Port4NooJ

Linguistic Resources Overview

Anabela Barreiro
barreiro underscore anabela at hotmail dot com

[last updated on 18 November 2008]

*

Port4NooJ is a set of linguistic resources developed on NooJ linguistic environment for the automated processing of the Portuguese language. They integrate a bilingual extension and can also be used in Portuguese to English machine translation. The linguistic resources are: the electronic dictionaries, the inflectional and derivational rules to formalize and document the Portuguese morphological structure, and the different types of grammar: morphological (for contracted forms), disambiguation, syntactic-semantic, multiword expressions, and translation grammars. The interaction between the different components and the application of the linguistic resources to text will be explained throughout this document.

*

1. Chronological Background

In any language processing application, the linguistic resources represent the foundation. Good linguistic descriptions lead to sophisticated resources that help improve systems. Associated with the resources, ontologies also play a very important role as descriptors of entities or events. The linguistic resources that will be described in this document combine lexical and ontological resources from the OpenLogos machine translation system (http://logos-os.dfki.de/) with new inflectional and derivational rules and distinct types of syntactic-semantic grammars developed within NooJ (http://www.nooj4nlp.net/). They also contemplate paraphrase and multiword dictionaries. The extraction process, the elements derived from this work and the new resources are detailed in the remainder of this document.

1.1. Original Sources

Port4NooJ linguistic resources were developed on two original sources: the *NooJ* linguistic environment and *OpenLogos* lexical and ontological resources. NooJ platform was already available for a large number of languages with extensive resources, including large coverage dictionaries and grammars for each individual language, but its Portuguese resources needed to build from scratch. *OpenLogos*, an open source derivative of the *Logos Machine Translation System* [Scott, 2003] became available at the beginning of 2006. Logos system's strength resides in its ontology and lexical resources, the bilingual dictionaries and the semantic-syntactic (Semtab) rules. The technology itself is outdated, but the object-oriented character of Logos rules lend itself to NooJ, and the linguistic data from *OpenLogos* is useful in launching the development of new language pairs.

As described in [Barreiro, 2008a], we used knowledge of the Logos system and experience in creating dictionaries and grammars to extract and select data from *OpenLogos* and use it for innovative research. A new mental linguistic model has been created. The new system's design is based on a synergistic approach, where the components of the Logos system that offered valuable, functional abilities are maintained alive and integrated in a flexible platform for further development. NooJ is linguist friendly and robust enough to absorb the Logos system and make use of its best features. The major issue in re-using the Logos data concerns the grammar rule base, namely RES-PARSE, TRAN, and target generation rules. RES and PARSE are two sets of parsing rules. RES concerns the macro-parse of the input sentence and PARSE is about the micro-parse. TRAN rules are the transfer rules. Target generation rules are related to the output in the target language. In contrast to the lexical resources and the Semtab rules, grammar rules are inaccessible, due to a lack of comprehensive documentation and unusual implementation of interesting linguistic strategies. These components of the system need to be built from the beginning. NooJ's *modus operandi* is suitable to leverage the linguistic phenomena that Logos grammar rules held, such as word order, syntactic transformations, inserts, etc. that are indispensable to the translation process and generation of the target languages.

1.1.1. **NooJ**

NooJ [Silberztein, 2004] is a freeware development environment for linguistic research and development. NooJ contains several modules that include large coverage lexical resources, dictionaries for specific purposes and local grammars for a dozen languages and is being extended to several other languages. Its tools support the development, testing, debugging, maintenance and gathering of other different types of linguistic resources, namely local grammars and assist the development of natural language processing applications. NooJ tools are also used to parse corpora, build sophisticated concordances, and apply its linguistic resources to texts for distinct purposes. Local grammars are language descriptions in the form of graphs containing an input entry (with linguistic information) and an output entry (with linguistic constraints to the output, or simply the binary information of the recognized or unrecognized sequence). In NooJ, these local grammars are represented by finite-state transducers, called (Extended) Recursive Transition Networks (RTNs), and are widely applied to texts. They are used for identification and analysis of local linguistic phenomena: location and annotation of morphological, lexical and syntactic-semantic patterns; identification and extraction of semantic units from texts, such as dates, named entities and terminological expressions; recognition and tagging of words and multiword expressions; identification of syntactic constituents such as noun phrases and other syntactic constituents; extraction of semantic relations, and disambiguation. Among these feasible applications, we created specific local grammars to recognize, paraphrase and translate support verb constructions, such as '*tomar uma decisão*' > '*make a decision*'. In future developments, we expect to explore the transformational aspects of translation and expand the capability of the local grammars to larger linguistic operations, indispensible for machine translation.

1.1.2. **OpenLogos**

OpenLogos source data, dictionary and rules employ a classification based on the so-called SAL ontology. SAL stands for Semantic-syntactic Abstract Language, a representation language, embodying both meaning (semantics), and structure (syntax). It is an interlingual-style hierarchical taxonomy comprising over 1,000 elements, distributed in supersets, sets and subsets, which are embedded in the

dictionary. It was designed in a way so that developers would expand and add to its capabilities (extensible system). It was initially developed for the English language, but most of its elements are universal and therefore applicable to Portuguese and other languages. Unlike other ontologies, it places semantics and syntax on a continuum. It may be not totally original, but it is eclectic in the categories included in the representation schema. Notwithstanding acceptable shortcomings, this ontology was designed to work in concert with other linguistic resources, namely lexical resources and a diverse set of linguistic rules, and it has already been used successfully for several decades in commercial machine translation. This is enough of a reason to be used by other systems. Furthermore, the abstraction echelon makes the ontology applicable at several levels and useful for applications other than machine translation. Despite its limitations, SAL is a good basis from which to work, like other researchers have used Wordnet.

OpenLogos system represents an immense original investment and a serious work effort expended for over thirty years, and it has much to offer. The opportunity to use the linguistic knowledge and intellectual hard work contained in OpenLogos should not be wasted. However, for the open source ontology to work, there needs to be a standardization process, so that the cooperative project will succeed. We now have the toolset and the linguistic resources to enable us to create more effective and coherent machine translation systems. By using open source technology, we hope to grow the system used in our work cooperatively, to improve and extend the SAL ontology and further develop grammar strategies not only for the language pairs already available in this system, but for new language pairs or simply for single language analysis.

2. Port4NooJ Resources

Port4NooJ resources are described in [Barreiro, 2008a] and they are freely distributed to the research community and publicly available at the NooJ and Linguateca websites. Port4NooJ uses OpenLogos lexical resources, bilingual dictionaries, ontology, and semantic-syntactic rules (Semtab). But, we have created a completely new inflectional system for analysis and generation of variable words, and the derivational system was also created from scratch. All FLX and DRV annotations in the dictionary are new (see §

3.1 for more details). Port4NooJ format is totally different from the original resources. There are several different dictionaries, which can operate as one or independently. There is a completely new inflectional and derivational system, new parsing, translation, and generation components. After converting Logos resources into NooJ format, we used Port4NooJ to create additional new resources for monolingual and bilingual/multilingual paraphrasing, and for the development of new machine translation systems.

Our annotation system is hybrid. At present, we are at a crossroads between Logos annotations, annotations used on previous works by LADL researchers and our own annotations. However, most of the bilingual dictionary entries and the list of syntactic-semantic annotations of the lexicon were drawn up by Logos linguists. These annotations are part of the SAL ontology. The original SAL codes were numeric. We maintained the superset, set and subset original schema, with separated annotations for each level. But, we converted them into mnemonics, which were also drawn up by Logos linguists, but were never integrated into the system. For instance, where Logos system had the numeric code 5 for the superset, we replaced 5 with a corresponding [AN], which means [ANIMATE] noun. Some mnemonics retained their original meanings; others have changed and adapted. Their form changed slightly. For example, mnemonic [INdata] became [IN+data], [TIday] became [TI+day], and so on and so forth. This subdivision allows searches or matches on either the set or the subset type. The dictionary format is completely new. The Logos dictionary was part of an Oracle relational database, which was all numeric except for source and target words. Our dictionary is in NooJ format. It is simple, more readable and easier to extract. Some of the Logos bilingual entries were old-fashioned, even archaic, idiomatic, or slang. We have added annotations for style, jargon and niche categories of this type, whenever suitable. We corrected entries that were wrong, deleted some entries that were automatically generated and did not suit our needs and added new entries, as we considered them necessary.

3. Electronic Dictionaries

This section describes the derivative and organization of the Port4NooJ large coverage electronic dictionary, and the development of supplementary dictionaries, such as those containing named entities (mostly proper names) and multiword expressions.

3.1. Large Coverage Bilingual Dictionary

Our preliminary dictionary was converted from the *OpenLogos* English-Portuguese dictionary. Figure 1 illustrates how this dictionary looked like when we started.

paraphrase	01		13	1		:	12	76	76
	N	63895	1577976		parafrase	01	2	100	1
		1							
paraphrase	02		2	1			7	68	185
	N	63896	1577966		parafrasear	02		4	1
	1								
paraphrase	02		2	1			7	68	185
	N	63896	1577967	1	parafrase	01	1	99	1
		1							
paraphrased	04		31	1			16	68	185
	N	63898	1577971	1	parafrasear	02		4	1
	1								
paraphrased	04		31	1			16	68	185
	N	63898	1577970		parafraseado	04		123	1
	1								
paraphrasing	04		31	1			15	68	185
	N	63900	1577975	1	parafrase	01	1	99	1
		1							
paraphrasing	04		31	1			15	68	185
	N	63900	1577974		parafraseante	04		127	1
	1								
paraplegia	01		13	1		:	6	40	736
	N	63901	1577977		paraplegia	01	2	100	1
		1							
parasite	01		13	1		:	5	51	121
	N	63902	1577978		parasita	01	1	99	1
		1							

Figure 1: Sample of the English-Portuguese Logos dictionary

The sample in the figure shows source word, numeric codes for source word-class, source gender, source pat (inflectional paradigm), source head-word, source auxiliary, source number, SAL codes superset, set, and subset, homograph code (it specifies if the word is an homograph or not), meaning id code, target id code, alternate code, and then the target word and the numeric codes for target word class, target gender, target pat, target head-word, target auxiliary, target number, target causative code and target reflexive code. Summarizing, the first numeric codes stand for part-of-speech, word-class in Logos terminology. For instance, 01 represents a noun; 02 represents a verb; 04 represents an adjective. The codes 12 76 76 corresponding to the

entry for the first instance of the word '*paraphrase*' represent the SAL superset, set, and subset. 12 stands for information noun type and 76 stands for recorded data noun type.

There are a few obstacles when using OpenLogos either to improve the language pairs already available in the system or to develop new language pairs. Logos system is an old system, which did not transition well to new technology environments. In addition, one of the disadvantages with the Logos system is the lack of documentation regarding the system internals. It would be a relatively straightforward task to develop new targets from the English source if the procedures were well documented, but they are not. Having experience with the system allowed us to reuse its resources. We were also aware that the only way to help others develop new target languages would be to do the work ourselves and create a new intelligible and plain system based on the best Logos could offer.

Our first step procedure consisted in inverting source and target languages in order to create a Portuguese source and an English target language pair, in NooJ dictionary format. This inversion worked considerably well because each dictionary entry in the Logos dictionary is semantically (and sometimes syntactically) disambiguated at the lexical level. The target word is the disambiguated translation of the source word. For example, the homograph word *rio* is classified as [PL+path] and [PL+nagcom], meaning *river* and the third person singular of the present indicative of the verb *rir* (*to laugh*). Words that have more than one translation are also lemmatized more than one. For example, the word *panorama* appears in the dictionary four times, with the following transfers: *landscape*, *panorama*, *lookout* and *scene*.

Our second step procedure replaced the obsolete numeric representation of grammar and word syntactic-semantic properties with a more effective representation. This meant replacing each SAL ontology numeric representation code, with a corresponding SAL mnemonic from the *LearnLogos* web application tutorial. SAL mnemonics consist of a list of alpha strings, which correspond to the superset, set and subset numeric codes of the Logos system. The numeric codes were superseded and user-unfriendly. The idea was to use mnemonics instead of numbers in writing rules. Even though this list of mnemonics was developed at Logos with the purpose of replacing the numeric codes and making the ontology easier to understand by linguists

working in the development teams and by users of the machine translation engine, this course of action was never implemented in the Logos system. They were set up as part of the dictionary development user interface, as part of noun prompts and deployed on a limited scale for training purposes only. The existing elements of the SAL ontology are well documented in the SAL tutorial.

The conversion involved some standardization and adaptation. We have tried to use normalized annotations whenever they existed and created new ones whenever they did not exist. We are aware that not all resulting annotations might be definitive. We believe that some efforts should be made to develop standardization norms.

After the major time-consuming conversion was finalized[1], we adapted some annotations to Portuguese. Our ultimate goal is to use these resources in our machine translation envisioned project. Machine translation process has two main distinct phases: analysis and generation. Analysis of language is slow, so this represents work in progress. We are aware that there are inconsistencies and errors that need to be corrected as the system evolves. We have already deleted outdated entries, such as verbal past and present participles. In our system, participles are analyzed and generated by means of inflectional rules. We have eliminated Logos separation between closed and open word classes, identified and separated multiwords for re-classification and processing, incorporated changes and modifications, and inputted new properties such as predicate noun, predicate adjective, and support verb. There are links between nominalizations and adjectivalizations and morpho-syntactic and semantically related lexical verbs and each nominalization and adjectivalization has specified which support verb occurs with them. There are also links between other semantically related words, as will be explained further ahead.

