Cross-language Semantic Relations between English and Portuguese*

Relaciones Semánticas entre los Idiomas Inglés y Portugués

Anabela Barreiro
L2F – INESC-ID
Rua Alves Redol nº 9, 1000-029
Lisboa, Portugal
anabela.barreiro@l2f.inesc-id.pt

Hugo Gonçalo Oliveira
CISUC, University of Coimbra, Pólo II
Pinhal de Marrocos 3030-290
Coimbra, Portugal
hroliv@dei.uc.pt

Resumen: Este artículo describe las relaciones semánticas conceptuales obtenidas de los recursos del sistema OpenLogos que fueron convertidos al formato NooJ. Estas relaciones están representadas simbólicamente en el léxico OpenLogos como un esquema taxonómico llamado abstracción semántico-sintáctica del lenguaje (SAL), que se utiliza para generar las relaciones jerárquicas de hiponimia e hiperonimia. El artículo también describe las relaciones acción-de, resultado-de, y sinonimia entre unidades multi-palabra y palabras sueltas, sobre todo donde existe una relación morfo-sintáctica y semántica entre las palabras de distintas categorías gramaticales. Las relaciones semánticas se generaron automáticamente a partir de la información lingüística asociada a cada entrada lexical en los diccionarios NooJ. Se desarrollaron gramáticas locales como mecanismo para leer esta información lingüística y generar las relaciones semánticas que se han utilizado en la producción de paráfrasis y en traducción automática. Los diccionarios y las gramáticas se pueden adaptar fácilmente a distintas lenguas y son útiles para diferentes tareas de procesamiento natural de la lengua, tanto monolingües como entre idiomas.
Palabras clave: relaciones semánticas, ontologías, diccionarios, gramáticas locales, relaciones entre idiomas

Abstract: This paper describes conceptual semantic relations obtained from OpenLogos resources converted into NooJ format. These relations were symbolically represented in the OpenLogos lexicon as a taxonomic scheme called semantico-syntactic abstraction language (SAL), used to generate hierarchical hyponymy and hypernymy relations. The paper also describes action-of, result-of, and synonymy relations between multiword units and single words, mostly where there is a morpho-syntactic and semantic relation between words of distinct parts-of-speech. The semantic relations were generated automatically, based on the linguistic information associated with each lexical entry in NooJ dictionaries. Local grammars were developed as a mechanism to read this linguistic information and generate the semantic relations, which have been used in paraphrasing and machine translation. Dictionaries and grammars can easily be adapted to distinct languages and are useful to various natural language processing monolingual or cross-language tasks.
Keywords: semantic relations, ontologies, dictionaries, local grammars, cross-language relations

1 *Introduction*

Lexical Semantics (Cruse, 1986) is the subfield of semantics that studies the words of a language and their meanings. It sees the lexicon as a finite list of lexical items (words or expressions) with a highly systematic structure that controls what words can mean. It can be seen as the bridge between a language and the knowledge expressed in that language (Sowa, 1999). The conceptual model of a language is structured around lexical items, their meaning (often referred as sense) and lexico-semantic relations held between

* Anabela Barreiro was partially supported by the UPV, award 1931, under the program Research Visits for Renowned Scientists (PAID-02-11). Hugo Gonçalo Oliveira is supported by the FCT scholarship grant SFRH/BD/44955/2008, co-funded by FSE.

the latter. To deal with the meaning of a language it is important to study these relations.

Semantic relations are crucial to understand and to structure the meaning of natural language. They are vital to communication overall, and highly employed in technical and specialized domains, where the most important content of texts is conveyed through the semantic relations between the terms that represent the domain's concepts, rather than by the meaning of the words alone (e.g., the semantic relations between BRCA1/*protein* and RNF53/*gene* in the biomedical field). Additionally, semantic relations are important for applications in the semantic web, mapping ontologies, text categorization, natural language understanding, etc., and a requisite for paraphrasing and machine translation, where words and expressions often must be substituted by semantic equivalents, such as synonyms between support verb constructions and single verbs (*make an operation* = *operate*; *say hello to* = *greet*), or other type of semantic alternates.

The most studied lexico-semantic relations are: (1) synonymy, when different lexical items have the same meaning (e.g. *car* synonym-of *automobile*); (2) homonymy, when lexical items have the same orthographic form but different meanings (e.g. *bank*, financial institution vs. *slope*); (3) hyponymy, when a lexical item is a subclass or a specific kind of another (e.g. *dog* hyponym-of *mammal*); and (4) meronymy, when a lexical item is a part, piece or member of another (e.g. *wheel* part-of *car*).

