Noun Sense Tagging: Semantic Prototype Annotation of a Portuguese Treebank

Eckhard Bick

Institute of Language and Communication, University of Southern Denmark
eckhard.bick@mail.dk, Rugbjergvej 98, DK-8260 Viby J

1. Introduction

There is a tendency towards ever-deeper treebank-annotation. While early treebanks, like the first version of the PENN treebank, only specified part of speech (PoS) and constituent hierarchies (bracketing), most treebanks today add grammatical-functional information, encoded as edge labels or token-based function tags (subject, object etc.). In addition, some treebanks have added a semantic layer. Thus, the tectogrammatical layer of the Prague Dependency Treebank (PDT 2.0) introduces semantically motivated dependencies and semantically motivated syntactic functors (Hajicova 2000).

The focus of this paper will be on noun semantics in Portuguese, in particular the dual role of semantic noun prototypes as a prerequisite for better automatic parses on the one hand, and a basic step towards semanticized treebanking on the other hand. Once present, semantic prototype tags implicitly handle noun sense disambiguation, and provide valuable context for other semantic task such as anaphora and referent resolution (Vieira et al. 2006), discourse structure or semantic role labeling.

The task at hand was to add a semantic layer to an existing Portuguese treebank, the Floresta Sintá(c)tica (Afonso et al. 2002), preparing it for work on anaphora and discourse, and at the same time to enhance the quality of the PALAVRAS parser (Bick 1996) used for the ongoing annotation of further material for the treebank[1].

2. Semantic prototypes

The semantic prototype system used here was first suggested by Bick (2000)

[1]The so-called *Floresta Virgem* ("jungle"), a large, but as yet unrevised, treebank of Portuguese.

and contains roughly 160 prototypes, that in turn can be regarded as *bundles* of 16 atomic semantic features. For instance, the vehicle prototypes (<V>, <Vwater> and <Vair>) can be characterized, in terms of atomic features, such as +concrete entity, +moving, +movable, -animate, -human, -location, -temporal, +countable, -mass etc.

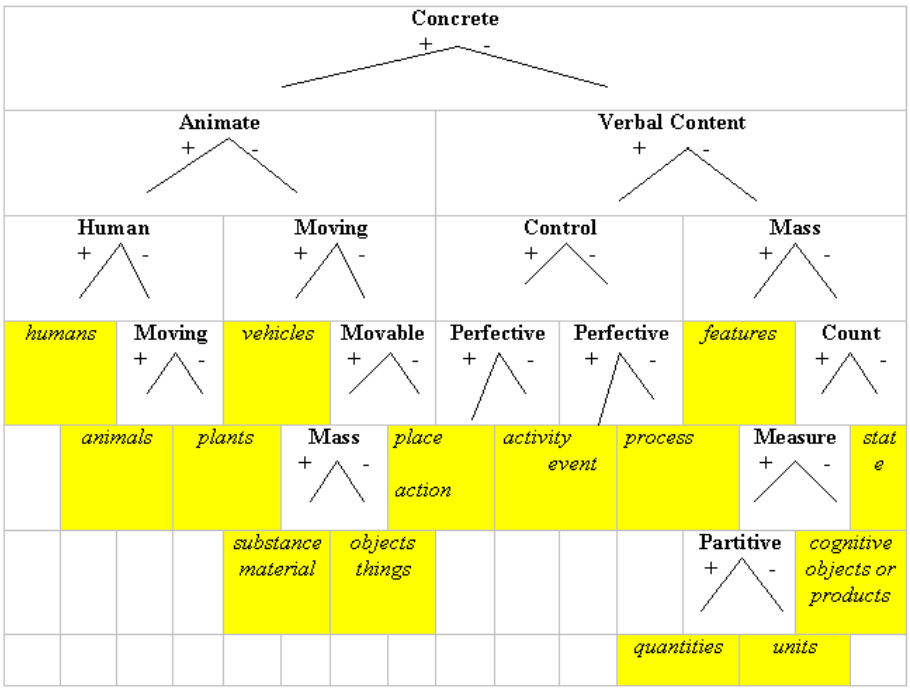

Figure 1: Atomic semantic features

The 160 protypes can be bundled into about 40 different feature-bundles. Some of the more common ones are presented in the figure below, in the order of columns (1) *human - professional / ideological etc., (2) acting places, (3) non-acting places, (4) concrete-countable objects, (5) concrete mass nouns, (6) vehicles - ground / air / water, (7) animals - zoological / ornithological etc., (8) semantic products - c=cognitive / w=watchable etc., (9) act nouns - including do-nouns and speechacts, (10) organized events/occasions, (11) processes.*

