Translated version of Amaro, Raquel & Sara Mendes (2016) "Lexicologia e Linguística Computacional", in Martins, A. M. & Carrilho, E. (ed.) Manual de Linguística Portuguesa, Manuals of Romance Linguistics, Mouton de Gruyter, pp. 178-199.

Lexicology and Computational Linguistics

Abstract: With as aim the automatic processing of the meaning of lexical units, the intersection of Lexicology and Computational Linguistics is an area of basic research in the field of Natural Language Processing. In this context, the lexical specification, i.e., the determination of the characteristics and the level of granularity pertinent to an adequate representation of lexical knowledge, is an essential aspect to be defined, since the modeling of this knowledge depends largely on this. Based on research developed for Portuguese, this chapter presents a description of lexical properties and phenomena essential to Computational Linguistics (e.g. the relation between semantic properties and syntactic behavior, ambiguity and creation of meaning in context), a brief discussion of empirical data that highlights the adequacy of lexical representation models and the integration of research results in WordNet.PT.

Keywords: lexical knowledge, modeling, wordnet, generative lexicon, Portuguese

1. The Nature of Lexical Knowledge

The study of the lexicon has always played a central role in the progressive discovery of the functioning of the language system and in the deepening of linguistic knowledge. Lexicology, a field of Linguistics dedicated by nature to the study of the lexicon, includes, in the definition we consider in this work, topics such as etymology, morphology, variation of meaning and lexical semantics, i.e., the study of synchronic phenomena related to the meaning of lexical items. Currently, in Lexicology, the study of the lexicon extends far beyond the chronological classification of meanings in the scope of Lexicography and the description of the laws that govern semantic change in the scope of etymology. In fact, today, the regularities identified are related to the actual use of language, on the one hand, and to the nature of the human mind and the conceptualization of knowledge, on the other, thus also associating with the field of Psychology (Geeraets 2010: 9).

In fact, the lexicon is an essential component of natural languages and plays a central role in the current linguistic theories and, of course, in the design and theory of systems for the processing

of natural languages since its inception (Nirenburg et al., 1987; et al., 1987, Guthrie et al., 1996, etc.). The intersection of Lexicology and Computational Linguistics is thus an area of fundamental research in the field of Natural Language Processing, since it has as general objective the automatic processing of the meaning of the lexical units (thus making possible, from the combination of these basic units of languages, also the processing of the meaning of the sentences in which they occur), considering the determination of the nuclear meaning of the lexical units, the interaction between their semantic properties and their syntactic properties and the formal representation of that knowledge, in a synchronic perspective of natural languages. Thus, the development of an ever deeper and more detailed knowledge about the lexical units and how they relate to each other within the lexical system, as well as the way in which the interface between this grammar module and the others, is an area of research in Linguistics with great impact, in particular on the performance of natural language processing systems.

1.1 Lexical units: complex and multifaceted units

The lexicon is par excellence the repository of information that allows the language system to deal with complex phenomena such as the creation of meaning in context or the different types of lexical ambiguity. To give an account of phenomena such as these, as well as of the fact that the lexicon is one of the most dynamic modules of the grammar – it is the one that changes most in very short periods of time, a fact that is largely due to the close connection between the lexicon and the representation of reality that is intended to be communicated and that is constantly evolving – implies a conception of the lexicon fundamentally different from that behind the construction of traditional lexical resources. These are typically characterized by a static conception of the lexicon, of enumerative nature, that is, in which the set of possible meanings of each lexical unit, as well as its contexts of use, is determined in advance, and in which there are no mechanisms that allow expanding it.

The observation that the construction of meaning in context presents regularities, and that these depend on particular properties that are common to certain sets of lexical items, results in alternative conceptions of lexicon that describe the lexical units no longer as atomic and indivisible units, but rather as units characterized by different aspects of meaning, or facets, by properties that are diverse and related to different levels of representation, which can be described individually and independently of each other, although they often relate to and influence each other.

Conceiving lexical units as complex and multifaceted units, encoding a range of aspects representative of lexical meaning, allows, on the one hand, a more adequate description of lexical phenomena and, on the other hand, a more comprehensive coverage of the observed data. In fact, in this type of lexicon design, these meaning components simply define the

semantic limits of their use. When considered in the context of other words, however, mutually compatible facets in the semantic information of each lexical unit become more prominent, determining specific interpretations.

Finally, this conception of the lexicon and lexical unit allows us to objectively describe aspects that preside over the constitution of lexical classes and that determine regularities in the syntactic-semantic behavior of sets of lexical items.

1.2 Ambiguity and construction of meaning in context

The traditional organization of lexical resources assumes that the meaning of lexical units can be defined exhaustively by a given number of meanings. The result of this approach is that whenever the interpretation of a statement is confronted with a situation of lexical ambiguity it is necessary to use mechanisms of disambiguation that seek to select from the list of enumerated meanings the most adequate definition available, based on the establishment of correspondences between the characterization of the meanings of a lexical unit and factors related to the context.

One of the main disadvantages of this type of approach stems from the need to specify in advance the contexts in which a given lexical unit can occur, and an incomplete description results in poor coverage of the system. Moreover, this conception of lexical resources reveals a static view of the lexicon, since the division of the meaning of lexical items into independent meanings blocks the permeability of meanings and does not allow us to account for the creative use of words in new contexts.

Alternative models of the lexicon, in which the characterization of the meaning of the lexical units is by nature open, allow overcoming the limitations of the traditional concepts of lexicon referred to above. This is achieved through schemes of explicit coding of lexical knowledge at different levels of generalization, in particular by making the resolution of lexical ambiguities an integral part of a uniform semantic analysis procedure. The mechanisms integrated in these procedures of analysis operate on a set of basic meanings, with more internal structure than what was traditionally assumed, allowing to widen the set of meanings of the individual lexical units when these are considered together in more comprehensive expressions. In this way, this type of lexical model does not presuppose a finite enumeration of meanings to represent lexical knowledge and allows accounting for the creative use of language.

Given the aspects covered in this section, it becomes clear that the nature of lexical knowledge is therefore complex and dynamic, involving conceptual properties (mental modeling of world knowledge), semantic properties (structure of meaning, thematic relations, combinatorial constraints and relations between the different units that make up the system of the lexicon) and syntactic properties (syntactic realization, (inter)dependence and relation between meaning and

context). In order to account for these properties of the lexicon, several computational models have appeared, whose main characteristics are presented below, in this chapter.

2. Computational models and lexicon representation

The computational models of representation of the lexicon have as purpose determining and describing the meaning and properties of lexical units in order to allow the use of this information in the automatic processing of languages. Based on structuralist principles, these models assume the independence of linguistic knowledge relative to world knowledge, and the idea that the formalization of such knowledge is possible, and can be generically grouped into two types of approaches: decompositional approaches and relational approaches.

