Complex Predicates annotation in a corpus of Portuguese

Iris Hendrickx, Amália Mendes, Sílvia Pereira, Anabela Gonçalves and Inês Duarte
Centro de Linguística da Universidade de Lisboa, Lisboa, Portugal
{iris, amalia.mendes}@clul.ul.pt

Abstract

We present an annotation scheme for the annotation of complex predicates, understood as constructions with more than one lexical unit, each contributing part of the information normally associated with a single predicate. We discuss our annotation guidelines of four types of complex predicates, and the treatment of several difficult cases, related to ambiguity, overlap and coordination. We then discuss the process of marking up the Portuguese CINTIL corpus of 1M tokens (written and spoken) with a new layer of information regarding complex predicates. We also present the outcomes of the annotation work and statistics on the types of CPs that we found in the corpus.

1 Introduction

Complex predicates are predicates composed of more than one element but functionally equivalent to a single predicate. Examples of complex predicates (CPs) are constructions of verb+noun, like *have a rest*, *take a walk*, and constructions verb+verb, like the constructions with a causative verb in Portuguese, like *mandar ler o livro a alguém* 'make read the book to someone'. These constructions raise interesting questions regarding the aspectual, semantic and syntactic properties which underlie the relationship between the elements of the CP. There are different theoretical perspectives on the compositional nature of CPs. For example, in the case of constructions of the type verb+noun, the verb is either considered a light verb (Jespersen, 1949) or a support verb (Gross, 1981), in the sense that it has lost part or all of its meaning and has no predicative value in the construction, or as an auxiliary verb with aspectual properties (Abeillé et al., 1998).

Our hypothesis is that both elements of the CP seem to contribute to the properties of complex predicates, in such a way that the argument structure and the attribution of thematic roles are determined by both constituents through the combination of their thematic structures (Grimshaw, 1988). One has to address several important questions: is there a systematic relationship between the syntactic and semantic selection properties of the two elements? How do the argument structure of the light verb and the derived noun combine and contribute to define the complex predicate? To study these questions we annotated the Portuguese CINTIL corpus (Barreto et al., 2006) with a new layer on CPs. By taking into consideration different types of CPs and by using corpus data for our analysis of their properties, the objective is to present a unified approach to CP formation, along which the CP constructions available in Portuguese may be accounted for, namely in what concerns their lexico-syntactic properties and their interpretation.

Here we focus on the corpus annotation of complex predicates. This paper is structured as follows. In section 2 we discuss related work on the annotation of CPs in other languages. In section 3 we present a typology of complex predicates. In section 4 we detail our the annotation schema and also focus on several specific cases of CPs and the annotation labels for these cases. In section 5 we give more information about the CINTIL corpus and in 6 we show the outcomes of the annotations and present statistics on the types of CPs that we found in the corpus. We conclude in section 7.

2 Related Work

For other languages, people have proposed different representations for CPs and for some languages there are corpora available enhanced with CP labeling. The Prague TreeBank for Czech, which is based on a dependency grammar, labels CPs explicitly. A complex predicate is represented

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 100–108,
Uppsala, Sweden, 15-16 July 2010.

by two nodes: the verb node is assigned a functor according to the function of the entire complex predicate in the sentence structure; the nominal node is assigned the CPHR functor, which signals that it is a part of a multi-word predicate, and is represented as an immediate daughter of the node for the verbal component (Mikulová et al., 2006; Cinková and Kolářová, 2005).

For German there is an example corpus annotated with verb phrases and light verbs (Fellbaum et al., 2006). However, only idiomatic expressions are labeled in this German corpus while we focus on non-idiomatic CPs. Calzolari et al. (2002) treat support verb constructions (verb+noun), and focus their attention, just like we did in our approach, on constructions where the subject of the verb is a participant in the event denoted by the noun. Their objective is however not corpus annotation, but the creation of a computational lexicon of MWEs with both syntactic and semantic information.