All nouns in the PALAVRAS lexicon carry at least one semantic prototype tag, information that has long been used to provide context for syntactic parsing rules. However, with the exception of a machine translation project (Bick 2000), no serious effort at disambiguating such tags had been made.

	Hprof Hnat Hfam Hideo ...	Lciv inst	Ltop Labs Lh ...	cc, cc-h tool clo ...	cm liq mat food ...	V, air wa-ter ...	A zo orn ich ent	sem -l -c -w -s	act -d -s	occ	pro cess
E (±Concrete)	+	+	+	+	+	+	+				
C (±Control)									+		
I (±Moving)	+					+	+				
J (±Movable)	+			+	+	+	+				
A (±Animate)	+						+				
H (±Human)	+	+									
M (±Mass)					+						
N (±Number)	+	+	+	+		+	+	+		+	
V (±Verbal)										+	+
P (±Perfective)									+	+	
S (±Measuring)											
D (±Partitive)											
X (±Human modifiable)	+	+						+	+	+	
F (±Feature)											
L (±Location)		+	+							+	
T (±Temporal)									+	+	+

Figure 2: Semantic prototype bundles

3. Semantic disambiguation

Using atomic semantic features allows the parser to do *feature inheritance reasoning:* On the one hand, the hierarchical structure and natural interdependence of feature, may allow rules to conclude that if sentence context guarantees the +hum feature, +anim and +entity will automatically ensue, or that moving things automatically also are movable, but not vice versa. On the other hand, bundles of semantic prototypes may be discarded in semantic disambiguation, if just one bundle-associated atomic feature can be ruled out, or if a forbidden feature can be proven present by contextual rules.

Typical rules are unification rules for selection restrictions (a, b), valency instantiation rules (c) and head-dependent association rules (d). As shown by the simplified examples below, all 4 rule types rely heavily on relational-structural and syntactic-functional information typical of treebanks

and deep parsers, and not provided by a simple PoS tagger.

(a) **subject-verb unification:** REMOVE (@=i) (0 @=I LINK 0 @¤SUBJ>) (*1 @MV LINK 0 V-MOVE) ;

(Remove -MOVE (i) for potentially +MOVE (I) words, if they function as subjects of a main verb (@MV to the right (*1) that is a move-verb *(go, run, swim* etc.)

(b) **passive agent selection restriction:** REMOVE (@=h) (0 @=H LINK 0 @P<) (*-1 PRP-POR BARRIER NON-PRE-N LINK 0 @A<PASS LINK -1 PCP LINK 0 V-HUM OR V-SPEAK LINK NOT 0 (<por^vtp>)) ;

(Remove -HUM (h) for potentially +HUM arguments (@P<) of the preposition *'por'* - if the latter functions as a passive agent dependent (@A<ACC) of the past participle (PCP) of a verb with a selection restriction for human agent subject (V-HUM).

(c) **valency instantiation:** REMOVE (@=E) (0 @=e LINK 0 (<+de+INF>) OR (<+de+que>)) (*1 PRP-DE BARRIER @NON-N< LINK *1 INF OR ("que" KS) BARRIER NON-ADV) ;

(Remove +ENTITY for a potentially abstract (e, -ENTITY) word, if it has the valency to bind finite or non-finite de-clauses (<+de+que>, <+de+INF>), and if this valency potential can, indeed, been instantiated by there being a *'de'* with an infinitive or *que*-conjunction to the right.

(d) **head-dependent relation:** REMOVE (@=h) (0 @=H) (*1 @N< BARRIER @NON-N< LINK 0 <jh> LINK NOT 0 <jn>) ;

(Remove -HUM (h) for potentially +HUM nouns, if there is a postnominal (@N<) with the adjective-semantic class of "humanoid" <jh>, with a check for simultaneous "non-humanoid" potential <jn>).