In our large coverage bilingual dictionary, every entry is classified as a noun, verb, adjective, adverb, determiner, pronoun, preposition, conjunction, or numeric expression. The dictionary contains both variable and invariable word forms. Each invariable form has an inflectional paradigm assigned to it in the corresponding dictionary entry. For example, *mesa* (*table*) inflects in accordance with paradigm class *CASA*, and the verb *afirmar* (*affirm*) conjugates according to class *FALAR*. These

[1] Some original "sets" and "subsets" of the OpenLogos system still need to be converted from the numeric codes into SAL mnemonics.

inflectional paradigms are independent standard pattern models (prototypes) based on particular morphological suffixation rules covering variations in gender and number (adjectives and nouns), person (verbs and pronouns), tense (verbs), diminutives, augmentatives and superlatives (nouns, adjectives and some adverbs), and nominalizations. Words are attributed to different hierarchical ontology classes and subclasses, according to their linguistic attributes. Syntactic-semantic properties are also assigned to each entry. For instance, *cão* (*dog*) is classified as a common noun, warm-blooded vertebrate animal, mammal; *vestido* (*dress*) is classified as a concrete noun, clothing, soft thing made of fabric, leather, etc.; *cidade* (*city*) is classified as a common noun, agentive proper name denoting a geographic place, geographical entity, and geographical location; *sair* (*leave*) is classified as a motional intransitive verb; *português* (*Portuguese*) is classified as an adjective, related to a country; *feliz* (*happy*) is classified as a pre-clausal descriptive adjective; *adequadamente* (*adequately*) is classified as a non-locative adverb, manner-type. Figure 2 below shows a small sample of the main dictionary, with representation of all part-of-speech categories, variable and invariable entries (variable entries have specified the inflectional paradigm) and syntactic-semantic properties.

```
mesa,N+FLX=CASA+CO+surf+EN=table
cair,V+FLX=ATRAIR+INMO+IntoType+EN=fall
holandês,A+FLX=INGLÊS+AN+lang+EN=Dutch
actualmente,ADV+FLX=FACILMENTE+TEMP+punc+pres+EN=nowadays
alguém,PRO+IMPERS+INDEF+EN=somebody
porque,RELINT+WhyType+EN=why
e,CONJ+JOIN+EN=and
durante,PREP+TEMP+EN=during
cada,DET+IMPERS+INDEF+SG+EN=each
terceiro+NUM+ord+EN=third
```

Figure 2: General dictionary sample representing all parts-of-speech, variable and invariable forms

The properties illustrated in the sample for each entry are the following:

1. *mesa* is classified as a noun (*N*) which inflects like the word *casa* (*FLX=CASA*), where *casa* represents the morphological paradigm for feminine nouns ending in

–a, plural adding *-s*, with semantic properties defined as *concrete* (*CO*), *functional, bearing surface* (*surf*), corresponding to the English noun *table*;

2. *cair* is a verb (*V*) which inflects like the verb *atrair* (*FLX=ATRAIR*), where *atrair* represents the morphological paradigm for regular verbs ending in *–air*, vowel change *-i- > -í-* in some forms, with syntactic-semantic properties defines as *motional intransitive* (*INMO*), *preposition governance into-type* (*IntoType*), corresponding to the English verb *fall*;
3. *holandês* is classified as an adjective (*A*) which inflects like the adjective *inglês* (*FLX=INGLÊS*), where *inglês* represents the morphological paradigm for adjectives ending in *–ês*, feminine ending in *-esa*, with syntactic-semantic properties defined as *predicate* (*APred*), *animate* (*AN*) and *language* (*lang*), corresponding to the English adjective *Dutch*;
4. *actualmente* is an adverb (*ADV*) which inflects like the adverb *facilmente* (*FLX=FACILMENTE*), where *facilmente* represents the morphological paradigm for regular adverbs ending in *–lmente*, superlative in *–íssimamente*, defined as *temporal* (*TEMP*), *punctual* (*punc*), *present* (*pres*), corresponding to the English adverb *nowadays*;
5. *alguém* is classified as an invariable pronoun (*PRO*), *impersonal* (*IMPERS*), *indefinite* (*INDEF*), corresponding to the English pronoun *somebody*;
6. *porque* is classified as a relative and interrogative pronoun (*RELINT*), with the property *why-type* (*WhyType*), corresponding to the English pronoun *why*;
7. *e* is classified as a conjunction (*CONJ*), *conjoining* (*JOIN*), corresponding to the English conjunction *and*;
8. *durante* is classified as a preposition (*PREP*), no inflection, defined as *temporal* (*TEMP*), corresponding to the English preposition *during*;
9. *cada* is classified as an invariable determiner (*DET*), *impersonal* (*IMPERS*), *indefinite* (*INDEF*), *singular* (*SG*), corresponding to the English determiner *each*;
10. *terceiro* is classified as a numeric expression (*NUM*), *ordinal* (*ord*), corresponding to the English numeric expression *third*.

It is important to point out that the one-to-one correspondence seen in Figure 2, i.e., one Portuguese entry corresponding to one English transfer, already reflects the

disambiguated word. For instance, there might be an entry for *corredor=runner* and another for *corredor=hallway*, the first one classified as an animate noun denoting a profession or other human designation, [AN+des], and the second one defined as a place that has the general structure of a path, [PL+path]. As with Logos, in Port4NooJ the disambiguation is done at the dictionary level, by enriching the lexicon with added syntactic-semantic properties. Therefore, there will be as many entries as meanings for that word, which implies that there are as many translations for that word as there are meanings. The dictionary representation has to be complex because language is complex, and there is no way of building linguistically refined machine translation systems by following a simplistic approach.

Our large coverage dictionary also stores some uninflected compounds of general language, closed word classes (mostly grammatical words) such as adverbs, prepositions, pronouns, conjunctions and numeric expressions. Figure 3 below is a sample of the types of compound currently stored in the main dictionary.

```
a curto prazo,ADV+TEMP+EN=in the short run
a favor de,PREP+CAUS+EN=in favor of
cada um,PRO+INDEF+SG+EN=each one
de quem,INT+ThatType+EN=whose
quem quer que seja,REL+WhateverType+EN=whoever
além disso,CONJ+COOR+EN=besides
um quarto,NUM+frac+EN=one fourth
```

Figure 3: Sample of invariable compounds in the general dictionary

In the sample, the compound *a curto prazo* is a temporal adverb (TEMP) corresponding to the English adverb *in the short run*; *a favor de* is a cause and condition type preposition (CAUS) corresponding to the English preposition *in favor of*; *cada um* is an indefinite pronoun (INDEF), singular only (SG) corresponding to the English pronoun *each one*; *de quem* is an interrogative pronoun (INT), defined as that-type (ThatType), corresponding to the English pronoun *whose*; *quem quer que seja* is a relative pronoun (REL), defined as whatever-type (WhateverType), corresponding to the English pronoun *whoever*; *além disso* is a coordinating conjunction (COOR), corresponding to the English conjunction *besides*; *um quarto* is a numeric expression (NUM), fraction (frac), corresponding to the English numeric expression *one fourth*.

Whereas a NooJ dictionary has normally been related to a single input language, that is, that specified when creating a new dictionary, its internal structure enables its lexical entries to be easily associated with properties which can hold values in other languages. By including an additional field linked to the target language, in this case, EN for English, and adding syntactic and semantic value to the properties of each entry, a regular NooJ dictionary can be transposed into a more sophisticated bilingual resource. Further, by the expedient of adding properties such as FR, IT and SP, referring to French, Italian and Spanish, a multilingual dictionary can be developed, providing a starting platform for a multiple language pair machine translation system.

In order to process support verb constructions, our dictionary was enhanced with extended features and lexicon-grammar annotations were included. Beyond the commonly used part-of-speech and inflectional paradigm, each dictionary entry includes a description of the syntactic and semantic attributes (*SynSem*), as well as the associated distributional and transformational properties, such as predicate arguments, support verbs, aspectual verbs, stylistic variants of elementary support verbs, information about which determiners and prepositions occur with predicate nouns in "less variable" expressions, and derivational descriptions. Derivation is a very important issue, because it has implications not only at the lexical level, but also at the syntactic level. Derivational suffixes often apply to words of one syntactic category and change them into words of another syntactic category, while semantically they maintain their integrity. For example, the affix *–ção* changes the verb *adaptar* (*to adapt*) into the noun *adaptação* (*adaptation*) and the affix *-mente* changes the adjective *literal* (*literal*) into the adverb *literalmente* (*literally*). This is extremely important for support verb constructions because it permits the establishment of equivalence grammars that map (i) support verb constructions such as *fazer uma adaptação (de)* (*to make an adaptation (of)*) to the verb *adaptar* (*to adapt*), where the predicate noun *adaptação* (*adaptation*) has a semantic and morpho-syntactic relationship with the verb *adaptar* (*to adapt*) or (ii) support verb constructions such as *ter uma dilatação rápida* (*to have a quick dilation*) to the verbal expression *dilatar rapidamente* (*to dilate quickly*), where the autonomous predicate noun *dilatação* (*dilation*) has a semantic relationship with the verb *dilatar* (*to dilate*), and the adverb *rapidamente* (*quickly*) has a semantic and morpho-syntactic relationship with the

adjective *rápida* (*quick*). Thus, our verb entries contain the identification of derivational paradigms for nominalizations (annotation *NDRV*) and a link to the derived noun's support verbs (annotation *VSUP*), as in Figure 4 below. Nominalizations are followed by their inflectional paradigm properties. Any other lexical constraints, such as prepositions, determiners, specific arguments, etc., will be added. Autonomous predicate nouns (non-nominalizations), such as *favor* (*favor*) are lemmatized and classified with the annotation *Npred* and have associated with them support verb and other lexical constraints, such as a preposition (*NPrep*), and a lexical verb (*VRB*) with the same semantics. We have also classified predicate adjectives and established the link between them and the corresponding verbs (*ADRV*), such as between the verb *adoçar* (*to sweeten*) and the adjective *doce* (*sweet*). We have started the assignment of corresponding support verbs to these adjectives. Stylistic variants of the support verb constructions are annotated as *VSTYLE*. Aspectual variants are annotated as *VASP*. We added to the dictionary the syntactic and semantic arguments of a predicate. For example, in the entry for the verb *transplantar* (*to transplant*), the property *SUBJ=AG* means that a verb selects an agent as its semantic argument in the syntactic position of the subject. *SUBJ=PAT* means that a verb selects a patient as its semantic argument in the syntactic position of the subject. Syntactic argument *DO=ORG* means that the predicate selects a direct object that is an organ (subclass of body part). *IO=PAT* means that the predicate selects an indirect object that is a patient. *NPrep=de* means that the support verb plus predicate noun construction selects the preposition *de* (*fazer um transplante de – do a transplant of*). Nouns are classified semantically. For example, the noun *médico* (*doctor/physician*) is classified as an animate being denoting a profession or other human designation (*AN+des*), belonging to the medical field (*med*).

adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset132+EN=adapt+*VSUP=fazer*+*DRV*=NDRV00:CANÇÃO +*NPrep*=de
favor,N+FLX=MAR+Npred+AB+state+EN=favor+*VSUP*=fazer+*NPrep*=a+*VRB*=ajudar
literal,A+FLX=IGUAL+IN+symb+EN=literal+*DRV*=AVDRV05:RAPIDAMENTE
adoçar,V+FLX=COMEÇAR+Aux=1+OBJTRundif75+Subset604+EN=sweeten+*DRV*=ADRV11:VERDE+*VCOP*=tornar
transplantar,V+FLX=FALAR+Aux=1+RECTR26+Subset=504+BioMed+EN=transplant+SUBJ=AG+VSUP=fazer
+DRV=NDRV79:ANO+NPrep=de+DO=BP+IO=PAT+*VSTYLE=sofrer*+*VSTYLE=realizar*+*VSTYLE=efectuar*+*VASP=iniciar*
+*VASP=prosseguir*+*VASP=concluir*
médico,N+FLX=ANO+*AN+des+med*+EN=doctor
médico,N+FLX=ANO+*AN+des+med*+EN=physician

Figure 4: Sample of the dictionary

According to these linguistic constraints, we have created relationship properties at the dictionary level and then apply those properties in local grammars in order to recognize support verb constructions in corpora and generate paraphrases of them automatically for applications such as technical language writing and machine translation.

Our strategy to formalize idiomatic expressions and distinguish them from expressions with a more complex syntactic behavior is to lexicalize them. Therefore, semi-frozen expressions, where the verb is the only variable word in the whole expression, are listed in the dictionary of multiword expressions. For example, in *dar a mão à palmatória* (*to acknowledge being wrong*) or *fazer vista grossa* (*to ignore*), the verbs *dar* (*to give*) and *fazer* (*to make*) are assigned an inflectional paradigm and the rest of the words in the expression remain invariable.

As our electronic dictionaries provide enhanced meaning of single words, including contextual significance and increasingly more valuable tagging data, we also intend to enlarge and refine the role of a bilingual dictionary to include entries for multiword expressions that consider the understanding and analysis of each type of multiword expression, by beginning with support verb constructions and their paraphrases. The ability to give the machine translation user multilingual paraphrasing ability constitutes an important step towards achieving better quality machine translation.

3.2. Other Dictionaries

At the current stage, the dictionaries supplementing the large coverage dictionary include a dictionary of named entities and a dictionary of multiword expressions. We will describe these dictionaries in § 3.2.1 and § 3.2.2 below.

