Collocational networks and their application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci)

Araceli Alonso, Chrystel Millon, Geoffrey Williams
Equipe LiCoRN - Université de Bretagne-Sud
Faculté de Sciences Humaines et Sociales
4 rue Jean Zay 56321 Lorient CEDEX (France)
E-mail: araceli.alonso@univ-ubs.fr, chrystel.millon@univ-ubs.fr, geoffrey.williams@univ-ubs.fr

Abstract

The present article deals with a situation that lies between the needs of an advanced learner's dictionary and those of a specialised dictionary in attempting to build a pattern dictionary for verbs which are being used in scientific research papers. Current dictionaries do not necessary assist in the particular production environment of the scientific article. This can be tackled by building a bottom-up phraseological dictionary which will help both with decoding and encoding. The building method uses collocational networks in order to compile a dictionary which will demonstrate usage of individual verbs, grouping them into a natural classification system that will grow from the corpus data. This organic dictionary ultimately makes wide use of mind mapping technology to allow the user to navigate within the dictionary. It contains both individual entries containing phraseological information and super entries linking quasi-synonyms and writing assistance. The dictionary provides the environment which can link phraseological patterns to the corpus data so as to limit the information retrieval process whilst providing real examples of language in use in specialised contexts.

Keywords: learner's dictionary; specialised dictionary; organic dictionary; phraseology; collocational networks; verbal patterns

1. Introduction

In recent years, developments in technology have brought about some major changes in dictionary-writing. The ground-breaking work of Sinclair and the COBUILD team in the 1980s introduced a move in lexicographical practice towards the creation of corpus-based dictionaries on the basis that users need to know not only the meaning of the word, but the way the word is used in context. Many monolingual and, especially, learner's dictionaries have applied corpus-based techniques for representation of word uses by giving examples taken from a corpus. Although the corpus is now integrated as a source, most of these dictionaries, whether print or cd-rom in format, have not implemented the full potential of adopting a corpus-driven approach to what may be extracted from a corpus, such as the networks of relations between words.

The rise of electronic dictionaries due to the widespread use of computers and especially of Internet has also contributed to pushing lexicographical practice further, even if technology changes much more quickly than the dictionary-writing process. As a result, many on-line dictionaries do not take full advantage of the potential offered by web technology. In fact, many of them are just a copy of the paper dictionary. Some attempts at creating genuine on-line dictionaries, such as, the visual dictionaries, or *Wordnik* (http://www.wordnik.com), based on the web 2.0 or social web, and the like, have been made, but there is still a long way to go. New approaches to dictionary-making practice are needed. In a society of knowledge and technology, dictionaries must be updated and adapted to the users' needs.

In the case of science dictionaries, there is a real need for innovation. Most dictionaries of science are very traditional in outlook or simply take the form of terminological databases applying an onomasiological approach which supply the user with a definition and a context, and in some cases, relations between the units, but fail to give detailed information on the syntagmatic and paradigmatic relations between technical words, and technical words and 'general' words. In reality, specialised communication is not just about technical words. In most cases, scientists already know the definition of the technical word, but look up the 'specialised' meaning of a general word in the dictionary, for getting information of the behaviour of the word in a domain-specific context.

The wealth of language lies in semitechnical words and general words in specific contexts. As has been stated by many authors (Cabré, 1999; Meyer, 2000; Ciapuscio, 2003; Hunston & Sinclair, 2003; ten Hacken, 2008), the dichotomy between general and specialised languages must be seen in terms of a continuum; they are not clearly separable entities. It can be stated that there is a transfer of lexical units from one side to the other; processes of *determinologization* or *banalization*, *terminologization* and *pluriterminologization* take place (Meyer, Mackintosh & Varantola, 1997; Cabré, 1999). This passage of meaning potentials from general language to specialised language, and back, is particularly a problem for non-native speakers who need to communicate in scientific contexts.

Furthermore, most specialised dictionaries consider only nouns as entries of the dictionary, as according to a classical perspective of terminology, the noun was considered the only category to have a terminological value, and do not take into account the role that other categories, such as verbs, can play in specialised discourse. In order to produce a text, non-native speakers need to understand the characteristics of the specialised discourse and it is not only the noun which plays a relevant role. Verbs, for instance, can help to organize the discourse, to articulate and structure the text, to establish links between different referential lexical units, to express the point of view of the author, to interactuate with the reader, to understand the meaning of a word, etc. As Hanks states '*meanings are constructed around the verb, the pivot of the clause*' (Hanks, 2010a:3). Therefore, for a language learner, it is extremely important to get to know the behaviour and use of verbs in order to be able to produce and understand a specialised discourse. A dictionary for verbs used in the sciences can assist by helping users to overcome their basic communication problems.

The main objective of this communication is to present the potential of collocational networks for a new approach to an experimental dictionary conceived from the beginning as a virtual dictionary, the *E-Advanced Learner's Dictionary of Verbs in Science* (DicSci). Collocational networks for the building-up of dictionary entries will be discussed and exemplified with reference to the most frequent verbs extracted from a corpus related to BioSciences.

This paper shows how specialised learner's dictionaries have evolved. The article presents the initial premises of the lexicographical project *DicSci*, paying special attention to the 'organic' nature of the *E-Advanced Learner's Dictionary of Verbs in Science*, and describes the work methodology and building-up of the dictionary by showing the verb *to take* as an example. Finally, some conclusions and perspectives are outlined.