The so-called decompositional approaches (Wierzbicka 1972, 1996, Jackendoff 1990, Bierwish 1971, Talmy 1985, Pustejovsky 1995) base the description of the meaning of the lexical units in smaller units, all sharing the interest in the interaction between the lexicon and cognition in a wider sense, either through the search of cognitive foundations for the descriptions of meaning, either through the search for the interface phenomena between semantics and contextual and non-linguistic information. On the other hand, the so-called relational approaches (Melcuk 1988, 1998, Miller et al., 1990; Fellbaum 1998) are based on the meaning of the lexical unit as a whole and describe this meaning based on the relations that the lexical units establish between themselves or between them and units of other levels, such as syntagmatic units[1].

Among the computational models of lexical representation, we highlight the Generative Lexicon (Pustejovsky 1995), for the proposal of a formal representation of the meaning of the lexical units and by the clear computational objective that underlies the construction of the model and the different proposed mechanisms and the model of WordNet (Miller et al., 1990; Fellbaum 1998), for the appropriateness of large-scale computational lexical resources, making it usable in natural language processing tasks.

2.1 Generative models of the lexicon: generation of meaning in context from the properties of lexical units

The Generative Lexicon model (Pustejovsky 1995) sees the lexicon as a complex and dynamic system that allows the generation of meaning in context from the information associated with each lexical item and a small set of mechanisms that establish general principles to combine and select information or part of it. This model contrasts, therefore, in several aspects with the enumerative models of the lexicon.

In the first place, it is distinguished from these in that the basic unit of the model is the lexical item with defined but subspecified meaning, which can integrate different types of lexical

[1] For a more in-depth view of the evolution of lexical theories and models, see Geeraets (2010).

ambiguity. The model thus recovers the notion of the basic meaning of Bréal (1987). According to this author, the basic meaning corresponds to the knowledge shared by the linguistic community that allows the efficient use of the language. From the definition of this subspecified base meaning, which is concretized in each specific context, it aims at predicting and explaining the behavior of the lexical units in context and their use, whenever such properties come directly from the information associated with a given lexical item.

Secondly, and although, as mentioned above, the basic unit of the Generative Lexicon model is in fact the lexical item, this is conceived as a complex unit, composed of several smaller elements that together form the meaning of each lexical unit. This way of conceiving the lexical items allows, on the one hand, to explain the sharing of characteristics between different lexical units, in particular regular alternations of meaning, and, on the other hand, to make lexical coding more economical, since it is possible to specify certain types of information only once and then share it between different lexical items. In this sense, the Generative Lexicon model is a decompositional model, in which the representation of lexical units as informational structures – whose content is broken down into features and values – follows a defined set of rules, allowing to formalize the creation of meaning in context and the interface between syntax and semantics, thus making evident common regularities and principles at the level of the behavior of sets of lexical items.

The Generative Lexicon uses a declarative and systematic approach to represent the lexical units and to account for their semantic and syntactic behavior. The lexical information is represented by attribute-value matrices, whose values can be atomic or, recursively, constituted by other matrices, structured in 4 levels of representation: argument structure, event structure, qualia structure and lexical inheritance structure.

The **argument structure** includes the definition of the semantic properties of the logical arguments of a given lexical item, as well as information about their syntactic mapping[2]. The **event structure** is constituted by the declaration of the events that constitute the represented event, by the stating of the temporal order constraints between the listed subevents and by the specification of the most prominent subevent. The **qualia structure** is composed of four attributes – qualia roles – and their respective values: constitutive (CONST) (that expresses the relation between a given object and its constituent elements); formal (FORM) (establishing the

[2] The generative lexicon includes 4 distinct types of attributes: **proper arguments** (P-ARGn) (parameters of the semantic content of the lexical item that correspond to the semantic object denoted by the lexical item in question), **true arguments** (ARGn) (parameters of the semantic content of the lexical item whose syntactic omission is only allowed when retrievable by the context), **default arguments** (D-ARGn) (parameters that enter the logical expressions of the qualia structure of a lexical item, but that are not necessarily syntactically expressed and can only be expressed through subtyping or speech specification operations) and **shadow arguments** (S-ARGn) (parameters incorporated in the semantics of a lexical item, which can only be expressed through subtyping operations or speech specification) (Pustejovsky 1995; Amaro, 2009).

stable properties that distinguish a given object within its semantic domain); telic (TEL) (that refers to the function or purpose of the object or event); and agentive (AG) (that determines the origin or causal chain of the object or event). Lastly, the **lexical inheritance structure** establishes the conditions for lexical inheritance, through qualia vectors, that is, links between nodes in the network of types established according to the qualia roles. Finally, it is important to mention that the model allows the representation of information sharing between structures through a unification mechanism.

However, the complexity and richness of the information associated with lexical items in the Generative Lexicon is not, in itself, sufficient to account for their dynamic behavior in context. Therefore, the Generative Lexicon proposes three generative mechanisms: **type coercion** (that consists of a semantic operation that converts an argument of a given type into the type expected by the argument selection of a predicate, without this implying a change in the syntactic realization of the argument (Pustejovsky 1995, 2001, 2007)); **selective binding** (to account for the relationship between modifiers and modified entities, allowing the modifier to select a specific argument from the set of declared values in the semantic content of the modified object); and **co-composition** (that enables the completion of the meaning of a predicate with values present in the semantic content of its arguments).

The Generative Lexicon also proposes the classification of lexical items into semantic types, which, in addition to allowing a organizing type structure for the lexicon, makes the defined mechanisms adequate to express the relations between the semantic objects and between them and the corresponding syntactic realizations.

Let us consider the sentences in (1) (adapted from Mendes 2009: 125) as an example and look at the (simplified) structures of the Generative Lexicon that represent them, in (2). These examples illustrate the representation levels of the Generative Lexicon and how representations in this model allow accounting, in a systematic and simple way, for the semantic restrictions that explain the co-occurrence restrictions of lexical items.

(1) a. A Ana escamou o peixe. (Ana scaled the fish.)
b. *A Ana escamou as escamas (do peixe). (*Ana scaled the scales from the fish)
c. *A Ana escamou a ave. (*Ana scaled the bird.)