3.2.1. Named Entities

Currently bilingual named entities are represented in two dictionaries: the proper names dictionary and the biomedical terms dictionary. The dictionary of proper names has mostly the names of people and toponyms that were extracted from the Logos dictionary. The dictionary of biomedical terms contains only abstract concepts, states or conditions [AB+state], most of them were extracted from REPENTINO [Sarmento,

2006][2]. The remaining terms were incrementally added by us, collected from corpora and annotated with the SAL-type tags. In § 3.2.1.1 and § 3.2.1.2 below we will describe these two dictionaries.

For machine translation, named entities is a key area of research. The better the named entities are represented, the greater the possibility of improving the translation results. So, we intend to extend the dictionary of named entities with other relevant instances in the list of REPENTINO or other open source repositories. Moreover, named entities are often connected with terminologies of specific fields of knowledge. The semantic properties of each term are more common in one field or another. For example, symptoms are normally related to diseases and therefore proper of a medical field. On the contrary, the syntactic-semantic properties of nouns that designate policy, directions, orders, commands, etc., such as *contrato* (*contract*), *directiva* (*directive*), *instrução* (*instruction*), *lei* (*law*), *manual* (*manual*), *norma* (*norm*), *aviso* (*notice*), *ordem* (*order*), *estatuto* (*statute*), classified as [IN+inst] are normally characteristic of instructional and legal subject matters and many can be part of a legal terminology. Words in this set tend to have an agentive character; e.g., [*o estatuto exige que …*] (*the statute requires that …*); [*o regulamento diz que…*] (*the regulation calls for ...*). Thus, these terms tend to have a hortatory or regulatory character rather than a mere descriptive or informational character. Terms like *linha de código* (*line of code*), *instrução do programa* (*program instruction*), or *hiperligação* (*hyperlink*) are coded as instructional data and not as [CO+soft], concrete noun, software. Terms like *página da web* (*web page*), and *sítio da web* (*web site*) are coded as [IN+data], recorded data under information superset, because they are nouns that denote information or knowledge recorded in symbolic form. Terms like *URL*, *ASCII*, *HTML*, *JAVA* are coded as symbolic data under information superset [IN+symb], because they represent nouns that denote information or knowledge that has been recorded. Still in the computer field, many nouns that designate places where data may be stored, such as *buffer* (*buffer*), *disco de arranque* (*boot disk*), *DVD*, *CD ROM*, *base de dados* (*database*), *disco* (*disk*), *ficheiro* (*file*), or *disco rígido* (*hard disk*) are classified as storage media for recorded data, [IN+stor]. The elements of each subset share, not

[2] The terms extracted from REPENTINO were classified as [ABS] and [ESTCOND], i.e., abstract states or conditions.

only semantic properties, but also some syntactic properties. They may occur with the same prepositions or share the same predicates.

3.2.1.1. Dictionary of Proper Names

The proper names that translate into the target language need to be lemmatized in the dictionary. The proper names that do not translate into other languages, do not need to appear in the bilingual dictionary. However, we consider that it might be helpful to list extremely common proper names, because they may help build more sophisticated grammars that help identify other proper names, namely full names, or names with more than one element and contribute to an easier process of annotation of them. Many proper names in our dictionary are names of human beings, classified as [AN+name], and they were inherited from the Logos system.

Figure 5 shows some entries that can be found in the dictionary of proper names. [AN+name] represents the group of the animates, proper name is a subset of designation; [AN+title] represents animates, titles of individual human beings; [PL+reg] stands for region, other non-agentive proper geographical entity; [PL+city] stands for agentive place, proper name, city; [PL+coun] stands for, agentive place, proper name, country; [PL+cont] stands for non-agentive place, proper name, continent; [PL+water] stands for non-agentive place, proper name, body of water; [PL+mtn] stands for mountain, other non-agentive proper geographical entity; [AN+org] stands for proper name of established organization. Other proper names associated with titles, places, etc. are formalized in local grammars and described in § 5.4.

```
Senhor Chanceler,N+AN+title+EN=Lord Chancellor
Amesterdão,N+PL+city+EN=Amsterdam
Estados Unidos da América,N+PL+coun+EN=United States of America
África,N+PL+cont+EN=Africa
Extremo Oriente,N+PL+reg+EN=Far East
Mediterrâneo,N+PL+water+EN=Mediterranean
Alpes Peninos,N+FLX=ALPES+PL+mtn+EN=Pennine Alps
ONU,N+AN+org+EN=UN
```

Figure 5: Sample of the dictionary of proper names

3.2.1.2. Dictionary of Biomedical Terms

We chose the biomedical field to make an experiment with support verb constructions [Barreiro, 2008c], so it made sense to us to start building a dictionary of biomedical terms. Also, the SAL ontology already included a syntactic-semantic class [AB+state], which was used to classify abstract nouns related to states, conditions or relationships, such as *cancro* (*cancer*), *coma* (*coma*), *condição* (*condition*), *doença* (*disease*), among others. These abstract nouns describe something about a thing or person that is not inherent to its nature. Being more extrinsic, these states, conditions, 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 [AB+prop], which represent abstract nouns that describe the inherent (intrinsic) nature of a person or thing, such as *clareza* (*clarity*), *cor* (*color*), *design* (*design*), *característica* (*feature*), *forma* (*form*), *formato* (*format*), *padrão* (*pattern*), *perfil* (*profile*), *forma* (*shape*), and *traço* (*trait*).

Figure 6 shows some entries of the dictionary of biomedical terms. The dictionary of biomedical terms is sub-classified in specialties of the biomedical field. So, for instance, [IMMUN] stands for immunology, [MH] stands for mental health, and [PULM] stands for pulmonology. Many of these terms are very common and used in general language.

```
HIV,N+AB+state+IMMUN+EN=HIV
doença maníaco-depressiva,N+FLX=MWE9+AB+state+MH+EN=manic-depressive disorder
doença bipolar,N+FLX=MWE7+AB+state+MH+EN=bipolar disorder
asma,N+AB+state+PULM+EN=asthma
```

Figure 6: Sample of the dictionary of biomedical terms

3.1.2 Multiword Expressions

The dictionary of multiword expressions currently comprises compounds of general language, some lexical bundles and other expressions. These cover nominal expressions such as *cabo de vassoura* (*broomstick*) or *luz solar* (*sunlight*); verbal expressions, such as *marcar pontos* (*score*) or *piscar o olho* (*wink*); adjectival expressions such as *fraco de espírito* (*feeble-minded*), *cor-de-rosa* (*pink*); and adverbial expressions such as *com entusiasmo* (*enthusiastically*) or *de parte* (*aside*). This dictionary is soon to be significantly expanded by the incorporation of several thousand nominal compounds (predicate nouns), which appear frequently in support verb constructions, such as *juízo de valor* (*judgment*) as in *fazer um juízo de valor* / *fazer juízos de valor* (*make a judgment* / *make judgements*) or *chamada telefónica* (*phone call*) as in *fazer uma chamada telefónica* / *fazer chamadas telefónicas* (*make a phone call* / *make phone calls*).

Figure 7 shows some of the entries that can be found in the multiword expressions' dictionary. The annotation [PL+encl] stands for enclosed spaces; [CO+tool] stands for concrete, functional tools/devices; [MA+liqu] stands for mass, liquids; [NAV+Apred+col] stands for non-adverbial, predicate, color; [AN+des] stands for animate, designations or professions; [LocTime+TEMP] stands for locative, time, temporal; [STAT+phr] stands for stative, phrase; [LocTime+TEMP+puncpast] stands for locative, time, temporal, punctual past; [COOR] is an annotation for a coordinating conjunction; [SUB] is an annotation for a subordinating conjunction; [ASSOC] stands for an associative preposition; [Loc+AT] stands for a locative, at-type preposition; [ALOG] stands for analogical preposition. The compound '*bebida alcoólica*' appears twice. One entry is translated by the neutral expression '*alcoholic drink*', which can be used to produce a more neutral translation (less marked); another entry is translated as '*booze*', marked as a slang word. By default, the machine translation system

translates the expression '*bebida alcoólica*' as '*alcoholic drink*', but in some texts the translation '*booze*', could be the most adequate.

```
adro da igreja,N+FLX=MWE6+PL+encl+EN=churchyard
cabo de vassoura,N+FLX=MWE6+COtool+EN=broomstick
bebida alcoólica,N+FLX=MWE4+MA+liqu+EN=alcoholic drink+UNAMB
bebida alcoólica,N+FLX=MWE4+MA+liqu+EN=booze+slang
cor de laranja,A+NAV+Apred+EN=orange
sul-americano,A+FLX=MWE11+AN+des+EN=South American
a curto prazo,ADV+LocTime+TEMP+EN=in the short run
fora de serviço,ADV+STAT+phr+EN=out of order
há muito tempo,ADV+LocTime+TEMP+puncpast+EN=a long time ago
isto é,CONJ+COOR+EN=i.e.
já não,CONJ+COOR+EN=no longer
mesmo assim,CONJ+SUB+EN=even so
juntamente com,PREP+ASSOC+EN=along with
à direita de,PREP+Loc+AT+EN=at the right of
em conformidade com,PREP+ALOG+EN=in congruence with
```

Figure 7: Sample of the dictionary of multiword expressions

Only words of general vocabulary that have a less variable character are stored in the dictionary. Frozen expressions are also stored in this dictionary, such as fully idiomatic expressions as *dar a mão à palmatória* (*acknowledge being wrong*), *fazer vista grossa* (*neglect*) or [*dar cabo dos nervos a NP*] (*irritate/ennerve NP*) in Figure 8.

```
dar a mão à palmatória,V+FLX=PHRDAR+EN=acknowledge being wrong
fazer o sangue subir à cabeça,V+FLX=PHRFAZER+EN=ficar tonto
ter o sangue nas guelras,V+FLX=PHRTER+EN=be alive
fazer vista grossa,V+FLX=PHRFAZER+EN=neglect
dar parte de fraco,V+FLX=PHRDAR+EN=give up
```

Figure 8: Sample of the dictionary of idiomatic multiword expressions

Other multiword expressions that are more variable or that come together with other elements are formalized in local grammars and described in § 5.4. That is the case of support verb constructions that maintain a certain flexibility in respect to determiners and prepositions, and permissibility of inserts, such as [*dar um passeio*] (*go for a walk*) or [*dar vários passeios por/em NP*] (*go for several walks to NP*).

4. Inflectional and Derivational Rule System

In NooJ, a dictionary is associated with a set of ".nof" files that are either textual or graphical. If their shape is similar to a description, they are called "inflectional/ derivational textual rules"; if they are represented by sets of graphs, they are called "inflectional/derivational graphical grammars". Port4NooJ formalizes both inflectional and derivational paradigms as textual rules, previously called ".flx" files. The connections between the various dictionaries (the dictionaries of lemmas and the dictionaries of inflected forms) are established by the complete inflectional and derivational rule system. This system is made up of descriptions that apply to both individual words and compounds. Inflectional rules formalize Portuguese morphological paradigms for simple words. For example, the word form *falamos* (*we speak*) is recognized, generated or annotated as the first person plural of the verb *falar* (*speak*) in the present tense. Derivational rules formalize nominalizations, adjectivalizations and adverbializations. Many derivative nouns and adjectives can be turned into action verbs and many adjectives can turn into adverbs and vice-versa. For example, *citação* (*quotation*) is recognized, generated or annotated as a nominalization derived from the verb *citar* (*quote*); *aplicável* (*applicable*) is recognized, generated or annotated as derived from the verb *aplicar* (*apply*); and *rapidamente* (*quickly*) is recognized, generated or annotated as an adverbialization of the adjective *rápido* (*quick*).

