Semantic Roles for Portuguese Verbs *

Rui Talhadas[1,3], **Nuno Mamede**[2,3] and **Jorge Baptista**[1,3]
[1] U. Algarve/FCHS, [2] U. Lisboa/IST, [3] INESC-ID LIsboa

Abstract. Semantic roles are a relevant layer of linguistic information, useful for many Natural Language applications. In the Lexicon-Grammar framework, it has not received much explicit attention, though many authors use them in a more or less formal perspective. This paper presents the tentative of defining in a precise way and systematically applying the concept of semantic role to the full extent of full verb constructions in European Portuguese. The purpose of this is double: firstly, to verify the consistency and applicability of this concept in a large-sized lexicon-grammar, and, secondly, to prepare the way for an effective SR labelling module, to be integrated in a fully-fledged NLP system, STRING[1].

Keywords: semantic role, semantic role labelling, lexicon-grammar, Portuguese, verb.

1. Introduction

Semantic roles (SR) (or *theta* or *thematic roles*) correspond, grossly, to the well-known notion of *lead*: "*Who did What to Whom, How, When and Where*" (PALMER *et al*. 2010). The concept of SR, though not exactly the term, is already present in linguistics, associated with the discussion on the semantic or syntactic values of *Case* and the development of Case Theory (ANDERSON 1999), particular with the work of FILLMORE (1968) and the notion of *Deep Case*, and later derived in Linking Theory (LEVIN 1993).

In languages such as Portuguese, whose order of constituents in the sentence is relatively stable, there is a great regularity between the function and the position of syntactic constituents, on the one hand, and the semantic role they play in relation to the operator on which they depend, on the other hand. Thus, for example, the subject of a verb is often the AGENT of the process, while the direct complement is, in most cases, its OBJECT:

(1) *O Pedro*/$_{\text{subject-AGENT}}$ *moldou o barro*/$_{\text{direct complement-OBJECT}}$ 'Peter shaped the clay'

The semantic role is often directly related to the syntactic function that the constituent plays in the sentence. However, it is not always possible to predict, from the syntactic function, the semantic role a constituent plays. For example, in the next sentence, the direct complement has a locative interpretation:

(2) *O Pedro*/$_{\text{subject-AGENT}}$ *atravessou a sala*/$_{\text{dir. compl.-PLACE}}$ 'Peter crossed the room'

Moreover, certain transformations modify sentences, changing the arrangement of its constituents relatively to the core of the predication, without, however, changing their respective thematic roles. This is the case of active-passive pair in verbal constructions:

(3a) *O Pedro*/$_{\text{subj-AGENT}}$ *já leu o texto da Ana*/$_{\text{dir. compl.-OBJECT}}$ 'Peter already read Ana's text'
(3b) *O texto da Ana*/$_{\text{subj-OBJECT}}$ *já foi lido pelo Pedro*/$_{\text{comp.-AGENT}}$ 'Ana's text has already been read by Peter'

or in the standard-converse tranformation, in nominal predicate constructions with support verbs (G.GROSS 1989, BAPTISTA 1997):

(4a) *O João*/$_{\text{subj.-AGENT}}$ *deu uma rápida leitura ao texto da Ana*/$_{\text{ind. compl.-OBJECT}}$
'John gave a quick reading to Ana's text'
(4b) *O texto da Ana*/$_{\text{subj.-OBJECT}}$ *levou uma leitura rápida por parte do Pedro*/$_{\text{compl.-AGENT}}$
'Ana's text got a quick reading from Peter'

The semantic roles express the relations between the predicate and its arguments (DOWTY 1991). This semantic level overlaps the syntactic level, captured by the parsing Natural Language Processing (NLP) task, that is, the analysis of the sentence and the identification of its immediate constituents (shallow parsing), as well as the extraction of the syntactic dependencies they hold with the main verb and between them. In other words, for numerous NLP applications, an adequate representation of the meaning of the sentence may just require the mere identification of the sentence's constituents and their syntactic dependencies. However, for more complex processes, a deeper semantic analysis may be useful or even necessary. In the context of natural language processing, the task of syntactic parsing (shallow

* This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011 and under FCT project CMU-PT/HuMach/0053/2008.
[1] https://string.l2f.inesc-id.pt/

parsing) and semantic analysis (deep parsing) are normally performed sequentially. SRL is part of this latter stage, directly built upon the former.

This paper is structured as follows: Next, in §2, we briefly overview how the problem of defining in a precise way a set of semantic roles has been addressed by several authors, in different perspectives, and then, in §3, we sketch the classification of semantic roles here used. The remainder of the paper reports on the systematic survey of the lexicon-grammar of European Portuguese verbs – ViPEr (BAPTISTA 2012) and presents the major results thus achieved. In section §4, we present the strategy and some preliminary experiments in devising a rule-based module for SRL, already implemented in a fully-fledged NLP system, STRING (MAMEDE *et al*. 2012).

2. Definitions of Semantic Roles

Though the general concept of semantic role is relatively clear, it is much more difficult to define precisely the requirements for a given constituent to be tagged as performing a certain role in the sentence. In fact, there is abundant discussion, in the vast literature already available, about the number of SR and the specifics on the definition of each SR (ALLEN, 1994; GILDEA and JURAFSKY, 2002; PALMER *et al.*, 2010; RUPPENHOFER, 2010; OLIVEIRA, 2010; Wikipedia[2], s/v). Before any SR labeling in undertaken, it is necessary to clearly define which semantic roles are to be used by a system and to define them in an unambiguous, clearly reproducible, way. Due to space limitations, we cannot comment here in detail all the definitions of the authors above[3]. Nevertheless, based on the authors cited above, the definitions and the adequacy of the designations given to the semantic roles presented in the litterature were analyzed critically in order to produce what we consider could be a more consensual, but especially a more reproducible, set of SRs.

To further justify the new set of SR, this was confronted with the main types of semantic predicates found in the European Portuguese lexicon-grammar of verbs (BAPTISTA 2012). This is a database containing more than 5,000 lemmas, distributed by more than 6,300 syntactic-semantic entries, each one consisting in a clearly reproducible semantic predicate, and organized in roughly 60 formal classes. Thus, for each ViPEr class, we checked for adequacy in the definition and naming of the relevant SRs involved in the semantic predicates there found. All the major semantic classes of verb predicates were considered: verbs of *communication*, verbs of *transfer*, *movement* verbs, *psychologic* verbs, *quantification* verbs, *transformation* verbs, *creation* verbs, etc., using as few SRs as possible but without losing sight of the need for precision in the definition and, at the same time, trying to encompass all the rich variety of the semantic relations a verb can hold with their arguments. In this way, a set of 38 semantic roles was devised. For some semantic roles, new designations, not previously found in the literature, were adopted, such as *addressee* or *message*, specific of communication predicates like *say*, which could be mapped onto already used SR types like *goal* and *object*.

Some designations (e.g. *goal*) were dropped on the basis of their poor and imprecise previous definition. Other semantic roles were deemed to be too general and insufficiently precise, so they were brokedown into several subtypes. In order to capture specific predication types, we have, for example, unfolded the traditional role of *agent* into five subtypes, namely: *agent-generic*, *agent-creator*, *agent-giver*, *agent-speaker* and *agent-taker*. We could have instead considered *agent* alone as a "macro-SR" representing all these subtypes.

In the case of some syntactic slots, two SR can be found, depending on the distributional nature of the words fiiling that slot, though the verb sense was considered to be the same and did not justify spliting it into two verb entries. This is the case of the well-known alternation between an *agentive* and a *causative* subject of *psychologic* verbs, e.g. *O João/esta notícia irrita imenso o Pedro* 'John/this news irritates Peter a lot'. Because of this alternation, a complex SR was devised (*SR-agent-cause*), to prepare the way for the selection of the appropriate SR when the verb is found in texts and the distributional nature of its subject is properly identified[4].

3. Encoding Semantic Roles in ViPEr

ViPEr (BAPTISTA 2012) is the lexicon-grammar of European Portuguese full (or lexical/distributional) verbs. It is a database in tabular format, that features 6,330 lines, corresponding to the full verbs' senses or constructions and 112 columns, indicating the corresponding syntactic, semantic and transformational properties. The full list of 6,330 ViPEr verbal entries was manually annotated for the SR in each of their syntactic slots (subject, object and other complement positions; only essential arguments are considered in ViPEr). A total number of 13,201 syntactic slots were classified for their SR. For example, the ambiguous verb *cheirar* 'to smell', appears in 4 ViPEr classes, and each verb entry may show different sets of SRs:

- in class **32C**: *O Pedro cheirou a flor* 'Peter smelled the flower', where it was given the SR features *experiencer-gen* and *object-gen*;
- in class **33**: *A Ana cheira a rosas* 'Ana smells of roses', where it was given the SR feature *object-gen* twice;
- in verb class **33MV**: *A flor cheira bem* 'The flower smells good', where it has the features *object-gen* and *manner*; and
- in class **05**: *Cheira-me que a Ana o sabe* 'I think that Ana knows it', where it was given the SR features *object-f* (subject) and *experiencer-gen* (indirect complement).

[2] http://en.wikipedia.org/wiki/Thematic_relation.
[3] For detailed description, see Talhadas (*in progress*).
[4] In the case of human subject, preference is given to an *agentive* interpretation.

Table 1 shows the results of this classification procedure, namely the number of SRs per syntactic slot in all verb constructions of ViPEr. The distribution of the semantic roles per syntactic slot is very asymmetrical: 10 SR features cover almost 90% of the argument positions in ViPEr. The remaining 28 features cover the other 10%. It can also be seen that many SRs are residual (less than 5 verbal entries) in ViPEr, such as *possessor* or *beneficiary*. This does not mean that these SRs will not be found, eventually with an expressive frequency, in real texts.

Table 1. Semantic Roles frequency in ViPEr verbs construction

Semantic Role	***N0***	***N1***	***N2***	**%**	**Cumul.%**
agent-gen	4,248	0	0	0,3218	0,3218
object-gen	392	2,682	164	0,2453	0,5671
patient	0	899	18	0,0695	0,6365
experiencer-gen	522	373	0	0,0678	0,7043
object-f	37	400	178	0,0466	0,7509
cause	536	1	0	0,0407	0,7916
locative-dest	0	165	242	0,0342	0,8258
locative-place	15	305	13	0,0252	0,8510
agent-speaker	281	0	0	0,0213	0,8723
object-cl	20	208	0	0,0173	0,8896
addressee	0	34	170	0,0155	0,9050
message	0	178	13	0,0145	0,9195
locative-source	0	29	131	0,0121	0,9316
occurrence	34	99	2	0,0102	0,9418
co-agent	0	106	20	0,0095	0,9514
recipient	4	1	106	0,0084	0,9598
co-object	0	12	75	0,0066	0,9664
agent-giver	60	0	6	0,0050	0,9714
*agent-cause**	62	0	0	0,0047	0,9761
object-q	0	52	2	0,0041	0,9802
locative-path	0	47	0	0,0036	0,9837
agent-creator	36	0	0	0,0027	0,9864
*patient-object**	0	32	0	0,0024	0,9889
experiencer-vol	26	0	0	0,0020	0,9908
*locative-source-locative-dest**	0	22	0	0,0017	0,9925
*agent-object**	13	5	0	0,0014	0,9939
agent-taker	18	0	0	0,0014	0,9952
topic	0	12	5	0,0013	0,9965
manner	0	8	1	0,0007	0,9972
co-patient	0	0	7	0,0005	0,9977
*co-agent-co-object**	0	1	5	0,0005	0,9982
time-duration	0	6	0	0,0005	0,9986
instrument	0	4	1	0,0004	0,9990
victim	0	0	4	0,0003	0,9993
co-experiencer	0	3	0	0,0002	0,9995
time-calendar	1	0	1	0,0002	0,9997
beneficiary	0	1	0	0,0001	0,9998
co-locative	0	1	0	0,0001	0,9998
co-occurrence	0	1	0	0,0001	0,9999
possessor	0	1	0	0,0001	1,0000
Total	6,305	5,688	1,164	1,0000	1,0000

Some compound features like *agent-cause* or *agent-object* (marked with '*') are used when a syntactic slot can be filled in by a human or non-human element, thus implying one of those two SRs depending on the specific distributional nature of that element. The semantic role attribution is made by rules that match the distributional class of the element to the appropriate SR. An example of an ambiguous syntactic slot is the subject of the verb *irritar*, that has the feature *SR-N0-agent-cause* feature, in order to cover the two situations in the following examples (5)-(6):

(5) *O Pedro irritou a Ana* 'Peter irritated Ana'
(6) *O artigo do jornal irritou a Ana* 'The newspaper article irritated Ana'

The rule applies the *agent-generic* SR, to the first sentence (5), if it matches the subject distributional feature as a human (a similar rule exists for the opposite case):

```
if (subj(#1[SR-N0-agent-object],#2[UMB-Human]) & EVENT[other](#1) & ~ EVENT(#1,#2))
EVENT[agent-generic=+](#1,#2).
```

Table 2. Macro Semantic Roles

Semantic Role	*N0*	*N1*	*N2*	% SR
agent-x	4,718	5	6	0,36
object-x	462	3,379	344	0,32
patient-x	0	931	18	0,07
experiencer-x	548	373	0	0,07
locative-x	15	568	386	0,07
co-x	0	124	107	0,02
time-x	1	6	1	< 0 ,01

As said before, a macro-SR is a construct that represents the set of all SRs of a certain type (*e.g. agent-x* includes all SRs that have an agentive nature). **Table 2** shows the representativity of the macro SRs in the entire set of 13,157 semantic roles that have been encoded in ViPEr. The complex SR *agent-cause*, *agent-object* and patient-object, were introduced in the *agent-x* and *patient-x*, respectively. Symmetric complements were kept apart from the corresponding SR. The most significant macro-SR in ViPEr is *agent-x* (36%), closely followed by *object-x* (32%). In spite of the breakdown of the semantic role of *agent* into several subtypes, the more generic agent SR (*agent-gen*) is significantly more representative on ViPEr (4,248 instances) than all the remaining, more specific *agent* SRs (481 instances). The next most productive configuration is *agent-patient*. Notice the two intransitive construction with *object* (5%) or *agent* (5%) subjects, and the *locative* structures (5%). 331 verb constructions occur 10 or less times, and 67 only once. This is not to say that these more specific SRs are not relevant to capture certain semantic relations expressed by those predicates that feature them, nor that their frequency in texts may prove to be significant.

Table 3. Most common Semantic Roles combinations

Verbs	Semantic Roles combinations	%
1,847	*SR-N0-agent-gen SR-N1-object-gen*	0,29
568	*SR-N0-agent-gen SR-N1-patient*	0,09
343	*SR-N0-experiencer-generic SR-N1-object-f*	0,05
330	*SR-N0-cause SR-N1-experiencer-generic*	0,05
289	*SR-N0-object-gen*	0,05
271	*SR-N0-agent-gen SR-N1-locative-place*	0,05
233	*SR-N0-agent-gen*	0,04
198	*SR-N0-agent-gen SR-N1-object-cl*	0,04
169	*SR-N0-agent-speaker SR-N1-message SR-N2-addressee*	0,03
151	*SR-N0-agent-gen SR-N1-object-gen SR-N2-location-destination*	0,03

Table 3 presents the 10 most common argument/SR combinations in ViPEr, from 177 different combinations. It is possible to observe that the combination *SR-N0-agent-gen SR-N1-object-gen* is, by far, the most productive construction, with 1,847 instances out of 6,632 (24%). In the top 10 constructions, the *N0* (subject) position is 70% of the times an *agent*.

4. A Semantic Role Labeling module for STRING

The features encoded in ViPEr are used to build the rules for the SRL module of STRING. The general strategy here adopted is the following: First, for each verbal construction in ViPEr, its class and all relevant features are processed into a specific format, in order to be integrated in the XIP parser lexicons, using specifically-built validation-conversion programs. The following example is the entry of one of the structures of verb *escrever* 'to write'. This verb is from class `32A` and its features are shown below:

```
"escrever-32A":"SR-N0-agent-creator SR-N0-Hum SR-N1-object-gen SR-N1-nHum SR-N1-cdir
SR-pass-ser SR-pass-estar"
```

The features here annotated represent the SR of the basic syntactic slots *N0* (subject) and *N1* (first complement) and their distributional features (human and non-human), including also the passive constructions allowed.

A word-sense disambiguation (WSD) module (TRAVANCA 2013), acting after the XIP parser, decides the most likely word sense (and the corresponding ViPEr class) for each verb instance in a text. The verb features from ViPEr associated to that verbal construction are then attributed to that disambiguated verb instance. For example, this is the set of features associated with an instance of the verb *escrever* in a text, after the disambiguation and after adding the ViPEr semantic roles' related information; the remaining features derive from other modules of STRING, for example the `32a` feature, indicating the verb's ViPEr class, resulted from the WSD module:

```
VERB(8-15)+[sr-n1-nhum:+, sr-n1-cdir:+,sr-pass-estar:+, sr-pass-ser:+, sr-n1-object-
gen:+, sr-n0-agent-creator:+, to-study:+, human_activity:+, mark_ger:+, 32a:+,
markviper:+, perf:+, ind:+, dicendi:+, 3p:+, sg:+, verb:+, hmmselection:+, last:+,
first:+]
```

Then, a set of rules attribute the SR to the specific syntactic slots depending on each verb features, at the same time that the system extracts the event structure associated to that verb. For example, if, in a given text, an already disambiguated verbal construction has the feature *SR-N0-agent-creator*, the output of the event extraction procedure will add that SR to the corresponding argument of the verb in the event structure it is associated to. Hence, given the rule:

```
if (subj(#1[SR-N0-agent-creator],#2)
        & EVENT[other](#1) & ~ EVENT(#1,#2) )
 EVENT[agent-creator=+](#1,#2).
```

the subject of any verb with the `SR-N0-agent-creator` feature, when the event expressed by this verb is extracted, will appear as an argument of that event, with the *agent-creator* semantic role assigned. The next sentence and its event structure illustrate the result of this procedure:

(7) *O Pedro escreveu um artigo* 'Peter wrote an article'

```
EVENT_AGENT-CREATOR(escreveu,Pedro)
EVENT_OBJECT-GENERIC(escreveu,artigo)
```

Naturally, these rules need to be more complex in order to take into account many different phenomena, for example, the word permutation involved in passive or in *verbum dicendi* (BATISTA 2010) constructions. Such description is already underway (TALHADAS, *in progress*). In future work, the macro SRs might be folded back to one generic SR. Only SRs related to full verbs were considered for this moment. Adjective-specific SR may yet prove to be necessary, e.g. *ser feito de <matéria>* 'to be made of <matter>'. One of the major difficulties already found is metonymy, as in *O Pedro leu Camões na escola* 'Peter read Camões at school', where an apparent violation of distributional constraints may perhaps be better solved through the use of SR information.

5. Conclusions and future work

This paper presented the distribution of a set of 38 semantic roles manually assigned to the argumental slots of over 6,300 entries of the Lexicon-Grammar for European Portuguese full verbs. About 13,200 SR were assigned. The fine-grained classification of SR, including the splitting of some major roles, like *agent* or *object*, into several subtypes did not proved productive and hinted at a simpler, but more effective classification. Regular correspondence was confirmed between some semantic roles and the verbs' syntactic slots in the verb basic structure. Initial steps towards a SRL module integrated in the STRING natural language processing chain were reported. Transformations and the argument slots of non-verbal predicates (nouns, and adjectives, mainly), as well as several semantic transfer phenomena (metaphor, metonymy), constitute a challenge to a precise SR labelling.

We intend to extend SR encoding to other non-verbal predicates, to further improve the event extraction task. For the time being, only locative prepositions allow for locative events and their corresponding argument SR labelling. In the case of predicative nouns the work is just commencing. Finally, the annotation of a corpus with SRs in on-going, and we expect to evaluate the SRL module in a near future (TALHADAS, *in progress*).

References

ALLEN, J. F. 1994. *Natural Language Understanding*. Benjamin Cummings.

ANDERSON, J. M. 1999. Case Theory. in Bown, Keith and Miller (eds.). *Concise Encyclopedia of Grammatical Categories*. Pages 58-65. Elsevier.

BAPTISTA, Jorge. 1997. *Sermão, tareia e facada*: Uma classificação das construções conversas *dar – levar*. *Seminários de Linguística* 1: 5-37. Faro: U. Algarve/CELL.

BAPTISTA, Jorge. 2010. *Verba dicendi*: a structure looking for verbs. Nakamura, Takuya; Laporte, Éric; Dister, Anne; and Fairon, Cédrick (eds.). *Les Tables. La grammaire du français par le menu*. Mélanges en hommage à Christian Leclère. *Cahiers du CENTAL* 6: 11-20. Louvain-la-Neuve: CENTAL/Presses Universitaires de Louvain.

BAPTISTA, Jorge. 2012. *ViPEr*: A Lexicon-Grammar of European Portuguese Verbs. *Proceedings of the 31st Conference on Lexis and Grammar*, Novè Hrady (Czeck Republic). Universitá degli Studi di Salerno (Italy)/University of South Bohemia in Nové Hrady (Czech Republic), *pp 10-16*.

DOWTY, David. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67: 547-619. Linguistic Society of America.

FILLMORE, Charles J. 1968. The case of case. *in* Holt, Rinehart and Winston (eds.). *Universals in Linguistic Theory*. pages 1-88.

GILDEA, Daniel & JURAFSKY, Daniel. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 245–288.

GROSS, Gaston. 1989. *Les construction converses du français*. Genève: Droz.

LEVIN, Beth. 1993. *English Verbs and Alternations: A Preliminary Investigation*. Chicago: UCPress.

MAMEDE, Nuno; BAPTISTA, Jorge; DINIZ, Cláudio and CABARRÃO, Vera. 2012. STRING: An Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. *in:* Caseli, Helena; Villavicencio, Aline; Teixeira, António & Perdigão, Fernando (eds.), *Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR 2012* (demos). Paper available at url: http://www.propor2012.org/demos/DemoSTRING.pdf

OLIVEIRA, Hugo; SANTOS, Diana & GOMES, Paulo. 2010. Extracção de Relações Semânticas entre Palavras a partir de um Dicionário: o PAPEL e sua Avaliação. *Linguamática*, 2:1, 77–93.

PALMER, Martha; GILDEA, Daniel & XUE, Nianwen. 2010. *Semantic Role Labeling*. Morgan and Claypool.

Ruppenhofer, Josef; Ellsworth, Michael; Petruck, Miriam R. L.; Johnson, Christopher R. & Scheffczyk, Jan. 2010. *FrameNet II: Extended Theory and Practice*. Technical Report. International Computer Science Institute.

TALHADAS, Rui (*in progress*). *Semantic Role Labelling for European Portuguese Verbs*. (MA dissertation). Faro: U. Algarve).

TRAVANCA, Tiago. 2013. *Verb Sense Disambiguation* (MSc thesis). Lisboa: IST-UL and INESC-ID Lisboa.