(2)
$$\begin{bmatrix} peixe\ (\text{fish}) \\ \text{ARG} - \text{ST} = \ [\text{P} - \text{ARG1} = \text{x: animal}] \\ \text{QUALIA} = \ [\text{CONST} = \text{has\_part(x, y: scale)}] \end{bmatrix}$$

$$\begin{bmatrix} ave\ (\text{bird}) \\ \text{ARG} - \text{ST} = \ [\text{P} - \text{ARG1} = \text{x: animal}] \\ \text{QUALIA} = \ [\text{CONST} = \text{has\_part(x, y: feather)}] \end{bmatrix}$$

$$\begin{bmatrix} escamar\ (\text{scale}) \\ \text{ARG} - \text{ST} = \begin{bmatrix} \text{ARG1} = \text{x: humano} \\ \text{ARG2} = \text{y} \\ \text{S} - \text{ARG1} = \text{z: scale} \end{bmatrix} \\ \text{EVENT} - \text{ST} = \begin{bmatrix} \text{E1} = \text{e1: process} \\ \dots \end{bmatrix} \\ \text{QUALIA} = \ [\text{FORM} = \text{remove}(\text{e1}, \text{x}, \text{y}, \text{z})] \end{bmatrix}$$

The statement of the property has_part `scale`, distinctive of *fish*, on the one hand, and the declaration of an argument (of the type shadow, that is, whose content is incorporated in the semantics of the predicate) of the semantic type `scale` associated with the verb *escamar (scale)*, on the other, correctly predicts the grammaticality of the sentence in (1)a and the grammaticality of the sentences in (1)b and (1)c.

2.2 Relational models of the lexicon: the lexicon as an organized set of interrelated units

As previously mentioned, the WordNet model (Miller et al., 1990; Fellbaum 1998) is currently the computational model of the lexicon with greater relevance in the creation of computational lexicons and with repect to the use of these resources in applications in the field of Computational Linguistics (Hanks 2003). In addition, insofar as it originated in research on the organization of the mental lexicon, the WordNet model also has a strong psychological motivation.

Wordnets are electronic databases structured as a network of relations between the nodes that constitute their basic units, the *synsets* (designation of sets of synonyms that lexicalize a given concept, and whose meaning is defined, in this model, by the relations established between them and the remaining nodes of the network). The relations that determine the meaning of a synset are lexical (synonymy), lexical-conceptual (hyperonymy / hyponymy, meronymy, etc.), of function or thematic role (agent relation, involved instrument relation, etc.), semantic opposition relations (antonyms, quasi-antonyms) and causal relations (is caused by / cause, etc.).

The major structuring relationship in the WordNet model is the relationship of hyperonymy / hyponymy, defined as

(3) X is a hyperonym of Y if Y is a type of X and X is not a type of Y.

The relation of hyperonymy / hyponymy is a lexicon-conceptual relation that simultaneously contemplates world knowledge, by its ontological character, and linguistic knowledge, as shown in the following examples, in which the hyperonym is used to refer a more specific entity (the hyponym) previously presented (Amaro 2009: 25):

(4) a. Ele comprou um pastor alemão$_{hyponym}$, mas o cão$_{hyperonym}$ não morde.
(He bought a German shepherd$_{hyponym}$, but the dog$_{hyperonym}$ does not bite.)
b. Ele rastejou$_{hyponym}$ pela floresta, movendo-se$_{hyperonym}$ assim para evitar ser visto.

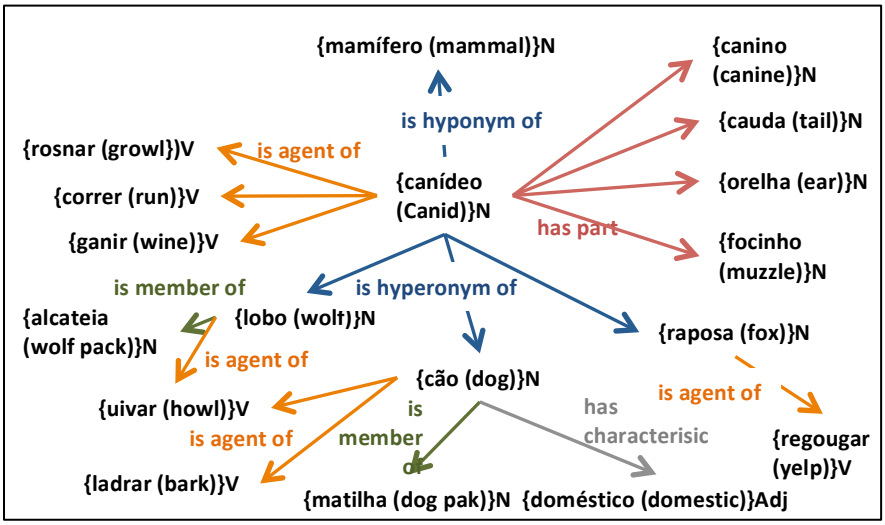
(He crawled$_{hyponym}$ through the forest, moving$_{hyperonym}$ like this to avoid being seen.)

c. Os animais marinhos$_{hyponym}$ são ameaçados pela proximidade de cidades, uma vez que os animais aquáticos$_{hyperonym}$ são facilmente afetados pela poluição dos esgotos.

(Marine$_{hyponym}$ animals are threatened by the proximity of towns, as aquatic$_{hyperonym}$ animals are easily affected by sewage pollution).

The relation of hyperonymy / hyponymy allows organizing the lexical items according to their type, since each hyponym has all the properties of its hyperonym plus some other aspect that characterizes it specifically and distinguishes it from all its cohyponyms[3]. In this way, the hyperonymy / hyponymy relationship simultaneously contemplates the definition of a monotonic inheritance mechanism – by contrast, for example, to what happens in the Generative Lexicon model described above – that allows the description of lexical items in a very economical way. Hyponyms inherit the conceptual properties of their hypernyms, as illustrated in the figure below (adapted from Marrafa et al. (2006)).

(5)



Through the relations of hyponymy / hyperonymy, in the figure in (5), the nodes {wolf}$_N$, {dog}$_N$ and {fox}$_N$ inherit all the properties established for their hyperonym, {canid}$_N$: have as part {canine}$_N$, {tail}$_N$, {ear}$_N$, {muzzle}$_N$; and are agents of {growl}$_V$, {run}$_V$ and {wine}$_V$. These three cohyponyms, on the other hand, distinguish themselves from each other and from their hyperonym by the relations that are particular to them: for example, {wolf}$_N$, in addition to being agent of {growl}$_V$, {run}$_V$ and {wine}$_V$, properties inherited from the hyperonym, is the only one that is member of {wolf pack}$_N$; {dog}$_N$ is the only one that is {domestic}$_{Adj}$, is agent of {bark}$_V$ and member of {dog pack}$_N$; {fox}$_N$ is the only agent of {regougar (yelp)}$_V$.

[3] By co-hyponyms we mean all nodes of a lexical-conceptual network that share a same direct hyperonym.

