Designing a Learner's Dictionary Based on Sinclair's Lexical Units by Means of Corpus Pattern Analysis and the Sketch Engine

Paolo Vito DiMuccio-Failla, Laura Giacomini[1]

[1] Heidelberg University/ Hildesheim University
E-mail: paolodimuccio@gmail.com, laura.giacomini@iued.uni-heidelberg.de

Abstract

This paper is part of a study for the design of an advanced learner's dictionary (in Italian) aimed at implementing Sinclair's vision of 'the ultimate dictionary' (see Sinclair et al., 2004: xxiv) and based on his conception of lexical units. Our present goal is to exhaustively portray the meaning profile of verbs, systematically distinguishing their meanings by their normal patterns of usage. To achieve this, we apply Hanks's Corpus Pattern Analysis by means of Kilgarriff and Rychlý's Sketch Engine.
The first chapter presents and discusses the theoretical background to our work. The second gives a description of our methodology, which is then exemplified by a case study of the Italian verb *seguire*. The final part of the paper draws a few conclusions on the feasibility and usefulness of Sinclair's 'ultimate dictionary' and hints at future steps of the dictionary-making process.
The dictionary project is in its design stage and is intended to be a platform for cooperation between the Italian publisher Zanichelli and a network of international universities and research institutes.

Keywords: learner's dictionary; Italian learner's dictionary; lexical units; Sinclair's thesis; Sinclair patterns

1. Methodological background

1.1 COBUILD's scientific revolution

At the start of the 1980s, lexicography, and linguistics in general, were undergoing a far-reaching paradigm shift thanks to the new availability of huge quantities of machine-readable text made possible by advances in computer technology. According to John Sinclair, a then leading linguist at the University of Birmingham, the situation was similar to that of the physical sciences in the first half of the 17th century, when they started to rely on empirical observation (Sinclair, 1991: 1). If the intuition of a single individual had been, up to that moment, the key to all linguistic investigation, lexicography finally had the possibility to utilize "objective evidence" (ibid.).

Given these premises, Sinclair (with funding from Collins publishers) founded the COBUILD (Collins Birmingham University International Language Database) project with the aim of producing innovative language reference works (Sinclair, 1991: 2). Together with his collaborators, he started building a large and representative

electronic corpus of contemporary English (Sinclair, 1987: 1), based on which, in the following years, "a completely new set of techniques for language observation, analysis, and recording" was developed (Sinclair, 1991: 2). Many consider this the very first study in corpus-driven lexicography (Tognini-Bonelli, 2001: 85), initiating the now thriving tradition of empirical lexical analysis (Hanks, 2008a: 222).

The main result was the compilation of the COBUILD English Dictionary (COBUILD 1987), the first dictionary based on evidence of actual contemporary usage, and the first to give a central role to the "spectacular" regularities of language patterning which had been displayed by corpus analysis (Sinclair, 1991: 4) and had lead Sinclair to conclude that "by far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns" (Sinclair, 1991: 108). This phenomenon goes far beyond that which the pioneer lexicographers like Palmer and Hornby had shown, since different senses of the same word present, in general, different characteristic patterns, as we will explain in the following subsection.

In order to display the "typical features" of the characteristic co-texts of words (Sinclair, 2004b: 5; see also Hanks, 1987), Sinclair systematically utilized full-sentence definitions, which he considered theoretically sounder and easier to understand (Hanks, 2008a: 221) than traditional ones (for a balanced discussion see Rundell, 2006). Furthermore, in the COBUILD dictionary, every observation about language was accompanied by at least one example, and all examples were taken from the corpus in order to obtain "genuine instances of language in use" (Sinclair, 1991: 4–5). All this was thought to help students to speak and write naturally and idiomatically (see Hanks, 2008a: 219).

An important point should be made about Sinclair's empirical corpus analysis. On the one hand, it proceeds along the standard scientific method of inspecting the data, discerning regularities, formulating hypotheses, and testing the hypotheses on the data (Sinclair, 2004a: 10 ff.). On the other hand, "intuition and introspection still play an important role, since perceiving meaning is a subjective experience, and descriptions in dictionaries need to satisfy intuition" (Moon, 1987). The role of introspection is to evaluate evidence rather than to create it (Sinclair, 1991: 39), whereas intuition is crucial exactly when introspection "is not in accordance with the newly observed facts of usage" (Sinclair, 1991: 4). Therefore, intuition, introspection, and data analysis must work together (Sinclair, 2004a: 115).

This is why Sinclair does not, in principle, discard traditional kinds of evidence, obtained for example by consulting other dictionaries or by testing native speakers (Sinclair, 1991: 38–39). Most of all, "ultimately... the lexicographic decisions will be personal evaluations by the lexicographer, giving due consideration to all evidence that he or she has amassed" (ibid.: 39). For these reasons, Sinclair takes a balanced stance in the debate between descriptive and prescriptive studies, stating that "a

purely objective description of text will not contain adequate generalization" (ibid.: 60) and that "prescriptive studies fall into disrepute only when they ignore or become detached from evidence" (ibid.: 61).

A second important point to be made for our purposes is that Sinclair distinguishes between typical language patterns on one side and extended, displaced, and distorted usages on the other side (Sinclair, 1991: 61). A synchronic dictionary of usage should be filled with norms (ibid.: 61), not with unusual language events, and should warn against specialized use (ibid.: 38).

1.2 Sinclair's thesis about lexical units

When lexical information began to be extracted from multi-million word corpora in the early 1980s, several long-accepted conventions in lexicography were called into question, for example the idea that a polysemous word could inherently, by itself, have several distinct meanings (Sinclair, 1998; Sinclair, 2004a: 132), and that any occurrence of such a word could signal any of those meanings (Sinclair, 1986: 60). Sinclair recognized that if this were actually the case, ambiguity would make communication virtually impossible (see Sinclair, 1998), because the meanings of polysemous words, though related, can be very diverse (this later became known as the *polysemy paradox* - see Falkum, 2011: 13 ff.). On the contrary, in continuous discourse, whether written or spoken, ambiguity is rare, except when intended (see Moon, 1987).

