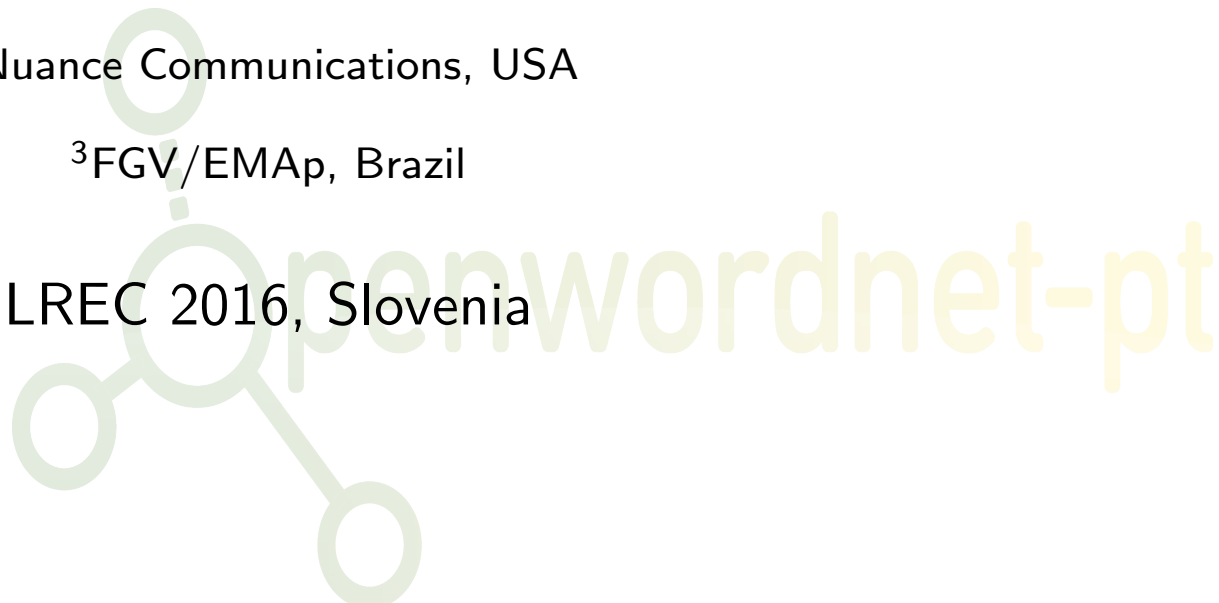
Semantic Links for Portuguese

Fabricio Chalub[1] Livy Real[1]
Alexandre Rademaker[1,3] Valeria de Paiva[2]

[1]IBM Research, Brazil

[2]Nuance Communications, USA

[3]FGV/EMAp, Brazil



LREC 2016, Slovenia

OpenWordnet-PT

http://wnpt.brlcloud.com/wn/

- Goal: not a simple translation of PWN. Based on PWN architecture, a true thesaurus and dictionary for the Portuguese language, based on lexical relations
- Three language strategies in its lexical enrichment process: (i) translation; (ii) corpus extraction; (iii) dictionaries.
- Freely available since Dec 2011. Download as RDF files, query via SPARQL or browse via web interface (above).
- Used by Google Translate, FreeLing, OMW, BabelNet, Onto.PT, etc.

OpenWordnet-PT and DHBB

Motivation

- Side project on historical information extraction from 2014.
- Using highly regarded by Brazilian historians "Dicionário Histórico-Biográfico Brasileiro" (DHBB).
- This is Brazilian Historical and Biographical Dictionary – entries on Brazilian History from 1930 onwards.
- long running project (since 1978) of Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) of the Fundação Getulio Vargas (FGV).
- Data available via `http://cpdoc.fgv.br`, `github.com/cpdoc`
- Previous publication on Digital Humanities Conference.

`http://wnpt.brlcloud.com/kb-extraction/search?db=dhbb&term=*`

Nominalizations

Nominalizations, nouns formed from other POS words, i.e. “construction” and “government”, are one of most well known polysemous and problematic issues of formal theories in Linguistics.