This paper describes the first attempt to extract cross-language semantic relations between English and Portuguese from the lexical resources of the OpenLogos machine translation system described by Scott (2003) and Barreiro et al. (2011). In combination with the former resources, new resources were created, namely derivational rules and grammars to recognize and generate morpho-syntactic and semantically related words and multiword units. Semantic relations, obtained by means of local grammars developed within NooJ linguistic environment (Silberztein, 2007), cover a larger number of items and can be extracted in a simple and easy way. This paper aims at showing how these resources combined can be used in cross-language tasks. Section 2 describes the state of the art in lexical semantics and automatic acquisition of distinct types of lexico-semantic relations. Section 3 presents the base linguistic resources used to attain semantic relations. Section 4 describes the relations of synonymy, hyponymy, action-of, and result-of. Section 5 presents the method for the extraction of the semantic relations. It describes, in particular, the morpho-syntactic and semantic relations established in the dictionary, how the grammars read this linguistic information, and how they use it to generate semantic pairs. This latter section also shows how to expand from monolingual to cross-language relations with minimal change in the local grammars. Section 6 presents some preliminary results. And finally, section 7 presents the conclusions and guidelines for future research work.

2 State of the Art

Dictionaries are probably the main source of lexico-semantic knowledge, as they are repositories of words, which include the description of several word senses. However, as definitions are written in natural language, dictionaries are not completely ready for being used as computational lexical resources.

Common representations of lexico-semantic knowledge, ready for being used in natural language processing tasks, include thesauri, taxonomies, as well as lexical ontologies or lexical knowledge bases. For example, the Roget Thesaurus (Roget, 1852) is one of the most well-known and complete thesaurus that is available in a machine readable format. Also, Princeton WordNet (Fellbaum, 1998) is a public domain lexical knowledge base, widely used in the natural language processing community. It is a handcrafted resource based on synsets, which are groups of synonymous words that may be seen as natural language concepts. Each synset has a gloss, which is similar to a dictionary definition, and several types of semantic relations between synsets are represented.

As the manual creation of lexical knowledge bases is typically an extensive and time-consuming task, there are several works where lexico-semantic relations are extracted automatically from text, and then used either to create new knowledge bases from scratch or to enrich existing knowledge bases. Due to their structure, dictionaries are an obvious target for the extraction of lexico-semantic relations (see, for example, (Chodorow, Byrd, and Heidorn, 1985) or (Richardson, Dolan, and Vanderwende, 1998)). Corpora and the Web have as well been exploited in the automatic acquisition of several types of lexico-semantic relations, including hyponymy (Hearst, 1992), meronymy (Berland and Charniak, 1999), causal relations (Girju and Moldovan, 2002), as well as in the discovery of new concepts (Lin and Pantel, 2002).

For Portuguese, in the latest years, semantic relations have also been a subject of increasing research interest. Santos et al. (2010) provide a review of the existing Portuguese lexico-semantic resources. Briefly, there are two handcrafted wordnets for European Portuguese, namely WordNet.PT (Marrafa, 2002) and MWN.PT[1], and an electronic thesaurus for Brazilian Portuguese, TeP (Maziero et al., 2008). There have also been attempts to the automatic acquisition of semantic relations, including: hyponymy extraction from corpora (Freitas and Quental, 2007); the extraction of several relations from a dictionary and the creation of the lexical resource PAPEL (Gonçalo Oliveira, Santos, and Gomes, 2010); and Onto.PT (Gonçalo Oliveira and Gomes, 2010), an ongoing project on the automatic creation of a lexical ontology for Portuguese, where several textual resources (thesauri, dictionaries, encyclopedias) are being exploited in the automatic acquisition of lexico-semantic relations.

Still, to the best of our knowledge, no research has been published on the automatic generation of cross-language semantic relations by using a linguistic method to map syntactic and semantically related words. This method can be extended to the type of relations that set equivalence between a word and a multiword unit (e.g. *take a look = look*), with a relative clause (*that was corrected = corrected*), with complex compounds (*bottle made of plastic = plastic bottle*) or even with a more complex construction, such as a possessive construction or a passive, by exploiting the morpho-syntactic and semantic relations pairs described in the dictionaries. The method has the advantage of being systematic, expandable, holding an unlimited possibility to grow and improve in observance of natural language complexity and compliant to distinct languages and across languages. This is the novel aspect of the work presented in this paper in relation to the state of the art.

3 Resources

In this section, we will describe the English and Portuguese resources used to achieve cross-language semantic relations.

Eng4NooJ and Port4NooJ (Barreiro, 2007) are sets of resources developed with the NooJ linguistic environment (Silberztein, 2007), aiming at the processing of the English and Portuguese languages. Both Eng4NooJ and Port4NooJ resources include lexica and grammars which are used for different tasks, including morphological and semantico-syntactic analysis, disambiguation, paraphrasing and translation. Both include a morphological system, contextual rules, different types of grammars (disambiguation, multiword units, etc.), and domain-specific dictionaries.

The Port4NooJ resources are publicly available[2] and, at the moment, are being used in tools such as Corpógrafo, a corpora tool (Maia and Sarmento, 2005; Sarmento et al., 2006; Maia and Matos, 2008), ParaMT, a paraphraser for machine translation (Barreiro, 2008a; Barreiro, 2008b), and eSPERTo[3], a system of paraphrasing for text editing and revision, currently being integrated in a cyber-school pedagogical program. Port4NooJ resources have not been reviewed, but they were made available to the Portuguese natural language processing (NLP) community because of their novelty aspects, which we hope are evocative for further pioneering research, including exploitation to other languages and cross-language tasks. The semantic relations included in the

[1] See http://mwnpt.di.fc.ul.pt/

[2] Port4NooJ can be found at the NooJ website under Portuguese module (http://www.nooj4nlp.net) and its resources are also available at Linguateca since October 2008 (http://www.linguateca.pt/Repositorio/Port4NooJ/).