The same kind of rules can, of course, be crafted directly for specific prototypes, rather than their atomic features. Thus, nominal valency for specific prepositions/pp's is used to select specific prototype readings for specific words. Valency for *'sobre' (about),* for instance, favours - for the word *painel (panel)* the occasion prototype <occ> (here: panel discussion) over the plate prototype <cc-board> (here e.g.: solar panel). In other cases, valency instantiation can even be generalised over a whole prototype class or even prorotype bundle. For instance, it is a fairly safe assumption that

sobre/about context should favour semantic products in general (<sem-c> plans/ideas, <sem-w> "watchables", <sem-l> "listenables" etc.). Another, unification based, example are deverbal nouns of the action and activity prototypes (both +CONTROL), that should be selected if the word occurs as object of a verbs like *cancelar (cancel), projectar (plan)* or *sugerir (suggest).*

Our system also exploits context features that only make sense for the more specific semantic prototypes, not the more general atomic semantic features, the most extreme case being the use of domain and usage markers. Here, the rules will chose a given prototype reading for a polysemouns noun simply because it occurs in a sentence (or with more stringent rules, clause) with one or more other words with matching prototypes, or - alternatively, for other word classes - with matching domain tags from the lexicon. For instance, the prototypes musical instrument <tool-mus>, "listenable" work of art <sem-l> and dance <dance> will so to say strengthen each other, as will "musical" words like "symphony", "concert", "C-flat" etc., that are marked as <D:mu> by the parser's lexicon[2]. For *RARE*-marked readings, such matching can even be performed in a negative way, at the heuristic level, discarding prototypes that are not enhanced by kindred context. Similar context-dependent strengthening or weakening is also applied to noun senses only present in either Brazilian (<U:B>) or Lusitan/European Portuguese (<U:L>), and to readings marked as slang usage (<U:S> or <U:SF>), and the latter will be discarded at the heuristic level, if no context or other rule-based support is found.

Finally, local, morphology-based sense disambiguation is also used, selecting or discarding prototype readings dependent on singular or plural inflexion. For inflexionally ambiguous words, the number feature will be "propagated" from possible singular or plural modifiers.

The growing semantic disambiguation grammar of PALAVRAS now contains about 3000 CG rules - 200 rules targeting atomic semantic features, 750 feature inheritance rules and 2015 prototype targeting rules. It must be born in mind that many of these rules target *sets* of tags or words rather than *individual* tags or words. For instance, the sets targeted by the 15 rules exploiting prepositional valency, and the 95 rules checking domain context (i.e. salient occurrence of words typical of e.g. finance or agriculture), contain 2275 prototype-lexeme tag combinations as set elements, about 50 for each of the rules, allowing the lumping together of what would otherwise have been 50 * (15+95) = 5500 individual rules.

---

[2] Domain sets can also be appended - or even defined from scratch - in the grammar.

4. Enhancing the treebank

The Floresta Sintá(c)tica comes in 2 basic formats - Constituent Grammar and Constraint Grammar, the former being bracketing based, the latter dependency based, originally with flat dependency (dependency direction markers), but with full numbered dependencies currently being added to the data. In both formats, each token carries the following tag types: Word form, lexeme (base form), PoS, inflexion, syntactic function. In addition, a small number of secondary tags have been maintained from the parsing stage[3], mainly concerning semantically motivated PoS subclasses (e.g. <rel> relative, <interr> interrogative, <quant> quantifier). The new semantic annotation of nouns is expressed through such secondary tags, rather than introducing new word senses at the lexeme level. Thus, maintaining lexeme-unity for senses with the same gender and inflexional paradigm, we add a semantic prototype tag instead. The basic polysemy of the Portuguese word *papel,* for instance, can be captured with the following prototypes:

papel (noun, male gender)

(1) <mat> (material) *'paper'*
(2) <cc-r> (read-object) *'piece of paper, document'*
(3) <ac> (abstract countable) *'role'*