This brief presentation of the main characteristics of these two lexical models makes evident the great expressive capacity of these alternative approaches to the traditional models of the lexicon. Sections 3, 4 and 5 below discuss some concrete empirical phenomena that raise issues in terms of lexical coding and present proposals for modeling within the framework of these two models, demonstrating their ability to adequately represent complex linguistic phenomena.

3. Lexical modeling and specification

As mentioned in the brief description of the characteristics and relations between Lexicology and Computational Linguistics presented so far, the modeling of lexical knowledge aims at determining the nuclear meaning of lexical units and the representation of the semantic and syntactic properties associated with them. Thus, the lexical specification, that is, the determination of the characteristics and level of granularity pertinent to an adequate representation of the lexical knowledge, is an essential aspect and, in turn, inseparable from the lexical representation models.

3.1 Identification of facets of meaning and definition of lexical classes

The division of the lexicon into semantic fields or domains (eg, food, living beings, health, communication, etc.) can be seen as a first example of the use of meaning for the establishment of subsets of lexical items with shared properties. However, the semantic domain approach (Trier 1931, 1934, Wittgenstein 1965; Vassilyev 1974; Lyons 1977), that is, the definition of sets of lexical items with related meanings, differs from the definition of lexical classes insofar as in the first, unlike what happens with the second, there can be objectively no property, semantic or syntactic, shared between members of the same domain. The case of *cough* (violently expell air through the throat with noise) and *forceps* (a tweezer instrument used to pull the baby's head out of the uterus of the parturient)[4] illustrate this case, since these two lexical units, although belonging to the field of health, do not share any kind of linguistic property.

The definition of lexical classes, in the perspective considered here, implies sharing of semantic and syntactic properties between the items of a same class that facilitates the process of their representation and the prediction of properties and generalized behaviors.

Based on this assumption, Mendes (2009) presents an exhaustive analysis of Portuguese adjectives and proposes the definition of adjective classes based on the identification of facets of meaning. The first aspect to be considered relates to the type of denotation that adjectives can assume: **property ascribing adjectives**, which denote states and whose contribution to the delineation of the denotation of the nominal phrase results from the addition of restrictions to

[4] Examples adapted from Marrafa *et al*. (2006).

the denotation, as exemplified in (6); and **non-restrictive adjectives**, which denote a semantic function (negation, probability, etc.) and whose contribution to the denotation of the noun phrase operates at the level of the intention of the modified name, functioning as modal operators, as exemplified in (7).

(6) a maçã luzidia (x é maçã e x é luzidio) (the shiny apple (x is an apple and x is shiny))

(7) o diamante falso (x não é diamante) (the fake diamond (x is not a diamond))

The class of property ascribing adjectives contemplates most of the adjectives of Portuguese, containing several subclasses that reflect distinctive semantic properties. For example, it is possible to group **descriptive adjectives**, adjectives assigning state-denoting properties that correspond to a single property, that select and modify a certain aspect of the meaning of the name, such as color or size (see (8) and 9)). However, even within the class of descriptive adjectives, it is possible to distinguish two subclasses distinct as to how the denotation of the noun phrase is established: in the case of adjectives such as *red*, **absolute** adjectives, the denotation of the noun phrase is the result of an intersection of sets (*red dress* = red ∩ dress, ie, an object that belongs simultaneously to the set of dresses and the set of entities that are red); in the case of adjectives as *small*, **relative** adjectives, the denotation of the noun phrase consists of a set that is included in a more comprehensive one, which functions, in essence, as a comparison class (*small elephant* = small elephant ⊂ elephant, ie, an object that belongs to the group of entities that are small in the universe of elephants). The elements of the class of relative adjectives thus share the property of retrieving from the content of the modified name the universe or the class of comparison that allows the complete determination of its denotation.

(8) o vestido vermelho (o vestido cuja cor é vermelha)
(the red dress (the dress whose color is red))

(9) o elefante pequeno (o elefante cujo tamanho é pequeno (para um elefante))
(he small elephant (the elepnaht whose size is small (for an elephant)))

Finally, within the property ascribing adjectives we can still consider the subclass of relational adjectives. These are distinguished from the descriptive ones in that they are not associated with an individual property but rather with sets of properties. As with names, relational adjectives denote more complex properties and usually establish with the name modified more diverse semantic relations.

(10) a. o orçamento municipal (x é um orçamento que tem uma relação com o município)
(the municipal budget (x is a budget that has a relationship with the municipality)

b. o animal marinho (x é um animal que pertence/é parte do mar)
(the marine animal (x is an animal that belongs / is part of the sea)

In fact, relational adjectives introduce sets of properties that typically correspond to properties that characterize a name (see examples in (10)) and determine more complex semantic relationships that may be of different nature. Thus, in (10) a, the adjective determines a subspecified relation (R1) between the modified name and *municipality* (budget x) ∧ R1 (x, municipality), heavily dependent on the context: *municipal budget* (available for the municipality), *municipal constructions* (accomplished/funded by the municipality). In (10) b, the relationship determined by the adjective is a relation of belonging or quasi-meronymy (animal (x) ∧ x ∈ sea), which is different, for example, from the quasi-synonym relational adjective *maritime* (R1 (x, sea)).

This brief empirical description of the case of adjectives thus exposes the possibility of establishing lexical classes based on the identification of a small set of semantic properties susceptible of being formalized. It is also important not to lose sight of the fact that lexical items that share these properties show similar linguistic behavior, particularly in terms of syntactic distribution, as presented in detail in section 3.2.

3.2 Relationship between facets of meaning and syntactic behavior

The meaning facets, which correspond to semantic properties, are also closely related to the syntactic behavior of the lexical units. The argument selection restrictions, for example, can be though of as the case that first illustrates this relation: *write* selects a human type argument as an agent, for example, explaining the grammaticality of sentences in (11)a and the semantic malformation of sentences in (11)b.

(11) a. O rapaz/padeiro escreveu a carta. (The boy / baker wrote the letter.)
b. #A lombriga/bicicleta escreveu o testamento. (#The worm / bicycle wrote the will.)

Going back to the adjectives described in the previous section, it is possible to verify that the properties that allow defining lexical classes are directly related to syntactic properties. For example, non-restrictive adjectives do not occur with adverbial modifiers or in comparative constructs, because they do not denote states that correspond to properties (examples in (12) taken from Mendes (2009: 49)). Relational adjectives distinguish from descriptive ones because they hardly occur in predicative contexts (cf. (13)a), do not occur in a pre-nominal position (cf. (13)b) and do not occur with adverbs of degree cf. (13)c), except in exceptional cases[5].

[5] There is a margin of acceptability for constructions in which relational adjectives co-occur with adverbs of degree, related to the possibility of reinterpreting these adjectives as descriptive. For an in-depth

(12) a. *o diamante muito falso (*the very fake diamond)

b. *um diamante mais falso do que o outro (*a diamond more fake than the other)

(13) a. *As casas são rurais. (*The houses are rural.)