2. Learner's dictionaries of science

Dictionaries of science or specialised dictionaries are usually terminologically based which have been elaborated taking into account terminological theoretical and methodological framework rather than those of lexicography, particularly those of advanced learner's dictionaries. In many cases, they are terminological databases. Most of these terminological dictionaries or databases are based on an onomasiological perspective, that is, the different entries are organized by means of the concepts. The terminological units are just means of the linguistic expression of the conceptual organisation of a particular domain. The focus is on explaining the concept, with the terminological unit is only observed as a way of designing the concept. Therefore, no attention is being paid to the different senses of a term, as the term is not considered as a lexical unit. More recent approaches to terminology advocate a semasiological approach to terminology — see L'Homme (2005) for more detailed information —, considering the term as a lexical unit which can have the same characteristics of other lexical units of general language. Despite this fact, little progress has been made in specialised dictionary practice. As Williams (2003:94) states, most of these dictionaries, whether multilingual or bilingual terminologies, mostly address the translator, whereas the monolingual encyclopaedic dictionaries principally address the subject specialist. The latter are prescriptive dictionaries whose main aim is to fix and explain terms for native-speakers of the language. They have not been compiled with the foreign learner's needs in mind. They do not explain use of terms in context and, therefore, are of little help for encoding purposes. In the 80s Moulin (1983:151) already considered that existing specialised dictionaries were of little use to foreign learners. Not much progress has been made since then.

Even though, some attempts have been made during these years to answer foreign learner's needs, learner's dictionaries of science are not really satisfactory. Some authors (Bergenholtz & Tarp, 1995; Fuertes-Olivera, 2009, 2010; Tarp, 2008) have defended a functional approach to lexicography, usually referred to as the *Function Theory of Lexicography*, considering lexicography as an area of social practice where the dictionary must take into account users' specific types of problems and situations and satisfy user's needs. From this perspective some specialised dictionaries for foreign learners have been compiled. And even though more attention has been paid to the linguistic characteristics of the terminological units, many problems of grammar and usage have received only minimal attention.

On the other hand, learner's dictionaries of English as a foreign language have a strong tradition, but aim at general usage with little coverage of the sciences. Over the years, learner's dictionaries of English as a foreign language have increased in number and variety. Since the first learner's dictionaries much work has been done for giving more information — see Cowie (2002, 2009) for a detailed history of English dictionaries for foreign learners —, paying special attention to the linguistic features of language. However, most advanced learner's dictionaries have paid little attention to the representation of specialised lexical units, being primarily aimed at learners of the language for general purposes. A similar situation can be found with standard bilingual dictionaries which essentially provide decontextualized equivalents with a minimum of encoding assistance. Consequently, many scientists have to rely on 'native English speakers', hopefully with an awareness of genre specificities, to correct their texts.

Learner's dictionaries of English as a foreign language deal with grammatical and usage aspects of lexical units, as the learner of the language need information not only

for understanding texts, but also for producing texts in the foreign language. For instance, many dictionaries have made an attempt to introduce information on collocations, "lexical co-occurences of words" (Sinclair, 1991:170), in order to give more information about the use of words in context, taking into consideration information extracted from corpora. This has also been recognized as a useful addition to specialised dictionaries, especially in relation to the user's needs for encoding. However, as explained by L'Homme & Leroyer (2009:259) "there does not seem to be a general agreement as to what types of word combinations should be listed, nor as to how they should be presented in specialised reference works."

A learner's dictionary of science must be a tool for an ongoing learning process where specific collocations and lexical patterns can help non-native speakers who need to produce scientific texts in English. To do this, we propose to use a bottom up model to create an experimental dictionary dealing with verbs used in scientific texts. We pay our attention to the verbal category, as verbs are the centre of the clause which link nodes of specific terminology and are of phraseological interest.

From a classical perspective of terminology, verbs were not considered of interest as they were not proper terminological units. Recent approaches to terminology have shown that not only does the nominal category can have a terminological value, but that other categories, such as adjectives or verbs, can also be domain-specific lexical units — see Lorente (2007, 2009) for more information. According to Lorente (2009:59) verbs are not *per se* terminological units, but can acquire a 'specialised value' in context when their immediate environment also provides specialised knowledge. Lorente (2007) establishes a classification of verbs used in scientific texts: a) *verbos casi-términos* ('near-term verbs'), such as *to ionize*; b) *verbos fraseológicos* ('phraseological verbs'), such as *to codify* (i.e *codify a protein*); c) *verbos de relación lógica* ('verbs of logic relation'), such as *to present*; d) *verbos performativos del discurso* ('verbs performative of discourse'), such as *to conclude*.

As it can be observed by Lorente's classification, in most cases, the 'specialised value' of a verb is determined by the company it keeps. As Hanks (2010b) establishes, taking into consideration Sinclair's distinction (Sinclair, 1991) between the *open-choice principle* and the *idiom principle*, many units have both a terminological tendency (*open-choice principle*) and a phraseological tendency (*idiom principle*). Verbs have mainly a phraseological tendency. It is impossible to know the meaning of some of these verbs without knowing the phraseological context in which the verb is used. This phraseological context is the information to which a learner of the language needs to pay particular attention. The difficulty of the learner of science is in the phraseology being used and not in the designation of a concept. An advanced specialised learner's dictionary must pay special attention to those units with a phraseological tendency. In order to write scientific texts in a foreign language, the learner of the language needs to know the meaning of the specific words used in specific contexts and, as it has been mentioned before, there are many words whose meaning can only be understood by knowing the environment where the word is used.

The DicSci is an advanced learner's dictionary of verbs whose main aim is to give account of the functioning of a verb in an scientific context, showing its phraseological behavior, taking into account its collocates and its textual environment.

3. DicSci – An E-Advanced Learner's Dictionary of Verbs in Science

The lexicographical project *DicSci* starts off from ongoing work that is both theoretical and practical in nature related to two research projects on science corpora coupled with and analyse of the place of scientific usage in advanced learner's dictionaries and the application of the methodology of *collocational networks* and *collocational resonance* (Williams, 1998, 2002, 2003, 2006, 2008a, 2008b, 2008c; Williams & Millon, 2009, 2010) and that of the technique developed by Patrick Hanks — see Hanks (2004, 2006), Hanks & Ježek (2008) for more detailed information —, named *Corpus Pattern Analysis* or CPA and supported by the *Theory of Norms and Exploitations* or *TNE* (Hanks, forthcoming).