Also the field of semantic or thematic role labeling investigates constructions of verb+noun, but it focuses on predicate-argument structures in general, while we focus on a specific type of relations. FrameNet uses frame semantics theory to represent such predicate-argument structures which also includes handling complex predicates (e.g. (Johnson and Fillmore, 2000)). For German, there exists a fully annotated corpus with semantic frames (Erk et al., 2003). The basis of the Framenet semantic annotation are conceptual *frames* expressing an event or object and the semantic arguments (*frame elements*) that are (obligatory or optional) parts of the frames. They also specifically address support verbs and observe that support verbs often occur with nouns expressing an event (Johansson and Nugues, 2006). In a Framenet semantic annotation, support verbs are not considered as parts of frames or as part of the frame elements, they are annotated with a specific 'support verb' label. We, on the contrary, view CP as one semantic and syntactic unit.

In Nombank, a distinction is made between idioms (which in principle are not marked) and light verb plus noun combinations, which are to be annotated, and criteria are given to make such distinction (English (Meyers, 2007), Chinese (Xue, 2006)). In (1) we show a NomBank annotation example of the sentence with a complex predicate.

Usually, CPs of the type verb+verb are treated as infinitive dependent clauses and are not annotated as CPs (cf. the Penn Treebank (Marcus et al., 1993) and the Portuguese treebank CordialSIN (Carrilho and Magro, 2009)).

(1) 'The campaign takes advantage of the eye-catching photography.'
SUPPORT = takes
REL = advantage
ARG0 = the campaign
ARG1 = of the eye-catching photography

3 Typology of complex predicates

We consider CPs as constructions sharing certain properties defined in Butt (1995). A complex predicate has: a multi-headed and complex argument structure; more than one lexical unit, each contributing part of the information normally associated with a single predicate and a grammatical functional structure equal to the one of a simple predicate. Several types of constructions are in accordance to this definition of CPs: (i) two main verbs, forming a restructuring construction, like *querer estudar* 'to want to study' (ii) two main verbs in a causative construction, like *fazer rir* 'to make laugh'; (iii) a light verb followed by a noun: *dar um passeio* 'to take a walk', *ter medo* 'to have fear'; (iv) a light verb followed by a secondary predicate: either an adjective, like *tornar a história credível* 'make the story believable', or a prepositional phrase, like *fazer x em pedaços* 'to make x into pieces'; (v) two concatenated verbs (serial verb constructions), like *O Pedro pegou e despediu-se* (lit: 'Pedro took and said goodbye'). This last construction is mostly restricted to the informal spoken register. Regarding constructions (i) and (ii) with two main verbs, it is generally assumed that these CPs include at least two verbs which behave as a single constituent under local phenomena such as Clitic Climbing or Long Object Movement (Kayne, 1975; Gonçalves, 2002; Gonçalves, 2003). Each one of the verbs preserves its own argument structure.

In the case of constructions (iii) involving a light verb and a noun derived from a verb, one of the most frequently referred property is the possibility of being paraphrased by the main verb from which the noun is derived (see example 2), although this is not a necessary condition.

(2) (a) *dar um contributo /contribuir*
'to give a contribution' / 'to contribute'
(b) *ter um desmaio / desmaiar*
'to have a blackout' / 'to faint'

Light verbs occurring in these constructions have a rather similar semantics across different languages and involve mostly verbs like *have*, *take* and *give* in English (Bowern, 2006) and *ter* 'to have', *dar* 'to give', *fazer* 'to make' in Portuguese. Furthermore, both the light verb and the derived noun contribute to predicate information and argument structure and theta-role assignment appear to be determined simultaneously by the two constituents. It is important to determine the exact nature of the semantic contribution of light verbs to the whole predicate and the similarities and differences between the light verb construction and its lexicalized verbal counterpart, if it exists.

4 Annotation system

The corpus annotation focused on four of the types of CPs listed in the previous section, excluding type (iv): constructions where a main verb is followed by a secondary predicate, due to time limitations. Constructions with a light verb (type (iii)) were consequently restricted to verb+noun. We only annotated constructions in which the subject of the CP controlled the event denotated by the noun. For example, constructions like *Mary gave a talk* where Mary is the one who is presenting, and not any other entity. We excluded cases where the subject does not seem to obligatorily control the event (e.g. *dar um título* 'to give a title').