[3] eSPERTo (in Portuguese, stands for Sistema de Parafraseamento para Edição e Revisão de Texto). It is a derivative of ReEscreve, proposed by Barreiro (2008a), and also described in (Barreiro and Cabral, 2009). The English version of eSPERTo is called SPIDER, standing for a System of Paraphrasing In Document Editing and Revision (formerly ReWriter). SPIDER uses Eng4NooJ resources and is described in (Barreiro, 2011).

Port4NooJ and Eng4NooJ resources resulted from the application of simple local grammars to the semantico-syntactic properties in the lexical entries and the use of derivational rules that link semantically related words of different parts-of-speech.

Eng4NooJ and Port4NooJ lexica were inherited from the OpenLogos system and enhanced with several new properties, which will be described in detail in Section 5.

The OpenLogos lexical entries are classified with more than 1,000 distinct categories, based on a taxonomy called SAL (*Semantico-syntactic Abstraction Language*)[4]. In the OpenLogos model, SAL is a meta-language that represents natural language, in effect, an ontology that represents things, ideas, relationships, dispositions, conditions, processes, etc., as well as the elements of grammar such as articles, prepositions, conjunctions, etc. In terms of natural language processing, the meta-language represents both syntax and semantics. SAL is an actual language, not a set of linguistic markers or primitives. This implies that natural language can be readily mapped to SAL. The granularity of the representational ontology is sufficient for translation purposes only, i.e., the ontology does not need to be especially fine-grained.

SAL elements are divided in a hierarchical scheme of supersets, sets and subsets, distributed by all parts-of-speech. SAL comprises 12 supersets for nouns: Concrete (CO), Mass (MA), Animate (AN), Place (PL), Information (IN), Abstract (AB), Process intransitive (PI), Process transitive (PT), Measure (ME), Time (TI), Aspective (AS), and Unknown (UN). For example, the concrete nouns superset consists of countable physical things, either man-made or natural, including parts of the human body. Concrete (count[5]) contain both sets and subsets. The principal sets of concrete nouns are functional things and agentive things. Other sets are: natural things (COnat); impulses/lights (COlight); marks/blemishes (COblem); edibles non-mass (COednm); edibles/color (COedcol); classifiers (COclass); amorphous (COamorph); and atomistic (COatom). For example, the set of natural things (COnat) includes subsets such as: minute flora (COflora) (e.g. *algae*, *spore*); plants (COplant) (e.g. *rose*, *weed*); trees (COtree) (e.g. *apple*, *willow*); trees/wood (COtrwd) (e.g. *oak*, *maple*); and miscellaneous natural things (COmnat) (e.g. *pebble*, *iceberg*).

The SAL meta-language is semantico-syntactic in nature, representing natural language at a second-order abstractions (common nouns are first-order abstractions). Syntax and semantics are seen as a continuum. This semantico-syntactic continuum is always taken into account when classifying each lexical entry within SAL. The classification was done through the years by trial and error. For example, when classifying elements into the functional (COfunc) or agentive (COagen) of the concrete noun superset, the following reasoning is taken into consideration: functional things tend to be passive, i.e. typically do not act of their own accord and generally require an agent to use them. Hence, they are more instrumental in nature. Agents typically do work in and of themselves. This distinction may sometimes seem arbitrary. For example, *hinge* is a fastener under functional things and clearly does work of itself, but is not coded as an agent. *Airplane*, on the other hand, obviously does require an agent and yet is coded under agentives as a vehicle. As a rule, agentives have a source of power or energy in themselves, while functionals do not. Parts of the human/animal body are also classified as concrete. Words like *heart*, *brain*, *digestive tract*, *stomach*, and organs in general are machines/systems under agentives. Words like *teeth*, *fingernail*, *toes*, *lips*, *tendons*, *ligaments*, *bones*, etc. belong to various subsets under functionals.

SAL categories contain domain-independent ontological (lexical-contextual) and semantico-syntactic relations (the same word form can be mapped to different concepts) are assigned to general language words or domain-specific terms. The general language dictionary contains many lexical entries which are broadly classified, which could be considered to pertain to a more specific domain. For example, the lexical entries

[4] The full description of the multiple SAL categories can be found at the Logos System Archives (http://logossystemarchives.homestead.com/) and all the resources (and descriptions) are downloadable from OpenLogos website at DFKI (http://logos-os.dfki.de/).