Figure 9 illustrates the inflectional paradigm for regular masculine nouns with regular plurals (adding an –s). The word *ano* (*year*) is an example of this paradigm. Characters before the slash sign (/) represent the word endings. Therefore, *s*/ means add an –s to recognize or generate the masculine (m) plural (p) form of the word. The singular (s) form has an empty string <E>, signifying that nothing should be added to or removed from the lemma because the masculine singular form remains the same as the lemma. The inflectional rules cater for the fact that Portuguese nouns have diminutive (Dim), and augmentative (*Aum*) forms. The backspace operator *<B>* signifies that one character is to be deleted from the paradigm example word as part of the morphological rule in order to derive the diminutive and augmentative forms.

```
ANO = <E>/m+s + s/m+p +
<B>inho/Dim+m+s + <B>inhos/Dim+m+p +
<B>ão/Aum+m+s + <B>ões/Aum+m+p ;
```

Figure 9: Noun inflection paradigm

Figure 10 illustrates the inflectional paradigm for regular verbs ending in *–ar* (first conjugation). The word falar is an example of this paradigm. The rules cover all verb tenses: infinitive (*INF*), inflected infinitive (*INFI*), past participle (*PP*), gerundive (*G*), present (*PR*), simple past (*PS*), imperfect (PI), pluperfect (*PMP*), future (*F*), conditional (*C*), present subjunctive (*PRS*), imperfect subjunctive (*PIS*), future subjunctive (*FS*) and imperative (*IMP*). The number of forms associated with each tense is different. The uninflected infinitive and the gerundive have only one form each in this paradigm. The past participle has four forms corresponding to the masculine singular (*m+s*) and plural (*m+p*), and feminine singular (*f+s*) and plural (*f+p*). The imperative has no form for the first person singular. All other verb tenses have three forms for the singular and three for the plural, corresponding to the first (1), second (2) and third (3) persons. The backspace operator *<B>* signifies that one character is to be deleted from the paradigm example word as part of the morphological rule. *<B2>* signifies that two characters are to be deleted, and so on and so forth. The morpheme characters of each inflected form follow the backspace operator. The operator *<E>* signifies that the word form following that paradigm remains the same as the lemma, represented by the paradigm example word. In Portuguese morphology, not all verbs behave like the ones that follow the paradigm illustrated in Figure 10. Some verbs are much more irregular, others are defective, i.e., they have fewer forms, etc. We have over 100 different verb paradigms.

```
FALAR = <E>/INF +
<E>/INFI+1+s + es/INFI+2+s + <E>/INFI+3+s + mos/INFI+1+p + des/INFI+2+p + em/INFI+3+p +
<B>do/PP+m+s + <B>dos/PP+m+p + <B>da/PP+f+s + <B>das/PP+f+p +
<B>ndo/G +
<B2> (o/PR+1+s + as/PR+2+s + a/PR+3+s + amos/PR+1+p + ais/PR+2+p + am/PR+3+p) +
<B2> (ei/PS+1+s + aste/PS+2+s + ou/PS+3+s + ámos/PS+1+p + astes/PS+2+p + aram/PS+3+p) +
<B> (va/PI+1+s + vas/PI+2+s + va/PI+3+s + <B>ávamos/PI+1+p + <B>áveis/PI+2+p + vam/PI+3+p) +
<B2> (ara/PMP+1+s + aras/PMP+2+s + ara/PMP+3+s + áramos/PMP+1+p + áreis/PMP+2+p + aram/PMP+3+p) +
ei/F+1+s + ás/F+2+s + á/F+3+s + emos/F+1+p + eis/F+2+p + ão/F+3+p +
ia/C+1+s + ias/C+2+s + ia/C+3+s + íamos/C+1+p + íeis/C+2+p + iam/C+3+p +
<B2> (e/PRS+1+s + es/PRS+2+s + e/PRS+3+s + emos/PRS+1+p + eis/PRS+2+p + em/PRS+3+p) +
<B2> (asse/PIS+1+s + asses/PIS+2+s + asse/PIS+3+s + ássemos/PIS+1+p + ásseis/PIS+2+p + assem/PIS+3+p) +
<B2> (ar/FS+1+s + ares/FS+2+s + ar/FS+3+s + armos/FS+1+p + ardes/FS+2+p + arem/FS+3+p) +
<B2> (a/IMP+2+s + e/IMP+2+s + emos/IMP+1+p + ai/IMP+2+p + em/IMP+3+p) ;
```

Figure 10: Verb inflection paradigm

Figure 11 shows the inflectional paradigm for regular adjectives ending in -o. The word *alto* (*tall*) is an example of this paradigm. The inflectional rules cater for the fact that Portuguese adjectives inflect in gender and number, and have diminutive (Dim), augmentative (*Aum*), and superlative (*Sup*) forms.

```
ALTO = <E>/m+s + s/m+p +
<B>a/f+s + <B>as/f+p +
<B>inho/Dim+m+s + <B>inhos/Dim+m+p +
<B>inha/Dim+f+s + <B>inhas/Dim+f+p +
<B>ão/Aum+m+s + <B>ões/Aum+m+p +
<B>ona/Aum+f+s + <B>onas/Aum+f+p +
<B>íssimo/Sup+m+s + <B>íssimos/Sup+m+p +
<B>íssima/Sup+f+s + <B>íssimas/Sup+f+p ;
```

Figure 11: Adjective inflection paradigm

Figure 12 describes the inflectional paradigm for adverbs ending in *-velmente*. The word *amigavelmente* (*amicably*; *in a friendly way*) is an example of this paradigm. Whilst adverbs ending in *-mente* are usually invariable and of regular (*Reg*) form, they may occasionally take a superlative form (*Sup*), and therefore this representation has been included in the interest of precision.

```
AMIGAVELMENTE = <E>/Reg +
<B8>bilíssimamente/Sup ;
```

Figure 12: Adverb inflection paradigm

Figure 13 shows three different inflectional paradigms for grammatical words, such as determiners and pronouns. Grammatical words (closed word classes) in most Romance languages, including Portuguese, adapt to paradigms where they are followed by a much smaller number of words compared to open word classes (nouns, verbs, adjectives, adverbs). Example word *o* (definite article) (*the*) represents the paradigm for determiners, which can be variable in gender and number. Example word *esse* represents the paradigm for demonstrative pronouns, such as *esse* (*that*), *este* (*this*) or *aquele* (*that*). Example word *qual* represents the paradigm for interrogative pronoun *qual* (*which*), or demonstrative pronoun *tal* (*such*).

```
O = <E>/m+s +
<B>a/f+s +
s/m+p +
<B>as/f+p ;

ESSE = <E>/m+s + s/m+p +
<B>a/f+s + <B>as/f+p ;

QUAL = <E>/m+s + <B>is/m+p +
       <E>/f+s + <B>is/f+p ;
```

Figure 13: Inflectional paradigms for determiners and pronouns

Nominalizations are verbs that have been transposed into nouns or, less frequently, nouns that have been derived from adjectives. For example, the noun *apresentação* (*presentation*) was derived from the verb *apresentar* (*present*) and the noun *importância* (*importance*) was derived from the adjective *importante* (*important*). Work has commenced to formalize derivational paradigms for other transpositions such as adjectivalizations, and verbalizations. Adjectivalizations are adjectives derived from nouns or verbs and verbalizations are verbs derived from adjectives or nouns. As transposition is a very common linguistic phenomenon, it is possible to develop rules

to cover a large number of occurrences, but many others are unique and are equivalent to a single dictionary entry.

The derivational paradigm for nominalizations such as *nomeação* (*nomination*) is described in Figure 14. The corresponding verb to *nomeação* is *nomear*. The derivational rule operator <*B*> instructs the system to backspace one character from the verb infinitive form and adds the suffix *-ção* to recognize and produce the nominalization.

```
DRV00 =
<B>ção/N+Npred+Nom;
```

Figure 14: Derivational paradigm for nominalizations in *-ção*

Multiword expressions have their own inflectional paradigms. As point out by [Ranchhod et al., 2004] they are normally made up of a combination of simple words, but their meaning is not compositional, i.e., it does not result from the meaning of the individual words that constitute the expression. They have morphological, combinatorial, and other linguistic constraints. Some of them are invariable while others inflect in gender and number. For the variable expressions, the inflectional pattern is not always predictable. Figure 15 illustrates a few representative inflectional paradigms for multiword expressions and phraseology. For example, idiomatic and semi-frozen expressions such as *dar a mão à palmatória* (*to acknowledge being wrong*) follow inflectional paradigm PHRDAR. The rule specifies that the verb *dar* (*to give*) inflects like the paradigm :DAR. That means that the expression uses the same rule as the one used to inflect simple verbs. The operator <PW> moves the cursor to the end of the first word in the multiword expression (*dar*). The remaining words of this expression do not inflect and therefore do not follow any inflectional paradigm. The multiword expression *ano civil* (*calendar year*) follows the inflectional paradigm NA01. According to this paradigm, the first component of the multiword expression, the word *ano* (*year*) inflects the same way as the simple noun *ano*. The second component of the multiword expression, the word *civil* (literally *civil*; in the multiword expression, *calendar*) inflects in the same manner as the simple adjective *civil*. The operator <NW> moves the cursor to the end of the next word form in the multiword expression (*civil*).

The multiword expression *célula sanguínea* (*blood cell*) follows the inflectional paradigm NA02. According to this paradigm, the first component of the multiword expression, the word *célula* (*cell*) inflects the same way as the simple noun *casa* (*home*), which is the example word for that paradigm. The second component of the multiword expression, the word *sanguínea* (*blood*), whose lemma is the *sanguíneo*, follows the same inflectional paradigm as simple adjective *alto*. Finally, the multiword expression, *código de barras* (*bar code*), follows the inflectional paradigm NdeN01. In this expression, the only component that inflects is the first one, corresponding to the word *código* (*code*). It inflects according to paradigm :ANO. The other components of the expression remain invariable.

```
PHRDAR = <PW> :DAR ;
NA01 = <PW> :ANO <NW> :CIVIL ;
NA02 = <PW> :CASA <NW> :ALTO ;
NdeN01 = <PW> :ANO ;
```

Figure 15: Inflectional paradigms for multiword expressions

5. Grammars

NooJ grammars are graphical forms of representing different linguistic phenomena. They can be used to annotate text or filter out annotations from text. These grammars are based on finite state transducers technology. Local grammars can deal with frozen or semi-frozen phenomena or morphological phenomena, but they can also be used more broadly as in the case of syntactic grammars, which disambiguate words, make active to passive transformations or semantic agreement checks, describe the structure of phrases and sentences, and tag their syntactic constituents.

Port4NooJ holds several types of grammars: morphological grammars, syntactic-semantic grammars, disambiguation grammars, grammars for multiword expressions, and translation grammars. Grammars for multiword expressions are used to formalize support verb constructions and other multiword constructions. In addition to these applications, as above mentioned, grammars can be used to perform semantic analysis, to represent named entities, to create paraphrases, and for translation purposes. The different types of grammars used in Port4NooJ are described below.

5.1. Morphological Grammars

Our system uses inflectional and derivational textual rules to represent morphology, but it processes contracted forms by means of a morphological grammar. Contracted forms result from the combination of two words, such as a preposition and a determiner or demonstrative pronoun. Examples are *das* (*of the*) derived from the preposition *de* (*of*) and the determiner *as* (*the*), and *neste* (*in/on this*) derived from the preposition em (*in/on*) and the demonstrative pronoun *este* (*this*).

The meta-graph illustrated in Figure 16, represents a finite state transducer grammar for the analysis of all Portuguese contracted forms, where the main nodes are the non-contracted prepositions, which can be concatenated with other part-of-speech elements from the sub-graphs.

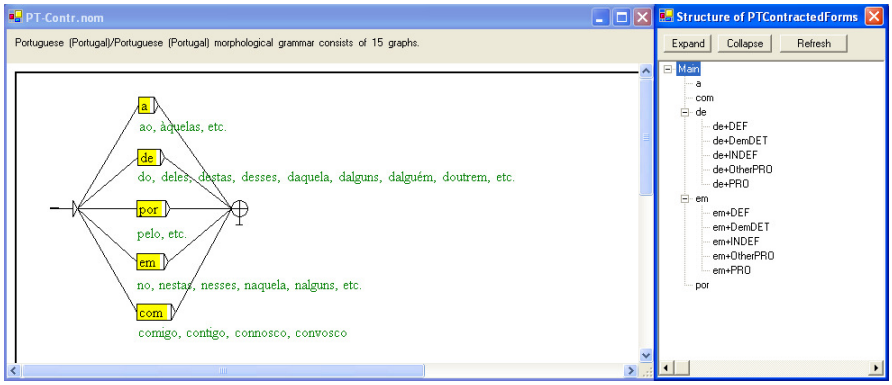


Figure 16: Morphological grammar for Portuguese contracted forms

One of the sub-graphs of the morphological grammar that formalizes contractions of the preposition *por* with Portuguese definite articles is contained in Figure 17.

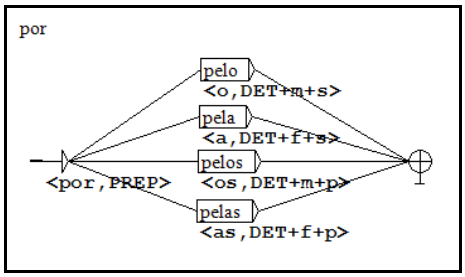


Figure 17: Graph for contracted forms resulting from preposition *por* with definite articles

Figure 18 illustrates another sub-graph of the morphological grammar that formalizes contractions of the preposition *em* with demonstrative and personal pronouns.

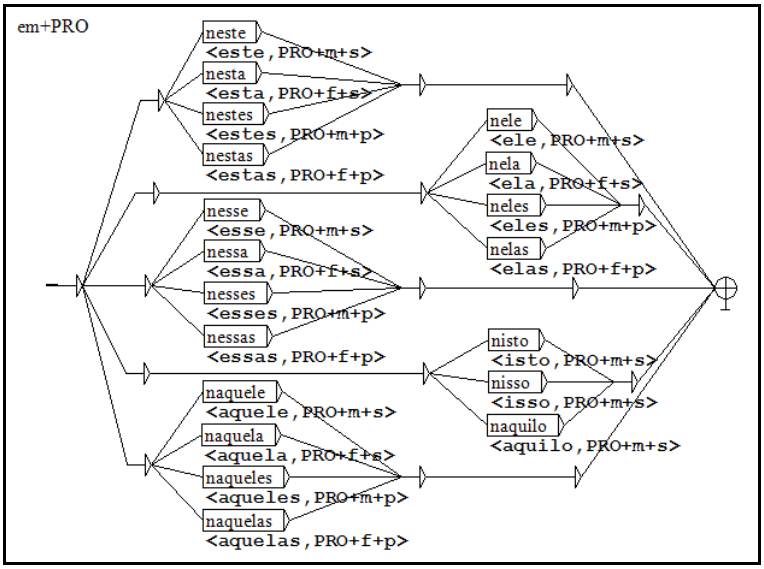


Figure 18: Graph for contracted forms resulting from preposition *em* with demonstrative and personal pronouns

During the normalization phase, words can be de-constructed to their basic constituents by applying the contracted forms grammar to the text, during the normalization phase. This is illustrated in Figure 19 below. The annotations of each word in that text can be seen beneath in the grey box. Clicking on the first sentence, *Declaração Universal dos Direitos Humanos*, enables the annotations for each word of that sentence (lexical ambiguities) to be scrolled through. The resolution of the contraction *dos* as *de*, *PREP* plus *os*, *DET+m+p* and of *do* as *de*, *PREP* plus *o*,*DET+m+s,* can also be observed.