We developed a smaller lexical resource, a lexicon of nominalizations in Portuguese called NomLex-PT, embedded into OpenWordnet-PT, with aprox. 4,240 pairs verb/noun.

Semi-automatically translated the original English NomLex, the French Nomage, the Spanish AnCora-Nom and manually verified.

Worrying about the missing truly Portuguese deverbals, we also used Portuguese corpora (the AC/DC corpora) to complete our collection of nominalizations.

Nominalizations

Cont.

- Nominals have a clear semantic relation with the verb, but their meanings are not automatically derivable from the meaning of the base verb.
- . . . nor are they directly obtainable from the composition between the meaning of the base verb and its suffix.
- *Government*, i.e., has suffix *-ment* which, in general means "the event of doing X", but *government* (and the Portuguese *governo*) has several meanings: the event of governing, the result of governing, the period of time some governing happened, the people that govern, etc.
- We want the nominalization meanings encoded in the lexicon, as their formation can provide more semantic information.
- We started Nomlex without knowing about the PWN semantic links.

Nominalization

Polysemy

Polysemy of the verbs (rows) and nouns (cols):

	1	2	3	4	5	6	7+
1	315	141	63	27	22	8	12
2	246	113	56	29	20	10	12
3	162	93	41	30	9	15	15
4	90	70	32	22	21	13	13
5	61	44	24	18	7	7	11
6	44	28	30	12	2	9	10
7	34	21	11	7	5	3	5
8	27	17	4	15	2	7	5
9	21	10	11	9	5	6	8
10+	62	40	30	17	16	12	33

Morphosemantic links

- Semantically typed relations between verbs and derivationally related nouns.
- *afinar-afinador* (*tune-tuner*) – linked through an *agent* link.
- Many traditional dictionaries include morphological derivations, but simply list without meaning.
- PWN has 17k manually checked typed links but they have not been made part of the official distribution of WordNet.

Nominalization

Cont. Polysemy

One major issue here is the polysemy of both verbs and nouns, related by morphosemantic links.

Number of monosemous (a single sense) in each resource:

	nomlex (OWN-PT)	morphosemantic links (PWN)
verb	963 (22.7%)	1,208 (7.1%)
noun	1,202 (28.3%)	2,832 (16.6%)
both	315 (7.4%)	717 (4.21%)
total	4,238 (100%)	16,995 (100%)

Morphosemantic links from PWN

Relation	Example
agent	*employ-employer*
body-part	*abduct-abductor*
by-means-of	*dilate-dilator*
destination	*tee-tee*
event	*employ-employment*
instrument	*poke-poker*
location	*bath-bath*
material	*insulate-insulator*
property	*cool-cool*
result	*liquefy-liquid*
state	*transcend-transcendence*
undergoer	*employee-employ*
uses	*harness-harness*
vehicle	*kayak-kayak*

Morphosemantic Links to Portuguese

From the 14 examples from PWN only three would work easily in Portuguese.

- The pair *employ-employer* (*empregar-empregador*) also has the *agent* relation in Portuguese.
- We also have the *undergoer* relationship in this case *employ-employee* (*empregar-empregado*) and
- *result* link for *liquefy-liquid* (*liquidificar-líquido*).

Morphosemantic Links to Portuguese

Cont.

- Some do not have direct translation (not specific verb for "teeing" in Portuguese) or the verb is rarely used.
- "*abduct-abductor*" can be used in the muscular sense in Portuguese, but is commonly used in its kidnapping sense. Role would be *agent*.
- Looking up pairs such as "*dilate-dilator*" it help us to complete empty synsets in Portuguese, especially for verbs that are not very commonly used.

This work (ongoing project)

Given a nominalization pair e.g *empregador-empregar* in PT would like to find their senses in English PWN. but too hard task, given the fine-grained nature of Princeton's synsets and the (not fully curated) OpenWordnet-PT.