[5] Concrete nouns are always count nouns and, unless in the plural, generally cannot occur without a preceding article or quantifier. For example: *Computers are effective. *Computer is effective.*

dog IS_HYPONYM_OF *animal* *cão* É_HIPÓNIMO_DE *animal*
dog IS_HYPONYM_OF *mammal* *cão* É_HIPÓNIMO_DE *mamífero*
dog IS_HYPONYM_OF *non-human being* *cão* É_HIPÓNIMO_DE *ser não humano*
dog IS_HYPONYM_OF *invertebrate* *cão* É_HIPÓNIMO_DE *ser vertebrado*
dog IS_HYPONYM_OF *animate being* *cão* É_HIPÓNIMO_DE *ser vivo/animado*

Table 2: Hyponymy relations for the noun *dog* - *cão*

for *HIV* (immunology), *manic-depressive disorder*, *bipolar disorder* (mental health) and *asthma* (pulmonology) are all classified under the superset Abstract and subset State (also for conditions and relationships). This subset corresponds to abstract nouns that describe something about a thing or person that is not inherent to its nature (e.g. *cancer*, *coma*, *circumstance*, *condition*, *disease*, *fatherhood*, *inequality*, *insolvency*, *loneliness*, *parity*, *poverty*, *status*). Being more extrinsic, these states, conditions or relationships could conceivably change without altering the nature of the thing or person. This is not a strict rule but is indicative of the difference between this subset and the properties/qualities/nature subset.

The information noun superset is comprised of nouns that denote data, information, or knowledge, which might be considered more specific to certain domains. But, this category also includes the medium on which the information is recorded, represented or communicated; i.e., spoken, written, dramatized, sung, etc. Table 1 presents a list of terms classified as Instructional/legal (INinst) under the information noun superset (IN).

4 Semantic Relations for English and Portuguese

Both in Eng4NooJ and Port4NooJ, each lexical entry is described with semantico-syntactic properties, which represent relations between words or expressions. These relations can be synonymy, hyponymy, action-of, result-of, process-of, made-of, property-of, member-of, among others. Table 2 illustrates several semantic relations for the concrete English and Portuguese nouns *dog* and *cão*, respectively. These relations were inferred from the SAL hierarchical categories.

ACTION	abolishment IS_ACTION_OF abolish aboliçao É AÇAO DE abolir
	abuse IS ACTION OF abuse abuso É AÇAO DE abusar
	happening IS ACTION OF happen acontecimento É AÇAO DE acontecer
	agreement IS ACTION OF agree acordo É AÇAO DE acordar
RESULT	lit IS RESULT OF light aceso É RESULTADO DE acender
	stuffed IS RESULT OF stuff embalsamado É RESULTADO DE embalsamar
	rotten IS RESULT OF rotten podre É RESULTADO DE apodrecer
	interdicted IS RESULT OF interdict interditado É RESULTADO DE interditar

Table 3: Action-of and result-of semantic relations

In addition to the taxonomical classification inherited from OpenLogos, which allowed the establishment of hyponymy relations, both Eng4NooJ and Port4NooJ resources include regular derivational, morpho-syntactic and semantic relations, such as synonymy, action-of, and result-of. The morpho-syntactic and semantic relations are established between words of a different part-of-speech, as for example, between an adjective and its derived adverb (e.g. *quick* > *quickly* - *rápido* > *rapidamente*), between a noun and an adjective (e.g. *enthusiasm* > *enthusiastic* - *entusiasmo* > *entusiasmado*), or between a noun and an adverb (e.g. *imagination* > *imaginatively* = *with imagination* - *imaginação* > *imaginativamente* = *com imaginação*).

Table 3 illustrates action-of and result-of semantic relations. Action-of relations are established between a noun and a verb, where the noun is a morphological derivation of the verb. Result-of relations are established between an adjective and a verb, where the adjective is morphologically derived from the verb.

5 Methodology for the Extraction of Semantic Relations

In order to obtain hyponymy relations from the OpenLogos properties in Port4NooJ and Eng4NooJ dictionaries, we created a local grammar that matches on the SAL code and presents, as an output, one or more words from the description of that specific SAL code. For the examples in Table 1, the NooJ local grammar recognizes the property [SAL=ANmamm], standing for Ani-

```
intimacão,N+FLX=CANCÃO+INinst+EN=summons
arrendamento,N+FLX=ANO+INinst+EN=lease
autorizacão,N+FLX=CANCÃO+INinst+EN=fiat
autorizacão,N+FLX=CANCÃO+INinst+EN=license
autorizacão,N+FLX=CANCÃO+INinst+EN=permit
autorizacão,N+FLX=CANCÃO+INinst+EN=warrant
cânone,N+FLX=ANO+INinst+EN=canon
clausula,N+FLX=CASA+INinst+EN=clause
condicão,N+FLX=CANCÃO+INinst+EN=proviso
contrato,N+FLX=ANO+INinst+EN=contract
credo,N+FLX=ANO+INinst+EN=credo
declaracão,N+FLX=CANCÃO+INinst+EN=affidavit
decreto,N+FLX=ANO+INinst+EN=decree
diretiva,N+FLX=CASA+INinst+EN=guideline
estatuto,N+FLX=ANO+INinst+EN=bylaw
estatuto,N+FLX=ANO+INinst+EN=statute
garantia,N+FLX=CASA+INinst+EN=guarantee
garantia,N+FLX=CASA+INinst+EN=warranty
lei,N+FLX=CASA+INinst+EN=law
licenca,N+FLX=CASA+INinst+EN=license
mandato,N+FLX=ANO+INinst+EN=mandate
moratoria,N+FLX=CASA+INinst+EN=moratorium
norma,N+FLX=CASA+INinst+EN=norm
norma,N+FLX=CASA+INinst+EN=standard
ordem,N+FLX=MARGEM+INinst+EN=order
ordem,N+FLX=MARGEM+INinst+EN=ordinance
pacto,N+FLX=ANO+INinst+EN=pact
patente,N+FLX=CASA+INinst+EN=patent
renuncia,N+FLX=CASA+INinst+EN=waiver
testamento,N+FLX=ANO+INinst+EN=will
tratado,N+FLX=ANO+INinst+EN=treaty
veredicto,N+FLX=ANO+INinst+EN=veredict
```