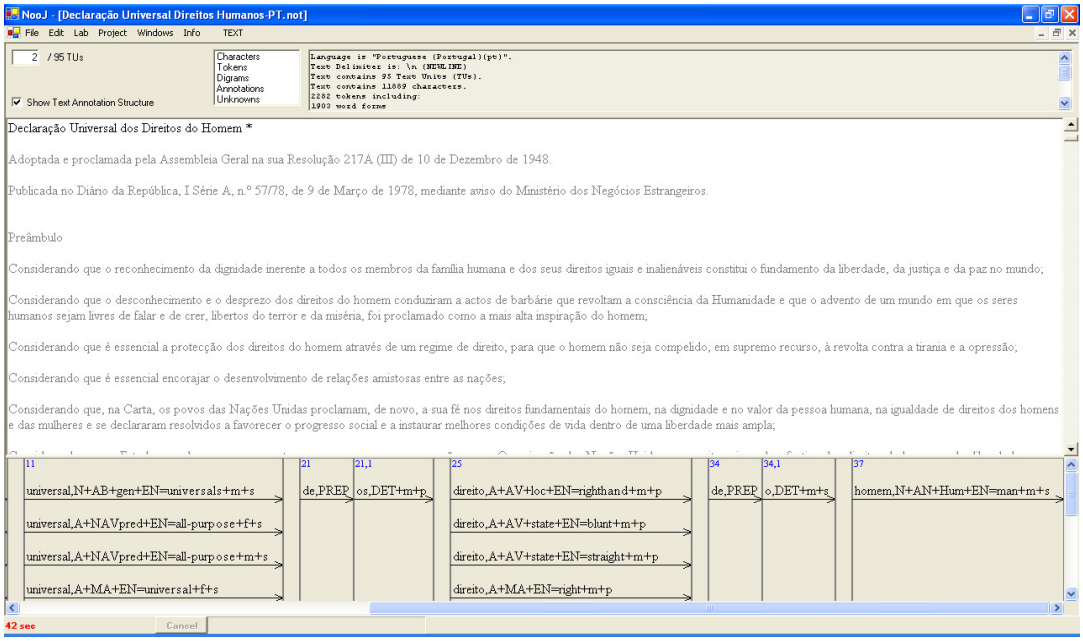


Figure 19: Annotated text with decomposed contractions

5.2. Syntactic-Semantic Grammars

The so called 'syntactic grammars' in NooJ are more than just syntactic grammars. They are used to identify and annotate syntactic patterns but also semantic units either separately or as part of the same analysis. They reflect the fact that there is often no clear separation of syntax and semantics in languages. For this reason, we have termed them, syntactic-semantic grammars. Syntactic-semantic grammars can be used, among many other things, for the identification and annotation of dates. Figure 20 shows a simple example of this. The meta-graph references three sub-graphs, Week, Month, and Year, each representing a category of the type Date.

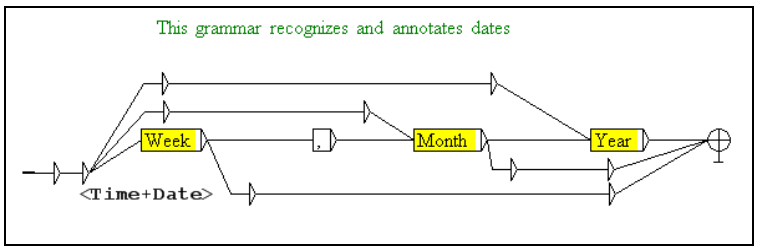


Figure 20: Graph to recognize and annotate dates

Figure 21 details the concordance resulting from the application of the grammar in Figure 20 to text, and the recognition and annotation of dates, such as *sexta-feira* (*Friday*); *em 1831* (*in 1831*); *meados de agosto* (*middle August*); *mês de Abril* (*month of April*); *17 deste mês de julho* (*17 of this month of July*); *ano da graça de 1843* (*year of grace of 1843*), among others.

São	17 deste mês de julho/ <Time+Date>	, ano da graça de 1843, uma Segunda feira, dia sem nota
e mês de julho,	ano da graça de 1843/ <Time+Date>	, uma Segunda feira, dia sem nota e de boa estréia. Seis
graça de 1843, u...	Segunda feira/ <Time+Date>	, dia sem nota e de boa estréia. Seis horas da manhã a
a um sol português	de julho/ <Time+Date>	! A doçura que mete na alma
nos primeiros dias	de abril/ <Time+Date>	, ondulando lascivamente com a brisa temperada
hora da rega, por	meados de agosto/ <Time+Date>	, a ver-se-lhe pular os caules com a água que lhe anda p
ão, da revolução	de julho/ <Time+Date>	(isto era em 1831), de M. de Lafayette, de Luís Filipe,
de julho (isto era	em 1831/ <Time+Date>	), de M. de Lafayette, de Luís Filipe, de Chateaubriand -
raiar uma alvorada	de maio/ <Time+Date>	!... Se haverá ali quem a aproveite,
Era	no ano de 1832/ <Time+Date>	, uma tarde de verão como hoje calmosa, seca, mas o
Joaninha o riso. -	Sexta feira/ <Time+Date>	de aziago.
va habilitado, e	em 1825/ <Time+Date>	, do lugar de corregedor do Ribatejo, em que já
Isto fora numa	sexta-feira/ <Time+Date>	; daí por diante em todas as sextas-feiras de cada s
. Andava ele já	no último ano/ <Time+Date>	de Coimbra e ia formar-se em leis, quando Frei Din
meio o memorável	ano de 1830/ <Time+Date>	, e Carlos, que se formara no princípio daquele verão, t
, e só por fins	de agosto/ <Time+Date>	voltara para a sua família. E veio triste, melancólico
e chegou era uma	sexta feira/ <Time+Date>	, dia de Frei Dinis vir ao vale.
Na	sexta feira/ <Time+Date>	depois da partida de Carlos, Frei Dinis veio ao vale
o, chegando outra	sexta-feira/ <Time+Date>	e estando a avó e a neta à espera do frade, este lhe
Chegou a	sexta-feira/ <Time+Date>	; e as horas desse dia, sempre desejado e sempre te
ras!... Leiam, e	sexta feira/ <Time+Date>	que vem... me dirão...
-	Sexta feira/ <Time+Date>	que vem - continuou Frei Dinis, sem ouvir ou sem
der a pergunta; -	sexta feira/ <Time+Date>	que vem eu tomarei conta da resposta, e lha farei ch
cair, quando uma	sexta-feira/ <Time+Date>	, ao pôr do sol, Frei Dinis aparecia no vale mais cu
Era a retirada	de 11 de outubro/ <Time+Date>	. - Deus tenha compaixão de
nte era chegado o	mês de abril/ <Time+Date>	, estávamos em plena e bela primavera.
m um triste dia	de novembro/ <Time+Date>	com o raio do sol transiente e inesperado que lhe rompe
Mas um dia	de abril/ <Time+Date>	é imenso; interminável. E as últimas horas pareciam
bem que hoje não é	sexta-feira/ <Time+Date>	, senão não vinha eu cá. - Por qu
por pecado, que é	sexta-feira/ <Time+Date>	. - Não te v

Figure 21: Concordance with annotations for dates

Syntactic-semantic grammars can also be used for named entities recognition, currently a very popular field with NooJ users [Mota & Silberztein, 2007]. Figure 22 illustrates a simple local grammar for recognition and annotation of named entities of the type PERSON, such as *Ana*, *rainha Isabel II de Inglaterra* (*Queen Elisabeth II of England*), *D. Maria I*, *papa J. Paulo II* (*Pope John Paul II*), *Sr. João*, etc. The meta-graph invokes four sub-graphs, namely Title, Name, Place and Organization, each representing a category of the entity type *PERSON*.

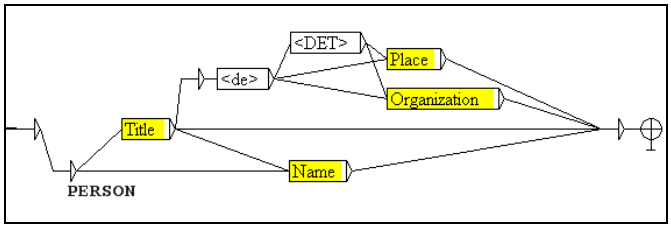


Figure 22: Graph for recognition and annotation of PERSON-type named entities

Figure 23 details the concordance following application of the grammar of Figure 22. The named entities are recognized and annotated by the category applying. For example, *Dom Quixote* and *Doutor Fausto* (*Doctor Fausto*) are annotated as PERSON+Title+Name because both *Dom* and *Doutor* are person titles, and *Quixote* and *Fausto* are person names; *Duquesa de Abrantes* (*Duchess of Abrantes*, where *Abrantes* is a Portuguese town) and *El-Rei de Dinamarca* (*King of Denmark*) are annotated as *PERSON+Title+Place* because they refer to people using the name of the place where they fulfill their title; *doutor* (*doctor*) and *enfermeiro* (*nurse*) are annotated as *PERSON+title*.

certo ponto, o	Dom Quixote/ PERSON+Title+Name	da sociedade velha.
o. O que eu fiz na	Dona Branca/ PERSON+Title+Name	é pouco e mal esboçado à pressa. O grande mago lusitano
A	Dona Branca/ PERSON+Title+Name	três, Frei Soeiro, Frei Lopo e S. Frei Gil - faz quatro
esus, logo o das	Donas/ PERSON+Title	, depois o de S. Domingos, célebre pelo jazigo do no
- S. Frei Gil e o	Doutor Fausto/ PERSON+Title+Name	. - De como o A. foi ao túmulo do santo bruxo e o achou va
antes e depois do	Doutor Fausto/ PERSON+Title+Name	. Mas sem Homero ou Goethe é que se não chega à rep
amei já o nosso	Doutor Fausto/ PERSON+Title+Name	: e é com efeito. Não lhe falta senão o seu Goethe.
vê que o nosso	doutor/ PERSON+Title	de bivaque, o soldado que lhe chamou maluco ao pensad
impressão era a	Duquesa de Abrantes/ PERSON+Title+Place	. Mas em meia hora de conversação, de trato, descobriam-
que lhe deram. - A	Duquesa de Abrantes/ PERSON+Title+Place	. - Chega- se enfim ao Vale de Santarém.
adíssimo bobo de	el-rei de Dinamarca/ PERSON+Title+Place	, o que alguns anos depois ressuscitou em Sterne com tã
e e o Condestável,	el-rei D. Fernando/ PERSON+Title+Name	e a Rainha D. Leonor, Camões desterrado aqui, Frei Lu
meiro-ministro de	el-rei D. José/ PERSON+Title+Name	? Por onde está Ixião e Tântalo, por onde demora Sísifo
seu augusto amo	el-rei de Dinamarca/ PERSON+Title+Place	. Por pouco mais que se generalize o princípio,
s promessas - como	el-rei de Prússia/ PERSON+Title+Place	prometeu uma constituição; e não faltou ainda, porque,
o.- De como S. M.	El-Rei de Dinamarca/ PERSON+Title+Place	tinha menos juízo do que Yorick, seu bobo. - Doutrina deste
a, em que domina	el-rei Sancho/ PERSON+Title+Name	. Depois há
ssível. Desenganado	el-rei/ PERSON+Title	de que um poder sobre-humano não permitia que ele s
ho, beijou a mão à	el-rei/ PERSON+Title	, e dai tomou um dia o caminho de Santarém, chegou àquel
nosso Carlos; e o	enfermeiro/ PERSON+Title	que o velava, uma bela mulher de estatura não acima de
emada elegância do	enfermeiro/ PERSON+Title	que o velava. O quarto era com efei
- O hospital. - O	enfermeiro/ PERSON+Title	. - Georgina.
parte da nossa	Espanha/+Place	é, geologicamente falando, já tão África, tão pouco
rase didática das	Espanhas/+Place	chamou romance em endechas. Eu, adotando para ele,

Figure 23: Concordance with annotation for PERSON-type named entities

5.3. Disambiguation Grammars

Disambiguation grammars perform the essential task of filtering out lexical or syntactic annotations in the text. Any particular word form can usually correspond to more than

one lexical entry, given that most words are ambiguous. For instance, the word form *a* has four entries in our main dictionary, each with a different part-of-speech. Without disambiguation grammars, this ambiguity would remain as part of the text as illustrated in Figure 24. The word *a* in *beijou-a* (*kissed her*) is annotated as:

(1) a noun (N), masculine, singular, classified as alphanumeric (alnum) and corresponding to the English word *a*, that is, the letter of the alphabet;
(2) a preposition (PREP), locative (Loc), corresponding to the English preposition *at*;
(3) a pronoun (PRO), personal (PERS), direct object (OD), third person, singular, feminine, corresponding to the English pronoun *her*;
(4) a determiner (DET), definite article (DEFART), feminine singular and whose canonical form is o, corresponding to the English article *the*.