In the course of the survey leading up to the publication of the COBUILD dictionary, evidence gradually accumulated for an alternative hypothesis which, at first, had been ridiculed (Sinclair, 1991: 10): that of a general correspondence between observable patterns of words and distinctions of meaning. In fact, Sinclair came to the conclusion that not single, isolated words, but rather *words in their contextual patterns* are the true bearers of meaning, and that every such pattern has only one meaning (not considering sub-meanings given by trivial generalization or specification - cf. for example Sinclair, 1991: 55–56). This claim can be stated in a more rigorous fashion, which we might call 'Sinclair's thesis':

> In general, each (major) (normal) sense of a word can be associated with a distinctive pattern of usage (see Moon, 1987: 89 ff.; Sinclair, 1991: 6 ff.; Sinclair, 2004b: 5; Sinclair, 2004c: 281; Sinclair et al., 2004: xxiv) determined by the following features (see also Sinclair 1996; Sinclair, 1998; Sinclair, 2003: 145 ff.; Sinclair, 2004a: 39 and 141):

1) collocation, i.e., the co-occurrence of particular words (with the given word);
2) colligation, i.e., the co-occurrence of particular grammatical patterns;
3) semantic preference, i.e., the co-occurrence of words with particular meanings;
4) semantic prosody, i.e., a co-text implying a particular connotation of the described

state of affairs or a particular attitude of the speaker[1].

Take for instance the word *put*. It can be part of a phrasal verb, in which case its meanings are co-determined by other parts of speech, or it can be a non-phrasal verb, in which case its senses mostly correspond to the (choices of the) semantic types of the referents associated with its arguments (i.e., its selectional preferences). As an illustration of this correspondence, we look at the first three senses of *put* in the corresponding entry of the latest edition (2014) of the COBUILD dictionary (see also Moon, 1987: 91):

1. "When you **put** <u>something</u> in a particular <u>place or position</u>, you move it into that place or position"
2. "If you **put** <u>someone</u> ... [in a particular <u>place or position</u>], you cause them to go there and to stay there for a period of time"
3. "To **put** <u>someone</u> or something in a particular <u>state or situation</u> means to cause them to be in that state or situation".

Sinclair's analysis even allows the finding of hidden senses of words. Consider for example the word *feeling*. No corpus analysis is needed to know that it frequently co-occurs with the adjective *true* in the phrase *true feelings*. Such a collocation would not be considered idiomatic and hardly given any special treatment in a traditional dictionary (Sinclair, 1996: 89). An accurate pattern analysis (cf. Sinclair, 2003) will in fact show statistical restrictions on the choice of its co-text. *True feelings* is usually preceded by a possessive adjective, which is in turn preceded by a verb synonymous with *express*, *show*, or *hide*. This constitutes a syntactic tendency, a colligation, but also a semantic preference for verbs of expression. In the case of semantically 'positive' expression, there is usually an even broader context, i.e. a semantic prosody, hinting at a reluctance or difficulty in expressing those true feelings. Hence the actual lexical unit here can be presented by

"to hide one's true feelings or show them with/after some reluctance/difficulty".

Thus, Sinclair arrived at the conclusion that the true units of meaning of a language are largely phrasal and that, as a consequence, phraseology is due to become central in the description of language (cf. Sinclair, 2004a). Sinclair used the term '(extended) canonical form' to refer to the most explicit, full and unambiguous presentation of a lexical unit (Sinclair, 2004c: 298), like the one we just proposed for *true feelings*. The shortest unambiguous presentation of the lexical unit (in our case, simply *true feelings*) he called 'short canonical form' (Sinclair et al., 2004: xxiv). In the final years of his career, he was convinced that a new kind of dictionary based on the canonical forms of lexical units "would be the ultimate dictionary" and would allow students to truly master a language (ibid.).

---

[1] The notion of semantic prosody was implicitly introduced by Sinclair (1987: 155; 1991: 75) and first defined by Louw (1993). It is actually rather controversial (see Whitsitt, 2005; Stewart, 2010) and hard to work with.

1.3 Hanks's analysis of corpus patterns

Patrick Hanks, one of the main collaborators of John Sinclair at COBUILD, has since been a committed supporter of the corpus-driven approach to lexicography and of Sinclair's thesis about lexical items (Hanks, 2004a: 87; Hanks & Pustejovsky, 2005; cf. also Krishnamurthy, 2008: 239). Hanks's focus on NLP has lead him to develop and standardize a technique, which he dubbed 'Corpus Pattern Analysis' (CPA), to analyze large corpora and find the "normal patterns of usage" associated with each word, with the aim "to link word use and word meaning in a machine tractable way" (Hanks & Pustejovsky, 2005: 64). The main result will be a dictionary for use in NLP (ibid.) and in language teaching (cf. PDEV website). In the Pattern Dictionary of English Verbs (PDEV), a pilot study currently in development under the supervision of Hanks, many verbs are being analysed, having priority over nouns (cf. Hanks, 2008b; Hanks, 2004a: 92).

Hanks' "semantically motivated syntagmatic patterns" (Hanks, 2004a: 88) are simplified and strictly formalized versions of Sinclair's word patterns. In the case of a verb, they consist of an argument structure, assigned together with the most general semantic types (and possibly semantic roles[2]) which the arguments of the verb *normally* refer to (ibid.: 87–88). The last bit is a tricky one: identifying the right semantic types as selectional preferences, in particular not leaving out normal usage on one side and not generalizing into abnormal usage on the other side, requires linguistic and ontological expertise: "Among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished" (cf. Hanks, 2004a and PDEV website). In general, "the identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art" (Hanks, 2004a: 88).

In CPA, one starts with concordance lines and groups them into patterns, whereas "associating a 'meaning' with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns" (ibid.). "The 'meaning' of a pattern is expressed as a set of basic implicatures" (ibid.). Let us look for example at the syntagmatic patterns of the verb *lead* according to the PDEV (cf. PDEV website)[3]:

1. Pattern: [Eventuality]$_1$ **leads** TO [Eventuality]$_2$
Implicature: [Eventuality]$_1$ is the cause of [Eventuality]$_2$

---

[2] Roles are not considered types by Hanks (cf. Hanks et al., 2007: 5). We will discuss the use of semantic roles in word patterns in the following section.