On the theoretical side the objective of the lexicographical project is to show how collocational networks, collocational resonance and lexical patterns can assist with understanding not just meaning change, but the carry-over of aspects of meaning from changing contextual environments, and also the relations between the technical and the general lexical units. The practical final outcome is an *E-Advanced Learner's Dictionary of Verbs in Science* (DicSci) built bottom-up using corpus-driven methodologies both for selection of headwords, semantic organisation of the data, representation of norms and exploitations, word syntagmatic and paradigmatic relations and movement of meanings between contexts.

The working methodology is based on the use of collocational networks and collocational resonance. This can be further enhanced by applying *Corpus Pattern Analysis* or CPA. In previous studies (Alonso, 2009; Williams, 1998, 2002, 2006, 2008a, 2008b, 2008c; Williams & Millon, 2009, 2010), these statistically based chains of collocations have been used to demonstrate thematic patterns in texts, as well as means for selecting the lexis for a specialised language dictionary, for observing the movement of meanings between contexts,

establishing syntagmatic and paradigmatic relations between units and determining the difference between the 'specialised' and 'general' language use.

The methodology proposed is influenced by John Sinclair's insights into collocation and the idiom principle (Sinclair, 1991), Wittgenstein's approach to prototypes (1953), the work on scientific texts developed by Roe (1977) and the later studies of the phraseology of scientific texts developed by Gledhill (2000), the work on pattern grammar by Hunston & Francis (1999), the study on semantic prosody by Louw (1993, 2000|2008), the theory of Lexical Priming proposed by Hoey (2005). Finally, as has been shown in previous studies (Alonso, 2009; Renau & Alonso, in press), the application of *Corpus Pattern Analysis* proposed by Hanks (2004) for building-up a *Pattern Dictionary of English Verbs* (PDEV)[1] seems to be useful for analysing the normal use of the lexical units in scientific contexts and establishing differences between the general and specialised use of a lexical unit, as well as it can help to improve the dictionary entry as it provides a systematic and very fine analysis of language in use. CPA is a technique which can complement the information given by the collocational networks.

Collocational networks, proposed by Williams (1998, 2002), are statistically based chains of collocations, a web of interlocking conceptual clusters realised in the form of words linked through the process of collocation. The idea that collocations "cluster" forming interwoven meaning networks comes from Phillips (1985). Phillips's aim was the study of metastructure within texts and the notion of 'aboutness'. Williams (1998) considered Phillip's work and hypothesised that "the patterns of co-occurrence forming the collocational networks will be unique to any one sublanguage and serve to define the frames of reference within that sublanguage" (Williams 1998:157). From a high frequency lexical unit, considered as node of the network, the collocates are calculated using a statistical measure, mainly MI or Z-Score, even though other statistical measures can be considered. The collocates are then treated as nodes and the collocates of each collocate is then calculated. The network will be allowed to extend through collocational chains until a point is reached where either no more significant collocates are found or where a word-form that has occurred earlier in the network is encountered. A detailed description of the procedure for the creation of collocational networks is shown in Williams (1998).

It must be taken into account the importance of the statistical measure selected for calculating the more significant collocates of a lexical unit, as different measures will give different results. For instance, Mutual Information displays more rarer items whereas Z-score gives more general collocates — see Church & Hanks (1990) for more information on measuring word association norms. It is also important to bear in mind that the collocational network can vary depending on the form of the lexical unit. For instance, in texts related to Molecular Biology, the environment developed by the use of 'gene' in singular is quite different to the environment of the form in plural:

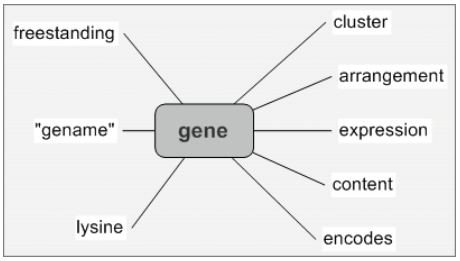

Figure 1: First level of the collocational network of 'gene' extracted from Williams (2008c:140)

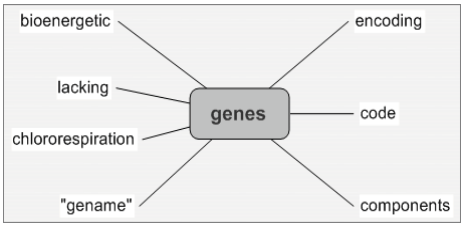

Figure 2: First level of the collocational network of 'genes' extracted from Williams (2008c:140)

Despite the different collocates associated to each form of the lexical unit, the lemmatised network must be also considered in order to have a complete panorama of the total environment of a word.

On the other hand, collocational resonance is also a tool being used at DicSci to show how elements of meaning are carried over from on textual environment to another. The mechanism of collocational resonance has been described in Williams (2008b) and Williams & Millon (2009). The notion of collocational resonance is based on the assumption that language users carry aspects of meaning from previously encountered usage, consciously and subconsciously, subcategorised for topic and genre, coloning the meanings and prosodies in use. This can be mapped by using lexicographical prototypes. For instance, if we consider the word 'culture', one of its meanings is that of *farming*. When 'we culture children', there are pieces of meaning that still carry a resonance of the meaning of 'culture' as *farming*. A detailed explanation of resonance with reference to the word 'probe' can be found in Williams & Millon (2009). Collocational resonance is used to explain particular patterns of usages. It can assist in understanding the movement from general to specialised usage of language, or from specialised to general. It can also help to build up the definition of dictionary entries. In the present

[1] The *Pattern Dictionary of English Verbs* (PDEV) is an ongoing project whose first results are free available on the Internet (http://deb.fi.muni.cz/pdev).

study we concentrate on collocational networks rather than on collocational resonance, as collocational network is the primary tool for building-up the dictionary DicSci.