(1a) Os argelinos, que desempenham um **papel** preponderante na equipa de Josip Skoblar, chegaram anteontem ...
(2a).. duas cadeiras , uma secretária , uma empregada e **papéis** A4 em a parede , escritos a a mão , com os preços ...
(2b) Com muitos dos **papéis** com a cotação interrompida em consequência do período de pagamento de dividendos, ...
(3a) ... remeter envelopes com folhas de **papel** higiénico ao ministério

The three prototype tags used do not define the different senses glossed by the English translations, but they do *distinguish* between them, so given a (fixed) dictionary of senses, turning semantic prototype tagging into noun sense annotation can be regarded as a mere lookup-procedure. Note that a single sense may have more than one translation in another language, such as the 'read-object' sense (2), which apart from the 'piece-of-paper' meaning (2a)

[3] At the parsing stage, hundreds of secondary tags are used internally by PALAVRAS, covering, for instance, verbal and nominal valency potential.

also covers both ID papers *(papéis de identificação)* and stock market shares (2b). Conversely, the English 'essay'-meaning of 'paper', has no equivalent in Portuguese.

One way to capture such subdistinctions is to introduce a second, orthogonal axis of disambiguation, such as domain or usage. For instance, the parser uses a <D:fin> tag for the financial domain that can be used as context to establish the 'share'-reading of *papel.* Likewise, information on etymology, regionalisms and sociolect are used to weed out contextually rare meanings.

Below, an example is given of a sentence in dependency treebank format with semantic prototypes added in bold face (e.g. *<Hnat> Human national*, *<HH> Human group / organisation).* PoS and inflexion tags are in upper case, the @-sign marks syntactic function tags, and numbered dependency links are prefixed by the #-sign.

```
Os          [o] <artd> ART M P @>N                  #1->2          The
argelinos   [argelino] <Hnat> N M P @SUBJ>          #2->15         Algerians
$,                                                  #3->0
que         [que] <rel> INDP M P @SUBJ>             #4->5          who
desempenham [des...nhar] <mv> V PR 3P IND @FS-N<PRED #5->2 play
um          [um] <arti> ART M S @>N                 #6->7          a
papel       [papel] <ac> N M S @<ACC                #7->5          role
preponderante    [preponderante] ADJ M S @N<        #8->7          prominent
em          [em] <sam-> PRP  @<ADVL                 #9->5   i      n
a           [o] <artd> ART @>N                      #10->11        the
equipa      [equipa] <HH> N F S @P<                 #11->9  t      eam
de          [de] PRP @N<                            #12->11        of
Josip=Skoblar [Josip=Skoblar] <hum> PROP M S @P< #13->12  Josip. Sk.
$,                                                  #14->0
chegaram    [chegar] <mv> V PS 3P IND @STA          #15->0         arrived
anteontem   [anteontem] ADV @<ADVL #16->15  the day before yesterday
a           [a] <sam-> PRP  @A<                     #17->16        at
a           [o] <artd> ART @>N                      #18->19        -
noite       [noite] <temp> N F S @P<                #19->17        night
a           [a] PRP @<ADVS                          #20->15        in
Famalicão [Famalicão] <civ> PROP M/F S @P< #21->20          Famalicão
```

We developed the semantic tagging and disambiguation as a separate module

of the PALAVRAS parser, so in principle we expected it to run autonomously on the (morphosyntactically revised) treebank. However, some differences in tokenization and lemmatization caused semantic annotation failures for some words. For instance, a polylexical like *'cabeça de cartaz' (top-of-the-list, star)* was not recognized as one token in the treebank, so the semantic class for the polylexical *(<ac> - abstract countable)* was not assigned, and the erroneous *<anmov> (movable part of anatomy)* was used on the isolated *cabeça (head)* instead. Differences between Brazilian and European Portuguese were also a problem, since PALAVRAS' first, morphological, stage normalizes European forms into Brazilian ones to find their semantic types, before re-europeanizing them in the final output. Bypassing morphological analysis by inputting analyzed treebank data forced us to handle orthographical differences in more heuristic ways. Not solved in the current project was the treebank treatment of quotes as tokens, rather than as tags on the token quoted, as internally done by the parser. Therefore, som semantic CG adjacency rules were blocked by the quote tokens.

