

[Smith and O'Boyle, 1992] F.J. Smith and P. O'Boyle. The N-Gram Language Model. In: The Cognitive Science of Natural Language Processing Workshop, 51-58. Dublin City University, 1992.

[Zipf, 1949] G.K. Zipf. Human Behaviour and the Principle of Least Effort. Reading, MA: Addison-Wesley Publishing Co., 1949.

Authors' Information

Tran Thi Thu Van *– Lecturer, Hochiminh City University of Technology HUTECH, 41/32 Le Duc Tho, Ward 16, Go Vap District, Hochiminh City, Vietnam; e-mail: Nlp.Sr@Shaw.ca*

Major Fields of Scientific Research: Natural language processing, Speech recognition

Le Quan Ha *– PhD Research Assistant, PhD, School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom; Vice-Head of Computer Science, Faculty of IT, Hochiminh City University of Industry, Ministry of Industry and Trade, Vietnam; contact: 8 Bermondsey Court N.W., Calgary, Alberta T3K 1V7, Canada; e-mail: lequanha@hui.edu.vn, lequanha@fit-hui.edu.vn*

Major Fields of Scientific Research: Natural language processing, Speech recognition

TABLE OF CONTENT