We further restricted our annotation to a particular set of nouns:

- nouns derived from a verb, like *dar um passeio* 'to take a walk' (lit: 'to give a walk');
- nouns expressing an emotion, i.e., psych-nouns like *ter medo* 'to be afraid' (lit: 'to have fear');

Nouns derived from a verb are very common. For example, half of the nouns in the English Nombank corpus that have semantic frame elements are actually nominalizations from verbs as stated on the NomBank homepage[1].

The restrictions on the type of noun occurring in CPs lead to the exclusion of constructions with idiomatic meaning (like *dar a mão* 'to give a hand')[2].

The annotation guidelines follow the results of our study of CPs under a generative grammar framework, and are consequently theory-oriented. We didn't include for the moment semantic and aspectual information in our annotation of CPs. We have undertaken some work on the aspectual information conveyed by both light verb and noun and on the aspectual restrictions that hold between the two elements (Duarte et al. 2009) and we plan to latter partially integrate those findings in our annotation system.

We divided the annotation of the CPs in two main groups: verb+verb constructions (type (i), (ii), (v) as described in section 3) and verb+noun constructions (type (iii)). The verb+verb constructions are denoted with the tag [CV] and the noun+verb constructions with [CN]. Furthermore, inside the verb+verb category, we make distinctions between restructuring constructions (tagged as [CVR]), causative constructions ([CVC]) and constructions with coordinated verbs ([CVE]). Example 3 gives an illustration of each of these subtypes. For the verb+noun constructions we distinguish contexts with bare nouns ([CNB]) and contexts where a determiner precedes the noun (just tagged as [CN]) (cf.example 4).

(3) (a) porque nos [CVR]*queriam convidar*
because [they] us wanted to invite
'because they wanted to invite us'

(b) veio abalar estes alicerces espirituais
[CVC]*fazendo traduzir* ao rapaz
"Pucelle" de Voltaire
he shacked these spiritual foundations
by making translate to the boy
"Pucelle" by Voltaire
'he shacked these spiritual foundations
by making the boy translate "Pucelle"
by Voltaire'

(c) e [CVE]*vai* um *e conta* ao outro
and goes one and tells to the other
'and he tells the other'

(4) (a) Facto que leva a CGD a considerar que
não [CNB]*tem obrigações* em relação
aos trabalhadores.
'The fact that leads the CGD to believe
that it doesn't have obligations towards
the workers.'

(b) o erro de [CN]*fazer uma interpretação*
literal
'the error of making a literal
interpretation'

[1] http://nlp.cs.nyu.edu/ meyers/NomBank.html

[2] These are currently under study in the scope of a project on multi-word expressions in Portuguese.

There is also information on the typical position of the element inside the CP (position 1, 2, etc.), as well as on its contextual position in the corpus (B=Beginning, I=Intermediate, E=End). With typical position we refer to the ordering of elements of the CP in its canonical form, corresponding to the descriptions and examples given in section 3. The typical and contextual position can differ as is illustrated in example 5.

(5) depois de *um*[CN2_B] *aviso*[CN3_I]
dado[CN1_E]
'after a warning was given'

The elements forming the CP may not be contiguous and in that case only the elements pertaining to the CP are annotated. In example 6 the adverb *logo* 'immediately' is not a part of the CP and consequently is not annotated. Also, only the main verb is annotated and not the auxiliary verbs which might occur (cf. the auxiliary *tinha* 'had' is not tagged in 7).

(6) *dar*[CN1_B] logo *uma*[CN2_I]
ajuda[CN3_E]
give immediately an help
'give help immediately'

(7) tinha *dado*[CN1_B] *uma*[CN2_I]
ajuda[CN3_E]
had given an help
'had given help'

The categories and tags which compose our annotation system provide an overview of different contexts of CP constructions encountered in authentic data, which is a major goal of this annotation project.

The process of annotation was based on concordances extraction using lists of verbs entering restructuring constructions (type (i)), given in 8 and lists of causative verbs (type (ii)), shown in 9. Considering the large candidate list of possible CPs with light verbs, the annotation first focused on constructions with verbs *ter*, *dar* and *fazer* followed by a noun. For CPs with coordinated verbs (type (v)), a list of typical verbs entering the construction was elaborated, shown in 10, and applied to a search pattern (two verbs separated by a conjunction and possibly by some other lexical element). Concordances retrieved were then manually evaluated.

