Automatic Semantic-Role Annotation for Portuguese

Eckhard Bick

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

Abstract. *The paper presents and evaluates a parsing system for the automatic annotation of Porguguese text with semantic role tags. All in all, 38 different categories, like agent, patient, location etc. are distinguished. The annotater uses a grammar of 500 hand-written Constraint Grammar rules and exploits syntactic dependency links as well as semantic prototype classes and syntactic function. The interaction with lexical information like verb-argument frames and the ensuing boot-strapping problem are discussed. The system achieved promising results in an evaluation with treebank input from the Floresta Sintá(c)tica, with an average recall of 86.6% and a precision of 90.5%.*

1. Introduction

Assigning semantic roles to the arguments of a verb (or to the arguments of a proposition in general) is an obvious way of adding deep, semantic structure to the syntactic analysis of a sentence, allowing different surface-syntactic functions, like subject and object, to "code" for one and the same semantic role, e.g. agent or patient depending on the arguments-slots of the governing verb and whether the syntactic structure is active, passive, reflexive or attributive. The idea of semantic roles has a long linguistic history, originally inspired by the concept of case roles (Fillmore 1968), later to be termed thematic or theta roles in Government & Binding theory (e.g. Jackendoff 1982). At the highest level of abstraction, universal semantic functors are postulated (Foley & van Valin 1984, Dowty 1987), obeying general laws of argument precedence based on features like control and animateness.

Within automatic annotation of running Portuguese text, robust results can be achieved for both morphology, syntactic function and even structural relations such as dependency relations. In the semantic area, notable progress has been made in the field of named entity classification (cp. the HAREM competition, Santos & Cardoso 2007), but no automatic systems have yet been published that would assign semantic functions to *all* types of arguments/constituents in a systematic way. In this paper, we will present and evaluate a method to add semantic roles to Portuguese sentence analyses at the treebank level. Input to the semantic role annotator is provided by the PALAVRAS parser (Bick 2000), using recent extensions regarding dependency and semantic prototype annotation (Bick 2006).

The semantic role annotator presented in this paper can be seen as an independent add-on module to PALAVRAS, but since the system uses the same rule-based methodology, likewise adhering to the Constraint Grammar (CG) paradigm (Karlsson 1995), and maintaining the same token-based annotation style (cp. appendix), it can maximally exploit the descriptive conventions and tagging information contained in its input.

2. Semantic roles

Semantic roles are different from semantic prototypes in that the latter bundle stable lexical features of a given lexeme, while the former are functional can only be assigned *in context.* Semantic prototype ambiguity does exist, of course, as words may be polysemic, and an automatic parser will use context to resolve such ambiguity, but even one and the same semantic prototype, with one and the same syntactic function, say subject, may fulfill a number of different semantic roles. Thus, the *civitas* prototype (towns, countries etc.) may fulfill the thematic roles of location, origin or destination of movement, but also non-place roles like that of human agent or patient. Here, rather than hypothesise different senses or lexical types for these cases, a role annotation level can help to build a bridge between syntax and semantics. In our approach, methodology closely matches descriptive issues, by letting the argument slots of predicators (mostly verbs) project a certain semantic interpretation (the roles) onto the slot-fillers (mostly nouns or noun phrases).

We use a set of about 35 semantic roles covering the major categories of the tectogrammatical annotation layer of the Prague Dependency Treebank (Hajicova et al. 2000), as well as those of the Spanish 3LB-LEX/3LB-SEM project (Taulé et al. 2005). In contrast to the latter, no separate level of argument hierarchy (ARG0, ARG1, ARG2, ARG-M) is used, mainly because the combination of syntactic function and semantic role tags allows the later addition of ARG-attributes on demand, without loss of information.