[3] According to CPA conventions (cf. Hanks, 2004a: 93), double square brackets indicate semantic types and curly brackets (braces) indicate sets of specific lexical items. The keyword is written in bold letters. For readability reasons, we have slightly modified the convention regarding types by using simple square brackets.

2. Pattern: [Eventuality]$_1$ **leads** UP TO [Eventuality]$_2$
Implicature: [Eventuality]$_1$ precedes but may not be the cause of [Eventuality]$_2$

3. Pattern: [Eventuality] **leads** [Human]/[Institution] TO-INFINITIVE
Implicature: [Eventuality] causes, enables, or encourages [Human]/[Institution] TO...

4. Pattern: [Human]/[Institution]$_1$ **leads** [Human group]/[Institution]$_2$
Implicature: [Human]/[Institution]$_1$ organizes, directs, or provides a model for activity of [Human group]/[Institution]$_2$

The choice of appropriate selectional preferences can be hard not only because of the inherent difficulty in building a coherent ontology compatible with everyday language, but also because it is not always immediately clear what semantic types normal usage can possibly refer to. Take for instance the English verb *toast* in the sense of "cook food by exposure to a grill or fire" (as in Hanks, 2004a: 91 and Jezek & Hanks, 2010). A quick look at the word sketch of *toast* on the Sketch Engine (see Kilgarriff et al., 2004) shows that the most frequent direct objects of *toast* are *bread, almonds, marshmallows, buns, walnuts, pecans, coconut, bagels, nuts, hazelnuts, sandwiches, baguettes, brioche, muffin...* In such cases, Hanks proposes to either use a general semantic type (as in the PDEV), like

> [Human] **toasts** [Food],

or, when possible, to insert directly into the pattern (see Hanks, 2004a: 91) the paradigmatic lexical set of the most frequent collocates. In our case, this results in

> [Human] **toasts** {*bread, almonds, marshmallows, buns, walnuts, pecans, coconut, bagels, nuts, hazelnuts, sandwiches, baguettes, brioche, muffin*}.

However, in the first case the type [Food] can be seen as too general and uninformative, whereas in the second case the list was truncated at *muffin* for no statistical reason: the actual progression of collocates slowly fades into statistical insignificance without any apparent discontinuity. This raises a semantic issue (see for example Jezek & Hanks, 2010), which we will try to resolve in the following section.

As an ontology for CPA, Hanks uses a shallow hierarchy of types selected for their prevalence in the manual identification of patterns (Pustejovsky et al., 2004). The number of types is kept to a minimum, as perfect ontological coherence is required. "New types are added occasionally, but only when all possibilities of using existing types prove inadequate" (Pustejovsky et al., 2004). Currently, there are 253 types in the PDEV.

Corpus Pattern Analysis hinges on the Theory of Norms and Exploitations (see Hanks, 2013), which makes a strict (conceptual) distinction between normal and

abnormal usage of language (Hanks, 2013: 3; Hanks 2004a: 89), be it because of anomalous syntactic structures, anomalous semantic arguments, or figurative uses (El Maarouf, 2013: 125; see also Hanks, 2004a: 92). When abnormal usage is intentional, it is called an 'exploitation' of a norm (Hanks, 2013: 8). This theory led Hanks to conclude that "attempts to account for all possible meanings [of words] are misguided. Projects with this aim tend to produce impractical results, because normal usage becomes buried in a welter of remote possibilities" (Hanks & Pustejovsky, 2005: 64). On the contrary, "the number of normal combinations is remarkably small and computationally manageable" (Hanks & Pustejovsky, 2004: 15).

2. Our purpose and method

We are convinced that Sinclair's concept of fundamental lexical units is the right one. We know this is still a controversial issue: many linguists do not even agree on the existence of objective criteria for correctly lumping/splitting the senses of polysemous words (see for example Kilgarriff, 1997: 100). However, by comparing the results of our present research with ItalWordNet, the Italian wordnet (see Roventini et al., 2003) created in the framework of the EuroWordNet project (see Vossen, 2002), we discovered a stunning overlap of the meanings of Sinclair's lexical units with the single senses of words implicit in the synsets of ItalWordNet. Such senses result from a completely different approach and it is hard to see how this could be a coincidence. We will explain our findings in detail for the Italian verb *seguire* in the following section.

We also share Sinclair's opinion that a dictionary extensively describing the canonical forms of each lexical unit would be 'the ultimate dictionary', because it would potentially contain all semantic information about word usage. This is why we started investigating the feasibility and actual utility of implementing Sinclair's vision. The two main problems we encountered are the following:

1) Extracting lexical units from a corpus and accurately studying their canonical forms can be difficult and time consuming.

2) It is not easy to present the extended canonical form of a lexical unit without overloading its entry with information of various degrees of importance. This is exactly the problem we have mentioned about the full-sentence definitions found in the COUBILD dictionary.

We will explain how we coped with both problems in a series of papers. For now, we will concentrate on the first one, showing in particular how we adapted Hanks's CPA and applied it by means of Kilgarriff and Rychlý's Sketch Engine (see Kilgarriff et al., 2004) to find all senses of the Italian verb *seguire*.

2.1 Building an ontology

We need to build an ontology not only because Sinclair's word patterns refer to semantic types, but also because ontologies facilitate homogenous definitions and a clean overview of any lexical domain. Our approach to the upper part of the hierarchy is similar to that of EuroWordNet (see Vossen et al., 1998), which distinguishes, along the lines of Lyons (1977), the category of concrete objects and substances (first-order entities) from that of properties, relations, situations, and events (second-order entities). We will discuss the details in a future paper. Concrete entities can be further classified into types according to the four independent criteria advocated by Pustejovski (1995): origin, form, composition, and function. Second-order entities can be classified into types according to more sophisticated criteria, which will also be examined in a future paper.