The third element of the work methodology in compiling DicSci is the use of *Corpus Pattern Analysis* to give a more accurate account of the normal uses of each of the significant collocates which form the collocational network. CPA is a work-in-progress corpus-driven methodology developed by Hanks for 'mapping meanings onto use' (Hanks, 2002). According to Hanks (2010c:590), "a corpus does not show directly what a word means, but it provides evidence on the basis of which meanings can be inferred." It provides evidence on the word use. Most of these uses are highly patterned. Each unique pattern is usually associated with a specific meaning. CPA is a methodology for identifying prototypical syntagmatic patterns with which words in use are associated. As Hanks (2006:1165) explains "a pattern consists of a verb with its valencies, plus semantic values for each valency and other relevant clues, and is associated with an implicature that associates the meaning with the context rather than with the word in isolation." A pattern is based on the structure of English clause roles described in systemic grammar (Halliday, 1961) — subject, predicator, object, complement, adverbial. Each clause role or argument is 'populated' by a set of collocations. The more significant collocates of a verb are usually nouns which share a semantic aspect of meaning. The meaning of a group of collocates is expressed by a semantic type. Using Hanks' words, semantic types represent 'folk concepts.' All semantic types are stored in a hierarchically structure shallow ontology which is continuously under review. The CPA ontology is corpus-driven. There are cases in which the argument slot is populated by one or more lexical items which cannot be grouped together into semantic types; these are considered as lexical sets. In other cases, the semantic type is complemented by a semantic role. The semantic type is an intrinsic property of the collocate, while a semantic role is an extrinsic property assigned by context. For instance, if we consider the verb *to filter*, one normal pattern would be [[Human]] *filter* [[Liquid]]. However, the corpus can show cases in which not all kinds of liquids are filtered but only some specific ones, such as *water*. The pattern in this case would be [[Human]] filter [[Liquid=Water]]. The organisation of semantic types and semantic roles is not easy and it is only by corpus evidence that this task can be achieved. For more detailed information on the general principles of CPA, see Hanks (2004, 2006, 2010a); for an explanation of the CPA ontology, see Ježek & Hanks (2010).

As mentioned before, the DicSci is a corpus-driven dictionary which takes into account the use of words in scientific texts. Therefore, it is obvious that a corpus of scientific texts is needed. To begin with the building-up of the dictionary a corpus was compiled, the BioMed Central corpus (BMC). The BMC is a 33-million-word English language built as part of the Scientext initiative. The Scientext initiative was a project for the creation of comparable corpora carried out by a consortium of three French universities led by the Université de Grenoble 3. The BMC corpus, which is now freely online at the Scientext website[2], stands at 33 million words drawn from 8945 scientific texts from 137 different journals, made freely accessible online by the independent publishing house BioMed Central[3]. The texts have been selected from a number of journals dating from 1997 to 2005. All texts have been formatted according to the TEI guidelines and have been part-of -speech tagged and lemmatized using *Treetagger*[4]. The texts in the BMC corpus encompass a large number of topics and genres, all related to two main areas: biology and medical research. Each text has been informed with XML-TEI annotation to which topic(s) and to which genre is belonged.

The corpus cannot be considered as fully representative of published scientific research, as it is focused on articles related to Biosciences. The distribution of topics and genres is not well-balanced, as stated in Williams & Millon (2009). In the present work, however, the subcategorisation of the corpus has not been exploited. Despite the limitations of the corpus, due to its size the BMC corpus provides adequate data for work on an experimental dictionary such as DicSci. More details about the corpus can be found on the Scientext website.

Finally, the experimental dictionary presented is considered an 'organic' dictionary. It is 'organic' in the sense that it refers to a living dictionary that will organised itself in a natural way thanks to the links between words shown by means of collocational networks. Collocational networks are used for headwords selection, for structuring and classifying verbs together into classes and as means of navigation. This dictionary will ultimately make wide use of mind mapping technology to allow user navigate within the different entries. The dictionary will provide the environment which can link phraseological patterns to the corpus data whilst providing real examples of language in use in specialised contexts. In the following chapter the use of collocational networks for building-up our dictionary is illustrated through the exploration of the verb *to treat*.

4. Collocational networks and dictionary making: the verb to treat

To treat is the 49th most frequent verb in the BMC corpus with 13018 occurrences. The collocational network was created by measuring the most significant collocates of the verb. Due to space restrictions, Figure 3 below shows

[2] http://scientext.msh-alpes.fr/scientext-site/?article30
[3] http://www.biomedcentral.com
[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

only the first level of the collocational network of the verb *to treat*, as the main aim in this paper is to demonstrate the principles, not to expose full networks. This network contains the eight most statistical significant noun collocates of *to treat*, namely *animal*, *rat*, *mouse*, *patient*, *intention*, *control*, *vehicle* and *cell* — showed in red in Figure 3 —, and the first ten most statistical significant verb collocates of each of the nouns. The collocates are calculated by means of Z-score in a span of 5:5, and the collocations that have less than 3 occurrences are kept out. Yet, five verbal collocates were removed from the network, that are *deciduoma-bearing*, *coimmunized*, *frequency-matched*, *transfected*, and *exhaust*. The first four are word-forms not recognized by the *Treetagger tool*, and the last one *exhaust* was removed because in the noun-verb collocation 'vehicle exhaust', *exhaust* correspond to a noun which belongs to the syntagmatic lexical unit *motor vehicle exhaust*.

In total, 54 verb collocates have been considered for the network. Among them, seven are amongst the 100 more frequent verbs in the BMC corpus: *compare*, *express*, *grow*, *include*, *receive*, *stain* and *use*. Moreover, there are eight verbs (without counting *treat*) that are shared by some of the seven noun collocates, namely *anesthetize*, *compare*, *feed*, *immunize*, *inject*, *receive*, *sacrifice*, and *stain* — marked in green in Figure 3.