Role	***definition***	***example***
§AG	agent	*alg.* come ac.
§PAT	patient	alg. come *ac,* *X* caiu, *X* foi PAS
§BEN	benefactive	dar ac. *a alg.*
§EXP	experiencer	*X* teme ac.,
§TH	theme	ver *ac., X* está doente, Y surpreende *X*
§RES	result	produzir *ac.*
§ROLE	role	Y trabalha *como guía*
§COM	co-argument	concorrer *com alg.,* reunir-*se*
§REFL	reflexive	manifestar-*se*
§MED	medial	derubam-**se** casas
§ATR	static attribute	X está *doente,* um anel *de ouro*
§ATR-RES	resulting attribute	tornar alg. *nervoso*
§POS	possessor	o carro *do pai, X* possui Y
§CONT	content	uma garrafa *de vinho*
§ID	identity	a cidade *de Itatiaia,* a empresa *NN*
§VOC	vocative	tranqüilo, *João!*
§LOC	location	morar *em X, aqui, onde ...*
§ORI	origin, source	fugir *de X,* carne *da Argentina*
§DES	destination	mandar *para X,* um vôo *para X*
§PATH	path	*ao longo de X*
§EXT	extension, amount	marchar *7 kilômetros,* pesar *70kg*
§LOC-TMP	temporal location	*em 2007, hoje, antes de X, há1ano*
§ORI-TMP	temporal origin	*desde janeiro*
§DES-TMP	temporal destination	*até domingo*
§EXT-TMP	temporal extension	*mais duas semanas*
§FREQ	frequency	*de vez em quando, 10 vezes*
§CAU	cause	*porque ..., a causa de X*
§COMP	comparation	melho *do que nunca*
§CONC	concession	*embora ...*
§COND	condition	*se ..., nesse caso*
§EFF	effect, consequence	foram tantos *que ....*

Role	***definition***	***example***
§FIN	purpose, intention	*para* se *instalar* em, destinado *a X*
§INS	instrument	governar *por,* pagar *em,* cortar *com*
§MNR	manner	*desta maneira, -mente* (most)
§COM-ADV	accompanier	*junto com, com X na mão*
§META	meta adverbial	*segundo X, talvez, obviamente*
§FOC	focalizer	*só, também*
§ADV	dummy adverbial	many gerund clauses: *admitindo ...*
§EV	event, act, process	permitir/iniciar *ac., X* termina/começa
§PRED	(top) predicatior	main verb in main clause
§DENOM	denomination	lists, headlines
§INC	verb-incorporated	ter *lugar*

Table 1: Semantic role categories

The notational convention is that semantic role tags are assigned to tokens, Constraint Grammar-style, alongside syntactic and dependency tags. Complex (non-terminal) consituents are marked on the semantic head, i.e. the noun in an np, or the dependent of a preposition. As the latter case (pp's) indicates, the semantic head is not necessarily the syntactic head, so a constituent's dependency-function and its role function may reside on different tokens. The semantic role of a subclause is marked on its main verb. The current system annotates all clause-level arguments, but so far restricts the role annotation of phrase-level arguments to certain cases, such as de-verbal nouns, appositions and valency-bound dependents.

3. The Grammar and lexicon of the system

In order to contextually map semantic roles onto tokens, we have developed a Constraint Grammar of about 500 mapping rules and a small number of disambiguation rules. Rule (a) in the example below maps an agent role (§AG) onto right-pointing subjects (@SUBJ>) if they are human (0 HUM) and followed - without interfering clause boundary (BARRIER CLB) - by a mainverb (@MV) that is a movement verb (MOVE/TR) or belongs to a transitive valency class (VT-ALL). The rule has one exception context: the verb chain in question must not be in the passive voice (PAS).

```
(a) MAP (§AG)
        TARGET @SUBJ>
        (0 HUM)
        (*1 @MV BARRIER CLB
        LINK 0 VT-ALL OR V-MOVE/TR
        LINK NOT 0 PAS) ;

(b) MAP (§FIN) TARGET @P<
        (0 N-DEVERBAL)
        (NOT 0 N/PROP-LOC OR N/PROP-HUM)
        (*-1 PRP LINK 0 PRP-PARA) ;
```

Rule (b) maps the purpose role (finality, §FIN) onto arguments of the preposition *'para',* if the argument (@P<) in question is a deverbal noun and not ambiguously carrying a human or place prototype tag, in which case the destination or benefactive roles would have been more appropriate. Note that the NOT condition for semantic prototypes here is a safety measure, since PALAVRAS should, in a best case scenario, already have disambiguated semantic prototypes in the input.