Fortunately, for the purpose of monolingual learner lexicography, hierarchies of types only have to be as systematic and coherent as normal language usage. Hence, in principle, we accept the possibility that semantic types assigned in different word patterns might not be perfectly compatible. Furthermore, it is natural to add to the ontology not only any lexicalized semantic role, like [Patient] or [Monarch], but also a distinct type of entity for every nominal lexical unit (of the language in question), like [Means of public transport] and [Job creation scheme].

Not having to use a limited, perfectly coherent ontology can make things a lot easier. Consider the example of the verb *toast* in the sense of "cook food by exposure to a grill or fire", which we have discussed in the previous section. Allowing for relatively uncommon concepts like [Breadstuff], a word pattern can be assigned which is easier for a human to read and understand:

> [Human] **toasts** [Breadstuff]/[Marshmallow]/[Nut]/[Seed]

Differentiating between prototypical, common, and possible usage of words is also an option:

> [Human] **toasts** *prototypically* [Bread]/[Sandwich]
>   *usually* [Breadstuff]/[Marshmallow]/[Nut]/[Seed]
>   *possibly* [Food]

2.2 Identifying lexical units

In general, we employ a bottom-up, empirical strategy to identify the semantic types selected by a word for its argument slots, following the clues provided by the word sketches of the Sketch Engine. Notice, for example, that the paradigmatic lexical set of collocates found in a particular argument slot of a particular word sense can be partially ordered, in a mathematical sense, according to the hyponym-hypernym relation. If it presents a maximum, i.e. a hypernym of all other words in the set, such

a hypernym denotes the needed semantic type. For instance, a paradigmatic lexical set of nouns associated, as subjects, with the Italian verb *fermare* (to stop) is

> {*bus, autobus, corriera* (coach), *tram, treno* (train), *metro, metropolitana, mezzo pubblico* (means of public transport)}.

Clearly, *mezzo* (*di trasporto*) *pubblico* is a hypernym of all other words in the set, and thus identifies the most appropriate semantic type for the subject slot of the corresponding lexical unit.

It must be stressed that Sinclair's patterns do not come out of a corpus by themselves: they must be properly looked for by means of the scientific method, as we mentioned in the previous section. Consider for instance the Italian word *braccio* (arm). We analysed the word sketches of *braccio* by taking into account all its possible syntactic constructions. Two of them turned out to be particularly informative: (N + Adj) and (N + *di* + N). After finding many lexical units, like *braccio di un essere umano*, *braccio di un carcere*, *braccio di terra/mare/fiume*, and others, we were left with what we thought to be a paradigmatic lexical set of a single remaining unit:

> {*mobile, meccanico* (mechanical), *flessibile* (flexible), *regolabile* (adjustable), *articolato* (jointed), *snodabile* (hinged), *estensibile* (extendable)}

Since *braccio mobile* (mobile arm) is a hypernym of *braccio meccanico*, *braccio flessibile*, and so on, we selected it as a candidate. However, in trying to confirm the hypothesis, we indeed falsified it when we found out that *braccio fisso* (fixed arm) also exists and that it refers to the same kind of objects: supporting arms of devices. Most of the adjectives in the set were confirmed to be, in fact, collocates of the lexical unit *braccio di sostegno di uno strumento* (supporting arm of a device). The remaining adjective, *mechanical*, must hence build a separate lexical unit: *braccio meccanico* (mechanical arm).

2.3 Formalizing word patterns

Our final objective is to compile an Italian learner's dictionary. Hence, an accurate adherence of word senses to actual normal usage is of paramount importance. However, since we are convinced that the best way to achieve this goal is by means of Sinclair's patterns of word usage, we do not want to exclude *a priori* an application of the dictionary for NLP, like the PDEV. Therefore, we will adopt a semi-formal approach: our patterns will have in general a formal part adapted from CPA and an informal expansion for human readers.

Now consider the first sense of the verb *follow* in the COBUILD dictionary (2014):

> "If you follow someone, who is going somewhere, you…"

The phrase "who is going somewhere" predicates a necessary stage-dependent (cf. Kratzer, 1995) condition for the action of following to take place. Such semantic prerequisites are often not needed for the disambiguation of a polysemous word because it is constant in all of its senses. However, they will inevitably be part of the semantic preference of any given word, and therefore we will always make them explicit, as they are in the COBUILD dictionary:

> "If you repair something that… is not working…"
> "When you unzip something which is fastened by a zip…"
> "If you find something that you need or want…"

In some cases, semantic conditions are essential for disambiguation. Suppose, for example, that you were just told to follow a man who is standing. If he is talking, you were probably told to listen to him. The prerequisite for the literal sense of the verb *follow* to be activated is that the person to be followed must be going somewhere. Its formalized canonical form could be

> [Human]$_1$ **follows** [Human]$_2$ *SUCH THAT* ([Human]$_2$ *IS A* [Goer]).

Notice that [Goer] is a rather unusual semantic role. To avoid cluttering our ontology with unnatural concepts, we prefer a different approach to the formalization of Sinclair patterns, allowing formulas to refer to meanings of predicates defined in the dictionary itself, as long as this does not result in a circular definition. The previous pattern can thus become more readable:

> [Human]$_1$ **follows** [Human]$_2$ *SUCH THAT* ([Human]$_2$ goes TO *SOME* [Place]).

Let us confront this pattern with the first sense found in the PDEV (we are ignoring the presence of the type [Animal] for the sake of clarity):

> [Human]$_1$/[Vehicle]$_1$ **follows** [Human]$_2$/[Vehicle]$_2$

This sense is not disambiguated from the second one in the same entry:

> [Human]$_1$ **follows** [Human]$_2$

Furthermore, the type [Vehicle] was needed because it was not possible to rely on the general regular alternation substituting people moving in vehicles for the vehicles themselves (when describing their motion). Incidentally, we conjecture that Hanks's question as to why semantic types do not seem to match well with paradigmatic lexical sets (see Hanks et al., 2007; Hanks & Jezek, 2008; Jezek & Hanks, 2010) can be at least partially answered by taking into consideration stage-dependent semantic conditions, which are not always easy to identify.

As a final remark, we will conform to the standard lexicographic practice of using in general (with a few natural exceptions) the type [Person] instead of both [Human] and [Animal], as this distinction is rarely needed for word sense disambiguation, and action verbs are principally thought to apply to any real or imaginary person. Similarly, if the type selected by a verb sense for the subject slot is [Person] we will omit it.