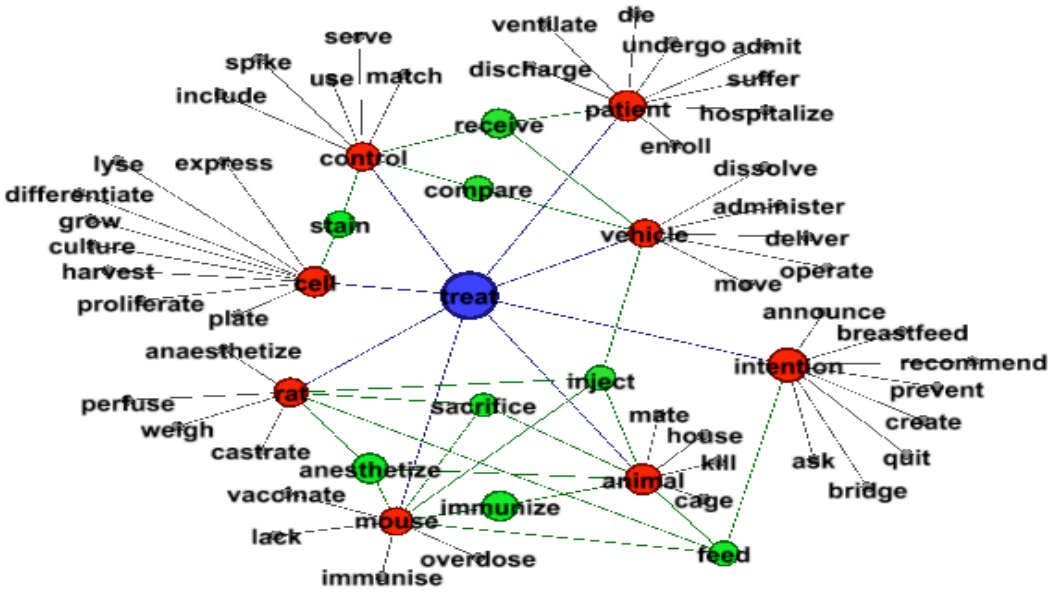

Figure 3: Collocational network from the verb *to treat*

Through the collocational network, verbs that are not in the top 100 verbs list are then introduced. In our illustration, this concerns 47 verbs of the network. Naturally, amongst this set of 'new' verbs, some could have been already enter in the dictionary, as they may have been introduced in a previous analysis. However, not all verbs present in the network will be selected as headwords and considered as entries of the dictionary. Indeed, this depends as well on the frequency.

This brief exemplification illustrates the organic nature of the constitution of the dictionary, which will grow in a natural way, by selecting what is statistically significant in the textual environment of the words. It is through the study of the 100 more frequent verbs that other verbs attested in the BMC corpus will in turn be enter in the dictionary. The constitution of the dictionary follows thus an iterative process: the analysis of one verb of the top-100 verb list leads to the consideration of verbs that are not in this list, and the analysis of one of them leads to the consideration of new verbs, and so on. As mentioned in the previous chapters, collocational networks are a mechanism for headwords selection. It also give a first picture of the environment of scientific texts, showing the most significant lexical units which are 'pivots', — using Hanks' terminology — of the clauses or are the main cognitive nodes that form the texts' framework.

The collocational network brings about a global picture of the node of the network, in this case the verb *to treat*. A lexicographical analysis of the network also show that

collocates can be grouped in different conceptual classes. In previous research (Williams & Millon, 2009), Levin's classification of verbs was considered (Levin, 1993). However, this classification does not suit all cases as it has not been built taking into account corpus data. Another option would have been that of using a vast hiercharchical ontology such as WordNet, but as Hanks (2006) points out not all lexical items fit into a hierarchical ontology. The relations between lexical units are not always of the same kind. Indeed, Hanks' point of view has been an inspiration for getting a way to group the different collocates into classes. Moreover, an analysis of the different collocations observed in the network brings about different semantic patterns of usage. These different lexical patterns are determined by using CPA.

In relation to our example, in general texts, four are the CPA patterns established by Hanks, as shown in Figure 4:

No.	%	Pattern / Implicature	
1	69%	**[[Human 1 | Institution 1 | Animal 1]] treat [[Human 2 | Animal 2 | Entity | Event]] [Adv[Manner]]**	conc.
		[[Human 1 | Institution 1 | Animal 1]] behaves toward [[Human 2 | Animal 2 | Entity | Event]] in the [[Manner]] specified	exploit.
2	17%	**[[{Human 1 = Health Professional} | {Process = Medical} | Drug]] treat [[{Human 2 = Patient} | {Animal = Patient} | Disease | Injury]] [NO ADVL]**	conc.
		[[Human 1 = Health Professional]] applies a [[Drug]] or [[Process = Medical]] to [[Human 2 =Patient]] for the purpose of curing the patient's [[Disease | Injury]]	exploit.
3	5%	**[[Human]] treat [[Inanimate]] (with [[Stuff]] | by [[Process]])**	conc.
		The chemical or other properties of [[Inanimate]] are improved or otherwise changed by [[Process]] or the application of [[Stuff]]	exploit.
4	5%	**[[Human 1]] treat [[Human 2 | Self]] {(to [[Eventuality = Good]])}**	conc.
		[[Human 1]] gives or pays for [[Eventuality = Good]] as a benefit for [[Human 2 | Self]]	exploit.

Figure 4: CPA patterns of the verb *to treat* extracted from Hanks' *PDEV*[5]

As can be inferred, the four patterns stand for different meanings of the verb *to treat*. The percentages assigned to patterns show the distribution of the four patterns within the corpus. At first sight, pattern 2 and 3 seem to be more thematically marked, the first related to Medical and the latter related to Chemistry domain. It could be thought that these patterns would also be commonly used in scientific texts. However, by analysing the BMC corpus applying CPA, differences of usage are brought about. The collocational network already shows that not all patterns are always coincident to those patterns distinguished in general texts.