A pivotal issue in the assignment of semantic roles is, of course, the lexicon serving the annotation stage as such. On the one hand, role-annotated text will allow to construct an inventory of Portuguese verb-sense argument frames in the style of the PropBank (Palmer et al. 2005), on the other hand exactly this kind of lexical information is necessary to assign semantic roles in the first place. In the absence of a hand-annotated gold ressource, we opted for a boot-strapping solution, where we did not manually construct full verb frames, but exploited ordinary (syntactic) valency information - such as <vdt> (ditransitive verb), <ve> (ergative verb), <por^vp> (prepositional valency with por), <vk> (copula-verb), <vta> (transobjective adverbial valency) etc. - to construct so-called *CG sets* of verb lexemes that typically *allow* (not *demand)* a given role as subject, object, complement or prepositional argument. All in all, about 80 sets with 1100 verb lexemes were defined in this way - moving part of the lexical information into the grammar.

One such set (VP-EM-TH) contains verbs allowing the theme role (§TH) with the preposition *'em'* - crer em, votar em, enganhar-se em, gastar ac. em. The set is then used in rules like the following, loking for a prepositional object (@PIV) referring to a VP-EM-TH class main verb:

```
(c) MAP (§TH) TARGET @P< OR @ICL-P<
        (*-1 PRP LINK 0 PRP-EM LINK 0 @<PIV
        LINK *-1 @MV LINK 0 VP-EM-TH) ;
```

4. A new CG formalism for dependency references

It is an obvious efficiency problem for rules like the above that head-daughter dependency relations have to be expressed in indirect - in fact topological - terms, using lots of unbounded (*1) and LINKed contexts with safety BARRIERs to locate the head of a dependency arc somewhere far away in the sentence. Furthermore, all head searches have to be conducted both left and right to cover arguments both left and right of their head, basically doubling the number of certain rules. CG's traditional topological coinage, a strength robust disambiguation, becomes, alas, a disadvantage where exact syntactic relations are, in fact, already known and *could* be exploited. Our research group has tried to remedy this problem by designing a new CG rule compiler[1] allowing direct context reference to dependency links. At present, three types are used: (p) parent, (c) child and (s) sibling, all allowing the usual combination with NOT and C (safe) operators. Rule (d) exploits the new formalism to assign the theme role (§TH) to direct objects (@ACC), if it has a dative/indirect object (@DAT) ambong its siblings, while rule (e) selects an agent role (§AG) for a subject with a speech-verb as parent.

```
(d) MAP (§TH) TARGET @ACC (s @DAT) ;
(e) SELECT (§AG) (0 @SUBJ) (p V-SPEAK LINK NOT 0 PAS) ;
```

The current system is hybrid, running the existing rule body in the old formalism *after* a preprocessing stage using new, dependency-based rules. Currently, this first stage is used to handle valency instantiation (concerning tags like the above-mentioned <vdt> or <ve>) and propagation of function labels from first to following conjuncts, as well as propagation of semantic prototype labels from nouns to pronouns and relative clauses.

Of course, dependency contexts will only work, if parsing input contains reliable dependency links. In the case of PALAVRAS, these links are given in a special field of the type *#id->head,* added by a special attachment rules which are not themselves formulated in the CG formalism,

[1] The new compiler formalism also supports the use of regular expressions, probabilistic or other "mathematical value" conditions, dynamic meta variables, as well as reference and linkage to non-local sentences and flexible rule section administration.

but exploit CG function tags, as well as special tags from a CG attachment module handling otherwise underspecified coordination close/long attachment.