3. Case study: the Italian verb *seguire*

On the Sketch Engine, we selected the 2010 itTenTen corpus (see Jakubíček et al., 2013) and set out to identify and study the Sinclair patterns of the Italian verb *seguire* (to follow).

3.1 Patterns

We analysed the first 500 concordances of *seguire* chosen as "good examples" by the Sketch Engine (cf. Kilgarriff et al., 2008). It quickly became clear that the main distinction to be made was between transitive and intransitive patterns. The intransitive patterns could then be distinguished according to their argument structure, and the transitive ones according to their semantic preference. We progressively classified the instances of *seguire* according to those criteria and also, subordinately, depending on whether we deemed them to be normal or abnormal. Regular alternations as described by Pustejovsky in his Generative Lexicon Theory (see Pustejovsky, 1995) were classified as normal usage, whereas *ad hoc* metaphors, metonyms and other figures of speech were considered exploitations.

One by one, we identified the following lexical units, here arranged in an order which facilitates an overview:

T1) Seguire qu. presente che sta andando da qualche parte (to follow sb. present who is going somewhere)

This is the most basic pattern of *seguire*, used as a transitive verb with the literal meaning of "andare dietro a qu." (to move along behind sb.). As already mentioned, in our approach, we attempt to identify semantic types by finding the most general semantic restriction which disambiguates the present sense from the other senses. In this case, however, the only such restriction is that, normally (excluding occasional extensions to small objects), in Italian you follow persons (possibly alternating with animals, as in the case of many other verbs of motion). As previously discussed, the disambiguating information for this sense is actually a stage-level semantic condition, i.e., the fact that the followed person is (present and) going somewhere.

T2) Seguire un certo tragitto o una certa descrizione di un tragitto (to follow a particular route or a particular description of a route)

This pattern has the meaning of "andare lungo un certo tragitto" (to move along a particular route). It displays a metonymical alternation between routes and descriptions of routes (*indicazioni*). By means of the word sketches provided by the Sketch Engine, we found a large number of collocates in the direct object position which refer to types of routes: *percorso, corso, traccia, sentiero, strada, itinerario, pista, via, cammino, tracciato, rotta, traiettoria.* Since *tragitto* (route) is a hypernym of all members of the lexical set in question, we selected it as the name of the associated type. We did not choose *percorso* (path), because its most common meaning is concrete, whereas, as confirmed by standard dictionaries (e.g., TRECCANI and DE MAURO), the basic meaning of *tragitto* is abstract. Definition no. 8 of *follow* in the COBUILD (2014) dictionary perfectly matches our pattern:

"If you follow a path, route, or set of signs, you go somewhere using the path, route, or signs to direct you."

T3) Seguire qu. presente che sta svolgendo una sequenza di azioni (to follow sb. present who is performing a sequence of actions)

This pattern has the meaning of "fare ciò che si vede/ sente fare a qu., imitare qu." (to do what you see/ hear sb. do, to imitate sb.). Definition no. 13 of *follow* in the COBUILD dictionary loosely corresponds to our pattern:

"If you follow what someone else has done, you do it too because you think it is a good thing or because you want to copy them."

T4) Seguire una certa linea di condotta o una certa descrizione di una linea di condotta (to follow a particular course of action or a particular description (of a course of action))

This pattern has the meaning "agire secondo una certa linea di condotta" (to act according to a particular course of action). Typical collocates we found are *dieta, esempio, moda, metodo, modello, tendenza, trend.* Definitions no. 17 and 12 in the COBUILD dictionary loosely correspond to our pattern:

"If you follow a particular religion or political belief, you have that religion or belief."
"If you follow advice, an instruction, or a recipe, you act or do something in the way that it indicates."

T5) Seguire con lo sguardo qu. che si sta spostando (to follow with your eyes sb. who is moving)

This pattern has the meaning "mantenere lo sguardo su qu. che si sta spostando" (to keep one's eyes on sb. who is moving). It is disambiguated by the prepositional phrase "con lo sguardo", which is an idiomatic argument of *seguire*. Definition no. 10 in the COBUILD dictionary corresponds to our pattern:

"If you follow something with your eyes, or if your eyes follow it, you watch it as it moves or you look along its route or course."

T6) Seguire una certa scena in corso (to follow a particular scene in progress)

This pattern has the meaning of "fare attenzione e percepire/ capire il progredire di una certa scena in corso" (to pay attention and perceive/ understand the progression of a particular scene). Typical collocates are *partita*, *concerto*, *trasmissione*, *discussione*, which may refer to actual shows or, more in general, to collective activities progressing with time (jumping in place would not qualify as one) and in which the perceiver does not take part. As a spectator, she or he may witness the activity in person or via a medium, for instance the TV.

T7) Seguire una certa attività remota/ regolare in corso (to follow a particular remote/ regular activity in progress)

The meaning is "tenersi aggiornati sul procedere di una certa attività remota/ regolare" (to keep up to date on the progress of a particular remote/ regular activity). Typical collocates are *sport*, *calcio*, *vicenda*, *movimenti di qu.* Definition no. 16 of the COBUILD dictionary corresponds to our pattern:

"If you follow something, you take an interest in it and keep informed about what happens."

T8) Seguire qu. che sta narrando, spiegando o argomentando (to follow sb. who is telling a story, explaining, or making an argument)

The meaning is "fare attenzione e capire lo svolgimento della narrazione, della spiegazione o dell'argomentazione di qu." (to pay attention and understand the progression of sb.'s story, explanation, argument). Notice that here the activity in progress is not only perceived, but must be interpreted.

T9) Seguire una certa narrazione, spiegazione o argomentazione (to follow a particular story, explanation, or argument)

We found several typical collocates for the direct object, such as *lezione*, *logica*, *filo*, *ragionamento*, *argomentazione*, *racconto*, *spiegazione*. In the COBUILD, definition no. 15 corresponds to our pattern:

"If you are able to follow something such as an explanation or the story of a film, you understand it as it continues and develops."

