Extending a Lexicon of Portuguese Nominalizations with Data from Corpora

Cláudia Freitas[1] Valeria de Paiva[2] Alexandre Rademaker[4,5]
Gerard de Melo[3] Livy Real[4] Anne Silva[1]

PUC-Rio

Nuance Comm.

Tsinghua University

IBM Research

FGV/EMAp

October 8, 2014

Getulio Vargas Foundation (FGV)

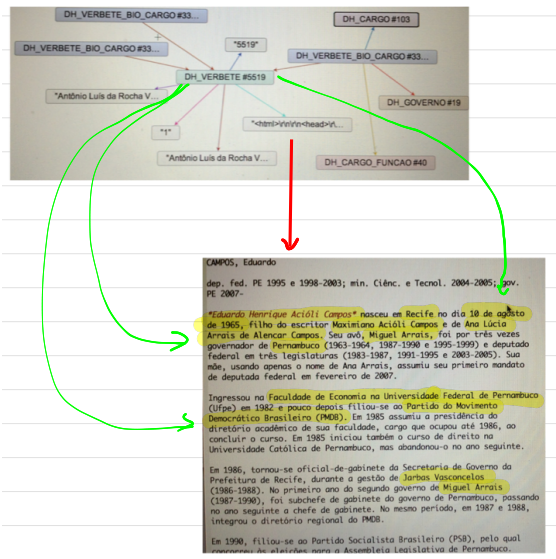
Brazilian higher education and research institution founded in 1944. It offers regular courses of Economics, Business Administration, Law, Social Sciences and Applied Mathematics. Its original goal was to train people for the country's public and private-sector management. Considered a top-5 policymaker think-tank worldwide.

`http://portal.fgv.br`

The Long Run Project

- Joint project between CPDOC (Centro de Pesquisa e Documentação de História Contemporânea do Brasil, School of Social Science) and EMAp (School of Applied Mathematics);
- Enrich the structure (semantics) of CPDOC data;
- Open and expose CPDOC's data and architecture making it more maintainable and dynamic;
- Uniform and integrated data treatment (standards and interlinks between collections).

Project Motivation: CPDOC's DHBB

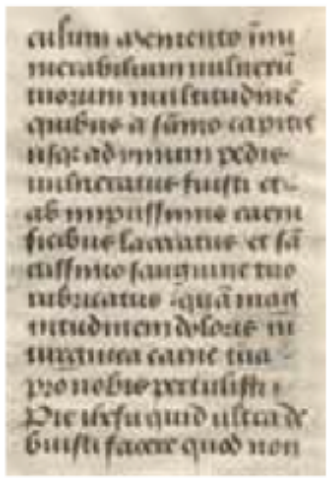


- 7.5K entries Brazilian Historical Biographic Dictionary (DHBB).
- Enrich the structure (semantics). Uniform data treatment (standards and interlinks between collections).
- NLP of DHBB entries: (1) word sense disambiguation with openWordnet-PT; and (2) named entity recognition to make links. (133K proper names)

We need grammars, lexical resources, ontologies, KBs, automated theorem provers etc to reason about knowledge extracted from text. This will empower QA, KE, MT, personal assistents and other systems.