*Adoro as casas que são rurais. (*I love houses that are rural)

b. *Adoro as rurais casas. (*I love rural houses.)

c. ?*Adoro as casas muito rurais. (?*I love very rural houses.)

Semantic properties related to the internal structure of events and with meaning facets embedded in the semantic content of the lexical units may also explain different behaviors of semantically related lexical items. Take the Portuguese verbs of movement as an example. Verbs that express movement (i.e., change of location in a given time interval) share, though with more or less specific constraints, their argument structure (Amaro, 2009). For example, the verbs *mover-se* (move) (change one's own location) and *regressar* (return) (move back to the place of departure) select an argument of animated type. However, they do not necessarily co-occur with the same adverbial phrases:

(14) a. Ele moveu-se durante meia hora./*Ele moveu-se em meia hora.

(He moved for half an hour./*He moved in half an hour.)

b. ?*Ele regressou durante meia hora./Ele regressou em meia hora.

(?*He returned for half an hour./*He returned in half an hour.)

This different syntactic behavior can be directly related to the internal structure of the event denoted by each one of the considered verbs, which, in turn, can be explained by their semantic content. Retrieving the notion of semantic incorporation proposed by Talmy (1985), it is possible to describe the semantic content of the verb *regressar* (return) assuming that it incorporates in its meaning the semantic element GOAL, defined as the final location of the object that experiences / participates in the event, a description that is consistent with the intuitive definition of *regressar* (return). Thus, *regressar* (return) incorporates a final location, which determines a final state that, added to the event denoted by *mover-se* (move), results in a transition event with different *Aktionsart* properties (Vendler 1967; Móia 2000) [6].

These properties allow to predict the different syntactic behavior of these verbs illustrated in (14): an activity, event not delimited temporarily, does not allow the specification of a time

---

discussion and explanation of the exceptional conditions that allow relational adjectives to occur in this type of contexts, see Mendes (2009).

[6] There are four classes of Aktionsart: **state** – event not temporally delimited, homogeneous and simple (eg, *high*); **activity** - simple event not temporally delimited relatively homogeneous (i.e., with repetition of subevents at regular intervals) (eg *run*); **accomplishment** - a event not temporally delimited event, heterogeneous and complex, consisting of a preparatory process, a completion point and a consequent state (which, in the previous type typology, roughly corresponds to a transition) (eg *build*); **achievement** - a point event, heterogeneous and complex, consisting of a culmination point and a consequent state (eg *die*).

interval that presupposes a final state (*in half an hour*) (see (14)a); on the contrary, an accomplishment, being an event delimited by the culmination point and consequent final state, does not allow the specification of a time interval that does not presuppose this same final state (*for half an hour*) (see (14)b). Another example of the relationship between facets of meaning and syntactic behavior can be seen in the co-occurrence constraints between co-hyponyms, shown here in (15).

(15) a. Ele subiu a rua correndo. (He ascended/went up the street running.)

b. Ele subiu a rua descendo. (*He ascended/went up the street descending/going down.)

The verbs *ascend*, *run* and *descend* are subtypes (hyponyms) of the verb *move*: to *ascend* is to *move* up; *run* is to *move* on the ground, quickly, using the limbs; to *descend* is to *move* down. The relationship of hyponymy between lexical items reflects the lexicalization of different semantic components, which also distinguishes sibling nodes (Fellbaum 1998). According to Mendes & Chaves (2001) and Amaro (2009), this distance of meaning explains the incompatibility of some co-hyponyms. Thus, hyponym items are compatible if they do not lexicalize different values for the same semantic aspect. In the example above, the verbs *ascend* and *descend* lexicalize different values (up and down, respectively) of the same aspect, DIRECTION, and therefore are not compatible. For its part, *run* lexicalizes a value for the MODE aspect, being therefore compatible with *ascend* and *descend*.

3.3 Lexical coding and construction of meaning in context

In addition to the relation between the facets of meaning and the definition of classes, on the one hand, and the syntactic behavior of lexical units, on the other hand, lexical modeling and specification have to account for the polymorphic properties of lexical units, particularly the construction of meaning in context. This implies that the models of representation of the lexicon allow an adequate lexical codification of the properties of lexical items, also taking into account the sharing of these properties, in order to guarantee the economy of the process of lexical knowledge representation and the utility and viability of the resulting lexical resources.

Let us return to the case of adjectives, which, by definition, tend to exhibit great plasticity of meaning, due to their semantic subspecification. For example, the adjective *bom* (good), depending on the noun it occurs with, appears to have different meanings, as illustrated in (16) (Mendes 2009):

(16) a. um bom professor = um professor que ensina bem (a good professor = a professor that teaches well)

b. uma boa faca = uma faca que corta bem; uma faca bem construída. (a good knife = a knife that cuts well; a knife that is well built)

The lexical coding of the properties of these adjectives in the framework of the Generative Lexicon allows for, with a single lexical entry, and through the mechanism of selective binding, accounting for the relationship established between the modifier and the modified entities. Selective binding, as the name implies, consists of a semantic operation that allows the modifier to select a specific argument from the set of values declared in the semantic content of the modified object. In this way, and considering the lexical entries represented in (17), it is possible to verify that the adjective *good* has a subspecified semantic content (positive evaluation), whose particular meaning is defined in context, from the connection established with an argument present in the structure qualia (represented by index 1) of the modified noun. Thus, when modifying *knife*, *good* can positively evaluate *build* or *cut*; when modifying *teacher*, *good* evaluates positively *teach*.

(17)
$$\begin{bmatrix} bom\ (good) \\ \text{ARG} - \text{ST} = \left[\text{ARG1} = \left[\overset{\ldots}{\text{QUALIA}} = [1]\right]\right] \\ \text{EV} - \text{ST} = [\text{E1} = \text{e1: state}] \\ \text{QUALIA} = [\text{FORM} = \text{positive\_evaluation(e1, [1])}\ ] \end{bmatrix}$$

$$\begin{bmatrix} faca\ (knife) \\ \text{ARG} - \text{ST} = [\text{P} - \text{ARG1} = \text{x: tool}] \\ \\ \text{QUALIA} = \begin{bmatrix} \text{AG} = \text{build(e1, y, x)} \\ \text{TEL} = \text{cut(e2, y, z, x} \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} professor\ (teacher) \\ \text{ARG} - \text{ST} = [\text{P} - \text{ARG1} = \text{x: human}] \\ \\ \text{QUALIA} = [\text{TEL} = \text{teach(e1, x, y]} \end{bmatrix}$$

The case of relative adjectives, described in section 3.1, is another example of creation of meaning in context, in the sense that these adjectives need to recover from the content of the modified name the comparison class that allows the complete determination of their denotation. The attribute-value matrices in (18) (Mendes 2009: 153) illustrate how this plasticity is codified in the Generative Lexicon, making use of the structuring of the information in the lexical entries and the representation levels of the model, on the one hand, and of the mechanism of co-composition[7], on the other, to adequately represent the 'real' meaning of *small elephant*.