I1) A un primo periodo/ situazione/ evento SEGUE un secondo periodo/ situazione/ evento (a second period/ situation/ event follows a first period/ situation/ event)

This pattern has the meaning "un secondo periodo/... viene immediatamente dopo un primo periodo/... in ordine temporale" (a second event/... comes immediately after a first event/... in time order). Collocates appearing as arguments were quite easy to identify and extremely heterogeneous: *caduta*, *dissoluzione*, *rielezione*, *proclamazione*, *bocciatura*, *sconfitta*, *dichiarazione*, *crollo*, *scoppio*, *terremoto*, *tracollo*, *sisma*, *ritrovamento* and many others. Definition no. 4 in the COBUILD dictionary corresponds to our pattern:

"An event, activity, or period of time that follows a particular thing happens or comes after that thing, at a later time."

I2) A una prima persona/ oggetto SEGUE una seconda persona/ oggetto (a second person/ object follows a first person/ object)

This pattern has the meaning "una seconda persona/ oggetto viene immediatamente dopo una prima persona/ oggetto in un ordine spaziale/ convenzionale" (a second person/ object comes immediately after a first person/ object in a spatial/ conventional order). The word sketches revealed no typical collocates. Hence we chose very general semantic types by introspection. Definition no. 7 in the COBUILD dictionary corresponds to our pattern:

"If you refer to the words that follow or followed, you are referring to the words that come next or came next in a piece of writing or speech."

I3) Un evento SEGUE DA un altro evento (an event follows from another event)

This pattern has the meaning "un evento è effetto di un altro evento" (an event is the effect of another event). Word sketches have not been particularly useful in this case. The only typical (idiomatic) collocation we could identify is "ne seguì" + [Evento] (an event followed from that), which hints at the fact that, in this pattern, *seguire* is just an abbreviation of *conseguire*, with precisely this meaning.

I4) Un'affermazione SEGUE DA un'altra affermazione (a statement follows from another statement)

This pattern has the meaning "un'affermazione è vera se è vera un'altra affermazione" (a statement is true if another statement is true). A statement is here the logic consequence of another. Also in this case, *seguire* seems to be an abbreviation of *conseguire* with the same meaning. Definition no. 6 in the COBUILD dictionary loosely corresponds to our pattern:

"If it follows that a particular thing is the case, that thing is a logical result of something else being true or being the case."

I5) Un testo SEGUE IN una parte di supporto testuale diversa dalla presente (a text follows in a different part of a textual carrier)

This pattern has the meaning "proseguire in un'altra parte di supporto testuale" (to continue in a different part of a textual carrier). The only typical collocate in the locative slot that emerges from the word sketches is *pagina*, indicating a 'textual place'. For the subject role we have chosen the semantic type [Testo], which covers all typical lexical items. In this case, *seguire* seems to be an abbreviation of *proseguire*.

3.2 Idiomatic sub-patterns and notable exploitations

We assigned the idiomatic expressions *seguire la corrente* and *seguire i passi/ le orme di qualcuno* to pattern T4. We did the same with a limited but significant number of similar figurative expressions, like *seguire il cuore/ l'istinto/ le inclinazioni/ gli impulsi* (to follow one's heart, instinct, inclinations, impulses)

In the concordance list found in Figure 1, we encounter the expression *seguire la voce di qu.* (to follow sb.'s voice). This is an exploitative alternation: *voce* (voice) → *narrazione* (story). In the same list, we also see an interesting example of a multiple exploitation. In the clause *lo sguardo segue la torre dall'alto in basso* (the eyes follow the tower from top to bottom), the tower is equated to a path along which the eyes can move. An interesting aspect of this exploitation is that a cognitive condition must be imposed on the tower for it to be compared to a path (it must have an elongated shape/ surface).

Figure 1: Excerpt from the Sketch Engine concordance list of *seguire*

3.3 Comparison with other resources

Following Sinclair's advice, we compared our results with those of existing resources, such as traditional dictionaries and ItalWordNet. What follows are the senses of *seguire* found on ItalWordNet, listed in exactly the same order but labelled according to our convention in order to highlight the similarities:

T1)   Synset: (seguire [1])
      Gloss: andare dietro a qlcu.

I1)   Synset: (seguire [2], succedere [3])
      Gloss: accadere successivamente o in conseguenza di qlco.

I3)   Synset: (avere_origine [2], conseguire [3], derivare [2], nascere [9], procedere [5], provenire [2], resultare [1], risultare [1], seguire [3], sorgere [6], uscire [11])
      Gloss: avere principio, essere causato (fig.); derivare, aver principio, origine, fondamento (fig.).

T2)   Synset: (seguire [4], tenere [7])
      Gloss: andare per un certo percorso.

T5)   Synset: (accompagnare [4], seguire [5])
      Gloss: seguire con lo sguardo, con il pensiero, ecc.

T4)   Synset: (conformarsi [1], seguire [6])
      Gloss: accettare un'idea, una dottrina e sim. "Seguire l'aristotelismo." "Seguire l'esempio di qc."

I4)   Synset: (conseguire [2], seguire [7], susseguire [2])
      Gloss: derivare come conseguenza, conseguire.

As aforementioned, the correspondence is remarkable: the main difference is that here senses T3, T6, T7, T8, and T9 seem to be missing. We think that this confirms the validity of our methodology.

As to the dictionaries, all problems lamented by Sinclair about traditional (pre-corpus) lexicography can be attested, e.g., the presence of long lost meanings (like sense C3 in ZINGARELLI: "accadere, avvenire: sono cose che seguono!"), abnormal examples, illogical splitting of meanings (like, in DE MAURO, senses '4a' vs. '5a': "mettere rigorosamente in pratica una regola, una norma, una convenzione" vs. "stare dietro all'evolversi di una tendenza uniformandosi ai suoi dettami"), illogical lumping.