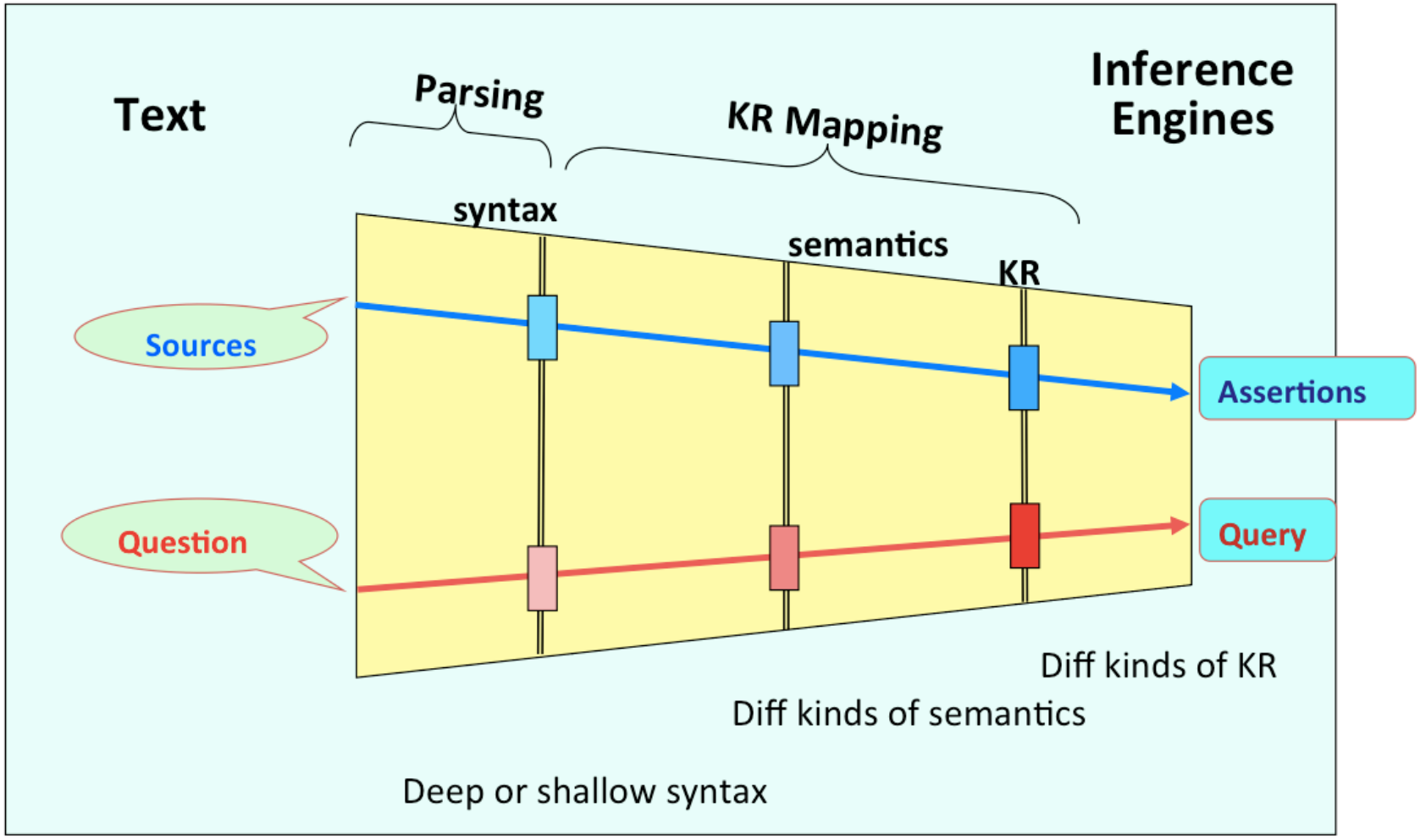
Why we started to build lexical resources for PT?

We need lexical resources for our work, but many previous projects are not openly available.



Simplifying the PARC's Bridge Architecture

Ontologies: SUMO, YAGO, Freebase, WordNet, etc. Logics: TIL, FOL, DL, etc.



The Collaborators

- Alexandre Rademaker: Logician, Proof Theory, Description Logics, KR, Reasoning, ATP.
- Cláudia Freitas: various Linguateca projects.
- Francis Bond: Japanese Wordnet, Malay and Indonesian Wordnets, Open Multilingual Wordnet.
- Gerard de Melo: UWN/MENTA, Lexvo.org, YAGO, FrameNet.
- Livy Real: thesis on nominalizations.
- Valéria de Paiva: PARC/Xerox, Cuil, Nuance. Logician.

Students and posdoc: Dario Augusto, Suemi Higuchi.

Nominalizations

- Lexical databases form an essential component of many modern Natural Language Processing (NLP) systems;
- Portuguese lacks many of the resources available in other languages, for example annotated collections of nominalizations, deverbals or deadjectivals;
- We're developing **NomLex-PT** a comprehensive lexicon of Portuguese nominalizations, integrated to **OpenWordNet-PT**, our open-source version of WordNet for Portuguese.

NOMLEX (Project Proteus)

Alexander's destruction of the city happened in 330 BC.

- a dictionary of English nominalizations, under Catherine Macleod.
- relate the nominal complements to the arguments of the corresponding verb.
- 1025 entries of several types of lexical nominalizations.
- first version on January 15, 1999, latest version October 2001 downloadable from `http://bit.ly/1aZWQmh`

NOMLEX (cont.)

```
(nom :orth "promotion"
     :verb "promote"
     :nom-type ((verb-nom))
     :verb-subj ((n-n-mod) (det-poss))
     :verb-subc ((nom-np :object ((det-poss)(n-n-mod)(pp-of)))
 (nom-np-as-np :object ((det-poss) (pp-of)))
 (nom-possing :nom-subc ((p-possing :pval ("of"))))
 (nom-np-pp :object ((det-poss) (n-n-mod) (pp-of))
    :pval ("into" "from" "for" "to"))
 (nom-np-pp-pp :object ((det-poss) (n-n-mod) (pp-of))
       :pval ("for" "into" "to") :pval2 ("from"))))
```

Using for NLP (IE)

- To write maps bettween IE patterns for active clauses to IE patterns for nominalizations.
- Active clause: "IBM appointed Alice Smith as vice president".
- Passive clause: "IBM's appointment of Alice Smith as vice president" and "Alice Smith's appointment as vice president".