Table 1: Sample of terms classified as Information + Instructional/legal (INinst)

mate, Mammal and retrieves, as its output, words that will be used as hypernyms of the words *dog* or *cão*, in English or Portuguese, respectively. These words are: animal, mammal, non-human being, invertebrate, animate being. If the description of the SAL category included more hypernyms, these could, of course, be easily added to the list of pairs of the semantic relation IS_HYPONYM_OF for *dog/cão*.

Table 4 shows distinct types of dictionary entries with implicit semantic information, namely the support verb construction that can be synonymous to a verb entry (*impressionar = causar impressão – impress = make an impression*; f*icar azedo = azedar – turn sour = sour*), the semantic relation between an adjective and a semantically related adverb (*aesthetic – aesthetically*), and the semantic relation between a noun and a semantically related adverb (*skepticism – skeptically*). These relations are established by means of grammar rules. We have focused on the most regular rules, which are the ones that allow transformation of part-of-speech through the process of derivation.

In the examples illustrated in Table 4, the properties in bold correspond to the derivational rule and inflectional paradigm. Accordingly, DRV=NDRV01:CANÇÃO is a dictionary property that calls the rule to derive (through the process of nominalization) the predicate noun *impressão* (*impression*) from the verb *impressionar* (*impress*) and assigns it the inflectional paradigm CANÇÃO (the noun *impressão* inflects in the same way as the noun *canção*; i.e., following the same process and using the same morphemes to form the plural, etc.); DRV=ADRV00:ALTO is a dictionary property that calls the rule to derive the predicate adjective *azedo* (*sour*) from the verb *azedar* (*sour*) and assigns it the inflectional paradigm ALTO (the adjective *azedo* inflects like the adjective *alto*). DRV=AVDRV03 is a dictionary property that calls the rule to derive the adverb aesthetically from the adjective aesthetic; and, finally, DRV=NAVDRV02 is a dictionary property that calls the rule to derive the adverb skeptically from the noun skepticism. The lexical entries for the verbs *impressionar* (*impress*), *adaptar* (*adapt*), *azedar* (*sour*), have the property VSUP, that is, the description of the support verb that occurs with the predicate nouns *impressão* (*impression*), *adaptação* (*adapt*) and with the predicate adjective *azedo* (*sour*), which derive from the corresponding cited verbs. The combination of the description in the properties VSUP and DRV allows the semantic association between these verbs and their equivalent support verb constructions, namely *fazer/causar impressão* (*make/cause impression*), *fazer adaptação* (*make adaptation*), and *ficar azedo* (*turn sour*).

Table 5 shows the transformational rules to associate morpho-syntactic and semantically related words of different parts-of-speech, extracted individually from the Eng4NooJ and Port4NooJ rule databases. Rules are indexed according to different types of transformation. NDRV transforms verbs into nouns, ADRV transforms verbs into adjectives, and AVDRV transforms adjectives into nouns. The rules of each type are numbered. For example, NDRV04 is the rule number 04 that transforms a verb into a noun. The slash (/) after each ending in-

```
impressionar,V+FLX=FALAR+SAL=PVPCpleasetype+EN=impress+VSUP=fazer+VSUP=causar+DRV=NDRV01:CANCÃO
adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset=132+EN=adapt+VSUP=fazer+DRV=NDRV00:CANCÃO
azedar,V+FLX=LIMPAR+Aux=1+OBJTRundif98+Subset=740+EN=sour+VSUP=ficar+DRV=ADRV00:ALTO
aesthetic,AFLX=NATURAL+SAL=AVstate+PT=estetico+DRV=AVDRV03
skepticism,N+FLX=BOOK+SAL=ABcause+PT=cepticismo+DRV=NAVDRV02
```

Table 4: General language dictionary entries with implicit semantic relations

troduces the part-of-speech of the derived word. The plus sign (+) introduces information about a specific noun or adjective. For example, Npred and Apred stand for predicate noun and predicate adjective, respectively. The capital letters between the less-than and the greater-than signs (<, >) correspond to commands. The command <B> means "backspace one character and add the string that follows the command, assigning it a new part-of-speech". The command <B2> means "delete the last two characters of the word from which the new word derives and add the string that follows the command", and so on and so forth. The strings that follow a command are the endings of the new generated words (e.g. *-ion* for the noun *acceleration*, *-tically* for the adverb *realistically*, etc.). The command <E> means that no character needs to be deleted. The command <A> means "delete the acute accent in the word from which the new word derives".

Eng4NooJ and Port4NooJ grammars are the devices used to recognize words or expressions and generate new ones, paraphrase or translate them. For example, the grammar in Figure 1, is used to recognize adverbial compounds in Portuguese and transform them into equivalent single adverbs. This grammar transforms multiword adverbs such as *de (um) modo rápido* (*in a fast/quick way*) into single adverbs such as *rapidamente* (*quickly*). This type of transformation is allowed by operations like the one represented in the first path of the graph. The box calls a new graph to recognize the strings *de (um) modo*, *de (uma) forma/maneira* (*in a (ADJ) way*), which make up the multiword adverbial. The output $A_ADV retrieves the adverb that is linked to the adjective $A. The adjective is transformed in the equivalent adverb by means of the derivational rules. The same grammar also recognizes multiword adverbs whose head is a noun, such as *por acidente* (*by accident*) or *com entusiasmo* (*with enthusiasm*), following the second and third paths.

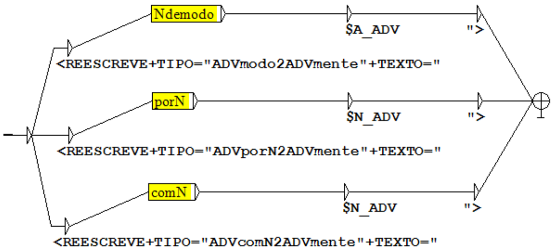

Ndemodo
$A_ADV
">
<REESCREVE+TIPO="ADVmodo2ADVmente"+TEXTO="
porN
$N_ADV
">
<REESCREVE+TIPO="ADVporN2ADVmente"+TEXTO="
comN
$N_ADV
">
<REESCREVE+TIPO="ADVcomN2ADVmente"+TEXTO="

Figure 1: Grammar to recognize multiword adverbials in Portuguese and transform them into equivalent single adverbs

The grammar in Figure 1 is monolingual, because there is no specification of the output for a different language. However, both Eng4NooJ and Port4NooJ resources contain Portuguese and English transfers for each lexical entry, i.e., they are in fact bilingual dictionaries. As a result, any grammar used to obtain monolingual transformations can be reused to generate bilingual (or multilingual) transformations. That is, the same grammar can be used to retrieve the output in English or in any other language (separately or together) as long as the words of that language are in the bilingual or multilingual dictionary and there are rules associated to the relevant dictionary properties. This means that, the grammar can generate translations from one to many languages, i.e., it can be used to create cross-language semantic relations. For monolingual transformations, no output language is specified. For bilingual or cross-language transformations, the parameter for the specification of the output language needs to be added. The parameter $EN for English, $IT for Italian, $SP for Spanish, etc. specifies the retrieval of the output in one of these languages or in all of them simultaneously. Similarly, the grammar presented in Figure 2, can be used for cross-language semantic relations. This grammar matches on a support verb construction of the type [Predicate Noun Construction] (*dar um abraço (a)* – *give a hug (to)*) (in the figure represented in a box that calls a sub-graph) and paraphrases it into a single verb (*abraçar* – *hug*).

Eng4NooJ	Port4NooJ
`NDRV04 = <B>ion/Npred`	`NDRV02 = <B>nca/N+Npred`
e.g. *accelerate > acceleration*	e.g. *mudar > mudança*
`ADRV02 = <B>icable/ADJ`	`ADRV02 = <B2>o/A+Apred`
e.g. *apply > applicable*	e.g. *azedar > azedo*
`AVDRV01 = <E>ly/ADV`	`AVDRV00 = <B>zmente/ADV`
e.g. *frequent > frequently*	e.g. *veloz > velozmente*
`AVDRV04 = <B>tically/ADV`	`AVDRV05 = <A> <B>amente/ADV`
e.g. *realism > realistically*	e.g. *rápido > rapidamente*

Table 5: Rules to transform morpho-syntactic and semantically related words of different parts-of-speech

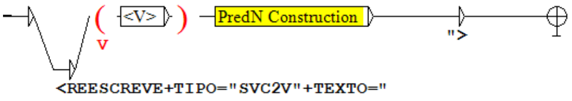
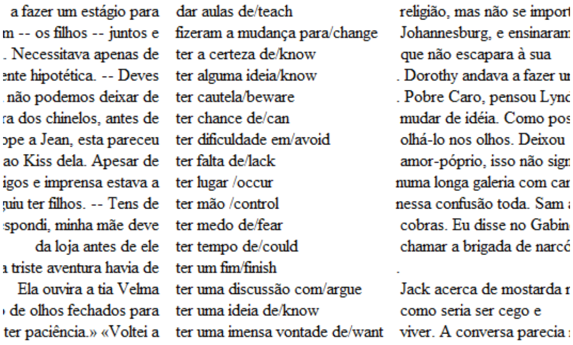