5. Evaluation

The semantic role annotator was evaluated on input from the European Portuguese part of the *Floresta Sintá(c)tica* treebank (Afonso et al. 2002). For the test run, a section of 2.500 running words was used. Prior to semantic role annotation, the following steps were performed, using different PALAVRAS modules: (1) dependency conversion, (2) semantic prototype tagging, (3) named entity classification. The advantage of being able to rely on manually revised function labels and phrase structure (as input to 1-2-3) was deemed to outweigh the problem of a certain amount of tokenization incompatibility between the treebank and live PALAVRAS runs. To facilitate evaluation, no role ambiguity was allowed.

All in all, the annotator module assigned 884 semantic role labels, of which 84 were wrong. In 38 cases, the label was missing altogether, corresponding to a recall of 86.8 %, a precision of 90.5 % and an F-score of 88.6. In order to identify problematic categories and to focus additions and corrections to the rule body, performance was also measured by category. In the table below, only those categories with at least 10 instances were included (18 categories), plus the average of *all* 38 categories, ranked by F-Score.

role label		***recall***	***precision***	***F-Score***
§FOC	t	97.4 %	97.4 %	97.4
§REFL	t	100 %	94.7 %	97.3
§DENOM	t	100 %	93.8 %	96.8
§PRED	t	97.4 %	96.1 %	96.7
§ATR	C, np	91.7 %	97.7 %	94.5
§ID	np	100 %	93.3 %	90.6
§AG	C	92.7 %	87.4 %	90.0
§PAT	C	91.5 %	86.6 %	89.0
§LOC	C	92.0 %	76.7 %	88.9
§ORI	C	100 %	80.0 %	88.9
all categories		**86.6 %**	**90.5 %**	**88.6**
§TH	C	81.6 %	86.6 %	84.0
§FIN	a	79.2 %	86.4 %	81.7
§LOC-TMP	a	87.1 %	72.8 %	79.3
§CAU	a	86.7 %	72.2 %	78.8
§RES	C	74.1 %	83.3 %	78.4
§BEN	C	80.0 %	72.7 %	76.2
§DES	C	84.6 %	68.8 %	75.9
§ADV	a	100 %	57.9 %	72.2

Table 2: Semantic role tagging performance

The data allows to identify certain performance patterns. First, most of the classical case roles (AG, PAT, TH and space roles, marked 'C') are lumped together in the middle of the field, while the best performing categories are "empty" top-level categories (PRED, DENOM), narrow one-word categories (REFL, FOC, marked 't'), or close-context np-internal attributes (marked 'np').

Adverbial roles ('a') perform below average, possibly because of their unclear distribution between different syntactic levels (clause types, verb-arguments and np-internal).

6. Conclusion and outlook

While encouraging as a proof-of-concept, our performance figures indicate, not all-too surprisingly, that semantic roles are a difficult topic to annotate, with error rates considerably higher than what has been published for CG-annotation of part of speech and syntactic function (Bick 2000), even on syntactically pre-analysed treebank data. However, given the tight integration of semantic role annotation with semantic prototype and name tagging, a synergistic trade-off can be expected from improvements in any one of these three areas, and all should be given due consideration in the future. We also hope that hitherto "CG-unexplored" annotation levels, targeted by our new rule compiler formalism, such as anaphora links, co-referent resolution and the integration of statistical data in the form of rule or context thresholds, will ultimately have a positive influence on semantic role annotation.

As a medium term goal, semantic roles could help to resolve the annotation conflict in NER between lexical/stable name classes and functional/contextual classes, by achieving the latter through a combination of the former with semantic roles (e.g. turning a "country" <civ> name into an administrative organisation <admin> by assigning it agent function (<civ> + §AG = <admin>). Long term, we suggest to annotate the Portuguese *Floresta Sintá(c)tica* treebank with semantic roles, in a combination of automatical meta-tagging and human revision. Such a ressource would ultimately support the construction of a Portuguese PropBank or FrameNet.

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

Bick, Eckhard (2005), Turning Constraint Grammar Data into Running Dependency Treebanks, In: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.), *Proceedings of TLT 2005*, *Barcelona, Spain,* pp.19-27

Bick, Eckhard (2006), "Noun Sense Tagging: Semantic Prototype Annotation of a Portuguese Treebank". In: Hajic, Jan & Nivre, Joakim (red.), *Proceedings of TLT 2006, Prague, Czech Republic*, pp.127-138

Dowty, D. (1987). Thematic proto roles and argument selection, *Language* **67:** 547-619.