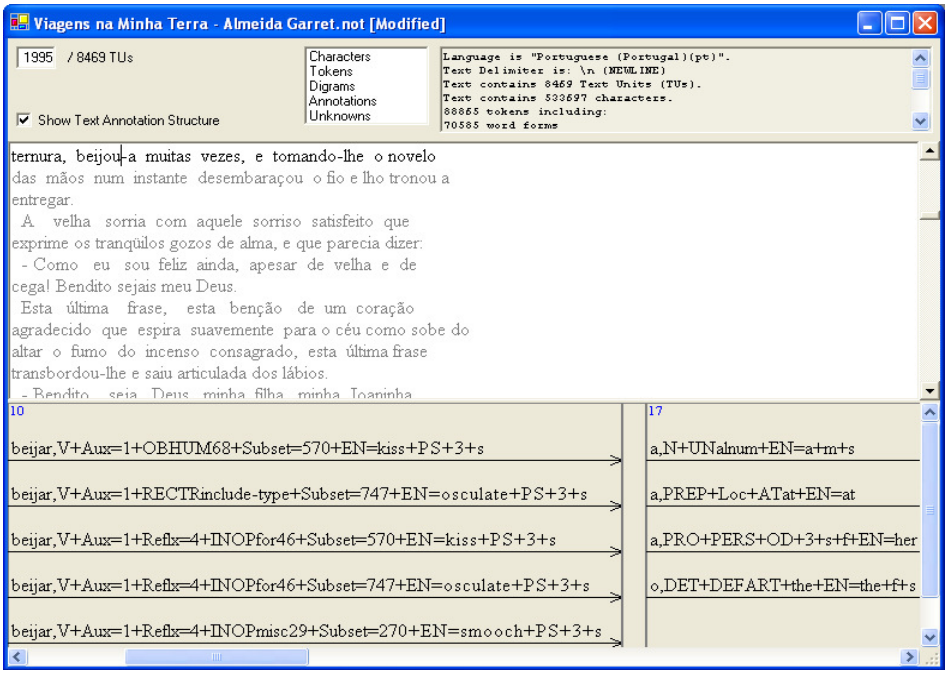


Figure 24: Ambiguity of the word form a

A simple disambiguation grammar such as that illustrated in Figure 25 can be used to resolve ambiguity of the type seen in Figure 24.

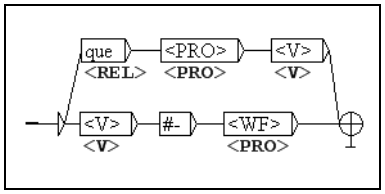


Figure 25: Graph to annotate and disambiguate pronouns before and after verbs following relative pronoun *que*

The grammar in Figure 25 stipulates that a word inserted between the relative pronoun *que* (*REL*) (EN *that*) and a verb (*V*) is a pronoun (*PRO*), and any word form (*WF*) coming after a verb (*V*) followed by an hyphen, is also a pronoun (*PRO*). This enable almost noise-free recognition and annotation of pronouns such as those illustrated in Figure 26 (all instances are correct in this particular concordance). Unwanted annotations may still arise in the case of larger concordances and more work will be required to eliminate them.

utor prestou-se a	dirigi-la/<V><PRO>	ele mesmo, corrigiu-a, aditou-a, alterou-a em muitas
gi-la ele mesmo,	corrigiu-a/<V><PRO>	, aditou-a, alterou-a em muitas partes, e a ilustrou
rrigiu-a, aditou-a,	alterou-a/<V><PRO>	em muitas partes, e a ilustrou com as notas mais in
to; e como resolveu	imortalizar-se/<V><PRO>	escrevendo estas suas viagens. Parte para Santaré
e um bom charuto.	Travam-se/<V><PRO>	de razões os ilhavos e os Bordas- d'Água: os da calç
cias de um amigo,	decidem-me/<V><PRO>	as tonteiras de um jornal, que por mexeriquice quis
por isso mesmo vou:	pronunciei-me/<V><PRO>	. São 17 deste mês de julho, ano
das populações, e	que se chama/<REL><PRO><V>	a si mesma por excelência a Sociedade, os seus passeio
ita reação antes de	completar-se/<V><PRO>	... No entreta
ossos charutos, e	deixe-mos/<V><PRO>	os precintos aristocráticos da ré; à proa, que é país d
u cigarro de papel,	que me vai/<REL><PRO><V>	emprestar lume. - Dou-lho eu, senhor
ribatejano.	Acenderam-se/<V><PRO>	os charutos, e atentamos mais devagar na companhia
raros pintos por	que se manifesta/<REL><PRO><V>	o sempre clamoroso e sempre vazio entusiasmo das multid
e forcado, assim	que o viu/<REL><PRO><V>	). Isto é um fidalgo como se quer. Nunca o vi numa ferr
i a nossa gente	que o sachou/<REL><PRO><V>	e plantou, e o fez o que é, e fez terra das areias da charn
, e as ricas terras	que lhes levam/<REL><PRO><V>	as enchentes. Mas nos , pé no barco, pé na terra, tã
- A força é	que se fala/<REL><PRO><V>	- tornou o campino para estabelecer a questão em t
estão em terreno	que lhe convinha/<REL><PRO><V>	. - A força é que se fala: um homem do campo que se deita
inha. - A força é	que se fala/<REL><PRO><V>	: um homem do campo que se deita ali à cernelha de um toi
um homem do campo	que se deita/<REL><PRO><V>	ali à cernelha de um toiro que uma companhia inteira de
	Declaram-se/<V><PRO>	típicas, simbólicas e míticas estas viagens. Faz o A.
civilização: e	mostra-se/<V><PRO>	como ela é dirigida pelo cavaleiro da Mancha, D.
r ao leitor, para	que ele esteja/<REL><PRO><V>	prevenido; não cuide que são quaisquer dessas rabis
moda germânica,	que se mete/<REL><PRO><V>	hoje em tudo e com que se explica tudo... quanto se nã
em tudo e com	que se explica/<REL><PRO><V>	tudo... quanto se não sabe explicar.
, e que pode bem	personalizar-se/<V><PRO>	, simbolizar-se pelo famoso mito do cavaleiro da mancha, D.
em personalizar-se,	simbolizar-se/<V><PRO>	pelo famoso mito do cavaleiro da mancha, D. Quixote;
utopias, pode bem	representar-se/<V><PRO>	pela rotunda e anafada presença do nosso amigo velho,
porque receio muito	que se esqueça/<REL><PRO><V>	. Somos chegados ao triste desembarc
	Parece-me/<V><PRO>	estar mais deserto e sujo, mais abandonado e em ruín

Figure 26: Concordance showing annotation of sequences <*V*> <*PRO*> and <*REL*> <*PRO*> <*V*>

Application of the grammar detailed in Figure 25 eliminates three of the ambiguities presented in the lexicon. As can be seen from Figure 27, the word *a* in *beijou-a* (*kissed*

her) is annotated only as a pronoun (*PRO*), personal (*PERS*), direct object (*OD*), third person, singular, feminine, corresponding to the English pronoun *her*, and comes immediately following the verb *beijar* (*kiss*), which is the correct annotation for the word in that context.

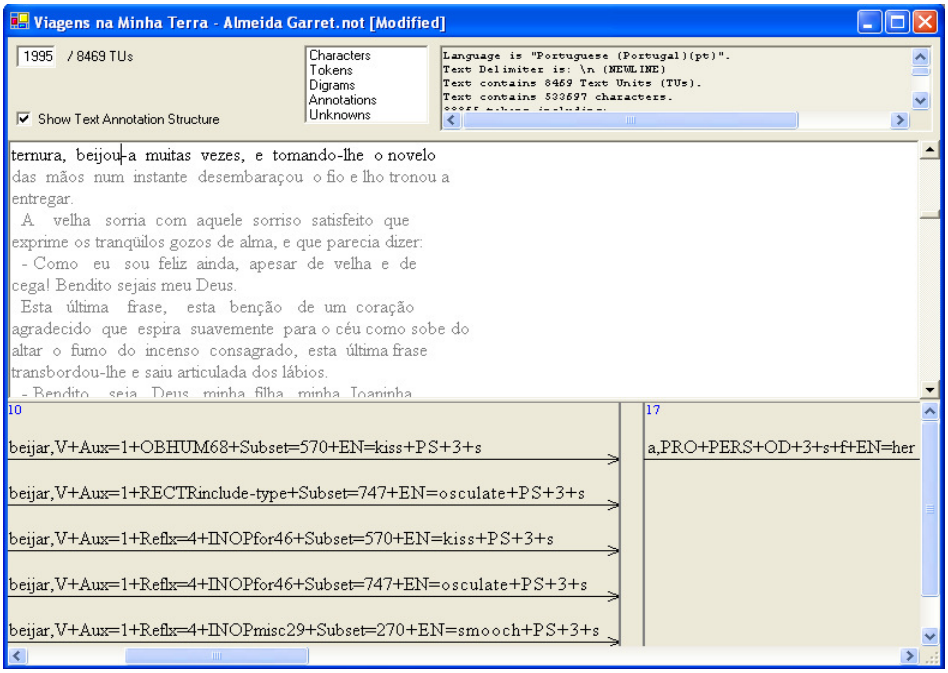


Figure 27: Word form a following a verb and separated by a hyphen is disambiguated as a personal pronoun

5.4. Grammars for Multiword Expressions

These grammars formalize different types of multiword expressions, such as phrasal verbs, support verb constructions, noun compounds and idioms. Figure 28 illustrates a very simple grammar, which recognizes and annotates support verb constructions and their predicates. The grammar checks for a support verb (*VSUP*), followed by any left modifier (LeftMod) and a nominalization (*N+Nom*), and annotates it as a support verb construction, identifying the contents of the variable *N* (in parentheses) as the predicate of that construction.

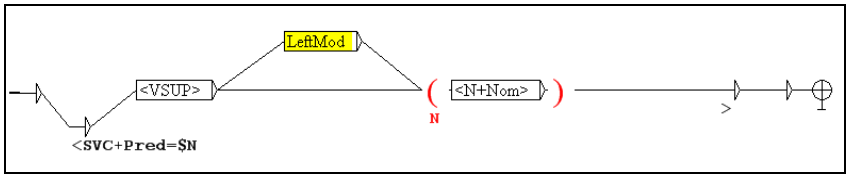


Figure 28: Grammar for recognizing and annotating support verb constructions and their predicates

To establish relations of equivalent morpho-syntactic predicates in the same language (Portuguese) or between two languages (particularly between Portuguese and English), as mentioned earlier, all predicate nouns in the dictionary have been classified as [NPred]. This lexical information can be used in a syntactic grammar to identify the predicate in a support verb construction and apply this grammar in corpora. Given the information in the lexicon, it is possible to identify and tag support verb constructions in a Portuguese text and identify the predicate noun for each support verb construction. After identifying the predicate noun, it is associated to a corresponding lexical strong verb, if it exists, and monolingual paraphrases are obtained. This is possible because there is a link in the dictionary between the nominalization and the support verb, with specification of the lexical verb that corresponds to the combination of those elements. For example, Portuguese the multiword expression *dar um beijo* (*to give a kiss*) is recognized as a support verb construction, whose predicate noun is *beijo* (*kiss*). Figure 29 illustrates a concordance that identifies and tags support verb constructions in text, and identifies the predicate noun for each support verb construction.

no bordado para lhe	dar um beijo/<SVC+Pred=beijo>	na cara e os nossos olhos se cr
uer sair. -- Está a	dar uma festa/<SVC+Pred=festa>	? -- perguntou. Talvez fosse do
a uma agulha me faz	dar um salto/<SVC+Pred=salto>	; e, quando não consegue encor
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>	-- o que é muito raro --, fica m
o ponto de me fazer	dar um grito/<SVC+Pred=grito>	. Toma conta da loja quando o
fazer dar um grito.	Toma conta/<SVC+Pred=conta>	da loja quando o pai está a dar
lições a iniciados.	Toma conta/<SVC+Pred=conta>	da loja quando o pai está a dar
regados, alunos que	fazem gazeta/<SVC+Pred=gazeta>	e mães com crianças de colo, c
ianças de colo, que	dão graças/<SVC+Pred=graças>	por terem este local acolhedor
que dão graças por	terem este local acolhedor/<SVC+Pred=local>	e alegre para passarem as tarde
primeira pedra, que	dá a outra/<SVC+Pred=a>	face, e por fora, mas não me c
jogador encantador,	dá gosto/<SVC+Pred=gosto>	vê-lo a defender, como se tives
a defender, como se	tivesse a bola atada/<SVC+Pred=bola>	aos pés, desviando os adversár
a maneira de evitar	fazer o pino/<SVC+Pred=pino>	, pôr-se a dar murros no peito c
er o pino, pôr-se a	dar murros/<SVC+Pred=murros>	no peito ou gritar de excitação.

Figure 29: Annotation of support verb constructions and identification of the predicate noun

Figure 30 shows the support verb construction *fez um esforço* (*made an effort*) in text before the application of the grammar in Figure 28. Figure 31 shows the same support verb construction already identified and annotated.