Figure 2: Grammar to generate cross-language relations between Portuguese support verb constructions and equivalent English single verbs



```
 a fazer um estágio para   dar aulas de/teach                religião, mas não se import
m -- os filhos -- juntos e fizeram a mudança para/change     Johannesburg, e ensinaram
. Necessitava apenas de    ter a certeza de/know             que não escapara à sua
ente hipotética. -- Deves  ter alguma ideia/know             . Dorothy andava a fazer u
 não podemos deixar de    ter cautela/beware                . Pobre Caro, pensou Lync
ra dos chinelos, antes de  ter chance de/can                 mudar de idéia. Como pos
ope a Jean, esta pareceu   ter dificuldade em/avoid          olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de    ter falta de/lack                 amor-póprio, isso não sig
igos e imprensa estava a   ter lugar /occur                  numa longa galeria com ca
uiu ter filhos. -- Tens de ter mão /control                  nessa confusão toda. Sam
spondi, minha mãe deve     ter medo de/fear                  cobras. Eu disse no Gabin
        da loja antes de ele ter tempo de/could              chamar a brigada de narc
a triste aventura havia de ter um fim/finish                 .
   Ela ouvira a tia Velma  ter uma discussão com/argue       Jack acerca de mostarda
 de olhos fechados para    ter uma ideia de/know             como seria ser cego e
ter paciência.» «Voltei a  ter uma imensa vontade de/want    viver. A conversa parecia
```

Figure 3: Cross-language relations between Portuguese support verb constructions and equivalent English single verbs

Figure 3 illustrates the output of a grammar that generates cross-language semantic relations between Portuguese support verb constructions and English single verbs. At present, the semantic relations included in Eng4NooJ and Port4NooJ are mostly used to generate paraphrases and integrated in the paraphrasing tools SPIDER and eSPERTo. However, cross-language relations such as those illustrated in Figure 3 can be used directly in machine translation and are fuelling the ParaMT bilingual paraphrasing tool. At the current stage of development, ParaMT translates mostly multiword units, performing well in the translation of Portuguese support verb constructions into English verbs, and vice-versa, the linguistic phenomena most researched when applying the current methodology.

Relation	Quantity
Hyponymy	14,963
Synonymy	10,395
between nouns	5,367
between verbs	20
between adjectives	34
between adverbs	5,014
Action-of	3,773
Result-of	283

Table 6: Relations in Port4NooJ v.2.0

6 *Preliminary Results*

In theory, the exploitation of the lexicon in combination with SAL allows the establishment of numerous relations between words and expressions. For the current paper, we focused only on a few of those relations which cover a larger number of items and could be extracted in a simple and easy way. The result of extraction for Portuguese (not yet reviewed) is publicly available[6]. Currently, Port4NooJ contains more than 30,000 morpho-syntactic relations between semantically related elements. Table 6 presents some preliminary results, which do not refer to paraphrasing capabilities, but simply to relations between lexical items. The total results for paraphrasing are significantly higher. Local grammars, applied to information (properties) described in the dictionary, enable the recognition and analysis of expressions such as *de (um) modo rápido*, *de (uma) forma/maneira rápida* (*in a fast/quick way*) (which could be considered as relations between an adjective and an adverb, but which were not counted), and also inflected forms such as *dar uns passeios* (*go for some walks*), etc.

Port4NooJ contains approximately 600 derivational rules, most of them transforming verbs into predicate nouns (587). 119 of

[6] See http://www.linguateca.pt/Repositorio/Port4NooJ/relacoes_semanticas_explicitas/

these rules are productive, covering nominalizations. 486 rules correspond to verb relations between verbs and autonomous predicate nouns. At this point in the research, rules were only superficially evaluated.

7 Conclusions and Future Research

This paper presented semantic relations, namely domain-independent semantico-syntactic and ontological relations, suitable for paraphrasing and cross-language tasks, including machine translation. We have demonstrated that given the appropriate linguistic resources, the generation of semantic relations can become very systematic. Any grammar to generate monolingual semantic relations can be reused to generate cross-language relations, rules can be standardized and often re-used across close languages, etc. Even though the methodology adopted was applied to the OpenLogos resources, it is compliant with the exploitation of other lexical resources with semantic relations, for any language besides English and Portuguese, studied in this research.

Future work would gather and combine open source available semantic resources, enhance properties on the existing resources, and enlarge the linguistic phenomena coverage.

References

Barreiro, Anabela. 2007. Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation. In Xavier Blanco, Max Silberztein, Xavier Blanco, and Max Silberztein, editors, *Proceedings of the 2007 International NooJ Conference*, pages 19–47. Cambridge Scholars Publishing, June 7-9.

Barreiro, Anabela. 2008a. *Make it simple with paraphrases. Automated paraphrasing for authoring aids and machine translation.* Ph.D. thesis, Universidade do Porto, Portugal.