Fillmore, C. (1968). The case for case, *in* E. Bach & R. Harms (eds), *Universals in linguistic theory*, Holt, Rinehart and Winston, New York.

Foley, W. & van Valin, R. (1984). *Functional syntax and Universal Grammar*, CUP, Cambridge.

Hajicova, E. & J. Panevova & P. Sgall (2000). A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. UFAL/CKL Technical Report *TR-2000-09*, Charles University, Czech RepubliJackendoff, R. (1972). *Semantic interpretation in generative grammar*, The MIT Press, Cambridge, Ma.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Antilla, A. (1995). *Constraint Grammar: A language-independent system for parsing unrestricted text*, Mouton de Gruyter, Berlin.

Palmer M, Kingsbury P, Gildea D. (2005) "The Proposition Bank: An Annotated Corpus of Semantic Roles". *Computational Linguistics* **31** (1): 71-106

Santos, Diana & Nuno Cardoso (2007). HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro. Linguateca

Taulé, M. et al. (2005). "Mapping Syntactic Functions into Semantic Roles". In: *Proceedings of TLT2005:* 185-194

Appendix: Tagging sample

```
O           [o]  <artd> ART M S @>N #1->3
primeiro              [primeiro] <NUM-ord> ADJ M S @>N #2->3
fabricante            [fabricante] <Hprof> N M S @SUBJ> #3->17 §AG
mundial  [mundial] ADJ M S @N< #4->3
de          [de] PRP @N< #5->3
«ratos»    [rato] <Adom> N M P @P< #7->5 §PAT
para        [para] PRP @N< #9->7
computador ,        [computador] <tool> N M S @P< #10->9 §FIN
a           [o] <artd> ART F S @>N #12->13
empresa  [empresa] <HH> N F S @APP #13->3 §ID
suíça       [suíço] ADJ F S @N< #14->13
Logitech,             [Logitech] <org> PROP F S @N< #15->13 §ID
apresentou            [apresentar] <mv> <vt> V PS 3S IND @STA #17->0 §PRED
esta        [este] <dem> DET F S @>N #18->19
```

```
semana	[semana] <dur> N F S @<ADVL #19->17 §LOC-TMP
em	[em] <sam-> PRP @<ADVL #20->17
uma	[um] <arti> ART F S @>N #21->22
feira	[feira] <occ> N F S @P< #22->20 §LOC
especializada		[especializar] V PCP PAS F S @N< #23->22
que	[que] <rel> INDP F S @SUBJ> #24->25 §TH
teve	[ter] <mv> <ve> V PS 3S IND @FS-N< #25->22 <vi> §ATR
lugar	[lugar] <L> N M S @<ACC #26->25 §INC
em	[em] PRP @<ADVL #27->25
Basileia	[Basileia] <civ> PROP F S @P< #28->27 §LOC
(Suíça)	[Suíça] <civ> PROP F S @N<PRED #30->27 §LOC
um	[um] <arti> ART M S @>N #32->33
equipamento		[equipamento] <cm> N M S @<ACC #33->17 §PAT
periférico		[periférico] ADJ M S @N< #34->33
denominado		[denominar] <mv> V PCP PAS @ICL-N< #35->33 §ATR
«Audioman»		[Audioman] <brand> PROP M S @<SC #37->35 §ATR-RES
que	[que] <rel> &hum INDP M S @SUBJ> #39->40  §AG
permitirá		[permitir] <mv> <vt> V FUT 3S IND @FS-N< #40->37 §ATR
dotar	[dotar] <mv> <vdt> V INF @ICL-<ACC #41->40 <vi> §EV
os	[o] <artd> ART M P @>N #42->43
computadores		[computador] <tool> N M P @<ACC #43->41 §BEN
de	[de] PRP @<PIV #44->41
«orelhas»		[orelha] <anmov> N F P @P< #46->44 §TH
.
```