Thus the work we describe here consists in adding to the pairs of translated morphosemantic links, pairs of senses of verbs/nouns in Portuguese, a label from Princeton's table and making such a triple, a link of the OpenWordnet-PT.

Princeton's morphosemantic links were used to:

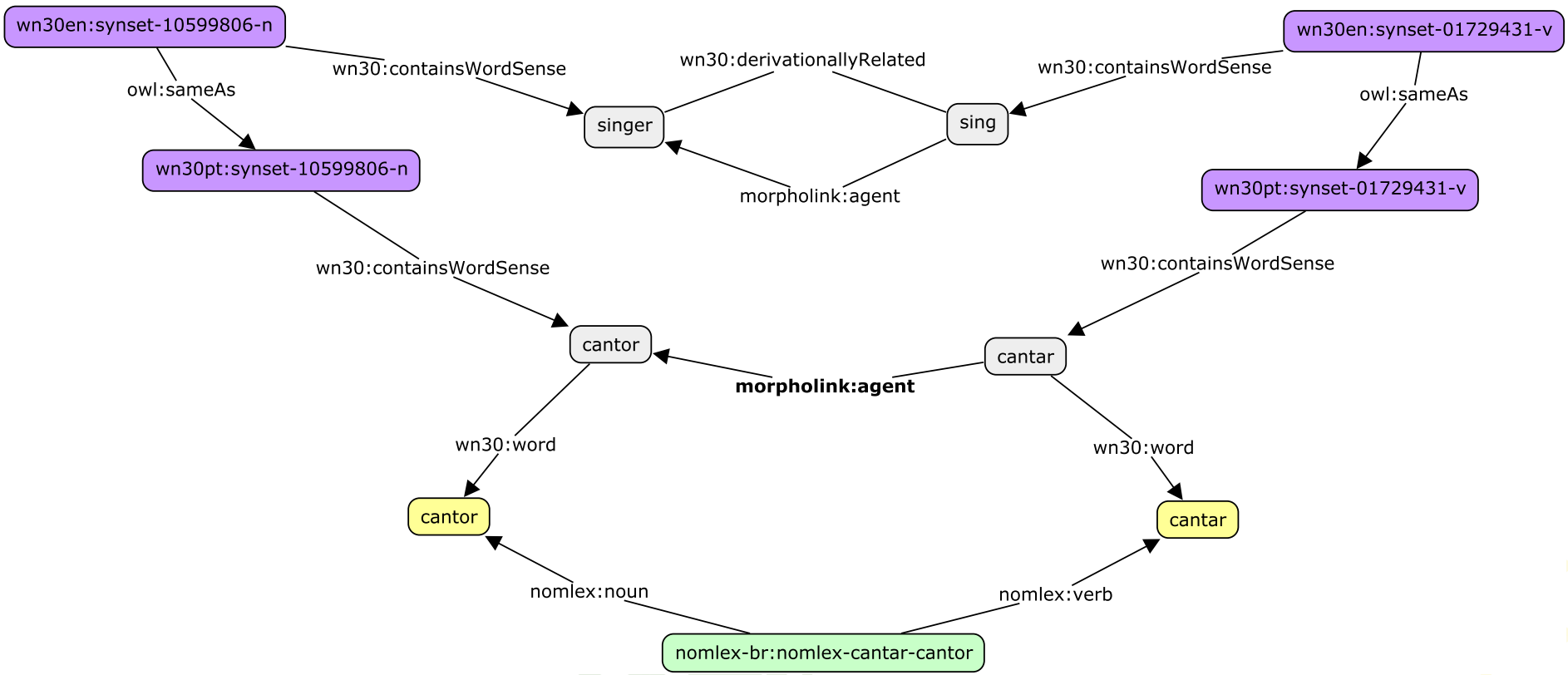
- help data to both discover issues with OpenWordNet-PT synsets
- to help to project the links from NomLex-PT (between words) to senses.