In illustrating our work methodology with the verb *to treat*, using to the BMC corpus, pattern 1 of the verb *to treat* is close to the second CPA pattern (see Figure 4), in that it refers to a medical context. Indeed, in the BMC corpus the following normal pattern is found:

- X treat Y *with* Z
  - [[Human 1 | Human Group]] treat [[Human 2 = Patient | Laboratory Animal = *Rat, Mouse* | Organism= *Cell*]] (with [[Drug= *Vehicle*]])

In this pattern, the different collocates are gathered in different semantic types, as in CPA. By trying to apply CPA to the BMC corpus is clearly not always possible to use the same ontology. The ontology being used in CPA is a corpus-driven shallow ontology created from a general corpus. Many semantic types are not necessary in our case; on the contrary, semantic types that are not considered in CPA ontology are needed for explaining specific uses of a word in Biomedical texts. It is in fact the selection of specific semantic types, semantic roles and lexical sets which makes the difference between the general and specialised use of a lexical unit. For instance, in the pattern shown above, not all animals are treated. The semantic type specifically refers to 'Laboratory Animals'. There is a restriction on what is being treated. In reality, the lexical sets that define a given semantic type change according to each verb. For example, we treat *rats* and *mice*, but we do not treat neither *lion* or *elephant*. Hanks & Ježek (2008) has referred to this change as 'shimmering lexical sets.'

By looking at the concordances of *treat*, a slightly difference between CPA pattern 2 and our pattern 1 can also be detected. Most occurrences of *treat* refer to medical research and not to medical practices. An animal is treated *not for the purpose of being cured*, but *for getting a cure to a disease*. The implicature is not exactly the same.

The collocational network shown in Figure 3, also shows that the collocate *vehicle* is polysemic. Indeed, in the collocational network, the verbal collocates of the nominal collocates of the central verb *treat*, do not necessarily collocates with *treat*, since the nouns have been taken in turn as word-nodes. Thus, collocational networks do not stand for one particular meaning of the verb from which they are built. If we consider the noun *vehicle* — see Figure 3 —, within the occurrences of the collocations (on the lemma level) *vehicle – operate* and *vehicle – move*, the noun *vehicle* denotes a means of transport, whereas within the syntagmatic lexical relations with the verb *treat*, or its other verbal collocates in the network, it is a medical term used to refer to an excipient. Hence the presence in the network of its verbal

[5] http://deb.fi.muni.cz/pdev/?action=patterns&id=treat

collocates *dissolve*, *deliver*, *administer*, *receive* and *inject*. These two meanings are therefore linked, because an excipient serves to 'transport' the active ingredients of a medication. This will lead us to draw two nominal semantic types to which the noun *vehicle* will be attached: 'Transport' and 'Drug'. The verbs *dissolve*, *deliver*, *administer*, *receive* and *inject* are in lexical relation with the semantic type 'Drug', gathering themselves in a verbal conceptual class that we could name 'Giving drugs'. Concerning the conceptual classes in which the verbs of DicSci will be gathered, Framenet is consulted, but, ultimately, the verbal clustering in the dictionary DicSci is based on the specialised contexts of the BMC corpus.

Using CPA has brought about the necessity of using a shallow ontology in order to explain the phraseological tendency of verbs used in science. Indeed, phraseology occupies a main place in language use, notably through the use of collocations. In the lexicon of a given language, there are strong syntagmatic links between words. The phraseology of a given language implies that speaker (or writer), especially a non-native one, could product unnatural speech if he/she uses a 'wrong' word even if it matches the idea to be expressed. Language use is mainly filled with conventional lexical combinations that a native speaker has unconsciously memorised because he/she has already met them during their life. Non-native speakers, who do not have this linguistic experience, would construct their speech according to the semantic compatibility between words, and not to the lexical compatibility between words. Thus, the speaker, especially the non-native one, has to know the phraseology in use within the language in order to produce natural speech. Naturally, inside the same language, lexical preferences may differ notably between the general language and specialised ones, as notably state L'Homme (1998) or Heid & Freibott (1991).

The mechanism used allows conceptual classes that semantically link verbs in the dictionary to grow naturally as new verbs are analysed, and thus eventually split in several sub-classes. This has been illustrated in Williams & Millon (2009). In addition to conceptual classes of verbs, nominal ones are also created, according to the collocational network of the verbs, and notably, through the shared collocates reported in them.

It is important to underline that although networks can be automatically built, the eye of the lexicographer is essential. What we are extracting are potential collocates, only through analysis of the concordance can potential definitions be made. The semantic groupings of verbs or nouns follows the same procedure as, although they do fall together naturally, their interpretation and naming is the work of the lexicographer. Nevertheless, we project to apply the word sense discrimination algorithm written by Millon (2011), as we believe that this processing would help us with this task.

The next step in the creation of DicSci is that of adding the information extracted from the collocational networks and verbal patterns to the entries of the dictionary. For that, the dictionary production software *TshwaneLex*[6] is being used. The *E-Advanced Dictionary of Verbs in Science* is conceived as a virtual dictionary. By using visualisation techniques, the idea is to enter the dictionary by means of the collocational networks and from there go into the verbal patterns, concordances and dictionary entries. The grouping of verbs into classes will also give more options for the user to visualise not only syntagmatic relations but also paradigmatic relations between different lexical units.

5. Conclusions

The first aim of the DicSci project is to build an organic online dictionary of verbs use in sciences which will reflect usage and assist non-native speakers of English with production. In doing so a work methodology based on *collocational networks*, *collocational resonance* and Hanks' *Corpus Pattern Analysis-CPA* is being developed.