(8) *querer* 'want'
desejar 'desire'
costumar 'use to'
tentar 'try'
pretender 'want'
tencionar 'make plan to'
conseguir 'succeed'

(9) *mandar* 'order'
deixar 'let'
fazer 'make'

(10) *ir* 'go'
agarrar 'grab'
pegar 'hold'

Information on the categories, tags, restrictions and special cases (discussed in section 4.1) were described in the annotation guidelines.

4.1 Special cases

The observation of corpus data pointed to a range of specific situations requiring new categories and tags.

4.1.1 Ambiguity

Some contexts in the corpus are clearly cases of CPs and are straightforwardly annotated as CPs, like restructuring constructions with clitic climbing (cf. 3a) and causative constructions with two internal arguments like in example 3b. Also example 11 is a clear case where the subject of the lower verb occurs as an indirect object (*aos cidadãos em geral*) and the that-clause which is the direct object of the lower verb (*que a fotocópia corresponde a um acto de pirataria inaceitável*) is reanalyzed as the direct object of the CP. Other clear cases of CPs are pronominal passives where the direct object of the second verb occurs as subject of the higher verb (Long Object Movement), producing subject-verb agreement (this construction was not encountered in the corpus, a possible example would be (12)).

(11) *fazer perceber* aos cidadãos em geral, que a fotocópia corresponde a um acto de pirataria inaceitável
'make understand to all citizens that a photocopy corresponds to an act of unacceptable piratery'

(12) *Querem-se estudar* os problemas.
'want-3PL.PASS study the problems'

Other contexts are clearly not instances of CPs and as such are not annotated. This is the case of constructions with a restructuring verb without clitic climbing, as in example 13.

(13) *querem perpetuá* -lo
'[they] want to perpetuate it'

But many CPs can have an ambiguous interpretation between a complex predicate construction and a construction with a main verb and an embedded infinitive clause, and we found it relevant to mark those constructions with the information of ambiguity (tag [_VINF]). For example, contexts similar to (12) but with a singular NP, as in example 14a, can receive two possible structural interpretations: the NP *justiça* 'justice' can be interpreted as the subject of the higher verb (a long object movement construction and consequently a CP construction) or as the direct object of the second verb (an impersonal construction). In (14b) we show how we annotated this example using a label expressing the ambiguity.

(14) (a) Pretende-se cometer justiça.
Aims-IMP to commit justice [IMP = Impersonal]
'One wants to commit justice'
(b) Pretende[CVR_VINF1_B]-se
cometer[CVR_VINF2_E] (...) justiça

4.1.2 Overlapping CPs

Beside these examples, the corpus includes constructions in which one of the elements of a CP (restructuring type) is also part of another CP (causative type), so that two CPs are in fact superposed. In these cases, the element which is part of both CPs receives a double tag (see the verb *deixar* in example 15).

(15) não o queriam[CVR1_B]
deixar[CVR2_E][CVC_VINF1_B]
fugir[CVC_VINF2_E]
not him want to let escape
'they didn't want to let him escape'

4.2 Coordination inside CPs

There are also occurrences of coordination inside the CP, possible when two CPs share the same higher verb (light verb, restructuring or causative verb). The coordinated elements of the CP are tagged with extra information on their first or second position in the coordinated structure (tags [CVR2_1] and [CVR2_2], cf. 16). The coordination is usually marked with a conjunction, like in example 16 with a restructuring construction, equivalent in fact to two CPs *querer ouvir* and *querer registar*. However, in the spoken subpart of the corpus there may be no overt connector and just a slight pause as in example 17 (the pause is marked by "/").