5. Quantifying noun sense ambiguity

The parser lexicon of PALAVRAS contains currently 36,771 nouns, with 43514 prototype bases sense distinctions, yielding a ratio of 1.183 readings per type. However, since frequent words tend to be more polysemous than rare words, and since "lexicon polysemy" is no measure for the amount of lecxical ambiguity in running text, token ambiguity can be expected to be considerably higher. In fact, in the treebank every second noun was polysemous, with an average of over 1.5 readings per noun:

Noun sense ambiguity ratio per token in F.S.	***Público (Portugal)***	***Folha de São Paulo (Brazil)***
lexicon-mapped onto revised treebank	1.57 (38052 : 24285)	1.50 (24238 : 16235)
treebank after automatic disambiguation	1.11 (26875 : 24285)	1.05 (17053 : 16235)
in live re-ananlysis of treebank texts	1.09 (26 467: 24277)	1.03 (16757 : 16228)

Table 1: Semantic ambiguity ratios (Readings : Tokens)

After automatic disambiguation, the ratio dropped to 1.05 for the Brazilian part, corresponding to a 90% reduction in ambiguity, and a similar disambiguation gain was achieved in a complete rerun on the unannotated text source of the treebank. By contrast, the disambiguation gain on the Portuguese part was somewhat less, most markedly for sense disambiguation of the revised treebank, with 11% remaining ambiguity. This figure can't be explained with the higher base ambiguity of the Porgutuese section alone (1.57), and is more likely to be caused by the fact that different versions of PALAVRAS were used in different periods of the treebank project (which has been ongoing since 2000), one of the relevant differences being the considerably lower number of noun tags in the revised Portuguese treebank (732, or 3% less than in the rerun), caused by linguistically motivated changes in the treatment of np-heading adjectives as either ADJ or N, and a rising lexicon coverage for the frequent (and productive) monosemous '-ista' nouns which may also function as adjectives (e.g. *comunista - communist*).

6. Name semantics

About 4% of the tokens in running newspaper text, such as used for the Floresta Sintá(c)tica treebank, consists of names, and semantic marking of nouns should therefore be matched by a corresponding marking of names, and since PALAVRAS-NER had recently won an open competition in named entity recognition, HAREM (organized by Linguateca), using the system on the treebank appeared to be a good idea.

However, the mark-up of names is a 2-stage process, with *identification* of names on the on side, and *classification* on the other, and since the former implies multi-word tokenization, using the parsers NER module on a ready-tokenized treebank was not straightforward. Thus, in its identification module, PALAVRAS-NER uses CG rules to mark tokens as 1. part or 2./later part of a name, while at the same time using semantic prototype information from potential noun parts to add preliminary semantic class markers to 1. parts. Unable to use this technique on the ready-revised name tokens of the treebank, the second, classification module of PALAVRAS-NER would lack the part-derived semantic information, having to depend only on context on its name lexicon.

Our solution was to perform 2 analyses in parallel, (a) running the second part of PALAVRAS-NER on the annotated treebank, and (b) running

the full PALAVRAS-NER on a raw-text version of the treebank. An arbiter program then inserted semantic name categories from (b) instead of (a), whereever a name-token match could be established in a sentence with the same ID. Inspection of the arbiter's results showed that in conflicting cases the classification with the full system (b) was almost always the correct one. A full quantitative evaluation showed that tokenization recall was 78-80% . Of these cases, raw-text classification overrides occurred in about a third of cases.

Name tokens	***Público (Portugal)***	***Folha de São Paulo (Brazil)***
in treebank	7,153	4,667
in live re-ananlysis	6877	4870
identical tokens	5580	3748
token "recall/precision"	78.0% - 81.1%	80.3% - 76.7%
semantic type arbitered	1920 (34.4 %)	1184 (31.6 %)

Tabel 2: Name token recall

About 40 semantic name categories were used, falling in 7 major categories: <hum> (person), <org> (organisation), <top> (place), <occ> (event), <tit> (work of art), <brand> (brand names), <common> (other object names). Many (sub)categories have a direct equivalent in the noun semantics scheme, such as <hum> (H in the noun scheme), <civ> (Lciv), <org> (HH), <top> (Ltop), <genre/school> (domain/genre) etc. However, a complete match would mean a complete overhaul of the parser, conflicting with the independent history and optimisation of the noun and name systems[4].