(18)
$$\begin{bmatrix} pequeno\ (small) \\ \text{ARG} - \text{ST} = [1][\text{ARG1} = \text{x: entity}] \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \text{relative to class(e1, [1])} \\ \text{FORM} = \text{reduced dimension(e1, x, [1]} \end{bmatrix} \end{bmatrix} [1] \begin{bmatrix} elefante(elephant) \\ \text{ARG} - \text{ST} = [\text{P} - \text{ARG1} = \text{x: mammal}] \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \text{has part(x, y: trunk)} \\ \text{FORM} = \text{big(e1, x)} \end{bmatrix} \end{bmatrix}$$

[7] Mechanism that allows to complete the meaning of a predicate with values present in the semantic content of its arguments.

4. WordNet.PT: from research in the framework of different lexical computational models to a computational lexicon for Portuguese

WordNet.PT (Marrafa et al., 2006), a lexicon-conceptual Portuguese network developed under the EuroWordNet approach (Vossen 2002), mirrors a model that differs from previous proposals, WordNet and EuroWordNet, regarding the extent of the set of relations used and the strategies for lexical coverage followed. WordNet.PT was initially developed favoring the appropriateness of the results to the detriment of the dimension. This, together with the integration of research on the properties of lexical items, motivated the option of manual selection, description and coding of the data in WordNet.PT, resulting in a smaller but more reliable database with greater density of relationships, when compared with automatically built wordnets. The increase of the database followed mainly the semantic fields approach, involving the integration of lexical items of different grammatical categories, which motivated the need to enrich the model with more information and with the codification of more relations, in particular, transcategorial relations (Mendes 2009, Amaro 2009, Amaro et al. 2010a, 2010b, 2013).

Currently, WordNet.PT contains about 19,000 lexical entries, covering the main parts of speech of several semantic domains (eg food, geography, health, living beings), and includes a set of more than 50 lexical-conceptual relations, grouped in relations of equivalence, opposition, general / specific, hole / part, categorization, participation in event and defining of the structure of the event.

Given the similarities to conceptual and ontological organization, the network of types proposed in the Generative Lexicon model can be replaced, with advantages, by the integration of the lexical entries themselves in a lexical-conceptual network, allowing the uniform formalization of lexical inheritance (Amaro 2009; al. 2010a), that is, the explication of the sharing conditions of semantic properties – formalized through links established in the network – between lexical structures related by hyperonymy. Based on this assumption, and bearing in mind that the relations established in the WordNet model are clearly and strictly defined relations, the information conveyed as values of the qualia papers is easily transposed into the relations available in WordNet.PT (Amaro et al., 2010a), maintaining the adequacy of an inheritance mechanism that allows only the correct inferences to be generated. WordNet.PT allows this integration with decompositional lexical models in that it replaces the monotonic inheritance generically assumed in the WordNet model by a default inheritance mechanism.

(19)
$$\begin{bmatrix} peixe\ (fish) \\ \text{ARG} - \text{ST} = \begin{bmatrix} \text{P} - \text{ARG1} = \text{x: } \begin{bmatrix} vertebrate \\ \text{ARG} - \text{ST} = [\text{P} - \text{ARG1} = \text{x: animal}] \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \text{has part(x, y: endoskeleton),} \\ \text{has part(x, z: nervous system),} \\ \text{has part(x, w: muscle)} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORM} = \text{has as characteristic(x, y: aquatic); is member of(x, r: schoal} \\ \text{CONST} = \text{ has part (x, w: scale); has part(x, t: gill); has part(x, u: fin)} \\ \text{TEL} = \text{ is agent/ cause of(x, z: egg); is agent of(x, s: swim)} \end{bmatrix} \end{bmatrix}$$

Going back to the example of *fish*, in (2), the property has_part (x, y: scale), value of the constitutive role, is expressed in WordNet.PT by the relation of meronymy (has as part) established between the synsets $\{\text{fish}\}_{\text{N}}$ and $\{\text{scale}\}_{\text{N}}$. Moreover, the statement of a subtype relation 'x: y → x is hyponym of y' as the value of the argument itself, in the argument structure, allows to define the inheritance structure through the integration of lexical items in a lexical-conceptual network. The proper arguments are thus linked to their hyperonym and respective informational structure, inheriting all properties, except those that specifically characterize the lexical item in question, as referred to in section 2.2. and as illustrated in (19) above.

In addition to the linguistic motivation of this type of lexical representation, mentioned above and further discussed in greater detail in the sections below, it is also important to mention the great potential of this type of lexical resource in terms of its computational uses thanks, to the wide range of codified information and to its structure. In this context, in addition to coverage, the density of relational lexicons is particularly relevant. In fact, wordnets have been used to solve basic obstacles to the functioning of reliable computational applications involving natural language processing and, in particular, to access information conveyed by language, such as information search and extraction systems, machine translation systems, summarization systems, natural language generation systems, or applications of word meaning disambiguation, among others, since all these different natural language processing systems need rich lexical information to function correctly.

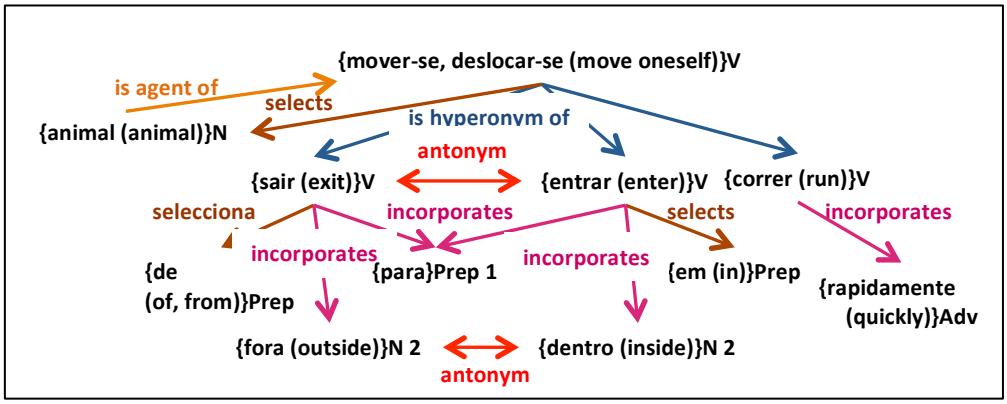
4.1 Meaning decomposition and verbal predicates properties

The semantic content of verbal predicates, by their intrinsic properties, is typically characterized by the argument constraints, which contribute significantly to the definition of their meaning. However, the relations available in the EuroWordNet model do not allow adequate representation of verbal predicates, namely in what concerns the distinction between argument selection vs. subcategorization properties – which do not always have direct correspondence –, besides not contemplating the representation of semantic elements incorporation, that allow to

express the facets of meaning that differentiate several cohyponyms, as well as hyponyms from their hyperonyms.