(16) para quem o quis[CVR1_B]
ouvir[CVR2_1_E] e eventualmente
registar[CVR2_2_E]
to whom him wanted to listen and eventually register
'to whom wanted to listen and eventually register him'

(17) nós temos[CN1_B] uma[CN2_1_I]
tristeza[CN3_1_E] / uma[CN2_2_I]
frustração[CN3_2_E] muito grande
'we have a sadness / a frustration very deep'

5 Corpus constitution

The CINTIL corpus[3] contains 1 million tokens and was compiled using different existing resources developed at the Centre of Linguistics of the University of Lisbon (CLUL): the written corpus Parole (Bacelar do Nascimento et al., 1998), the spoken corpus C-ORAL-ROM (Bacelar do Nascimento et al., 2005) and new written texts from the Reference Corpus of Contemporary Portuguese-CRPC (Bacelar do Nascimento, 2000), a large monitor corpus with over 300M words. One third of the corpus is composed of transcribed spoken materials (both formal and informal) and the remaining two thirds are composed of written materials.

This corpus has been previously annotated and manually revised (Barreto et al., 2006), in a joint project of NLX-FCUL[4] and CLUL. The CINTIL corpus has important features, compared to other resources for Portuguese, namely the depth of its linguistic information, its size, range of domains and sources, and level of accuracy. The annotation comprises information on part-of-speech (POS), lemma and inflection, multi-word expressions pertaining to the class of adverbs and to the closed POS classes, and multi-word proper names (for

[3]The CINTIL corpus is available for online queries (//cintil.ul.pt) through the use of a concordancer adapted to Portuguese.

[4]http://nlx.di.fc.ul.pt

named entity recognition), together with specific categories for spoken texts (like Emphasis (/EMP), Extra-linguistic (/EL), Fragment (/FRG)). Below is an excerpt of the POS annotation and lemmatization where tags follow the order [lemma/ POS category # inflected features [named entity]].

(18) pretende/PRETENDER/vpi#3s[O]
reconverter/RECONVERTER/inf-nifl[O]
o/O/da#ms[O]
centro/CENTRO/cn#ms[B-LOC]
de/de/prep[I-LOC]
Matosinhos/MATOSINHOS/pnm[I-LOC]

In the next section we present the results of the addition of a new layer of information on complex predicates to this corpus.

6 Annotation results

The annotation of the whole corpus was done manually by one MA student who was well familiar with the task. A concordancer was used to identify possible complex predicate structures. Difficult cases were picked out and discussed with two other persons to reach an agreement on the annotation. Several of such hard cases were then added to the annotation guidelines. After manual annotation, the annotations were checked with a script to check the consistency of the labels and to correct some minor errors.

To validate the annotations we performed a small experiment. A second person annotated a small sample of sentences independently of the first annotator. Next we compute the inter-annotator agreement on the two different annotations. This gives us some indication of the difficulty of the task and the consistency of the labeling of the first annotator. We computed the kappa statistics (Cohen, 1960) on the complex predicates labeled by the two annotators in 50 sentences. We acknowledge that this is just a very small sample, yet this gave us a kappa value of .81 which indicates a high overlap between both annotations.

In Table 1 we list the frequencies of the complex predicates found in the CINTIL corpus. In total we found 1981 CPs, the majority (1292 CPs) are combinations of a verb with a noun. For the verb predicates the table clearly shows that these cases are mostly ambiguous. We also looked at the occurrences of the more complex events described in section 4.1 presented in table 2. We encountered 28 cases of coordinated complex predicates

label	written	spoken	total
CV total	470	219	689
CVR	34	47	81
CVC	13	3	16
CVE	0	1	1
CVR_VINF	300	143	443
CVC_VINF	123	25	148
CN total	706	586	1292
CNB	353	213	566
CN_	353	373	726
total	1176	805	1981

Table 1: Number of annotated complex predicates in the spoken and written parts of the CINTIL corpus.

label	written	spoken	total
CV ambiguity	423	168	591
coordination	15	13	28
overlap	6	10	16

Table 2: Zooming in on the frequencies of the special cases (sec. 4.1) in the CINTIL corpus.

and 14 times a verb was part of two different CPs at the same time. The CPs with verb+verb constructions show a very high number of ambiguous occurrences. It is clear that in most cases the context of such a construction does not provide sufficient evidence to disambiguate it. We only found a handful of cases in which the context did resolve the ambiguity.

We also looked into the ordering of the CPs in the corpus. To what extent do the CPs occur in their canonical form? Table 3 shows the results. We found a change in ordering only for the verb+noun CPs. For the CPs with a bare noun we found only 9 cases of non-canonical order. For CPs with an NP with a determiner-noun combination we did see more variation in order, of the total number of 726 occurrences, 16.9% had a different word order.

We also wanted to see if all the verbs used to identify CP constructions (verbs listed in 8 9, 10 plus the 3 light verbs) were equally present in the CINTIL corpus or if there was any significant lexical difference. We present the results of the frequencies of the verbs of each CP type in Tables 4, 5, 7 and 6. When comparing the list in

label	written	spoken	total	% of occ
CN	86	37	123	16.9
CNB	7	2	9	1.6

Table 3: Number of complex predicates that do not follow their canonical form. The last column presents the percentage of the total number of CN or CNB occurrences that are not in their canonical form.

8 with the verbs in Table 4, we can see that the verbs *desejar* and *tencionar* were included for the query of restructuring predicates but do not occur in the corpus in CP constructions. Out of the five verbs, *querer* 'want' is clearly the most frequent in both written and spoken sub-parts of the corpus. Apart from *conseguir* 'succeed', the rest of the verbs have very low frequencies, and *costumar* 'use to' is only present in the spoken corpus, while the opposite is true for *pretender* 'want', a verb associated to a more formal register. In causative constructions with CPs (Table 5), the verb *fazer* 'make' is clearly prominent in the written corpus, although it does not occur in the spoken one. The only causative verb in CP constructions in the spoken corpus is *mandar* 'order'. In causative constructions, contrary to restructuring ones, the genre seems to influence the lexical choice of the higher verb of the complex predicate.

CVR	written	spoken
conseguir	6	7
costumar	0	3
pretender	2	0
querer	25	34
tentar	1	3
total	34	47

Table 4: frequencies of the main verb in CVR complex predicates.

The verb+noun constructions are divided in two different tables, according to our categorization in bare nouns (Table 6) and nouns preceded by a determiner (Table 7). The same three verbs enter the constructions although their frequencies are different in the two different structures: the verb *fazer* is clearly dominant when followed by a noun preceded by a determiner, while the verb *ter* is the more frequent light verb with bare nouns.

CVC	written	spoken
deixar	1	0
fazer	11	0
mandar	1	3
total	13	3

Table 5: frequencies of the main verb in CVC complex predicates.

CNB	written	spoken
dar	69	27
fazer	87	52
ter	197	134
total	353	213

Table 6: frequencies of the main verb in CNB complex predicates

CN	written	spoken
dar	79	34
fazer	193	231
ter	81	108
total	353	373

Table 7: frequencies of the main verb in CN complex predicates.

7 Final remarks

We presented the annotation process of complex predicates in the CINTIL corpus. We first explained our theoretical framework and gave a broad typology of CPs. Next we detailed the annotation schema that we used and zoomed in on some difficult cases. We presented the outcomes of the annotation work. We gave a first broad statistical analysis of the annotations, and next we zoomed in on some insights in characteristics of CPs in Portuguese that this new annotation layer has offered. This new resource provides diversified authentic data that will enable a general overview of CP constructions and can shed new light on the Syntax-Semantics interface. It is also an important part for forthcoming tasks of syntactic and semantic corpus annotation.

In the future we plan to further analyze the results of the verb+verb types of CPs. The large

number of ambiguous cases and the few contexts which give us definite clues for categorizing the sequence as a CP challenges our concept of complex predicates. The causative and restructuring constructions require more attention and further study. As to the verb+noun constructions, we want to examine the contexts with and without determiner to see if the same CP can occur in both structures. We also want to look further into the high frequency of specific light verbs with bare nouns and the possible relationship with the semantics of the light verbs. In this study we restricted the annotation to a particular group of light verbs. In a next step we would like to look at a broader list to try to establish the necessary properties to categorize a verb as a light verb. We plan to address, for example, certain contexts of psych-nouns like *sentir medo* 'feel fear', *experienciar uma profunda emoção* 'experience a deep emotion', where the predicative nature of the verb is unclear. We also plan to enlarge our description and annotation of CPs to include idiomatic expressions with light verbs.

References

A. Abeillé, D. Godard, and I. Sag, 1998. *Complex Predicates in Nonderivational Syntax*, volume 30 of *Syntax and Semantics*, chapter Two Kinds of Composition in French Complex predicates. San Diego Academic Press, San Diego.