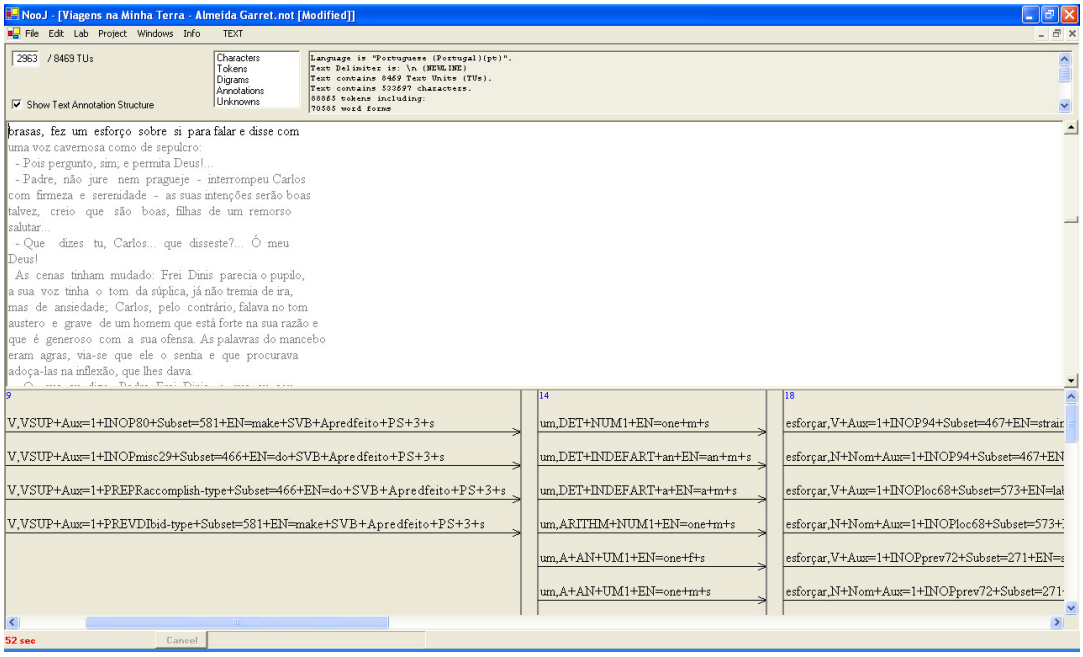


Figure 30: Annotation for the support verb construction *fez um esforço* before the application of a support verb construction grammar

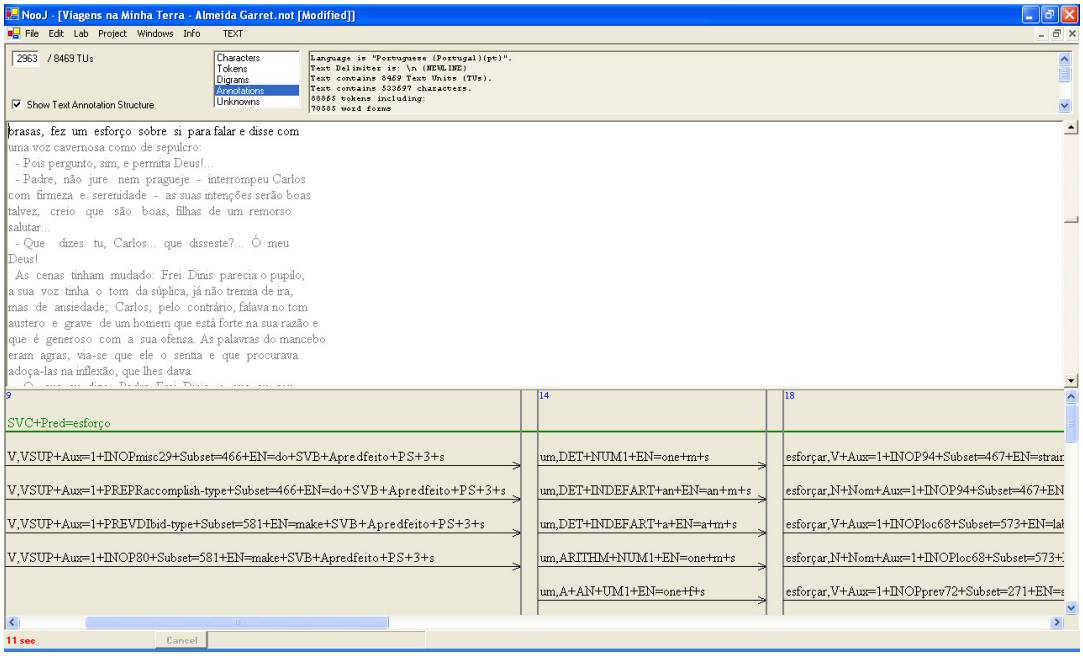


Figure 31: Annotation for the support verb construction *fez um esforço* after the application of a support verb construction grammar

Grammars that facilitate the recognition and annotation of multiword expressions, such as support verb constructions, may enable the whole expression to be paraphrased using another equivalent construction or replacing it by a strong lexical verb, and for this reason they are extremely important for translation. For example, the support verb construction *fez um esforço* in Figures 30-31 could be paraphrased by the lexical verb *esforçou-se* and translated into English as one single word, *tried*. Figure 32, extracted from [Barreiro, 2008b] represents a "naïve" local grammar used to recognize and generate support verb constructions and transforms them into their verbal paraphrases.

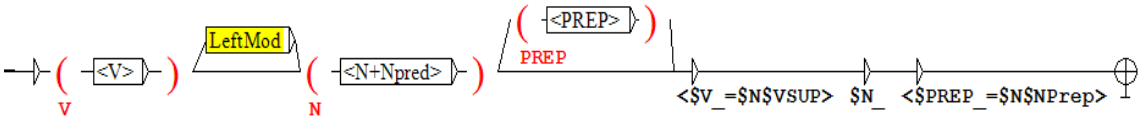
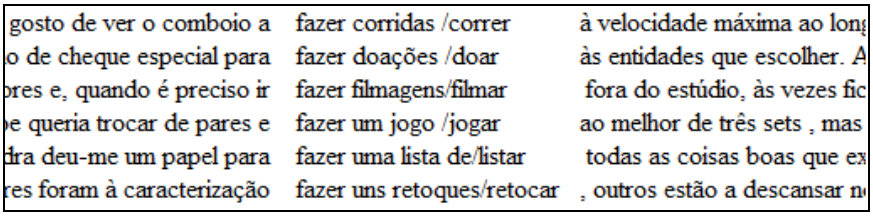


Figure 32: Grammar to recognize and paraphrase support verb constructions

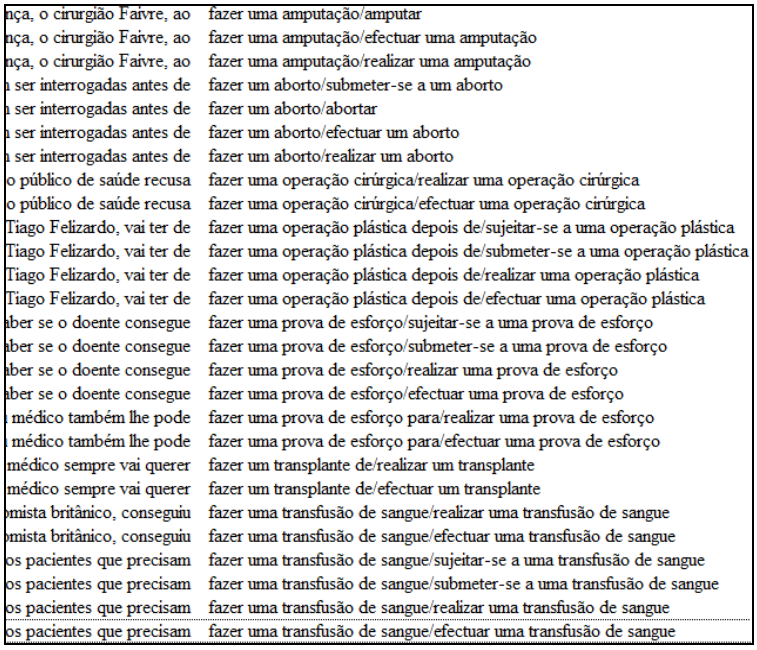
This grammar matches verbs, which are marked in the dictionary as support verbs that are followed by a left modifier (determiner, adjective or adverb or other quantifiers), a predicate noun and optionally a preposition. Since we have classified all predicate nouns in the dictionary as [NPred], we can now use this lexical information in a syntactic grammar to identify the predicate in a support verb construction and apply this grammar in corpora. The elements in parentheses ( ) are stored in variables V, N or PREP. If a dictionary entry has a lexical constraint, such as NPrep=*a* in the phrase [*dar um grande abraço a*] (*to give a big hug to*], the support verb construction will be recognized by the grammar and mapped to the verb *abraçar* (*to hug*), the lemma of the noun specified in the variable $N_. The elements in bold <$V_=$N$VSUP>, and $PREP_=$N$NPrep> represent lexical constraints that are displayed in the output, such as specification of the support verb or the preposition that belongs to a specific support verb construction. The predicate noun is identified, mapped to its deriver and displayed as a verb, the other elements of the phrase are eliminated. Figure 33 shows a concordance where Portuguese support verb constructions are recognized and paraphrased as lexical strong verbs.



gosto de ver o comboio a	fazer corridas /correr	à velocidade máxima ao lon
o de cheque especial para	fazer doações /doar	às entidades que escolher. A
res e, quando é preciso ir	fazer filmagens/filmar	fora do estúdio, às vezes fic
e queria trocar de pares e	fazer um jogo /jogar	ao melhor de três sets , mas
dra deu-me um papel para	fazer uma lista de/listar	todas as coisas boas que ex
res foram à caracterização	fazer uns retoques/retocar	, outros estão a descansar n

Figure 33: Recognition and monolingual paraphrasing of support verb constructions (Support verb construction / corresponding verb)

Figure 34 shows a concordance where Portuguese biomedical-related support verb constructions are recognized and paraphrased as lexical strong verbs or as stylistic variants. Stylistic variants *sujeitar-se a* and *submeter-se a* (lit. *to be submitted to*) are only allowed when the subject is a patient. Some lexical strong verbs are only allowed with agentive subjects. There is a strong connection between predicate-argument structure knowledge and the use of a particular stylistic variant.



nça, o cirurgião Faivre, ao	fazer uma amputação/amputar
nça, o cirurgião Faivre, ao	fazer uma amputação/efectuar uma amputação
nça, o cirurgião Faivre, ao	fazer uma amputação/realizar uma amputação
a ser interrogadas antes de	fazer um aborto/submeter-se a um aborto
a ser interrogadas antes de	fazer um aborto/abortar
a ser interrogadas antes de	fazer um aborto/efectuar um aborto
a ser interrogadas antes de	fazer um aborto/realizar um aborto
o público de saúde recusa	fazer uma operação cirúrgica/realizar uma operação cirúrgica
o público de saúde recusa	fazer uma operação cirúrgica/efectuar uma operação cirúrgica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/sujeitar-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/submeter-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/realizar uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/efectuar uma operação plástica
aber se o doente consegue	fazer uma prova de esforço/sujeitar-se a uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/submeter-se a uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/realizar uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/efectuar uma prova de esforço
médico também lhe pode	fazer uma prova de esforço para/realizar uma prova de esforço
médico também lhe pode	fazer uma prova de esforço para/efectuar uma prova de esforço
médico sempre vai querer	fazer um transplante de/realizar um transplante
médico sempre vai querer	fazer um transplante de/efectuar um transplante
mista britânico, conseguiu	fazer uma transfusão de sangue/realizar uma transfusão de sangue
mista britânico, conseguiu	fazer uma transfusão de sangue/efectuar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/sujeitar-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/submeter-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/realizar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/efectuar uma transfusão de sangue

Figure 34: Recognition and monolingual paraphrasing of biomedical-related support verb constructions (support verb construction / corresponding verb or stylistic variant)

5.5. Translation Grammars

Machine translation using NooJ is implemented by translation grammars, relating two languages by means of variables and a translation operator (*TRANS*). Figure 35 shows a graph to implement basic Portuguese-English translations. This simple example

demonstrates NooJ's suitability for machine translation and its capability of providing several appropriate translations of the same sentence.

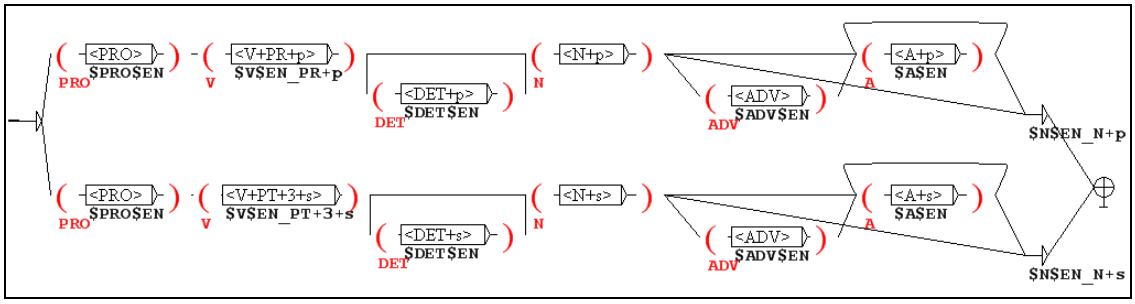


Figure 35: Graph to translate simple sentences

Figure 36 shows a concordance containing several different translations for the same sentence resulting from application of the graph in Figure 35 to text consisting of a sample of test suite sentences. This ambiguity results from the fact that there is more than one dictionary entry for the same word, because each word can have more than one transfer. For example, the Portuguese noun *casa* is listed twice in the dictionary, with the English transfers *house* and *home*, classified with the attributes for enclosed (*encl*) place (*PL*): *casa,N+FLX=CASA+PL+encl+EN=home* and *casa,N+FLX=CASA+PL+encl +EN=house*.