The HAREM category scheme also called for a mark-up of time and value expressions, turning numbers, units and date chains into names, while neither the treebank nor the parser as such use multi-word tokenization and the PROP word class in these cases. As a compromise, standard word class (NUM, ADV etc.) was maintained for non-upper case words, but the actual

[4] Application programs can work around this lack in symmetry by using filtering into supercategories on both sides, or by many-to-one mapping in one or the other direction.

semantic category, once established, was added to the head token of what would have been a name in HAREM.

7. Evaluation

Though polysemy resolutions for many nouns (in particular, from the lexicon secion S-Z) still rely on a few generalized rules rather than more specific lexeme-driven rules (for at least the more frequent nouns), a quantitative evaluation of treebank prototype tagging was performed on a 4,300 token chunk from the European Portuguese section of the treebank. Because the annotation system as such is still in flux and expected to change, a simple inspection method was used in stead of creating a formal gold standard version of the text, and results are therefore likely to be (slightly) biased by parser suggestions. Still, the figures provide a fair idea of how large a portion of suggested tags was compatible with human judgement. During the evaluation, 6 inflexion/lemma errors and 9 N/PROP categorization problems were found in the treebank section, and since these prevented the semantic parser to add tags, they were disregarded in the statistics below.

	Recall	***Precision***
Nouns (N) 744	93.1 %	87.8 %
Names (PROP) 240	88.8 %	88.8 %

Table 3: Semantic tagging performance

The data shows that nouns and names had similar F-Scores (90.5% for the former, 88.8% for the latter), but semantic noun tagging had a higher recall. The named entity classifier assigned exactly one reading, allowing running text readings to override treebank mappings where token identity could be established (85.8% of unambiguous treebank names). Of these, overrides occured in 33% of cases (28.6 correct, 4.4 incorrect).

8. Outlook

Nouns and names are, of course, only part of a full semantic mark-up, and other word classes should be treated later, as well as semantic relations between them (referents, anaphora, semantic roles etc.). We already

mentioned that the adjective lexicon of the parser specifies modifier preference for human or non-human heads. Likewise, verbs are categorized with respect to ±HUM subjects. However, preference is not absolute, it may be overridden in metaphorical usage, and it may not be (statistically) strong enough for a lexicon entry. Thus, about 30% of the 14000 adjectives in the lexicon were left with ambiguous ±HUM preference. The challenge then is to disambiguate such ambiguity, mark metaphorical usage, and exploit valency and dependency links to handle sense distinctions for verb and adjectives, too. Portuguese-Danish machine translation experiments (Bick 2000), using valency instantiation as a discriminator for translation equivalents, suggest a certain potential for such syntax-/dependency-based polysemi resolution, and we hope that the present work on the semantic tagging of nouns and names will provide useful contextual stepping stones for work on other word classes.

References

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos (2002). Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of LREC'2002, Las Palmas*. pp. 1698-1703, Paris: ELRA

Bick, Eckhard (1996). Automatic Parsing of Portuguese. In García, Laura Sánchez (ed.), *Anais / II Encontro para o Processamento Computacional de Português Escrito e Falado*. Curitiba: CEFET-PR.

Bick, Eckhard (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. Aarhus: Aarhus University Press

Bick, Eckhard (2006). Functional Aspects in Portuguese NER. In: Nuno J. Mamede et.al. (eds.) *Computational Processing of the Portuguese Language* (Proceedings of PROPOR 2006, Itatiaia, May 15th-17th, 2006), pp. 80-89. Springer

Hajicova, E. (2000). Dependency-Based Underlying-Structure Tagging of a Very Large Czech Corpus. In *T.A.L.*, vol. 41, n.1, pp. 47-66, 2000

Vieira, R., E. Bick, J. Coelho, V. Muller, S. Collovini, J. Souza & L. Rino (2006). Semantic Tagging for Resolution of Indirect Anaphora, In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (Sidney, July 15-16, 2006)*.