WordNet.PT, based on work developed on Portuguese verbs of movement (Amaro 2009), integrates *selects* / *is selected* (true arguments), *incorporates* / *is incorporated* (shadow-arguments) and *has as default argument* / *is default argument* (default arguments) relations, adapted from the Generative Lexicon model. These are different from the participation in events realtions already available (eg, the agent of) given that they express subcategorization properties (Amaro et al. 2010b, 2013). The figure below illustrates this proposal.

(20)



Like the figure in (5), the network represented in (20), which shows a network focused on verbal synsets, represents three hyponyms of {move oneself}$_V$. These share with their hyperonym the subcategorization of a nominal argument, {animal}$_N$, which in turn has the distinguishing feature of being a moving agent, a characteristic that differentiates it from its co-hyponym {plant}$_N$, for example. The three co-hyponyms in question are distinguished by their subcategorization properties (*exit* subcategorize a prepositional argument introduced by *of, from*; *enter* subcategorizes a prepositional argument introduced by *in*; *run* does not subcategorize any other argument) and by the elements that they incorporate (*exit* incorporates the semantic content of the node *outside*, *enter* incorporates the semantic content of the node *inside*, *run*, in turn, incorporates the semantic content of the node *quickly*). This modeling allows expressing in an objective and integrated way the lexicalization of different semantic components, which, as mentioned in section 3.2, has a direct relation with combinatorial constraints of lexical items.

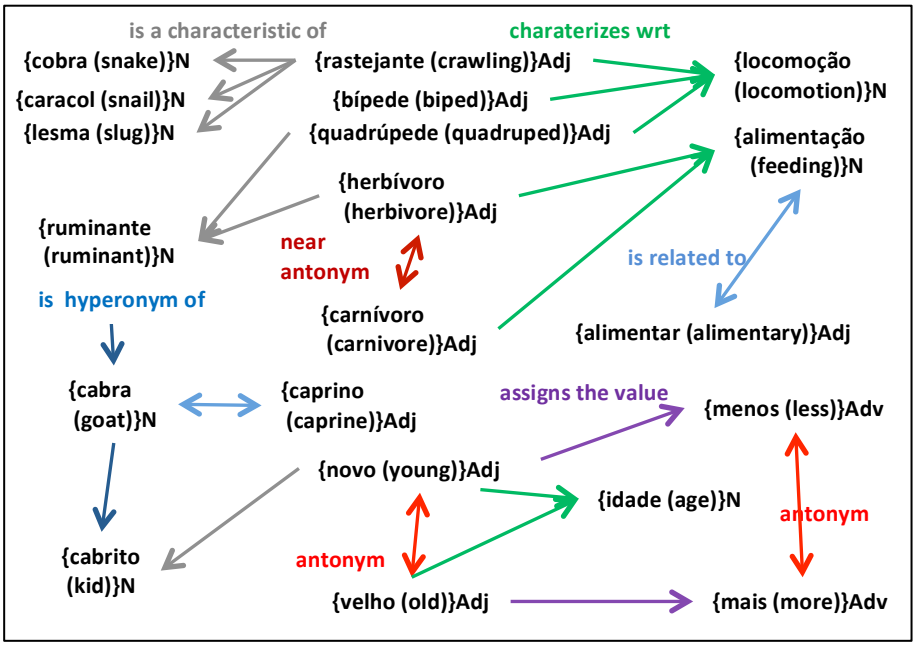
4.2 Definition of adjective classes based on structural and semantic properties

Research on the syntactic-semantic behavior of adjectives, partially discussed in section 3.1, and its modeling in relational lexicons (Mendes (2009)) has determined the definition of a small set of lexicon-conceptual relations to represent the elements of this category in the WordNet.PT. This section briefly presents the defined strategy of modeling, which allows to account for

several semantic phenomena in a generalized and systematic way, simultaneously reflecting and making evident the contrasting properties of the different classes of adjectives.

Although hyperonymy is the main structuring relationship in the WordNet model, the organization of adjectives is considerably different. In WordNet, descriptive adjectives are organized into clusters of synsets, reflecting the psychological principles of organization of this category in the mental lexicon (Miller 1998). However, in addition to this strategy raising problems at the level of implementation, as discussed in detail in Mendes (2009), this allows only to adequately represent one of the classes of adjectives, the descriptive adjectives. The different classes of adjectives show very salient distinctive properties (see section 3.1), which determine the need for coding strategies appropriate to their specificities.

(21)



In this context, a small set of lexical-conceptual relations[8] was defined in WordNet.PT to model the adjectives taking into account these different aspects: *characterizes with regard to*, which codifies the relation between the adjectives and the attributes that they modify; *assigns the value*, which determines the value of the attribute introduced by a given adjective; *is related to*, which establishes the relationship between a relational adjective and the name that lexicalizes the set of properties to which it is associated; and *is characteristic of*, which makes it possible to relate an adjective with a noun of which it reflects a distinctive property, i.e. a specific difference with respect to its co-hyponyms. The figure in (21) presents the relations in WordNet.PT for a small set of adjectives from different classes.

[8] Some inherited from the EuroWordNet model.

The network presents representative adjectives of different classes, highlighting the different modeling strategies used to adequately represent their meaning and their specific semantic properties. Thus, a descriptive adjective like {new}$_{Adj}$ relates to the lexicalization of the attribute that modifies, {age}$_{N}$, through the relation *characterizes with regard to*, as well as to the value that it specifies for this same attribute, {less}$_{Adv}$, in this case through the relation *assigns the value*. Moreover, precisely by lexicalizing frequently opposing values of the same attribute, the adjectives of this class also show semantic opposition relations, codified in the example presented through the relation of antonyms between {young}$_{Adj}$ and {old}$_{Adj}$. In addition, because they are associated with a unique property, these adjectives often lexicalize specific differences from other nodes in the network, such as {kid}$_{N}$ in our example, which lexicalizes a *goat* that is characterized by being of little age, that is, the aspect expressed in the network through the relation *is characteristic of* {kid}$_{N}$ and {young}$_{Adj}$. Relational adjectives, on the other hand, have a significantly less dense network of relations, typically presenting only the link to the noun that lexicalizes the set of properties to which they are associated (see relation between {caprine}$_{Adj}$ and {goat}$_{N}$). Thus, thanks to the representation of the most salient semantic properties of each adjective in this model, resulting in significant contrasts in terms of coding, it is possible to recover from the set of links expressed the different classes of adjectives.