M. F. P. Bacelar do Nascimento, P. Marrafa, L.A.S. Pereira, R. Ribeiro, R. Veloso, and L. Wittmann. 1998. Le-parole - do corpus à modelização da informação lexical num sistema-multifunção. In *Actas do XIII Encontro da Associação Portuguesa de Linguística, APL*, pages 115–134, Lisboa.

M. F. Bacelar do Nascimento, J. Bettencourt Gonçalves, R. Veloso, S. Antunes, F. Barreto, and R. Amaro, 2005. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, chapter The Portuguese Corpus, pages 163–207. Amsterdam/Philadelphia: John Benjamins Publishing Company, Studies in Corpus Linguistics. Editors: E. Cresti and M. Monegnia.

M. F. Bacelar do Nascimento, 2000. *Corpus, Méthodologie et Applications Linguistiques*, chapter Corpus de Référence du Portugais Contemporain, pages 25–30. H. Champion et Presses Universitaires de Perpignan, Paris. Editor: M. Bilger.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. F. P. Bacelar do Nascimento, F. Nunes, and J. Silva. 2006. Open resources and tools for the shallow processing of portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy.

C. Bowern. 2006. Inter theorical approaches to complex verb constructions: position paper. In *The Eleventh Biennal Rice University Linguistics Symposium*.

E. Carrilho and C. Magro, 2009. *Syntactic Annotation System Manual of corpus CORDIAL-SIN*. http://www.clul.ul.pt/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html.

S. Cinková and V. Kolářová. 2005. Nouns as components of support verb constructions in the prague dependency treebank. In *Insight into Slovak and Czech Corpus Linguistics*. Veda Bratislava.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measuremen*, 20:37–46.

K. Erk, A. Kowalski, S. Padó, and M. Pinkal. 2003. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 537–544, Sapporo, Japan, July. Association for Computational Linguistics.

C. Fellbaum, A. Geyken, A. Herold, F. Koerner, and G. Neumann. 2006. Corpus-based studies of german idioms and light verbs. *International Journal of Lexicography*, 19(4):349–360.

A. Gonçalves. 2002. The causee in the faire-inf construction of portuguese. *Journal of Portuguese Linguistics*.

A. Gonçalves. 2003. Defectividade funcional e predicados complexos em estruturas de controlo do português. In I. Castro and I. Duarte, editors, *Miscelnea de estudos em homenagem a Maria Helena Mira Mateus*, volume I. Imprensa Nacional-Casa da Moeda.

J. Grimshaw. 1988. Light verbs and marking. *Linguistic Inquiry*, 19(2):205–232.

M. Gross. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63:7–52.

O. Jespersen. 1949. *A Modern English Grammar on Historical Principles*. Londres: George Allen & Unwin; Copenhaga: Ejnar Munksgaard.

R. Johansson and P. Nugues. 2006. Automatic annotation for all semantic layers in FrameNet. In *Proceedings of EACL-2006*, Trento, Italy, April 15-16.

C. R. Johnson and C. J. Fillmore. 2000. The framenet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, pages 56–62, Seattle WA.

R. Kayne. 1975. *French Syntax: the Transformational Cycle*. The MIT Press, Cambridge, Mass.

M. Marcus, S. Santorini, and M. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M.Butt. 1995. *The Structure of Complex Predicates in Urdu*. Stanford, CA: CSLI Publications.

A. Meyers. 2007. Annotation guidelines for nombank – noun argument structure for propbank. Technical report, New York University. http://nlp.cs.nyu.edu/meyers/nombank/nombank-specs-2007.pdf.

M. Mikulová, A. Bémová, J. Hajič, E. Hajicková, and J. Havelka et al. 2006. Annotation on the tectogrammatical level in the prague dependency treebank annotation manual. technical report. Technical Report UFAL CKL Technical Report TR-2006-35, ÚFAL MFF UK, Prague, Czech Rep.

N. Xue. 2006. Annotating the predicate-argument structure of chinese nominalizations. In *Proceedings of the LREC 2006*, pages 1382–1387, Genoa, Italy.