The confirmation of Sinclair's thesis is indeed remarkable, and even the abstract semantic types which we identified are surprisingly robust (cf. Figure 2). The communication types, for example, correspond to Brinker's classification of texts (cf. Brinker, 2005). Furthermore, three out of four subtypes of [Comunicazione] immediately disambiguate to pattern T9, whereas [Indicazione/Descrizione] disambiguates to pattern T2 or T4 (and needs further disambiguation).
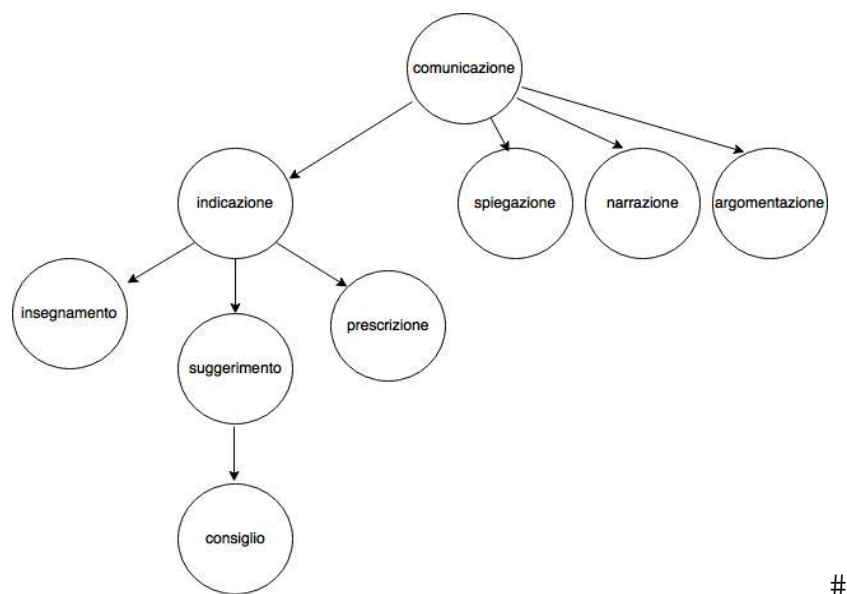
Figure 2: Communication types

4. Conclusions

As seen in our case study, Sinclair's legacy is more important than ever, most of all in those languages, such as Italian, where the corpus-driven approach is not yet mainstream. This is why the Italian advanced learner's dictionary we are currently designing with Zanichelli (which also aims at bridging an existing gap in Italian learner's lexicography) will be based on Sinclair's patterns of word usage.

The dictionary we are designing will have other important features, which we will introduce in upcoming articles. We will take into account the three mainstream approaches (cognitive linguistics, computational semantics, and lexical pragmatics) to the representation of polysemy in the mental lexicon and to its treatment in lexicography. Based on these, we will propose a user-oriented method for describing and differentiating word meanings. Disambiguators, as key microstructural items, will systematically apply in an ideal top-down procedure: ontological categories will distinguish lemmas and sub-entries (upper-level disambiguation), cognitive principles will determine word sense clusters (middle-level disambiguation) and Sinclair patterns will differentiate main word senses (lower-level disambiguation), whereas pragmatical principles will explain word sub-senses. In the enumeration and grouping of senses, we will prioritize semantic closeness criteria over frequency, since semantic closeness facilitates learning by association and is a key organising principle in our mental lexicon. Definitions will be created for each word sense by coherently employing a restricted defining vocabulary and by avoiding hidden circularities.

5. References

5.1 Dictionaries and lexicons

CALD: *Cambridge Learner's Dictionary.*
Accessed at: http://dictionary.cambridge.org/dictionary/english (22.05.2017)
COBUILD: *Collins COBUILD Advanced Learner's Dictionary* (2014). Harper Collins
COBUILD: *Collins COBUILD English Language Dictionary* (1987). Collins
DE MAURO: *Il Nuovo De Mauro.*
Accessed at: https://dizionario.internazionale.it (22.05.2017)
ItalWordNet. Accessed at: http://www.ilc.cnr.it/iwndb_php (22.05.2017)
LDOCE: Longman Dictionary of Contemporary English for Advanced Learners (2009). Pearson
MEDAL: MacMillan English Dictionary for Advanced Learner's. (2016). MacMillan
MultiWordNet. Accessed at: http://multiwordnet.fbk.eu (22.05.2017)
OALD: *Oxford Advanced Learner's Dictionary.*
Accessed at: http://www.oxfordlearnersdictionaries.com/ (22.05.2017)
PDEV: *Pattern Dictionary of English Verbs.* http://pdev.org.uk (22.05.2017)
TRECCANI: *Vocabolario Treccani.*
Accessed at: http://www.treccani.it/vocabolario (22.05.2017)
WordNet. Accessed at: http://wordnetweb.princeton.edu (22.05.2017)
ZINGARELLI: lo Zingarelli 2017 (2016).
Accessed at: http://dizionari.zanichelli.it/dizionarionline/online.php (22.05.2017)

5.2 Other resources

Bogaards, P. (2008). Frequency in learners' dictionaries. *Proceedings of the XIII EURALEX International Congress*, pp. 15-19.

Brinker, K. (2005). *Linguistische Textanalyse*, Erich Schmidt Verlag.

Carlson, G. (1977). *Reference to kinds of English*. Doctoral thesis, University of Massachusetts.

Carston, R. (2002). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind & Language*, *17*(1-2), pp. 127-148.

Cehan, A. (2015). The end of meaning-driven dictionaries?. *Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue. Section: Language and Discourse.*

El Maarouf, I. (2013). Methodological Aspects in Corpus Pattern Analysis. *ICAME Journal*, 37, pp. 119-148.

Falkum, I.L. (2011). *The Semantics and Pragmatics of Polysemy: A Relevance-Theoretic Account.* Doctoral Thesis, University College London.

Geeraerts, D. (2001). The definitional practice of dictionaries and the Cognitive Semantic conception of polysemy. *Lexicographica*, *17*, pp. 6-21.

Hanks, P. (2013). Lexical analysis: Norms and exploitations, MIT Press.

Hanks, P. (2008a). The lexicographical legacy of John Sinclair. *International Journal*

of Lexicography, 21(3), pp. 219-229.

Hanks, P. (2008b). Lexical patterns: From Hornby to Hunston and beyond. *Proceedings of the XIII EURALEX International Congress*, pp. 89-129.