In this article, special attention has been paid to the use of collocational networks and application of CPA for building-up the dictionary. Collocational networks provide a natural selection of the main cognitive nodes of scientific texts, show links between lexical units, demonstrate thematic patterns in texts, and facilitate observation of what it is the 'normal' use/s of a specific lexical unit in a scientific context. By taking each collocate at a time, a number of lexico-semantic patterns can be detected. For that, the procedure *Corpus Pattern Analysis* described by Patrick Hanks is used. CPA method allows us to show the central and prototypical uses of a verb in science. By looking at the own output from Patrick Hanks' CPA, the *PDEV*, differences between 'general' and 'specialised' uses can be highlighted. From the patterns, the meaning potentials of the verbs can be inferred in a second stage.

Furthermore, collocational networks and semantic patterns show similarities and differences between the different uses of a lexical unit. Both mechanisms facilitate sense disambiguation of polysemic words. The methodology proposed shows also differences and similarities between different lexical units. Words that that are semantically related can be clustered together naturally in a conceptual class. In this way, both paradigmatic and syntagmatic relations can be illustrated.

The work methodology permits different ways to structure, organise and access the DicSci entries. In this sense, the dictionary is structured and organized according to the collocational networks. Apart from the traditional alphabetically ordering of entries, in the DicSci each central node of a network, which

[6] http://tshwanedje.com/tshwanelex/

corresponds to a verb, is an access to the entries of the dictionary. Each verbal collocate can also be a central node of another network and, therefore, another way to enter the dictionary. At the same time, other collocates, such as nouns or adjectives, can also be a means of access. The groupings of verbs will also permit access to the main verbal lexical units. The dictionary is both semasiologically and onomasiologically conceived.

The DicSci is an ongoing bottom-up, corpus-driven dictionary which describes how verbs are used in science. It is an organic dictionary in the sense that it is being developed in a natural and continuous process. It is dynamic, a moving system. Each collocational network can bring about new uses and new relations between other verbs and lexical units which have been already included in the dictionary. The relations between the units are continuously in motion.

In this paper, we have explored the first stage of the building-up of the dictionary which affects the global organisation and structure of the dictionary, the selection of headwords, the establishment of classes and the demonstration of semantic patterns. Further development is needed in relation to the definition and naming of conceptual classes and the microstructure of each entry. In a second stage, it is also expected to apply the mechanism of collocational resonance to assist in a better understanding of the movement from general to specialised usage of language, or from specialised to general.

The final aim of the DicSci project is to compile a dictionary which provides a way to explain not only the terminological tendency of words used in science, but also the phraseological tendency. The information included will help non-native speakers of English who need to produce scientific texts in English to improve their communication skills at different levels.

6. Acknowledgements

This paper was carried out during a postdoctoral research stay by one of the authors at the Equipe LiCoRN-Laboratoire HCTI of the Université Bretagne-Sud, directed by Prof. Geoffrey Williams, in the framework of the National Mobility Programme of Human Resources of the R+D National Programme 2008-2011, financed by the Spanish Ministry of Education. This research has also been funded by the European Project Metricc, and the Spanish National Project HUM2009-07588/FILO, supported by the Spanish Ministry of Education and Science.

7. References

Alonso, A. (unpublished 2009). Características del léxico del medio ambiente y pautas de representación en el diccionario general. PhD Thesis. Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra, Barcelona.

Bergenholt, H., Tarp, S. (1995). *Manual of Specialised Lexicography*. *The Preparation of Specialised Dictionaries*. Benjamins Translation Library 12. Amsterdam/Philadelphia: John Benjamins.

Cabré, M.ª T. (1999). La terminología. Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra.

Ciapuscio, G. (2003). *Textos especializados y terminología*. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra.

Cowie, A.P. (2002). *English Dictionaries for Foreign Learners. A History*. Oxford: Oxford University Press.

Cowie, A.P. (2009). The Oxford History of English Lexicography. Volumen II. Oxford: Clarendon.

Church, K., Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational linguistics*, 16(1), pp. 22-29.

Fuertes-Olivera, P.A. (2009). Specialised Lexicography for Learners: Specific Proposals for the Construction of Pedagogically oriented Printed Business Dictionaries. *Hermes – Journal of Language and Communication Studies*, 42, pp. 167-188.

Fuertes-Olivera, P.A. (ed.) (2010). *Specialised Dictionaries for Learners*. Lexicographica Series Maior, 136. Berlin/New York: De Gruyter.

Gledhill, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.

Halliday, M.A.K. (1961). Categories of the Theory of Grammar. *Word*, 17, pp. 241-292.

Hanks, P. (2002). Mapping Meaning onto Use. In M.-H-Corréard (ed.) *Lexicography and Natural Language Processing: a Festschrift in honour of B. T. S. Atkins*. United-Kingdom: Euralex, Göteborg University, pp. 156-198.

Hanks, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.

Hanks, P. (2006). The Organization of the lexicon: Semantic Types and Lexical Sets. In C. Marello *et al*. (eds.) *Proceedings of the XII EURALEX International Congress*. Torino: Università di Torino, pp. 1165-1168.

Hanks, P. (2010a). How People Use Words to Make Meanings. In B. Sharp, M. Zock (eds.) *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2010*. In conjunction with ICEIS 2010, Funchal, Madeira, Portugal, pp. 3-13.

Hanks, P. (2010b). Terminology, Phraseology, and Lexicography. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden, Pays Bas: Fryske Akademy.

Hanks, P. (2010c). Compiling a Monolingual Dictionary for Native Speakers. *Lexikos* 20, 580-598.

Hanks, P. (forthcoming). *Lexical Analysis: Norms and Exploitations*. Massachusetts: The MIT Press.

Hanks, P., Ježek, E. (2008). Shimmering Lexical Sets. In E. Bernal, J. DeCesaris, J. (eds.) *Proceedings of the*

XIII Euralex International Congress. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra, pp. 391-402.

Heid, U., Freibott, G. (1991). Collocations dans une base de données terminologique et lexicale. *Meta,* 36(1), pp. 77-91.