Projecting the morphosemantic links

$$
\begin{aligned}
& \mathit{morpholink}(\mathit{sense}^{v}_{en}, \mathit{sense}^{n}_{en}, \mathit{type}) \\
& \wedge\ \mathit{nomlex}(\mathit{verb}_{pt}, \mathit{noun}_{pt}) \\
& \wedge\ \mathit{sense}(\mathit{sense}^{v}_{en}, \mathit{verb}_{en}, \mathit{ss}^{1}_{en}) \\
& \wedge\ \mathit{sense}(\mathit{sense}^{n}_{en}, \mathit{noun}_{en}, \mathit{ss}^{2}_{en}) \\
& \wedge\ \mathit{same}(\mathit{ss}^{1}_{en}, \mathit{ss}^{1}_{pt}) \\
& \wedge\ \mathit{same}(\mathit{ss}^{2}_{en}, \mathit{ss}^{2}_{pt}) \\
& \wedge\ \mathit{sense}(\mathit{sense}^{v}_{pt}, \mathit{verb}_{pt}, \mathit{ss}^{1}_{pt}) \\
& \wedge\ \mathit{sense}(\mathit{sense}^{n}_{pt}, \mathit{noun}_{pt}, \mathit{ss}^{2}_{pt}) \\
& \qquad \rightarrow \mathit{morpholink}(\mathit{sense}^{v}_{pt}, \mathit{sense}^{n}_{pt}, \mathit{type})
\end{aligned}
$$

Projecting the morphosemantic links

Cont.



Results

- We got 2,735 possible links to add to OpenWordNet-PT.
- Some of them are mistakes from the original wordnet directly, as the pair *confess-confessor* linked by an *agent* (priest that hears the confession, is not the agent).
- Some more mistakes were found when looking at the Portuguese senses. We want to use the incorrect links as opportunities to complete the correct ones.
- When we cannot find all the elements for the rule, say we miss a Portuguese verb sense, we use SPARQL queries to produce candidates.

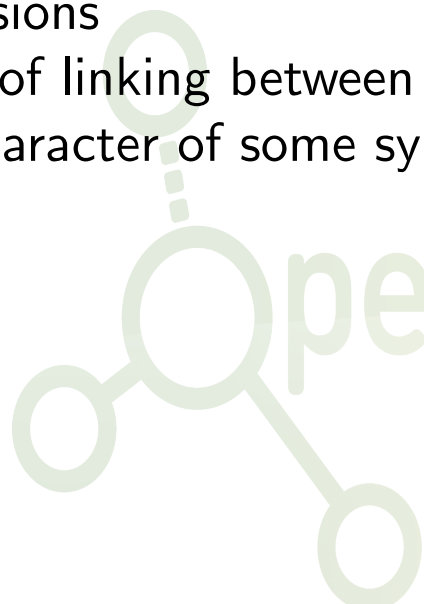
Results

Cont.

- i.e. the pair *bray-bray* (*zurrar-zurro*) for which we did not have the verb *zurrar* in OWN-PT. (Actually we had the word only in the metaphoric meaning of "bray", that is to laugh coarsely.) We miss the literal meaning *making the noise characteristic of donkeys*.
- Clever heuristics, perhaps using the number of senses of a verb or using words in glosses and definitions might be used to narrow down the number of candidate links reasonable to present to the evaluation team.

Conclusions

- Improving an automatically created resource, to make sure that meanings are not mangled and that bad translations are not solidified, is a hard task.
- Adding morphosemantic links can help wordnets to correct:
  - mistakes and omissions
  - failings of sparsity of linking between synsets
  - too fine-grained character of some synsets (GWA is working on the ILI)



Conclusions

Cont.

- Despite the relatively lower numbers of links added if compared with PWN links, the exercise helped us to improve the quality of OWN-PT. Quality is difficult to measure.
- We need to consider ways of evaluating the resource as a whole and the new additions.
- We will start to relate verbs and adjectives *redden-red* (*avermelhar-vermelho*) and pairs of adjectives and adverbs *fast-fast* (*rápido-rapidamente*).
- still debating how to best present information, as PWN info is not available in their interface and we reckon that showing is informative for users both in *en* and in *pt*. Following OWN for the time being.