5. In short

The brief itinerary for the generative and relational models of the lexicon presented in this chapter highlights their adequacy for modeling lexical knowledge, insofar as they account for complex lexical phenomena, essential for the processing of natural languages, thus reflecting the narrow relation between Lexicology and Computational Linguistics.

Based in the results from research on the distinctive properties of different part-of-speech in Portuguese, we presented lexical specification proposals for the lexical items studied and demonstrated their suitability through their integration in WordNet.PT.

6. References

Amaro, R. (2009), *Computation of Verbal Predicates in Portuguese: relational network, lexical-conceptual structure and context*, dissertação de doutoramento, Univ. de Lisboa.

Amaro, R., Mendes, S. & Marrafa, P. (2010), *Lexical-conceptual relations as qualia roles encoders*, in: Sojka P., Horák, A., Kopecek, I. & Pala, K. (eds.), *Text, Speech and Dialogue*, LNAI 6231, Berlim Heidelberg: Springer-Verlag, 29–36.

Amaro, R., Mendes, S. & Marrafa, P. (2010), *Encoding Event and Argument Structures in Wordnets*, in: Sojka P., Horák, A., Kopecek, I. & Pala, K. (eds.), *Text, Speech and Dialogue*, LNAI 6231, Berlim Heidelberg: Springer-Verlag, 21–28.

Amaro , R., Mendes, S. & Marrafa, P. (2013), *Increasing Density through New Relations and PoS Encoding in WordNet.PT*, International Journal of Computational Linguistics and Applications 4, Bahri Publications, India, ISSN 0976-0962, 1–12.

Bierwisch, M. (1971), *On classifying semantic features*, in: Steinberg, D. & Jakobovits, L. (eds.), *Semantics: An Interdisciplinary Reader in Philosophy,* Linguistics and Psychology, Cambridge: Cambridge University Press, 410–435.

Bréal, M. (1897), *Essai de sémantique: science des significations*, Paris: Hachette.

Fellbaum, C. (ed.) (1998), *WordNet: An Electronic Lexical Database,* Cambridge, MA: The MIT Press.

Geeraerts, D. (2010), *Theories of Lexical Semantics*, Oxford University Press.

Guthrie, L., Pustejovsky, J., Wilks, Y.& Slator, B. M. (1996), *The Role of Lexicons in Natural Language Processing*, Communications of the ACM, 30: 1, 63–72.

Hanks, P. (2003), *Lexicography*, in: Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 48–69.

Hobbs, J., Croft, W., Davies, T., Edwards, D., & Laws, K. (1987), *Commonsense Metaphysics and Lexical Semantics*, Computational Linguistics 13, 241–250.

Jackendoff, R. (1990), *Semantic Structures*, Cambridge, MA: The MIT Press.

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

Marrafa, P. (2001), *WordNet do Português - Uma base de dados de conhecimento linguístico*, Lisboa: Instituto Camões.

Marrafa, P. (2002), *The Portuguese WordNet: General Architecture and Semantic Internal Relations*, DELTA.

Marrafa, P., Amaro, R., Chaves, R. P., Lourosa, S., Martins, C. & Mendes, S. (2006), *WordNet.PT – Rede Léxico-Conceptual do Português 1.6*, CLG – CLUL, http://www.clul.ul.pt/clg/wordnetpt/index.html.

Melcuk, I. A. (1988), *Semantic description of lexical units in an Explanatory Combinatorial Dictionary*, International Journal of Lexicography 1, 165–188.

Melcuk, I. A. (1998), *Collocations and lexical functions*, in: Cowie, A. P. (ed.), *Phraseology. Theory, Analysis, and Applications*, Oxford: Clarendon Press, 23–53.

Mendes, S. (2009), *Syntax and Semantics of Adjectives in Portuguese: analysis and modeling*, dissertação de doutoramento, Universidade de Lisboa.

Mendes, S. & Chaves, R. P. (2001), *Enriching wordnet with qualia information*, in: *Proc. of the NAACL Workshop on WordNet and other Lexical Resources*, Pittsburgh, PA, 108–112.

Miller G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990), *Introduction to WordNet: An On-line Lexical Database*, International Journal of Lexicography 3:4, 235-244.

Miller, K. J. (1998), *Modifiers in WordNet*, in: Fellbaum, C. (ed.), *WordNet: an electronic lexical database*, Cambridge, MA: The MIT Press, 47–68.

Móia, T. (2000), *Identifying and computing temporal locating adverbials: with a particular focus on Portuguese and English*, dissertação de doutoramento, Universidade de Lisboa.

Nirenburg, S., Raskin, V. & Tucker, A. (1987), *The Structure of Interlingua in T upranslator*, in: Nirenburg, S. (ed.), *Machine Translation: Theoretical and Methodological Issues*, Cambridge: Cambridge University Press, 90–113.

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

Pustejovsky, J. (2001), *Type construction and the logic of concepts*, in: Bouillon P. & Busa F. (eds.), *The Syntax of Word Meanings*, Cambridge University Press.

Pustejovsky, J. (2007), *The Mechanics of Selection and Coercion in Grammar*, GL2007 Paris, École Normale Supérieure, http://www.issco.unige.ch/gl2007/talks/.

Talmy, L. (1985), *Lexicalization patterns: semantic structure in lexical forms*, in: Shopen, T. (ed.), *Language typology and syntactic description*, Cambridge University Press, vol. III, 57–149.

Trier, J. (1931), *Der deutsche Wortschatz im Sinnbezirk des Verstandes*, Heidelberg.

Trier, J. (1934), *Das sprachliche feld: Eine auseinandersetzung*, Neue Fachrbücher für Wissenschaft und Jugendbildung 10, 428-449.

Vassilyev, L. M. (1974), *The theory of semantic fields: a survey*, Linguistics 137:79-93.

Vendler, Z. (1967), *Linguistics in Philosophy*, Cornell University Press, Ithaca.

Vossen, P. (ed.) (2002), *EuroWordNet General Document*, EuroWordNet Project LE2-4003 & LE4-8328 report, University of Amsterdam.

Wierzbicka, A. (1972), *Semantic Primitives*. Frankfurt: Athenaeum.

Wierzbicka, A. (1996), *Semantics. Primes and Universals*, Oxford: Oxford University Press.

Wittgenstein, L. (1965), *Philosophical Investigations*, New York: The Macmillan Company.

RAQUEL AMARO AND SARA MENDES