Hanks, P. (2004a). Corpus pattern analysis. *Proceedings of the XI EURALEX International Congress*, Vol. 1, pp. 87-98.

Hanks, P. (2004b). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.

Hanks, P. (1987). Definitions and Explanations. In Sinclair, J. McH. (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary.* Collins COBUILD, pp. 116-136.

Hanks, P., & Bradbury, J. (2013). Why do we need pattern dictionaries (and what is a pattern dictionary, anyway)?. *DICTIONARY News*, *27*.

Hanks, P. & Jezek, E. (2008). Shimmering lexical sets. *Proceedings of the XIII EURALEX International Congress*, pp. 391-402.

Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, *10*(2), pp. 63-82.

Hanks, P., & Pustejovsky, J. (2004). Common Sense about Word Meaning: Sense in Context. Extended Abstract. In P. Sojka, I. Kopecek, I. & K. Pala (eds.) *Text, speech and dialogue*, Springer, pp. 15-17.

Hanks, P., Pala, K., & Rychlý, P. (2007). Towards an empirically well-founded semantic ontology for NLP. *Workshop on Generative Lexicon,* Paris, France.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The tenten corpus family. *7th International Corpus Linguistics Conference CL*, pp. 125-127.

Jezek, E. (2008). Polysemy of Italian event nominals. *Faits des langues*, *30*, pp. 251-264.

Jezek, E., & Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis. Journal in English Lexicology*, (4), pp. 7-22.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities* 31, pp. 91-113.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, *10*(2), pp. 135-155.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 The Sketch Engine. *Information Technology*, *105*, pp. 116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX International Congress,* pp. 425-431.

Kilgarriff, A., & Gazdar, G. (1995). Polysemous relations. In F.R. Palmer (ed.) *Grammar and Meaning: Essays in Honour of Sir John Lyons*, Cambridge University Press, pp. 1-25.

Kozaki, K, Sunagawa, E., Kitamura, Y. & Mizoguchi, R. (2006). Fundamental consideration of role concepts for ontology evaluation. *Proceedings of EON2006*,

Edinburgh.

Kratzer, A. (1995). Stage-level and individual-level predicates. In G. N. Carlson & F. J. Pelletier (eds.) *The generic book*, University of Chicago Press, pp. 125-175.

Krishnamurthy, R. (2008). Corpus-driven lexicography. *International Journal of Lexicography*, *21*(3), pp. 231-242.

Lakoff, G. (1999). Cognitive models and prototype theory. In E. Margolis & S. Laurence (eds.) *Concepts: Core Readings*, MIT Press, pp. 391-421.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*, University of Chicago Press.

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (ed.) *The Bloomsbury companion to lexicography*, Bloomsbury, pp. 284-302.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, John Benjamins Publishing, pp. 157-176.

Lyons, J. (1977). *Semantics*, Cambridge University Press.

Moon, R. (1987). The analysis of meaning. In J. McH. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, Collins COBUILD, pp. 86-103.

Nerlich, B., Todd, Z., Herman, V., & Clarke, D. D. (eds.). (2003). *Polysemy: Flexible patterns of meaning in mind and language* (Vol. 142), Walter de Gruyter.

Pustejovsky, J. (2001). Type construction and the logic of concepts. In: P. Bouillon & F. Busa (eds.) *The language of word meaning*, Cambridge University Press, pp. 91-123.

Pustejovsky, J. (1995). *The Generative Lexicon*, MIT Press.

Pustejovsky, J., Hanks, P., & Rumshisky, A. (2004). Automated induction of sense in context. *Proceedings of the 20th international conference on Computational Linguistics*, *COLING*, pp. 924-930.

Roventini, A. et al. (2003). ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian. *Computational Linguistics in Pisa, Special Issue XVIII-XIX* Pisa-Roma IEPI. Tomo II, pp. 745-791.

Rundell, M. (2006). More than One Way to Skin a Cat. Why Full-Sentence Definitions Have not Beeen Universally Adopted. *Proceedings of the XI EURALEX International Congress,* pp. 323-337.

Sinclair, J. McH. (2004a). *Trust the text: Language, corpus and discourse*, Routledge.

Sinclair, J. McH. (2004b). In Praise of the Dictionary. *Proceedings of the XI EURALEX International Congress*, pp. 1-12.

Sinclair, J. McH. (2004c). New evidence, new priorities, new attitudes. In Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching*, John Benjamins Publishing, pp. 271-299.

Sinclair, J. McH. (2003). *Reading concordances: An introduction*. Pearson Longman.

Sinclair, J. McH. (1998). The lexical item. In E. Weigand (ed.) *Contrastive Lexical Semantics (Current Issues in Linguistics Theory 171)*, John Benjamins

Publishing, pp. 1-24.

Sinclair, J. McH. (1996). The search for units of meaning. *Textus: English Studies in Italy* 9(1), pp. 75-106.

Sinclair, J. McH. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Sinclair, J. McH. (1987). Grammar in the Dictionary. In J. McH. Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, Collins COBUILD, pp. 104-115.

Sinclair, J. McH. (1986). First throw away your evidence. In G. Leitner (ed.) *The English Reference Grammar,* Max Niemeyer Verlag, pp. 56-65.

Sinclair, J. McH., Jones, S. & Daley, R. (2004). *English Collocation Studies: The OSTI Report,* Continuum, xvii-xxix.

Stewart, D. (2010). *Semantic Prosody: A critical Evaluation.* Routledge.

Tarp, S. (2001). Lexicography and the linguistic concepts of homonymy and polysemy. *Lexicographica, 17,* pp. 22-39.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (Vol. 6), John Benjamins Publishing.

Vossen, P. (2002) (ed.) *EuroWordnet. General Document*, Version 3. Accessed at: http://dare.ubvu.vu.nl/bitstream/handle/1871/11116/EWNG?sequence=1 (22.05.2017).

Vossen, P. et al. (1998). *The EuroWordNet Base Concepts and Top Ontology.* Document LE2-4003, D017, D034, D036, WP5. Accessed at: http://dare.ubvu.vu.nl/bitstream/handle/1871/11130/D017.pdf (22.05.2017).

Whitsitt, S. (2005). A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10(3), pp. 283-305.