Hoey, M. (2005). Lexical Priming: A New Theory of Words and Language. London: Routledge.

Hunston, S., Francis, G. (1999). *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Hunston, S., Sinclair, J. (2003). A local grammar of evaluation. In S. Hunston, G. Thompson (eds.) *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press, pp. 74-101.

Ježek, E., Hanks, P. (2010). What lexical sets tell us about conceptual categories. In *Corpus Linguistics and the Lexicon, Special issue of Lexis, E-Journal in English Lexicology*, 4, pp. 7-22.

L'Homme, M.-C. (1998). Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale. In *Proceedings of the VIII EURALEX International Congress*. Liège, Belgium, pp. 513-522.

L'Homme, M.-C. (2005). Sur la notion de «terme». *Meta: Translators' Journal* 50(4), pp. 1112-1132. On line: http://id.erudit.org/iderudit/012064ar

L'Homme, M.-C. & Leroyer, P. (2009). Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology* 15(2), pp. 258-283.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago: University of Chicago Press.

Lorente, M. (2007). Les unitats lèxiques verbals dels textos especialitzats. Redefinició d'una proposta de classificació. In M. Lorente et al. (eds.) *Estudis de lingüística i de lingüística aplicada en honor de M. Teresa Cabré Castellví. Vol. 2: De deixebles 2*. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra; Documenta Universitaria, pp. 365-380.

Lorente, M. (2009). Verbos y fraseología en los discursos de especialidad. In M. Casas, R. Márquez (ed.) *XI Jornadas de Lingüística: homenaje al profesor José Luis Guijarro Morales (Cádiz, 22 y 23 de abril de 2008)*. Cádiz: Universidad de Cádiz. Servicio de Publicaciones, pp. 55-84.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker (ed.) *Text and Technology.* Amsterdam: John Benjamins, pp. 157-76.

Louw, B. (2000|2008). Contextual Prosody Theory: bringing Semantic Prosodies to Life. In C. Heffer, H. Sauntson (eds.) *Words in Context: A Tribute to John Sinclair on his Retirement*. CD-ROM: English Language Research Discourse Analysis Monograph No. 18. Reprinted in online journal *Texto* (2008): http://www.revue-texto.net/index.php?id=124.

Meyer, I. (2000). Computer Words in Our Everyday Lives: How are they interesting for terminography and lexicography? In U. Heid *et al*. (eds.) *Proceedings of IX EURALEX International Conference 2000*. Stuttgart: Universität Stuttgart, pp. 39-57.

Meyer, I., Mackintosh, K., Varantola, K. (1997). Exploring the reality of *virtual*: on the lexical implications of becoming a knowledge society. *Lexicology,* 3(1), pp. 129-163.

Million, C. (unpublished 2011). Acquisition automatique de relations lexicales désambiguïsées à partir du Web. PhD Thesis. Université de Bretagne-Sud, Lorient.

Moulin, A. (1983). LSP Dictionaries for EFL Learners. In R. R. K. Hartmann (ed.) *Lexicography: Principles and Practice*. London: Academic Press, pp. 144-152.

Phillips, M. (1985). Aspects of Text Structure: An investigation of the lexical Organisation of Text, Amsterdam, North Holland.

Renau, I., Alonso, A. (in press). Using Corpus Pattern Analysis for the Spanish Learner's Dictionary DAELE (Diccionario de aprendizaje del español como lengua extranjera). In *Proceedings Corpus Linguistics Conference 2011*. Birmingham: University of Birmingham.

Roe, P. (unpublished 1977). The notion of difficulty in Scientific Text. PhD thesis. University of Birmingham, Birmingham.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-knowledge General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Lexicographica Series Maior 134. Tübingen: Max Niemeyer Verlag.

ten Hacken, P. (2008). Prototypes and discreteness in terminology. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress. Barcelona, 15-19 july 2008.* Papers de l'IULA. Sèrie Activitats. 20. Barcelona: Documenta Universitaria. Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra.

Williams, G. (1998). Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. *International Journal of Corpus Linguistics*, 3(1), pp. 151-171.

Williams, G. (2002). In search of representativity in specialised corpora: categorisation through collocation. *International Journal of Corpus Linguistics*, 7(1), pp. 43-64.

Williams, G. (2003). From meaning to words and back: Corpus linguistics and specialised lexicography. *Asp, la revue du GERAS* 39-40, pp. 91-106. On line: http://asp.revues.org/1320.

Williams, G. (2006). Advanced ESP and the Learner's Dictionary. In C. Marello *et al*. (eds.) *Proceedings of the XII EURALEX International Congress*. Torino: Università di Torino, pp. 795-801.

Williams, G. (2008a). Verbs of Science and the Learner's Dictionary. In J. DeCesaris, E. Bernal (eds.) *Proceedings of the XIII Euralex International Congress. Barcelona, 15-19 july 2008. Papers de l'IULA. Sèrie Activitats*. 20. Barcelona: Institut Universitari de Lingüística Aplicada – Universitat Pompeu Fabra; Documenta Universitaria.

Williams, G. (2008b). The Good Lord and his works: A corpus-based study of collocational resonance. In S. Granger, F. Meunier (eds.) *Phraseology: an interdisciplinary perspective*. Amsterdam: John Benjamins, pp. 159-174.

Williams, G. (2008c). Les corpus et le dictionnaire dans les langues scientifiques. In F. Maniez et al. (eds.) *Corpus et dictionnaires de langues de spécialité*. Grenoble: Presses Universitaires de Grenoble.

Williams, G., Millon, C. (2009). The General and the Specific: Collocational resonance of scientific language. In *Proceedings of the Corpus Linguistics Conference CL2009, 20-23 July 2009*. Liverpool: University of Liverpool.

Williams, G., Millon, C. (2010). Going organic: Building an experimental bottom-up dictionary of verbs in science. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden, Pays Bas: Fryske Akademy, pp. 1251-1257.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.