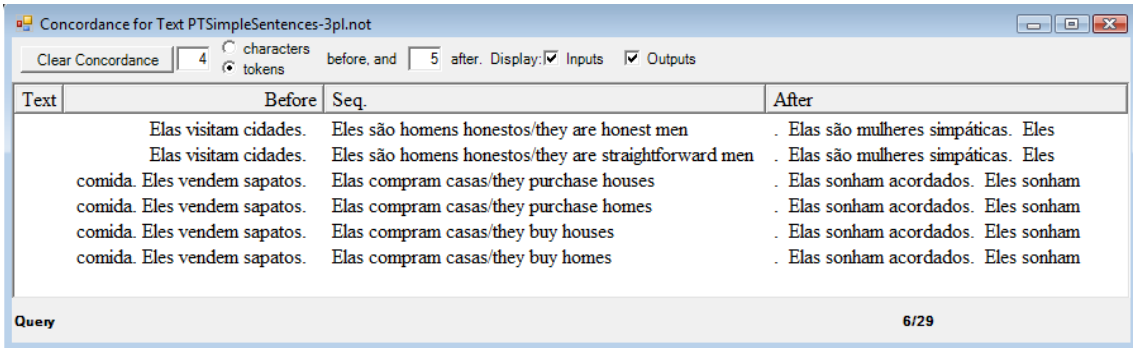
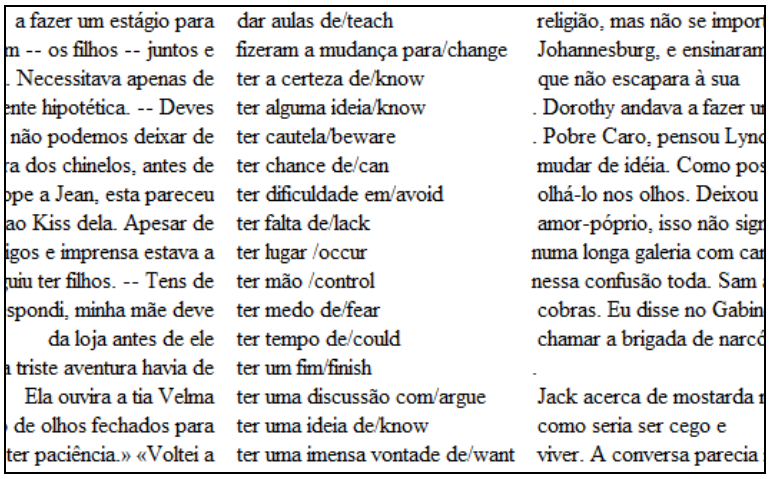


Figure 36: Translation and bilingual paraphrasing of simple sentences

These simple illustrative translations show that, like humans, machines can also produce more than one "acceptable" translation for the same sentence. Parallel translations of the same sentence are paraphrases. Like other machine translation

systems, NooJ enables multiple translation, with the advantage that the *+UNAMB* feature makes it possible to establish a default transfer to be used in general translations. This provides one translation, eliminating all other possibilities. As a result, the system can be customized according to the needs of the linguist or the user, provide multiple translations or one specific translation.

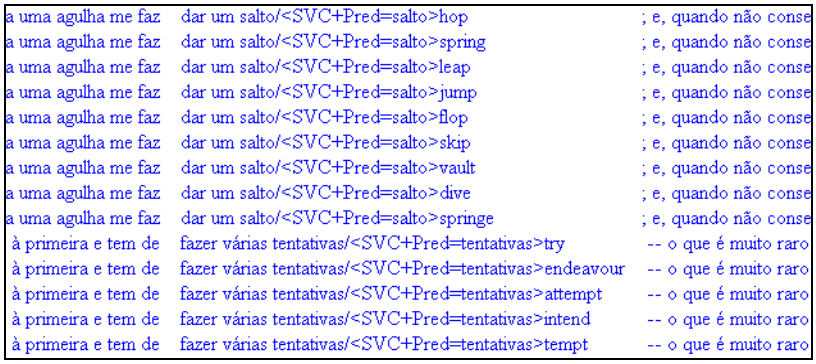
A similar grammar to the one in Figure 32 is used to generate English translations. The only difference is that the output is specified to be in English. Figure 37 shows a concordance where Portuguese support verb constructions are recognized in a text and converted automatically into an English lexical verbs.



a fazer um estágio para	dar aulas de/teach	religião, mas não se impor
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
nte hipotética. -- Deves	ter alguma ideia/know	. Dorothy andava a fazer u
não podemos deixar de	ter cautela/beware	. Pobre Caro, pensou Lyn
a dos chinelos, antes de	ter chance de/can	mudar de idéia. Como po
pe a Jean, esta pareceu	ter dificuldade em/avoid	olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de	ter falta de/lack	amor-próprio, 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 37: Recognition and bilingual paraphrasing of support verb constructions (Portuguese support verb construction/corresponding English verb)

The concordance illustrated in Figure 38 shows that the output produces several different English lexical verbs for each support verb construction. For instance, *fazer várias tentativas* (*make several attempts*) can be translated into five different lexical verbs *try*, *endeavour*, *attempt*, *intend*, and *tempt*. This ambiguity is related to the fact that there are five English dictionary transfers for Portuguese verb *tentar*, the lexical verbs *try*, *endeavour*, *attempt*, *intend*, and *tempt*. The support verb construction *fazer várias tentativas* could be further translated into *strive*, *aim*, *seek*, or *undertake*, if these verbs were dictionary transfers for the Portuguese verb.



a uma agulha me faz	dar um salto/<SVC+Pred=salto>hop	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>spring	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>leap	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>jump	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>flop	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>skip	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>vault	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>dive	; e, quando não conse
a uma agulha me faz	dar um salto/<SVC+Pred=salto>springe	; e, quando não conse
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>try	-- o que é muito raro
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>endeavour	-- o que é muito raro
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>attempt	-- o que é muito raro
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>intend	-- o que é muito raro
à primeira e tem de	fazer várias tentativas/<SVC+Pred=tentativas>tempt	-- o que é muito raro

Figure 38: Annotation of support verb constructions, identification of the predicate noun and translation into a single English verb

Semantic constraints can be used for meaning disambiguation and refinement, if necessary or preferable as long as we define those constraints in the grammars, as in Figure 39. Also, it is possible to establish a default transfer to be used in general translations so that the output does not show several possibilities. However, the interesting aspect here is to show that there are several valid possibilities for translating the same word or phrase, i.e., there are several bilingual paraphrases, which can be used either to simplify text before translation or to use in comparing legitimate outputs automatically. The grammar corresponding to this concordance has a constraint to tell NooJ that the Portuguese noun "*tentativa*" occurs with the support verb *fazer*. In other words, any graph that deals with support verb constructions indicates that it is not any support verb that can be used with any noun, but only the one specified in the dictionary. Compound variables are used so that the noun in the recognized sentence actually corresponds to the support verb in the dictionary for that noun. Figure 39 shows how the syntactic-semantic properties in the dictionaries are used in local grammars to paraphrase Portuguese support verb constructions for *fazer barulho* (*make a noise*) and translate them into English. This grammar recognizes the sequence of a support verb with a predicate noun of the type [measure + abstract + noise] with any pre- or post-modifiers and translates it into *make a noise*. The grammar filters support verbs or support-verbs-like, such as *fazer* (*make*), *produzir* (*produce*) or *criar* (*create*) and predicate nouns such as *barulho*, *ruído*, *barulheira*, *chinfrineira*, *chinfrim*, etc. as long as they are classified in the dictionary with the semantic property “noise”. This is the type of paraphrase from one support verb construction into another support verb construction in different languages. This

grammar shows how to recognize very specific support verb constructions in corpora with exact pattern recognition.

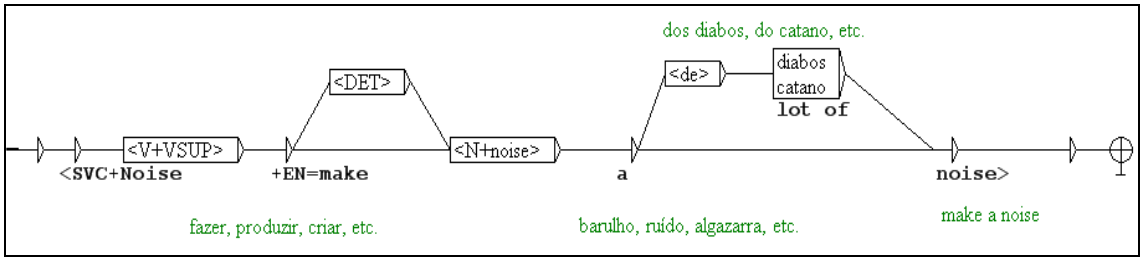


Figure 39: Local grammar to analyze, paraphrase and translate support verb constructions for Portuguese *fazer barulho* (*make a noise*)

Figure 40 shows the application of the previous grammar to text; i.e., the recognition and paraphrasing of Portuguese support verb constructions in text and their translation into English.

lo. Eles estão a	fazer barulho/<SVC+Noise+EN=make a noise>	. A criança fazia
a. A passarada	fazia uma algazarra/<SVC+Noise+EN=make a noise>	enorme. O carr
norme. O carro	fazia uma barulheira/<SVC+Noise+EN=make a noise>	enorme. A máq
arulho. O bebé	faz uma berraria/<SVC+Noise+EN=make a noise>	enorme. O beb
no. A multidão	fará um barulho/<SVC+Noise+EN=make a noise>	dos diabos. Os

Figure 40: Application of previous local grammar to text

Figure 41 shows how to use semantic constraints to filter out undesired translations. For example, the verb *pregar* is translated differently into English depending of the noun that follows it. If it is a noun type information (IN), instructional/legal, ritual, such as *missa* (*mass*) or *sermão* (*sermon*), the expression translates into the lexical verb *preach* or into *say* plus the noun transfer *mass* or *sermon*. If it is a noun type abstract (AB), general concept (gen), such as *ideia* (*idea*), *virtude* (*virtue*), *religião* (*religion*) or *verdade* (*truth*), the expression translates into the lexical verb *proclaim* or *advocate*. If it is a noun type concrete (CO), fastener (fast), such as *prego* (*nail*), the expression translates into the lexical verb *hammer* plus the noun transfer. If it is the noun type abstract (AB) concept, negative cause (negc), such as *susto* (*fright*), the expression translates into the lexical verb *scare*. In this case, it is required an argument N1, corresponding to the indirect object *a N* (preposition + N).

```
pregar Det N(missa,sermão) > preach = say N
pregar Det N(ideia,virtude,religião,verdade) > proclaim N > advocate N
pregar Det N(prego ,etc.) > hammer N(nail,etc.)
pregar N(susto) Prep(a) N > scare N
estabelecer Det N(negócio,empresa,loja,etc.) > open Det N
estabelecer N(regras,princípios) > lay down N
apresentar N(desculpa) > apologize
apresentar Det N(opinião,sugestão) > give Det N(opinion,suggestion)
apresentar Det N(moção,censura) >  bring N forward
prestar Det N(serviço) > offer Det N(service)
prestar N(atenção) > pay N(attention)
perseguir N(objectivo,propósito,etc.) > follow N
perseguir N(pessoa) > chase N = hunt after/down N
pedir N(desculpa,perdão) > apologize = say sorry
pedir Det N(esmola) > beg
observar Det N(lua) > observe N
observar Det N(lei) > obbey N(law)
provocar Det N(pessoa) > seduce N = make advances on N
provocar N(sarilho,confusão) > pick up N(trouble)
provocar N(tempestade) > cause N(storm)
representar Det N(papel) > play Det N(role)
representar Det N(contributo,etc.) > represent N
```

Figure 41: Sample of Portuguese-English translation rules

Translation grammars assign precision to the translation of expressions that cannot be translated literally. They also help improve meaning disambiguation and provide semantic refinement to the source language. The few Portuguese-English translation rules of Port4NooJ were adapted (inverted) from the Logos English-Portuguese Semtab rules available at Linguateca website: http://www.linguateca.pt/ - *Repositório*.

Acknowledgments

We would like to thank Max Silbeztein and Slim Mesfar for providing support on technical aspects of NooJ, and Paula Carvalho for reading and commenting on this document.

This work was partly supported by grant *SFRH/BD/14076/2003* from *Fundação para a Ciência e a Tecnologia* (Portugal), co-financed by POSI.

References

Anabela Barreiro, 2008a. "Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation". In *Proceedings of the 2007 International NooJ Conference* (Barcelona, Spain, June 7-9, 2007), Cambridge Scholars Publishing.

Anabela Barreiro, 2008b. "ParaMT: a Paraphraser for Machine Translation". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* 5190, (Aveiro, Portugal, 8-10 de Setembro de 2008), Springer Verlag, pp. 202-211.

Anabela Barreiro, 2008c (submitted) "Paraphrasing Biomedical Support Verb Constructions for Machine Translation".

Bernard Scott, 2003. The Logos Model: An Historical Perspective. In: *Machine Translation*, 18, pp. 1-72

Cristina Mota & Max Silberztein, 2007. "Em busca da máxima precisão sem almanaques. O Stencil/NooJ no HAREM". In Diana Santos & Nuno Cardoso (eds.), *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro da Linguateca*. 2007.

Elisabete Marques Ranchhod, Paula Carvalho, Cristina Mota & Anabela Barreiro, 2004. "Portuguese Large-scale Language Resources for NLP Applications". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)* (Lisboa, Portugal, 26-28 May 2004), pp. 1755-1758.

Luís Sarmento, Ana Sofia Pinto & Luís Cabral, 2006. "REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese". In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006*. Itatiaia, Brazil, May 2006 (PROPOR'2006) LNAI 3960, 13-17 de Maio de 2006, Berlin/Heidelberg: Springer Verlag, pp. 31-40.

Max Silberztein, 2004. "NooJ: A Cooperative, Object-Oriented Architecture for NLP". In *INTEX pour la Linguistique et le traitement automatique des langues*. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté.