DOI: 10.14393/DL52-v16n4a2022-7

PFN-PT: A Framenet Annotator for Portuguese

Anotação semântica automática: um novo Framenet para o português

Eckhard BICK[*]

ABSTRACT: This article presents PFN-PT, a robust system for the automatic semantic annotation of Portuguese, consisting of a new, parsing-oriented framenet and a rule-based frame- and role-tagger. The framenet provides almost 13,000 valency frames covering 7,300 verb lemmas with 10,700 senses. Frame and role tagging is achieved by iterated matching of syntactic structures and semantic noun types with slot-filler conditions in the framenet. We discuss design principles and present frame and role statistics. In an evaluation run on news data, the system achieved an overall F-score of 92.2% for frame senses.

KEYWORDS: Portuguese Framenet (PFN-PT). Semantic Role Labeling (SRL). Semantic Parsing. Constraint Grammar (CG). PALAVRAS.

RESUMO: Este artigo apresenta o PFN-PT, um sistema robusto para a anotação semântica automática de Português, consistindo numa nova framenet com foco em *parsing*, e um tagger para frames e papéis semânticos baseado em regras. A framenet contêm cerca de 13.000 padrões sintáticos cobrindo 7.300 lemas verbais com 10.700 sentidos. A etiquetagem é realizada por meio de um alinhamento iterativo de estruturas sintáticas e classe semântica de substantivos com as condições listadas no framenet para argumentos sintáticos. Discutimos princípios de desenho e apresentamos estatísticas de distribuição de categorias. Numa avaliação realizada com base em textos jornalísticos, o sistema alcançou 92,2% sentidos/frames corretos para verbos.

PALAVRAS-CHAVE: FrameNet português (PFN-PT). Etiquetagem semântica. Papéis semânticos. Gramática de restrições (CG). PALAVRAS..

1 Introduction

In modern corpus linguistics, automatic grammatical annotation of free text is a central task, and the usefulness of a corpus depends on the linguistic complexity of

[*] PhD. University of Southern Denmark. ORCID: https://orcid.org/0000-0002-5505-4861. eckhard.bick@gmail.com

its annotation. However, the performance of annotation tools decreases with increasing complexity. Thus, while lower-level annotation such as lemmatization, morphology and part-of-speech (POS) work reasonably well, syntactic-structural annotation (treebanks) is more difficult to achieve, and for most languages other than English, robust semantic annotation, with the possible exception of named-entity-annotation (NER), remains an unsolved challenge. The task can be broken down into semantic classification and disambiguation on the one hand, and semantic function and relations on the other. For nouns, the classification task is addressed with ontologies like WordNet (FELLBAUM, 1998), that have good coverage and operate on an effective classification principle (hyponymy) but fail to provide the structural-relational information necessary to disambiguate senses. Verb classification is even more difficult, because though troponymy can be used instead of hyponymy, classification is less "local" than for nouns, and intertwined with the second semantic task, assigning semantic structure and argument relations to a clause. Thus, for English, VerbNet (KIPPER et al., 2006), Berkeley FrameNet (BAKER et al., 1998; JOHNSON; FILLMORE, 2000; RUPPENHOFER et al., 2010) and PropBank[1] (PALMER et al., 2005) classify entire predications, assigning a semantic class to the core lexeme (typically, but not necessarily a verb) and semantic roles (also called case/thematic roles, FILLMORE, 1968) to its arguments and possibly adjuncts. The combination is constructed as a lexical knowledge representation called a frame (FILLMORE; BAKER, 2001), that can be triggered (evoked) by the presence of certain frame elements (FE). For instance, in Portuguese, depending on concept granularity, the frame of selling can

[1] Methodologically, VerbNet and the English Berkeley FrameNet were conceived as lexicographical projects, one frame at a time, while PropBank and the German SALSA framenet depart from corpus data, one sentence at a time. Coverage problems are therefore different in nature: In the former, a common sense may be missing in a verb with several rare senses assigned. In the latter, common senses are registered first, but rare lemmas may be missing entirely.

involve verbs like vender (sell), *exportar* (export), *liquidar* (liquidate), *aleiloar* (auction off) as well as nouns like *venda* (sale) and *exportação* (export). The involved FE's are the core argument roles of a 'seller', 'buyer' and 'goods' and the peripheral adjunct role of 'price'. These could be specified as such, but can also seen as implied - in the context of a sell verb - by the more general 'donor' (or just 'agent'), 'receiver', 'theme' and 'value'. It is a matter of framenet design - and possibly language-dependent - where to use a separate frame where the perspective changes, i.e. *buy sth from sb instead of sell sth. to sb*., or where transitivity or agency changes, e.g. *change (self/subject) and alter (sth. else/object)*. The fact that the English Berkeley FrameNet has inspired framenets in several other languages, such as German (BURCHARDT et al., 2006), Spanish (CARLOS; PETRUCK, 2010), Portuguese (SALOMÃO, 2009; TORRENT; ELLSWORTH, 2013) and Japanese (OHARA et al., 2004) has shown that such distinctional choices, though inspired by English, have a level of abstraction and universal validity that allows them to be ported across languages.

In his comparison of the WordNet and FrameNet approaches, Boas (2005) stresses the added level of abstraction provided by the latter - e.g. that frames are independent of part-of-speech (POS) -, as well as the systematic link between semantic information and lexical-syntactic patterns. Thus, both FrameNet and PropBank provide morphosyntactic restrictions to frame realization. In addition, the former also specifies ontological slot filler information. Because it is anchored in lexico-syntax, frame-based semantics has the potential of providing a bridge between classical NLP (Natural Language Parsing) and real-world applications within AI (Artificial Intelligence). Machine translation (MT), for instance, can profit from word sense disambiguation (WSD), which is an inherent "by-product" of frame assignment.

The framenet resource presented here, PFN-PT (Parsing Framenet for Portuguese), is meant to provide, for Portuguese, such a syntactic-semantic bridge at a practical level. It therefore has a methodological focus, meaning to support robust

automatic frame and role annotation of running text. This methodological primacy, as well as the frame inventory, has been borrowed from the Danish Framenet (BICK, 2011), that also focuses on automatic annotation (BICK, 2017). The complete system consists of two complementary parts, the framenet and the frame annotator, linked by a semantically informed[2,] valency-driven dependency representation provided by a morphosyntactic parser. For such a resource to allow automatic annotation, good and, above all, evenly distributed lexical coverage is crucial, and granularity should be kept at a realistic level. This homogeneity in coverage and granularity is what sets PFN-PT apart from the concept/domain-based and example-based creation methods of traditional framenets and propbanks, respectively.

2 Building the framenet

In a first round of bootstrapping, using tailor-made software, we identified Danish-Portuguese verb sense matches by harvesting the machine-translation (MT) dictionary used in the Portuguese-Danish section of the GramTrans[3] system (BICK, 2014), where polysemy is handled by providing syntactic argument and semantic slot-filler information in much the same way a frame entry would. Rather than using MT to match existing framenets in two languages (GILARDI; BAKER, 2018), we use it to match valency patterns as an anchor for frame transfer. For instance, if a Portuguese verb is allowed four different translations depending on the semantics of its subject and object, the software would look up the translations in the Danish FrameNet and choose a frame with the same slot-filler conditions. Similarly, but more heuristically,

[2] To work optimally, the method requires semantic type tags on nouns and possibly adjectives or adverbs, such as 'human', 'tool'', 'unit', 'food', 'feature' etc. These need not necessarily be disambiguated, but should be of medium granularity, so as to have distinctional potential as slot-fillers without being too fine-grained to have abstractional value.

[3] https://gramtrans.com

prepositional complements were harvested and matched to a Danish frame asking for a corresponding preposition in Danish (as suggested by the MT dictionary). All in all, the method came up with frame suggestions and slot-filler semantics for 9,414 valency patterns covering 7245 Portuguese verb lemmas and 8816 verb senses (defined as different frame names for the same lemma). In a second round of manual revision, the frame candidates were checked and, if necessary, corrected. Particular care was taken to check prepositional arguments and reflexives, where different syntactic realizations in the two languages (e.g. intransitive as reflexive, or transitive as pp argument and vice versa) sometimes caused errors in the matches suggested by the harvesting program. In a third step, PALAVRAS' parsing lexicon was used to add missing lemmas and valency realizations. In these cases frames, argument roles and their semantic selection restrictions had to be written from scratch, or adapted/expanded from another frame for the same lemma. Finally, noun and adjective frames were specified manually for all cases where the parser lexicon specified a (prepositional) valency argument.

Using the parser lexicon - rather than e.g. a corpus-based frequency list - to decide which lemmas to include has three advantages. First, the method ensures broad coverage, also for rare words. But as important, the parser lexicon provides valency patterns that can be used as syntactic skeletons for the frames both by the lexicographer and by a live frame tagger adding frames and roles to syntactically annotated input.

3 Lexicon size, coverage and granularity

The verb part of PFN-PT currently contains 7,273 verb lemmas, with manually revised or assigned frames. All in all, these cover 12,835 valency frames (i.e. with differences in either valency, argument roles or semantic slot-filler conditions) and

10,612 different lemma+frame combinations (verb senses, sense frames). On average this amounts to 1.77 valency patterns and 1.46 frame senses per lemma, with a Zipfian distribution, where some frequent construction verbs and polysemic verbs needed 10 or more entries, while 60% had only one valency type and 70% only one sense frame. The top-scoring lemmas were *dar* (30 entries), *fazer* (27), *passar* (22), *levar* (19), *ficar* (19), *ter* (18), *estar* (18), *tornar* (15), *ir* (15), *sair* (14), *contar* (14), *chamar* (14), *trabalhar* (13), *pegar* (13), or - in terms of verb senses - *dar* (21), *fazer* (20), *passar* (14), *levar* (14), *estar* (14), *ter* (13), *ficar* (13), *ir* (11), *tomar* (10), *tirar* (10). As to coverage, it is important to stress that virtually all verb lemmas (over 7,000) in the parser lexicon (as well as all valency-marked nouns and adjectives) were assigned at least one frame and that corpus evaluation (chapter 5) indicates a raw lexical failure rate as low as 1-4% for lemma types and 0.3% for tokens[4]. By comparison, the above-mentioned SALSA resource for German, albeit more refined, more revised and less "bootstrapped", only contains about 1000 unique lemma-frame types (REHBEIN et al. 2014).

For Portuguese, the obvious comparison reference is FrameNet Brasil (FB, TORRENT; ELLSWORTH, 2013). Judging from its online database, the project pursues an example-based and domain-driven approach, where coverage progresses in an uneven fashion, good in one area (sports/tourism), but insufficient at a general level. Many verbs have no frame at all, or - worse for running-text applications - a specialized but rare frame rather than the most frequent one. Thus, a lookup of a randomized 1% sample of PFN-PT's verb list in FrameNet Brasil revealed that 61 out of 72 verbs (84.7%) had no frame at all. Detailed inspection of one common frame, 'adquirir' (fn:obtain in PFN-PT), found 4 verbs and 1 adjective in FB, but 25 verbs and 2 adjectives in PFN-PT.

[4] Lexical coverage is otherwise an issue even for English. Thus, Palmer & Sporleder (2010), comparing SemEval data with Framenet data, found that 3.4% of lexical units and 12.1% of frames from the former were not found in the latter. In terms of training data, the gaps were even more pronounced, 6.9% missing senses and 26.0% missing verbs.

The prototypical verb 'obter' (obtain) was not in FB's 'adquirir' frame set, but had only the less general 'obter_documento', as well as 'Posse', the latter being problematic because, according to the website, it asks for a POSSESSOR subject, whereas 'adquirir' needs a RECEIVER subject, the latter being a better match for the core meaning of *obter*. Another coverage difference between FB and PFN-PT concerns valency variation, i.e. meaning differences triggered by syntax. Thus, the verb *recuperar* (reclaim) did make it into the 'adquirir' set, but FB did not have an entry for the health domain meaning 'to recover' linked to the verb's reflexive form *recuperar-se*. Finally, the very fact that FB has a specialized coverage of the sports domain makes it less useful for ordinary text, because coverage is less balanced than in a full-lexicon parser extension like PFN-PT. The verb *marcar* (mark), for instance, has only domain-specific frames in FB, 'Ações do árbitro', 'Infrações', 'Jogadas_interativas' and 'Jogadas_pontuadas', all of which would count as errors if used to annotate marcar when used with non-sports meanings, such as 'mark' (a surface), 'decide on' (a meeting), 'identify/show' (time). PFN-PT, for its part, has to make do with the general frames in sport texts, too, e.g. fn:obtain for 'scoring goals' or fn:accompany for 'tagging a player', but in a general annotation context, a lack of precision is preferable to wrong or missing labels. Also, apart from the domain information in the frame name, precision is not necessarily better with the specialized frames. Thus, the frame 'Jogadas interativas' lumps *marcar* with other special meanings of e.g. *atacar* (attack), *bater* (beat), *combinar* (combine) and *cruzar* (cross) that arguably cover meaning distances as large as the one between the general *acompanhar* (accompany) and the specific *marcar* (tagging a player).

Out of the 494 different verb frame categories[5] available in the Danish framenet scheme, almost all (482) also ended up being used for Portuguese, too[6]. On top of these, we introduced 5 new frames: fn:repeat, fn:return, fn:path, fn:enough and fn:alter_soc. In order to capture further nuances, 230 frame senses were used as secondary frames senses in combination with a primary frame sense, yielding about 640 different combinations of these "atomic" frames. In some cases, the two frame senses have almost equal weight, either because each could be used on its own (insistir [to insist] fn:declare&demand), because they combine different aspects such as method and domain (empanar [to bread / coat with breadcrumbs] fn:cover_ize&prepare_food), or because they address two parts of a complex action (desenterrar [to unearth] fn:poke&take). Complex frames can also be used to capture lexicalized information such as aspect/aktionsart or polarity (more/less, better/worse), where there is no separate frame in the inventory to make the distinction. Sometimes Portuguese prefixes indicate a secondary frame in a systematic way (*re-* = &repeat, &return, *des-* = &fail)

estrear [to debut] - fn:start&perform
reabrir [to reopen] - fn:open&repeat
desgarrar-se [to stray] - fn:orient&fail
baratear [to become/make cheaper] - fn:cost&decrease

In principle, the same frame inventory can be used for both verbal and nominal predications, but obviously the distribution will be different. PFN-PL contains two types of nominal frames, static and dynamic. Static frames exist for both nouns and

[5] A smaller set of 200 frame senses exists with a hypernym-mapping from the more fine-grained set. This smaller set was meant to facilitate cross-language comparisons, syntactic-contextual uses and parser training.
[6] For a full overview, definitions and examples, see https://framenet.dk/verbal_prototypes.pdf

adjectives and consist of valency-based, manual entries in the framenet lexicon. There are currently 849 noun frames (for 682 lemmas) and 407 adjective frames (for 376 lemmas). Given the much smaller number of lemmas, the spread of frames is smaller than for verbs, with 296 different frame senses for nouns, and 178 for adjectives.

Dynamic frames are computed on the fly based on the morphological analysis of de-verbal nouns and participle-based adjectives in an actual annotation run. In this process, prepositional arguments are taken over as-is from the corresponding verb frame. For noun frames, the subject of intransitive verbs or the object of transitive verbs is turned into an argument with the preposition "de".

> *fertilizar* [to fertilize sth.] (transitive) --> *fertilização de* OBJECT [fertilization of]
> *capotar* [overturn] (intransitive) --> *capotamento de* SUBJECT [rollover]
> *atribuir X a Y* [to attribute X to Y] --> *atribuição de X a Y* [atribution of X to Y]
> *categorizar X como Y* [categorize X as Y]-> *categorizado como Y* [categorized as Y]

The derived frames inherit their semantic roles from the verb frame, and in all cases semantic slot-filler restrictions are maintained - for instance, 'vehicle' AG(ent) for the subject of *capotar/capotamento*, or 'human' REC(eiver)for the Y in *adjudicar/adjudicação*.

4 Frame role distinctors: Valency, syntax and semantic class

The distinctional backbone of all frames in PFN-PT's frame is provided by syntactic valency patterns, such as <vt> (monotransitive), <vdt> (ditransitive), <com^vrp> (reflexive verb with a prepositional argument headed by the preposition “com”, e.g. *aborrecer-se com* [to abhor]). For any given verb lemma, each of these

valency patterns is assigned at least one, and possibly more[7], verb senses, each corresponding to a separate semantic frame. There can be different surface realizations of a sense/frame, so more than one syntactic pattern or semantic slot order may trigger the same verb sense / frame name, but two different verb senses will almost always differ in at least one syntactic or semantic aspect of at least one of the arguments governed. Therefore, almost all senses can in principle be disambiguated exploiting the tags and dependency links provided for each argument by a deep syntactic parser like Portuguese PALAVRAS (BICK, 2014).

Though the frame inventory and role granularity of PFN-PT is modeled on the Danish framenet, we decided to make an important change in notational conventions, extending the shorthand system suggested by Bick (2017) for noun frames to cover the main, verbal lexicon, too. Thus, for each of the almost 13.000 verb sense frames, a list of arguments is provided in a single, composite tag ready to be used by CG rules. For instance, it is possible to differentiate at least seven different meanings[8] of the verb *apontar*, each with its only valency and frame patterns (<FN:...).

meaning: *point out (a person/thing)* valency: <vt> (monotransitive)
<FN:identify/S§AG'H/O§TH'H>
meaning: *point out (a fact)* valency: <vt> (monotransitive)
point out (that ...) valency: <vq> (que-clause)
<FN:emphasize/S§SP'H|sem/O§SOA'fact|ac|f|fcl>
meaning: *sharpen (a pencil)* valency: <vt> (monotransitive)
<FN:shape/S§AG'H/O§TH'"lápis"> # spidse
meaning: *appear (moon, sun, dolphin)* valency: <vi> (intransitive)
<FN:appear/S§TH'Lstar|A>

[7] In PFN-PL, 890 cases multiple verb senses share the same valency frame - in other words, in 7.6% of the verb entries, verb senses cannot be disambiguated on the grounds of syntactic function and form alone, but need help from semantic (noun) classes.

[8] For etymological reasons, the English verb point shares part of this polysemy (albeit with its own phrasal particles). In other languages, however, there may be no overlap at all.

meaning: *point at* valency: <para^vp> (PP-monotransitive)
<FN:gesture&identify/S§AG'H/P-para§DES'cc>
meaning: *point (a gun) at* valency: <a^vtp> (PP-ditransitive)
valency: <para^vtp> (PP-ditransitive)
<FN:orient/S§AG'H/O§TH'cc|tool|V/P-a|para§DES'H|L>
meaning: *appoint (a person) as* valency: <como^vtp> (PP-ditransitive)
<FN:appoint/S§AG'H/O§BEN'H/P-como§ROLE'H>

In this scheme, frames match valencies in terms of arity, so a monotransitive frame will have two role arguments. Arguments are slash-separated ('/') and contain themselves three information fields:

1. Syntactic function (**S** - subject, **O** - direct object, **D** - dative object etc.)
2. Thematic role (e.g. **§AG** - agent, **§TH** - theme, **§DES** - destination)
3. Semantic slot filler conditions (e.g. '**H**' - human, '**food**', '**act**') for argument nouns

For prepositional arguments, the preposition in field 1, for instance P-em for *insistir em* (to insist on). Syntactic form restrictions for material other than 'np' (noun phrase), field 3 is used (e.g. 'fcl' - finite clause, 'icl' - non-finite clause, 'num' - numeral).

In our evaluation of PFN-PL, 29% of verb lemma types were sense-ambiguous in terms of frame names. In 890 cases, valency patterns were ambiguous with regard to verb frames (senses + arguments), with a maximum of 8 frames for one valency pattern. However, in some of these, the difference only concerned roles and slot fillers, not the verb sense. Discounting these cases, 731 valency patterns (covering 661 verb lemmas) were verb-sense ambiguous, with only 133 valency patterns (128 verb lemmas) being affected by more than one such ambiguity. This means that almost 90% of all verbs could in theory be assigned a unique sense using syntactic argument structure alone (i.e. syntactic function, form and dependency), that is, with input from

a simple morphosyntactic parser without lexical semantics[9]. For the rest, semantic slot-filler clues are needed.

The most common valency patterns (table 1) were monotransitive with accusative (S/O), intransitive (S) and monotransitive with a PP argument.

Table 1 – Valency patterns.

Argument inventory	of all frames in lexicon
S/O	55.0 %
S	16.7 %
S/R	10.8 %
S/P	5.9 %
S/O/P	4.4 %
S/R/P	4.3 %
S/SA	0.6 %
S/D/O	0.5 %
S/O/OC	0.2 %
S/O/OA	0.2 %
S/D	0.2 %
S/SC	0.2 %

(S=subject, O=object, P=pp argument, REFL=reflexive, D=dative object, SA=subject-related adverbial argument, OA=object-related adverbial argument, A-INC=verb-incorporated adverbial

PFN-PT uses 44 "atomic" semantic roles (or case/thematic roles, FILLMORE, 1968). The role inventory closely resembles that of the original PALAVRAS role annotator (BICK, 2007), with a few additions, such as §COG (cognizer), §ASS (asset), §POSS possesor, as well as a few roles targeting clausal arguments in particular: §SOA (state-of-affairs), §ACT (action) and §EV (event). The number of roles is similar to the one used in the semantic level of treebanks such as the Prague Dependency Treebank (BÖMOVÁ et al., 2003) and the Spanish 3LBLEX/3LBSEM project (TAULÉ et al., 2005). This kind of medium category granularity is big enough to allow useful distinctions,

[9] Note that these numbers refer to types, not tokens. Obviously, in running text, the most frequent verbs are the most ambiguous ones.

but small enough for generalizations and to avoid simply duplicating lexical information. Also, a limited level of granularity is easier to disambiguate and easier to define for human consensus, thus lending itself particularly well to automatic corpus annotation.

Where necessary, these roles can be combine, e.g. §AG-EXP (agent & experiencer) for the subject of "zuhören" (listen). All in all, 61 such combinations occur in the lexicon, the most frequently used being §AG-EXP (e.g. watching), §TH-COM (e.g. associating) and §AG-PAT (especially reflexives wit agents causing changes in themselves). In addition, a number of adverbial roles is used that are never valency-bound and therefore do not occur in any frame, but can be used by the frame tagger rules to specify the function of free adverbials and adverbial subclauses (such as §COND for conditional subclauses). The 44 roles are far from evenly distributed in running text. Table 2 shows role frequencies at the token level, for News texts[10].

Table 2 – Semantic roles.

	Semantic Role	surface verb arguments %	secondary verb args %	free (adjunct) adverbials %	noun arguments %
§TH	Theme	24.2	23.6	0.4	28.8
§AG	Agent	12.7	26.4	0.6	6.0
§ATR	Attribute	12.2	10.4	0.5	1.7
§PAT	Patient	5.7	5.4	-	11.4
§ACT	Action	5.0	3.1	1.4	4.1
§SOA	State-of-affairs	3.3	1.6	-	1.4
§SP	Speaker	3.3	3.6	0.1	1.8
§MES	Message	3.2	2.3	-	1.4
§COG	Cognizer	3.1	5.1	-	3.7
§EV	Event	2.9	3.5	1.4	3.2
§RES	Result	2.1	2.4	0.1	2.7
§REFL	Reflexive	2.0	-	-	

[10] 24483 random sentences from the Leipzig corpus collection (cp. chapter 6, evaluation).

§LOC	Location	1.9	1.5	16.7	5.9
§BEN	Beneficiary	1.7	1.5	2.0	2.8
§MNR	Manner	1.6	-	8.8	
§DES	Destination	1.6	0.7	3.8	2.0
§REC	Receiver	1.5	1.0	0.9	1.3
§INS	Instrument	1.0	0.4	0.9	0.7
§EXP	Experiencer	0.9	1.7	0.1	0.4
§CAU	Cause	0.9	0.6	2.6	0.8
§LOC-TMP	Temporal Location	0.7	-	21.3	1.7
§ORI	Origin	0.7	0.6	2.1	0.7
§VAL	Value	0.4	0.7	0.7	0.7
§PART	Part	0.2	0.7	-	0.1
§DON	Donor	0.2	0.6	-	-
§FOC	Focus marker	-	-	8.6	(4.1)
§CIRC	Circumstance	-	-	7.0	0.4
§FIN	Purpose	-	-	5.8	1.7
§DES-TMP	Temp. Destination	-	-	1.7	-
§EXT	Extension	-	-	1.5	0.1
§TP	Topic	0.6	0.3	0.7	2.7
§ASS	Asset	0.3	0.2	-	1.8

Other roles: §STI - Stimulus; §REFL - Reflexive; §PATH - Path; §EXT-TMP - Duration; §ROLE - Role; §CONC - Concession; §COMP - Comparison; §HOL - Whole; §POSS Possessor; §CONT - Content; §ID - Identity; §ATR-RES - Attribute-result; §COM - Co-role; §INC - Incorporated

At the clause level, §TH (theme), §AG (agent) and §ATR are the most common argument roles, while §LOC (location), §LOC-TMP (temporal location) and §MNR (manner) dominate adverbial adjunct roles, joined by §PAT (patient) at the noun phrase level (last column). For primary dependencies, i.e. with verb and argument in the same clause (column 3), §ACT (action) and §SOA (state-of-affairs) are relatively common because they are typically describe the function of object clauses relative to the main clause verb. For secondary, cross clause, dependencies (4th column), typical subject roles, in particular §AG, is more frequent than otherwise because of the non-surface subjects of infinitives and relative clause.

For some roles, there is a very tight link to a syntactic function. Thus §ID (identity) is closely linked to the function of apposition (@APP), while subject and object complements (@SC and @OC) carry exclusively attributive roles, §ATR and §ATR-RES (attribute-result). The same is obviously true of syntactic "dummy" roles without independent semantic information: §PRED (predicator, for top verbs), §REFL (reflexive) and §INC (verb-incorporated). The latter is used for binding particles in verb chains and in support verb constructions, where the semantic weight and - to a certain degree - valency reside in a nominal element, typically a noun that syntactically fills a (direct or prepositional) object slot, but semantically orchestrates the other complements. While our input parser, PALAVRAS, already marks verb chain particles at the syntactic level (@PRT-AUX<), it assigns noun incorporates an ordinary syntactic tag (@ACC), which therefore needs lexical treatment in the frame lexicon to allow assignment of a special sense to the support verb.

ter ***que/de*** + INF (must)	<fn:must>
botar ***fogo*** *em* (set s.th. on fire)	<fn:burn&start>
dar ***certo*** (succeed)	<fn:succeed>
dar ***cabo*** *de* (put an end to)	<fn:destroy>

One could argue that the real frame arguments (like the §PAT - patients - of *botar fogo em* and *dar cabo de* should be dependency-linked to the §INC constituent, and the frame class marked on the latter, but for consistency and processing reasons we decided to mark the frame name on the verbal element of such support constructions, with a corresponding semantic dependency link from the verb.

PP incorporates are handled in a similar way, with both the preposition and its argument listed in the frame pattern. Because of the above-mentioned principle of semantic dependencies, the §INC role with fall on the nominal part of the PP, blocking assignment of other (adverbial) roles.

*ter **a ver** com* (have to do with) <fn:relate>
*por **em causa*** (jeopardize) <fn:risk>

5 Automatic frame annotation

In our scheme, frame annotation is achieved by checking the elements of a potential frame against the feature sets of syntactic arguments, exploiting morphosyntactic and semantic information already assigned and disambiguated by the PALAVRAS parser. For instance, the verb *contar* has a number of different meanings, each sense corresponding to one or more different valency frames:

<fn:tell> *contar que ... contar uma história* (to relate that ... a story)
<fn:math> *contar ovelhas* (count sheep)
<fn:contain_have> *contar 300,000 habitantes* (have 300,000 inhabitants)
<fn:assume&rely> *contar com* (count on somebody or something happening)
<fn:plan> *contar viajar para ..., contar em viajar para ...* (plan to travel to ...)
<fn:matter> *a sua ajuda conta muito para mim* (your help means a lot to me)

Since only one of these - a "tell" frame - has a slot for a finite clause complement (fcl), the presence of an object clause will trigger this frame, harvesting a speaker role (§SP) for the subject (S), a message role (§MES) for the subclause and a receiver role (§REC) for a possible dative object (D), as specified by the following frame template example:

<FN:**tell**/S§SP'H/D**§REC**'H/O**§MES**'fcl>

The technical implementation of this annotation mechanism uses the Constraint Grammar formalism (CG, BICK; DIDRIKSEN, 2015), also used by the frame annotator's input parser, PALAVRAS. Apart from the obvious advantage of notational compatibility (tags and dependency), CG also has the advantage of allowing very

complex contextual rules including the use of tag set operators, variable unification, named relations and regular expressions, to name just a few. This makes it possible to unify frame constraints with actual tags and relations, while at the same time performing additional checks on the context.

As a preparatory step, all possible frames for a given verb lemma are mapped as template tags on all main verbs (cp. the 'tell' frame above). Also, some minor dependency adaptations are performed to make the syntax tree more "semantic", such as making sure that the subject is linked to a main verb rather than an auxiliary and marking certain nouns as dependency-"transparent" (e.g. *a maioria de*, *um monte de*, *uma copa de*, *parte de*, *um dos/das*, *uma sorte de*). Also, we introduce secondary, semantic dependency links for implicit arguments in relative or non-finite clauses, and add shadow pronouns to subject-less but person-inflected verbs, where a semantic category (especially ±HUM) can be deduced.

In the annotator grammar itself, four main rule types are used:

1. frame template tag selection
2. frame template tag removal
3. role instantiation
4. mapping of free roles

Template removal is a simpler task than template selection, because a single mismatch is enough to trigger the former, while the latter optimally needs as many frame element matches as possible. Therefore, matching rules are ordered in heuristicity batches with high-arity valency matches and full semantic slot matches coming before partial valency matches and hypernym or semantics-free slot matches. Safest are lexical matches, where word forms (e.g. verb incorporates or prepositions) are mentioned as such in a frame template. Thus, in the *contar* example above, the

'assume&rely' frame can be safely discarded or selected based on the presence of a PP argument with th preposition *com*:

<FN:**tell**/S§SP'H/D**§REC**'H/O**§MES**'fcl>

Another relatively safe method are syntactic mismatches - at least as long as the parser gets syntactic functions right, with correct dependency links. Thus, subject and object complements are obligatory arguments and the corresponding frames can be safely discarded in their absence. For direct objects, the situation is more complicated because of passivization and ellipsis, so more contextual checks are needed. Also, there can be a fail-safe condition asking for the existence of a competing, similar frame with lower valency (e.g monotransitive instead of ditransitive). In Germanic languages, impersonal verb frames could be chosen or discarded based on formal subjects, but in Portuguese, there is ambiguity here, because the subject-less 3. person singular form used for impersonality can also just mean an omitted subject. Therefore, while weather verbs (snowing, raining etc.) are safe, because they do not have other meanings, the verb ter and haver (to have) are quite difficult for the frame annotator, because without a subject, a lot of context and semantics is needed to make the distinction between on the one hand <fn:exist> (*tem/há* X - there is X) and, on the other, ordinary having frames ('have', 'have part', 'have attribute'), not to mention auxiliary and support verb constructions.

The most important rules, however, are the ones capable of differentiating frame templates with identical syntactic/valency skeletons. In this scenario, semantic slot filler information is used, exploiting the so-called semantic prototype tags that PALAVRAS assigns to nouns and proper nouns. Although these are lexical tags, and only partially disambiguated by the parser, ambiguities rarely overlapped with frame ambiguities found in a given verb. The noun ontology has about 200 categories, e.g. <tool>, <food>, <act> (action) or <mon> (money). Categories are organized in a shallow

hierarchy, with lower-case or hyphenated subcategories. 'H', for instance, means +HUM (human) and occurs in tags such as <Hprof> (human professional), <Hfam> (family member), <Hideo> (ideological human) etc., while <sem-r> (readable), <sem-c> (concept), <sem-s> (sayable) etc. all belong to the class of semantic products, sharing the 'sem-' prefix.

The frame templates specify - for all non-trivial function/role pairs (e.g. 'S§AG' for subject agent) - at least one semantic slot filler, drawing on the categories discussed above. The frame tagger's matching mechanism proceeds from safe to unsafe by trying rules with many, or more specific, conditions first, followed by underspecified matches, using heuristic defaults as a last resort to decide unresolved ambiguities. In order to handle category fuzziness or overlaps, creative language use or just incomplete slot filler information, the frame tagger grammar uses "umbrella category" matches. Here 17 semantic hypernym sets, e.g. 'HUMAN', 'THING', 'PLACE' etc. At this intermediate level, two categories (from a frame template and a sentence token, respectively) will be considered a match if they are part of the same hypernym category. Applying these principles, the following (simplified) scheme for rule ordering/prioritization was applied:

- 2 or more syntactic slots with a full semantic match
- 1 slot with a full match, 1 with an "umbrella category" match
- 2 or more "umbrella" matches
- 1 slot with a full match
- 1 slot with an "umbrella" match
- syntactic match, slot(s) marked <all> or <cc>
- longest syntactic match (e.g. ditransitive match beats monotransitive match)

As long as roles manifest as surface constituents with a direct dependency link to the frame-evoking verb, the method is quite robust, since it can simply draw on the existing dependency parse. However, in many cases, the role-carrying constituent has no, or only pronominal, surface representation in the clause itself. In Portuguese, this

is the case for e.g. subjects of infinitive clauses and antecedents of relative clauses. Thus, in the example sentence, the word Eltern (parents) functions - in two different subclauses - as both an object-theme in an 'exist' frame and as a subject-cognizer in an 'allow' frame. Here, in order to be able to check slot-filler conditions and to assign roles, we first introduced secondary dependency relations using special, relation-mapping CG rules. These additional dependency relations can then be drawn upon by variants of the ordinary frame-matching rules.

In a vertical, one-word-per-line CG notation, the frame-tagger adds <fn:sense> tags on verbs (red), and §ROLE tags on arguments (blue). The example contains four ordinary main verb frames with one or more role dependents each, one auxiliary frame and two noun frames with a single dependent. Roles relate to the next-coming frame upward in the dependency tree, bypassing prepositions and transparent nouns[11]. Primary dependency arcs are shown as #n->m ID-links, secondary dependencies are marked as R:c- (child) and R:p (parent) relations (green). The latter are matched by a secondary role link on the dependent, marked as R:sd- (semantic dependency) relations. In the example, *unidades de infantaria* has such a secondary subject link to the verb of the relative clause, *faltam*, with the role of §TH (theme).

```
O               [o] <artd> DET M S @>N #1->2
Secretário-geral        [secretário-geral] <Hprof> N M S @SUBJ> §AG #2->10
de              [de] <sam-> <np-close> PRP @N< #3->2
a               [o] <-sam> <artd> DET F S @>N #4->5
ONU             [ONU] <org> PROP F S @P< #5->3
$,              [$,] PU @PU #6->0
Butros          [Butros] <hum> PROP M S @APP §ID #7->2
Butros-Ghali    [Butros-Ghali] <hum> PROP M S @N< #8->7
```

[11] If desired by a corpus user with other annotation conventions (e.g. Universal Dependencies), or if needed for the use with a specific corpus search tool, it would take only a couple of rules to change this dependency convention into direct semantic links, with prepositions as dependents of nouns, and the role-carrying noun getting a direct dependency arc to its frame-carrying semantic head.

$, [$,] PU @PU #9->0
deixou [deixar] <fn:leave> <mv> V PS 3S IND @FS-STA §PRED #10->0
ontem [ontem] <atemp> ADV @<ADVL §LOC-TMP #11->10
Luanda [Luanda] <civ> PROP F S @<ACC §ORI #12->10
com [com] PRP @<ADVL #13->10
a [o] <artd> DET F S @>N #14->15
promessa [promessa] <act-s> <fn:promise> N F S @P< §CIRC #15->13
de [de] PRP @N< #16->15
que [que] <clb> <clb-fs> KS @SUB #17->25
até [até] PRP @ADVL> #18->25
a [a] <sam-> PRP @P< #19->18
o [o] <-sam> <artd> DET M S @>N #20->21
fim [fim] <fn:end> <temp> N M S @P< #21->19 §TMP-DES
de [de] PRP @N< #22->21
Agosto [agosto] <month> N M S @P< §TMP-LOC #23->22
deverão [dever] <fn:must> <aux> V FUT 3P IND @FS-P< #24->16
chegar [chegar] <fn:reach> <mv> V INF 3P @ICL-AUX< §ACT #25->24
a [a] PRP @<SA #26->25
Angola [Angola] <civ> <*> PROP F S @P< §DES #27->26
as [o] <artd> DET F P @>N #28->29
unidades [unidade] <HH> N F P @<SUBJ §AG #29->25 ID:29 R:c-subj:33 R:sd-TH:33
de [de] PRP @N< #30->29
infantaria [infantaria] <HH> N F S @P< #31->30
que [que] <clb> <rel> <hum> SPEC M S @SUBJ> #32->33
faltam [faltar] <fn:lack_itr> <mv> V PR 3P IND @FS-N< §ATR #33->29 ID:33 R:p-subj:29
para [para] PRP @<ADVL #34->33
se [se] PERS F 3S/P ACC @ACC>-PASS §REFL #35->36
completar [completar] <fn:complete_finish> <mv> <passive> V INF 3S @ICL-P< §FIN #36->34
a [o] <artd> DET F S @>N #37->38
Missão=de=Verificação [Missão=de=Verificação] <project> PROP F S @<SUBJ §PAT #38->36
$. [.] PU @PU #39->0

6 Evaluation and discussion

For a quantitative evaluation of the system, we automatically annotated 24,483 random sentences of news text (610,620 tokens12) from the Leipzig corpus collection[13]. The sample contained 2,222 different verb types[14], or 64,313 tokens[15]. 213 tokens with verb tags had no frame. However, manual inspection revealed that among these there were 28 misanalysed foreign words, 22 spelling errors and 27 unrecognized correct spelling (mostly "frequentar", that had been listed with the older, pre-reform 'ü'). 25 were morphological analysis errors and 21 morphological disambiguation errors. Only 90 (44 types) were regular frame failures, amounting to 0.1% of verb tokens, or 2% of verb types. A recount for a larger sample (2 million words) yielded, as expected, the same 0.1% for tokens, but a higher percentage (4%) for verb types. This can be explained by Zipf curve characteristics - as the corpus grows, it covers logarithmically more verb types, but in this so-called "tail" of the Zipf curve rare is made up of rare words, a phenomenon that will affect types, but not tokens.

At the type level, the annotated sample corpus contained 1.38 senses per verb, close to the ratio of 1.46 found in the lexicon. 75.8% of verb types in the corpus occurred with only one frame sense. However, frame sense ambiguity is very unevenly distributed, with a lot of ambiguity at the token level. Thus, in terms of running verb tokens, the grammar had a substantial disambiguation load, with 4.94 frame senses per verb, and only 23.4% 1-sense verb tokens. For the assignment of complete frames, i.e. including argument roles, the disambiguation load is even higher, because not just

---

[12] Raw count with space-separated tokens, including punctuation. Due to fusion of multi-word-expressions (e.g. names), PALAVARAS annotated token count had 38,096 fewer tokens.

[13] https://wortschatz.uni-leipzig.de/en/download/Portuguese

[14] This is the corrected count, after subtraction of 82 types caused by spelling errors etc. The passive auxiliary use of ser and *ficar* was excluded, because by matter of design there is never assigned a frame in this case.

[15] This count is the annotation count, which is only an approximation of the true count, due to possible annotation errors. The type count, on the other hand, is reliable, because the verbs either received frames from the lexicon or were checked by hand. It is unlikely that a verb token should be consistently, in all instances/contexts, sim-disambiguated as a different part-of-speech.

senses, but specific valency frames need to be disambiguated for this task. Here, token ambiguity was 7.36 (valency) frames per verb, and only 12.3% of verb tokens were unambiguous.

For the 12 most frequent verbs, together accounting for 17.4% of all verb tokens in the corpus, frame sense ambiguity is shown in table 3. The numbers show that most lexicon senses were also present in the corpus, but for some verbs (*dar*, *fazer*, *ficar* ...), there is an extra disambiguation load for rare senses or support verb constructions.

Table 3 – Frame sense ambiguity.

Verb lemma	frame senses in corpus	frame senses to be disambiguated (lexicon)
passar	14	14
dar	14	21
levar	13	14
fazer	12	20
estar	12	14
ter	10	13
ir	9	11
ficar	9	13
trabalhar	7	7
tornar	7	8
tomar	7	10
tirar	7	10

Since matching the semantic categories of arguments with slot-filler conditions is very important for frame sense disambiguation, missing surface arguments can be a problem. Salomão (2009, p. 177), for FB, distinguished between three types: (a) definite (subject recoverable from matrix clause), (b) indefinite (e.g. object missing or unclear) and (c) constructional (e.g. passives). Accordingly, infinitives, passives, relative clauses, ellipsis and incomplete sentences all get special treatment in our frame

annotator to increase the system's robustness[16]. For instance, secondary dependencies are created for noun antecedents of relative clauses and subject antecedents of infinitives[17]. However, not every slot-filler can be recovered. Thus, passive participles (type c), though their object slot can be filled with a subject or noun antecedent, usually have an empty subject slot, as the role of agent-of-passive is rarely specified. In order to assess the size of the empty-slot problem, we ran a count of surface expression for the main argument types (table 4).

Table 4 – Surface expression of slot-filler arguments.

Argument type	filled slots (primary dep.)	filled slots (secondary dep.)	filled slots (all dep.)
SUBJ (subject)	51.3 %	17.4 %	62.1 %
SUBJ + 1./2. person inflection			66.2 %
ACC (object)	77.0 %	8.6 %	82.5 %
SC (subject complement)	86.1 %	1.3 %	86.1 %
PP (prepositional argument)	100 %	0 %	100 %

As can be seen, almost half of all verbs had no direct surface subject (2nd column), and over a third had neither a direct or indirect link to a subject (4th column), even if 1./2. person inflection is counted as a +HUM (human) slot-filler for an otherwise missing subject. For direct objects, there are more filled slots, but still a fair percentage of frames had to be chosen without slot fillers. The 100% expression of prepositional slots is due to the fact that these were never the shortest frames for a given lemma, and could be safely discarded if not matched. In other words, surviving PP-frames were

[16] One might want to add coordination to this list, but the situation is really rather the opposite: By attaching conjuncts in parallel to the same head, the latter's frame slots will actually get 2 or more chances of slot filling rather than just one.

[17] Portuguese infinitives can, in fact, carry some person-number inflection, or even govern surface subjects, but subject-less, uninflected infinitives are still the most common.

chosen precisely because the PP was matched. For all argument types it should be noted that the percentage of real semantic matches is somewhat lower, since on average 9.2 % of all verb arguments were pronouns - a part of speech with only limited semantic content (±HUM for some). Again, this problem is more pronounced for subjects (15.9%) than for direct objects (12.7%) and PP arguments (4%).

To judge the performance of verbal part of the frame tagger, we evaluated a random set of 300 sentences (about 6000 raw tokens) of mixed Portuguese news from the Leipzig corpus collection, harvested in 2019 from tweets from the minority-filtered subcorpus, with 9054 tokens. All in all, the parser tagged 739 words as main verbs and 165 as auxiliaries, only three of which (0.7%) were POS errors. Our frame tagger knew all verb lemmas in the sample and assigned frames to all of them, with the exception of the 47 *ser* passive auxiliaries, that PFN-PT by design doesn't consider as frame carriers. The other auxiliaries were spread over 13 types[18] and 16 different frames. Since auxiliaries are fairly easy to annotate (only 1 frame error), evaluation concentrated on main verbs. The frame tagger found a correct frame sense for 92.4% of the correctly tagged verbs (table 5), with 3 spurious frames (92.0% precision). Inspection of the 56 errors revealed 9 cases, where the reason was faulty input - 2 spelling errors and 7 parser errors (1 POS error and 6 syntactic). This indicates that in a perfect world with perfect input, the performance of the frame tagger on its own could be a couple of percentage points higher.

Table 5 – Verb frames: Recall and precision.

	Recall	Precision	F-score[19]

[18] The Portuguese definition of auxiliary is complex and includes more than tense and diathesis, using criteria like pronoun movement. There is no 100% consensus on the matter in the literature. In addition to auxiliaries proper, PALAVRAS annotates other verb chains with light verbs, all of which get individual frames in PFN-PT.

[19] This is the F1 measure, defined as 2 * R * P / (R + P).

main verb frames	92.4 %	92 %	92.2
auxiliary frames	99.2 %	100 %	99.6 %

Automatic frame tagging is not yet a standard part of NLP pipelines, and systems and approaches are difficult to compare due to different tag granularity and domain, not least across languages. However, even without another Portuguese frame tagger to compare with, these figures constitute an encouraging result. Thus, although the "weak", inspection-based evaluation method makes a direct comparison impossible, performance compares favorably with e.g. an early English baseline for rule-based frame tagging, Shi & Mihalcea (2004), with an F-score of 74.5%. The task of word sense disambiguation (WSD) can also be a relevant comparison, e.g. the German SHALMANESER tagger (BURCHARDT et al., 2009), with 79% accuracy. More recently, in the machine learning camp (ML), Hermann et al. (2014) report F=70.1% for predicate frame identification, and Cai & Lapata (2019), also for German, performed (in-domain) semantic role labeling (SRL) with an F-score of 82.7%, using neural networks. It is interesting that both of the latter exploited syntactic features, as does our own frame tagger. Without using linguistic features, Do et al. (2018) reached a somewhat lower F-Score (73.5) for SRL on the same German test data (CoNLL 2009).

As described above, the PFN-PT tagger also performs SRL. For valency-bound arguments this task is intertwined with frame disambiguation, because the same dependency links and semantic slot filler matches can be used to harvest the relevant roles from a matching frame and assign them to the argument in question. In fact, SRL is often more robust than frame identification. For instance, one verb may have several frames all sharing the same subject role (e.g. §AGent). Time constraints for this paper did not allow a proper SRL evaluation, but an inspection of 15% of the test corpus suggests a recall of 95.4% and a precision of 96.9% for argument roles. This is better than the F-score of 88.6% reported by (BICK, 2007) for a syntax-only approach to

Portuguese SRL, though a direct comparison is impossible because the latter also included adjunct roles. In any case, SRL results can be expected to closely correlate with frame identification, according to Hartmann et al. (2017, p. 475), who compared SRL with automatic frame identification to SRL based on correct "gold" frames, finding that gold-based SRL was stable across text types (domains), while automatic frame identification and SRL varied considerably, but with a stable accuracy ratio of about 3:2 (i.e. SRL accuracy 33% lower than frame accuracy). Given a stable ratio, frame identification results can be used as a rough predictor for SRL, and improvements in the former should manifest in the latter. That said, it is interesting that the SRL accuracy of our system is actually higher than its frame identification accuracy, rather than lower, as suggested by Hartmann et al.'s experiments. Possible reasons for this could be a difference in the size of either the overall role tag set or the number of roles per frame type. Also, quality differences in the underlying parser could mean a lower SRL performance because of wrong argument attachments despite the gold senses. More likely, however, the difference is simply methodological - in our setup there is no separate, statistical step of role assignment. Rather, given an established verb sense and a correct syntactic analysis, role assignment is a 1-on-1 lexical lookup. Even where the system, due to rule hierarchies, has decided on one or more roles first, and deduced a frame sense from these in a second step, wrong roles will lead to wrong frames, too. So it is all but impossible for frame accuracy to be higher than role accuracy. The inverse, however, is possible because several of a verb's frames may share, e.g., the same subject or object role. Here, SRL will be counted as correct even if a wrong frame sense was chosen.

Hartmann et al.'s comparison of in-domain and cross-domain performance also highlights an important difference between our rule- and lexicon-based frame-annotator and other current systems, most of which rely on machine learning (ML). An ML system needs training data and therefore heavily depends on the availability

of human gold annotations for a particular domain. Without (rule-based) matching of syntactic slot information in the lexicon on the one hand and a syntactic-structural annotation on the other, there is no direct way of exploiting a framenet for annotation (beyond lemma-filtering of frame candidates), since lexicon coverage does not automatically translate into training data coverage. In fact, Hartmann et al. claim that a lexicon-free setup, based only on a vector space model of the frame-word-bundle-pairs found in the training data can perform on par with the lexicon-informed standard setup. One interpretation of this is that the lexicon-free setup compensates for coverage problems and unevenly distributed granularity in the frame lexicon, something that is not relevant for our own system, where frame annotation is not based on training data, but on direct structural matches.

7 Sense or metaphor?

Following the reductionist Constraint Grammar philosophy that category disambiguation gains robustness from removing readings incrementally as ambiguity decreases in the context, rather than selecting a category with one single rule, our frame tagger whittles down a lemmas list of potential senses (frames) step by step, and interactively with role assignment. The robustness of this method resides in the fact that it is not necessary to achieve a 100% match for all syntactic, lexical and semantic constraints of a given frame pattern. Instead, the most conflicting options will be removed first, and the least conflicting ones will survive longest. Ideally, the last surviving frame will be the correct one, even if an argument was missing or did not have the expected semantic type, or if it wasn't properly tagged or attached by the parser. Apart from being robust, the resulting "mismatching" frame annotation can be useful in its own right. First of all, mismatches can be semi-automatically exploited lexicographically - to hone the list of semantic selection restrictions for a given

argument or - if that is not enough - add a new sense (and frame) to the head lemma. But rather than adding more and more eclectic senses, there is also the alternative of keeping the sense and treating the mismatch as metaphor. Thus, the attribute '±metaphorical' foreseen by Torrent & Ellsworth (2013) in their discussion of frame annotation layers could be handled intra-sense rather than cross-sense, providing a mechanism for addressing the inherent fuzziness of word senses. For instance, the Portuguese verb *disparar* (generate) initially had three senses in PFN-PT - <fn:activate>, <fn:throw> and <fn:leave> with subject selection restrictions of 'human' for the first two and 'animal/human' for the third. Typical objects were listed as 'tools' and 'things' for the transitive senses, <fn:activate> and <fn:throw>, respectively. These are the senses found in GramTrans' MT lexicon and in the Portuguese online dictionary Dicio[20] and most lexicographers would probably not accept an etymologically motivated reduction to one core sense of 'letting loose' - but what about the opposite, increasing the number of senses? Should there be a fourth sense <fn:increase> for *preços disparando* (soaring prices) and a fifth sense <fn:start> or <fn:cause> for *disparar tensões* (trigger crises), because you can't literally turn on or throw a crisis, and because prices aren't entities that can 'run off'. And if not, which is the closest existing sense? Given the current granularity, our frame annotator will resolve the subject mismatch (a) as <fn:leave> and the object mismatch (b) as <fn:activate>, the former because of the intransitive use and lack of a +HUM subject, the latter heuristically by picking the top transitive meaning:

> (a) 'os preços dispararam' (prices soared)
> (b) 'A recessão internacional poderia disparar novas tensões e picos de crises financeiras' (The international recession could trigger new conflicts and peaks of financial crises)

[20] https://www.dicio.com.br (accessed 28. Oct. 2021).

The suggested new senses do exist in the frame inventory of PFN-PL, and obviously the frame annotator can make the distinction if given the relevant selection restrictions, i.e. <f-q> (quantifiable feature) for the subject in (a) and <event> for the object in (b). So once examples like the above are on the lexicographer's table, the framenet lexicon can be amended[21]. However, letting the system choose the "most matching" frame (rather than no frame) seems to be an acceptable solution where there is no precise frame entry (yet), lending robustness to the system and even preparing it for a future mark-up of potential productive metaphors. In this vein, we experimentally added rules to the annotator grammar that allow it to add a tag for "projected sense". The tag is assigned to mismatching arguments in the case of "fuzzy" or heuristic frame matches, i.e. if an argument matches syntactically, but not semantically, and it will contain the semantic class (or classes) specified for this argument in the last surviving frame. Thus, in example (a), preços will receive a tag <PROJ:A>, meaning the chosen fuzzy frame, <fn:leave>, projected the selection restriction of 'animal' (A), suggesting a metaphorical reading of prices running wild, like an unbridled horse[22].

8 Conclusion and future work

PFN-PT is meant to help closing the semantic gap in the rule-based parsing pipe of PALAVRAS, or - for that matter - any other morphosyntactic parser with capable of producing dependency trees. We have shown that this new framenet resource has the

[21] Some dictionaries, such as the online dictionary Priberam (https://dicionario.priberam.org, accessed 28. Oct. 2021), do list these senses. Still, the differences between dictionaries, as well as constantly evolving language usage, call for a robust backup solution, such as the one suggested here.

[22] The projection 'human' and 'tool' for the subject and object, respectively, in the recession/crisis example would be less illuminating. Though <fn:activate> can work as a hypernym for 'trigger', this is likely not a metaphor, but rather - if not a true sense - a case of to-be-amended slot fillers.

breadth and depth to provide good coverage of the Portuguese verb lexicon on running text in terms of both types and tokens. A preliminary evaluation of the system's frame tagger suggests a satisfactory performance for verbal frame sense disambiguation above the 90% correctness threshold, allowing meaningful automatic corpus annotation, as well as WSD for AI tasks such as machine translation.

However, in spite of its high lemma coverage for verbs, PFN-PT is by no means a finished task. Coverage and granularity should be improved, adding rarer senses and - not least - more verb constructions with incorporated nominal material. Also, the noun and adjective frame lexica have a far lower coverage, where performance could be improved by replacing ad-hoc frames with lexicon frames. Also, a clearer definition is needed of which nouns and adjectives can be frame carriers, e.g.:

- do you need surface dependents, or can a noun carry a frame on its own?
- how to treat participle-derived adjectives without dependents?

Last not least, a more thorough evaluation is needed, ideally after these improvements. This is relevant not least for semantic role labeling, where adjuncts/adverbials, as well as noun frames, need to be included in a future evaluation round. Finally, the theory-based claim of cross-domain robustness should be empirically corroborated.

Obviously, these tasks, as well as future maintenance of the resource, would greatly profit from a participation of the general research community, and hopefully PFN-PT will facilitate and inspire semantic annotation and research projects within the large and growing tapestry that is Portuguese language technology. To this end, the framenet will be open for collaboration and improvements, and the semantic

annotator version of PALAVRAS will be freely available[23] for academic research institutions making use of PFN-PT and its semantic annotation.

References

BAKER, C. F.; FILLMORE, J. C.; LOWE, J. B. The Berkeley FrameNet project. *In*: **Proceedings of the COLING-ACL** (Montreal, Canada). ACL, 1998. p. 86-90. DOI https://doi.org/10.3115/980451.980860

BICK, E. Automatic Semantic Role Annotation for Portuguese. *In*: **Proceedings of TIL 2007 - 5th Workshop on Information and Human Language Technology** / Anais do XXVII Congresso da SBC (Rio de Janeiro, July 5-6, 2007). Rio de Janeiro, 2007. p. 1713-1716.

BICK, E. A FrameNet for Danish. *In*: Proceedings of NODALIDA 2011 (May 11-13, Riga, Latvia). **NEALT Proceedings Series**, Vol 11. Tartu: Tartu University Library, 2011. p. 34-41.

BICK, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. de L. S. B. (ed.). **Working with Portuguese Corpora**. London/New York: Bloomsburry Academic, 2014. p 279-302.

BICK, E. Swedish-Danish Machine Translation in a Constraint Grammar Framework. *In*: PRZEPIÓRKOWSKI, A.; OGRODNICZUK, M. (ed.). Advances in Natural Language Processing, **Proceedings of 9th International Conference on NLP** (PolTAL 2014, Warsaw, Poland, September 17-19, 2014), Heidelberg: Springer, 2014. p. 216-227.

BICK, E.; DIDRIKSEN, T. CG-3 - Beyond Classical Constraint Grammar. *In*: MEGYESI, B. (ed.): **Proceedings of NODALIDA 2015** (May 11-13, 2015, Vilnius, Lithuania). Linköping: LiU Electronic Press, 2015. p. 31-39

BICK, E. From Treebank to Propbank: A Semantic-Role and VerbNet Corpus for Danish. *In*: TIEDEMANN, J. (ed.): **Proceedings of the 21st Nordic Conference on**

[23] If interested, please contact the author. Please note that unlike PFN-PT itself, the Constraint Grammar rules used by PALAVRAS and its semantic annotator module are not open for editing, and possible modifications should be made by output tag filtering or by adding a new, task-oriented rule set or machine learning module. For commercial use, a separate, paid license is required.

Computational Linguistics (NoDaLiDa 2017, Göteborg). NEALT Proceedings Series Vol. 29. Linköping University Electronic Press, 2017. p. 202-210.

BOAS, H. C. From theory to practice: Frame semantics and the design of FrameNet. *In*: LANGER, S.; SCHNORBUSCH, D. (ed.): **Semantik im Lexikon**. Tübingen: Gunter Narr Verlag, 2005. p. 129 - 159.

BÖHMOVÁ, A.; HAJIČ, J.; HAJIČOVÁ, E.; HLADKÁ, B. The Prague dependency treebank. *In*: **Treebanks**. Springer: Dordrecht, 2003. p. 103-127. DOI https://doi.org/10.1007/978-94-010-0201-1_7

BURCHARDT, A., ERK, K., FRANK, A., KOWALSKI, A; PADO, S.; PINKAL, M. Using FrameNet for the semantic analysis of German: Annotation, representation and automation. *In*: BOAS, H. C. (ed.): **Multilingual FrameNets in Computational Lexicography**: Methods and Applications. Mouton de Guyter, 2009. p. 209-244.

CAI, R.; LAPATA, M. Syntax-aware Semantic Role Labeling without Parsing. *In*: **Transactions of the Association for Computational Linguistics**, 7. ACL, 2019. p 343-356. DOI https://doi.org/10.1162/tacl_a_00272

DO, Q. N.; LEEUWENBERG, A.; HEYMAN, G.; MOENS, M. A Flexible and Easy-to-use Semantic Role Labeling Framework for Different Languages. *In*: **Proceedings of COLING 2018** (Demos). ACL, 2018. p. 161-165

FELLBAUM, C. (ed.). WordNet: An Eletronic Lexical Database. *In*: **Language, Speech and Communications**. MIT Press: Cambridge, Massachusetts, 1998. DOI https://doi.org/10.7551/mitpress/7287.001.0001

FILLMORE, C. J. The case for case. *In*: BACH; HARMS (ed.). **Universals in Linguistic Theory**. New York: Holt, Rinehart, and Winston, 1968. p. 1-88.

FILLMORE, C. J.; BAKER, C. F. Frame semantics for text understanding. *In*: **Proceedings of WordNet and Other Lexical Resources Workshop**. NAACL, 2001

GILARDI, L.; BAKER, C. F. Learning to Align across Languages: Toward Multilingual FrameNet. *In*: **International FrameNet Workshop 2018**: Multilingual FrameNets and Constructions (Miyaki, Japan). 2018. p. 13-22

HARTMANN, S.; KUZNETSOV, I.; MARTÍN-VALDIVIA, M. T.; GUREVYCH, I. Out-of-domain framenet semantic role labeling. *In*: **Proceedings of the 15th Conference of**

the European Chapter of the Association for Computational Linguistics. Vol. 1. Long Papers. ACL, 2017. p. 471-482. DOI https://doi.org/10.18653/v1/E17-1045

HERMANN, K.M; DAS, D.; WESTON, J.; GANCHEV, K. Semantic Frame Identification with distributed Word Representations. *In*: **Proceedings of the 52nd Annual Meeting of ACL** (Baltimore, Mariland). ACL, 2014. P. 1448-1458. DOI https://doi.org/10.3115/v1/P14-1136

JOHNSON, C. R.; FILLMORE, C. J. 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, Seattle WA). ACL, 2000. p. 56-62.

KIPPER, K.; KORHONEN, A; RYANT, N.; PALMER, M. A Large-Scale Extension of VerbNet with Novel Verb Classes. *In*: **Proceedings of the 12th EURALEX International Congress** (Turin, Italy, September 2006). 2006. p. 173-184

OHARA, K., FUJII, S.; OHORI, T.; SUZUKI, R.; SAITO, H.; ISHIZAKI, S. The Japanese FrameNet project: an introduction. *In*: **Proceedings of the satellite workshop "Building lexical resources from semantically annotated corpora"** (LREC 2004). ELRA, 2004. p. 9–11.

PALMER, M; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An Annotated Corpus of Semantic Roles. **Computational Linguistics**, 31:1., p. 71-105, 2005. DOI https://doi.org/10.1162/0891201053630264

PALMER, A.; SPORLEDER, C. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. *In*: **Proceedings of the 23rd international conference on computational linguistics**: posters. ACL, 2010. p. 928-936.

REHBEIN, I; RUPPENHOFER, J.; SPORLEDER, C.; PINKAL, M. Adding nominal spice to SALSA - frame-semantic annotation of German nouns and verbs. In: **Proceedings of KONVENS 2012** (Vienna, Austria). 2014. p. 89-97.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M. R. L.; JOHNSON, C. R.; SCHEFFCZYK, J. **FrameNet II**: Extended Theory and Practice. 2010. Available at: http://framenet.icsi.berkeley.edu. (accessed 28. Oct. 2021).

SALOMÃO, M. FrameNet Brasil: um trabalho em progresso. **Calidoscopio** 7. p. 171-182, 2009. DOI https://doi.org/10.4013/cld.2009.73.01

SHI, L.; MIHALCEA, R. Open Text Semantic Parsing Using FrameNet and WordNet. *In*: **Proceedings of HLT-NAACL 2004**, Demonstration Papers. 2004. p. 19-22. DOI https://doi.org/10.3115/1614025.1614031

SUBIRATS, C.; PETRUCK, M. Surprise: Spanish FrameNet!. **Estudios de Lingüística del Español**, n. 31, 2010.

TAULÉ, M. et al. Mapping Syntactic Functions into Semantic Roles. *In*: **Proceedings of TLT2005**. 2005. p. 185-194.

TORRENT, T.T.; ELLSWORTH, M.J. Behind the Labels: Criteria for defining analytical categories in FrameNet Brasil. **Veredas** 17(1), p. 44-65, 2013.

Artigo recebido em: 29.10.2021 Artigo aprovado em: 21.01.2022