Barreiro, Anabela. 2008b. ParaMT: A paraphraser for machine translation. In *Proceedings of Computational Processing of the Portuguese Language, 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS*, pages 202–211, Aveiro, Portugal. Springer.

Barreiro, Anabela. 2011. SPIDER: a System for Paraphrasing In Document Editing and Revision - applicability in machine translation pre-editing. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*, CICLing'11, pages 365–376, Berlin, Heidelberg. Springer.

Barreiro, Anabela and Luís Miguel Cabral. 2009. ReEscreve: a translator-friendly multi-purpose paraphrasing software tool. In Marie-Josée Goulet, Christiane Melançon, Alain Désilets, and Elliott Macklovitch, editors, *Proceedings of the Workshop Beyond Translation Memories: New Tools for Translators, The Twelfth Machine Translation Summit*, pages 1–8.

Barreiro, Anabela, Bernard Scott, Walter Kasper, and Bernd Kiefer. 2011. Openlogos machine translation: philosophy, model, resources and customization. *Machine Translation*, 25(2):107–126.

Berland, M. and E. Charniak. 1999. Finding parts in very large corpora. In *Proceedoings of 37th annual meeting of the ACL on Computational Linguistics*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of 23rd annual meeting on Association for Computational Linguistics*, pages 299–304, Morristown, NJ, USA. ACL Press.

Cruse, D. A. 1986. *Lexical Semantics.* Cambridge University Press, Cambridge.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication).* The MIT Press.

Freitas, Cláudia and Violeta Quental. 2007. Subsídios para a elaboração automática de taxonomias. In *XXVII Congresso da SBC - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 1585–1594.

Girju, Roxana and Dan Moldovan. 2002. Text mining for causal relations. In Susan M. Haller and Gene Simmons, editors, *Proc. 15th Intl. Florida Artificial*

Intelligence Research Society Conference (FLAIRS), pages 360–364.

Gonçalo Oliveira, Hugo and Paulo Gomes. 2010. Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium (STAIRS 2010)*. IOS Press.

Gonçalo Oliveira, Hugo, Diana Santos, and Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática*, 2(1):77–93, May.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. 14th Conf. on Computational Linguistics*, pages 539–545, Morristown, NJ, USA. ACL Press.

Lin, Dekang and Patrick Pantel. 2002. Concept discovery from text. In *Proceedings of 19th International Conference on Computational Linguistics (COLING)*, pages 577–583.

Maia, Belinda and Sérgio Matos. 2008. Corpógrafo v4: tools for researchers and teacher using comparable corpora. In *Proceedings of LREC 2008 Workshop on Comparable Corpora*, pages 79–82, Marrakech, Morocco. ELRA.

Maia, Belinda and Luís Sarmento. 2005. The corpógrafo - an experiment in designing a research and study environment for comparable corpora compilation and terminology extraction. In *Proceedings of eCoLoRe / MeLLANGE Workshop, Resources and Tools for e-Learning in Translation and Localisation*, pages 45–48, Leeds University, UK, March 21-23. Center for Translation Studies.

Marrafa, Palmira. 2002. Portuguese Wordnet: general architecture and internal semantic relations. *DELTA*, 18:131–146.

Maziero, Erick G., Thiago A. S. Pardo, Ariani Di Felippo, and Bento C. Dias-da-Silva. 2008. A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil. In *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 390–392.

Richardson, Stephen D., William B. Dolan, and Lucy Vanderwende. 1998. Mindnet: Acquiring and structuring semantic information from text. In *Proceedings 17th International Conference on Computational Linguistics (COLING)*, pages 1098–1102.

Roget, P. M. 1852. *Roget's Thesaurus of English words and phrases.* Available from Project Gutemberg, Illinois Benedectine College, Lisle IL (USA).

Santos, Diana, Anabela Barreiro, Cláudia Freitas, Hugo Gonçalo Oliveira, José Carlos Medeiros, Luís Costa, Paulo Gomes, and Rosário Silva. 2010. Relações semânticas em português: comparando o TeP, o MWN.PT, o Port4NooJ e o PAPEL. In A. M. Brito, F. Silva, J. Veloso, and A. Fiéis, editors, *Textos seleccionados. XXV Encontro Nacional da Associação Portuguesa de Linguística.* APL, pages 681–700.

Sarmento, Luís, Belinda Maia, Diana Santos, Ana Pinto, and Luís Cabral. 2006. Corpógrafo v3: From terminological aid to semi-automatic knowledge engine. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC 2006, pages 1502–1505. ELRA.

Scott, Bernard. 2003. The logos model: An historical perspective. *Machine Translation*, 18:1–72, March.

Silberztein, Max. 2007. An alternative approach to tagging. In *Proceedings of Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007)*, volume 4592 of *LNCS*, pages 1–11, Paris, France, June 27-29. Springer.

Sowa, John. 1999. *Knowledge Representation: Logical, Philosophical and Computational Foundations.* Thomson Learning, New York, NY, USA.