The Proteus Extraction System starts with:

```
np(C-company) vg(appoint) np(C-person) "as" np(C-position)
```

Meta rules to produce passive clause pattern:

```
np(C-person) vg-pass(appoint) "as" np(C-position) "by"
np(C-company)
```

When a pattern matches the input, the pieces corresponding to its constituents are used to build a semantic representation of the pattern (e.g. logical form).

Related Works

- Nominalizations have been studied for more than 4 decades (Chomsky, 1970).
- NomLex-Plus (Meyers et al., 2004). Extension of NOMLEX with 7.050 nominalizations.
- The NomBank Project (Meyer, 2007) `http://bit.ly/1d5G7L9`. "mark the sets of arguments that co-occur with nouns in the PropBank Corpus, just as PropBank records such information for verbs... firmly on the shoulders of NOMLEX..."
- Berkeley FrameNet (`https://framenet.icsi.berkeley.edu/`). 11600 lexical units based on frame semantics supported by corpus evidence. Deverbal nominalizations are annotated as events (in the frame of verbs) or entities/results (diff. semantic frame). FrameNet-Brazil, `http://www.ufjf.br/framenetbr/`.

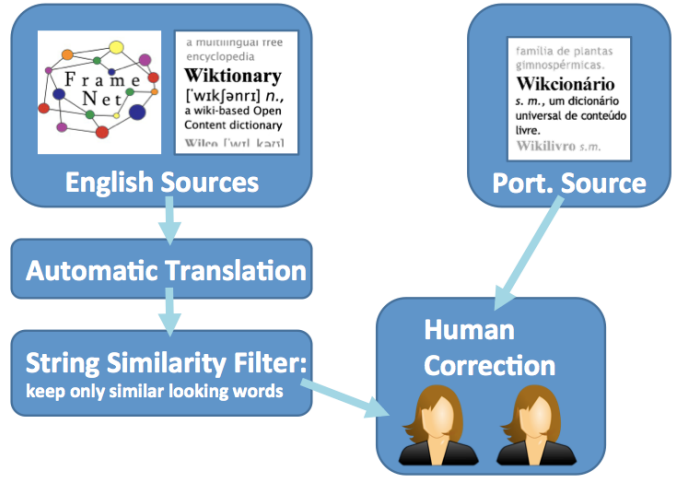
We did not considered NomLex-Plus and NomBank since they were not manually verified.

NomLex-PT

- Extension of OpenWN-PT aims at incorporating links to connect deverbal nouns with their corresponding verbs.
- For English, NOMLEX (Macleod et al., 1998) has provided extensive descriptions of nominalizations via extensions of initial core.
- NOMLEX was constructed starting out with nominalizations with the suffixes -ion, -ment and -er, taking samples of the most frequent words first in a list of nouns from a combination of the Brown Corpus and the Wall Street Journal (about 1 million words of each).
- NOMLEX-BR Translation of initial core, plus French Nomage.
- Overall, we created originally over 2,000 entries. These have been integrated into OpenWN-PT to facilitate its use for linguistic research as well as information extraction.

Growing NomLex-PT

- Starting from the seed of the original NOMLEX in English translated, we grew our lexicon via steps.
- First we added the translation of French collection of nominals NOMAGE
- Then we considered a collection of nominals extracted from Wiktionary and Wikcionario, manually verified.
- Lucky to meet in 2013 Livy Real, who was already working on a PhD thesis on nominalizations.



NomLex-PT (cont.)

- Incorporating NOMLEX-PT data into OpenWN-PT has shown itself useful in pinpointing some issues with the coherence and richness of OpenWN-PT.
- The word **abasement** corresponds in NOMLEX to the verb **abase**, and thus we would like a similar correspondence between the Portuguese noun **aviltamento** and the verb **aviltar** (our suggested translations). OpenWN-PT simply has two synsets {humilhar, abaixar} and {humilhar, rebaixar}. The more common verb **humilhar** is repeated, while the uncommon **aviltar** was left out.
- How can we make sure that NOMLEX-PT has all the most used nominalizations?
- Use Portuguese corpora!

AC/DC project – Linguateca

- more 1 billion words, distributed over genres: general newspaper text, narrative fiction, specialized newspaper text, and other (e-mail spam, EU calls, business letters, legal documents and web texts, especially blogs).
- Candidates could be filter since AC/DC was annotated with PALAVRAS.
- Considered the following five suffixes in the corpus:
  - *-ção* (and its allomorph *-ização*), for example *padronizar - padronização/standard - standardization*;
  - *-ncia* (as in *dominar - dominância/dominate - dominance)*;
  - *-agem* e.g. does not change form in English *(reciclar-reciclagem/recycle-recycle)*;
  - for agentives *-or* (as in *trabalhar-trabalhador/work-worker)*; and *-nte* as in *(estudar-estudante/study-student)*.
- We also considered regressive nominalizations, or zero-derivation nominals, using PAPEL. It is well-known that these make up a great number of Portuguese nominals (for example *atacar - ataque / attack - attack*;*lutar - luta / fight - fight*).

Classification of nominals

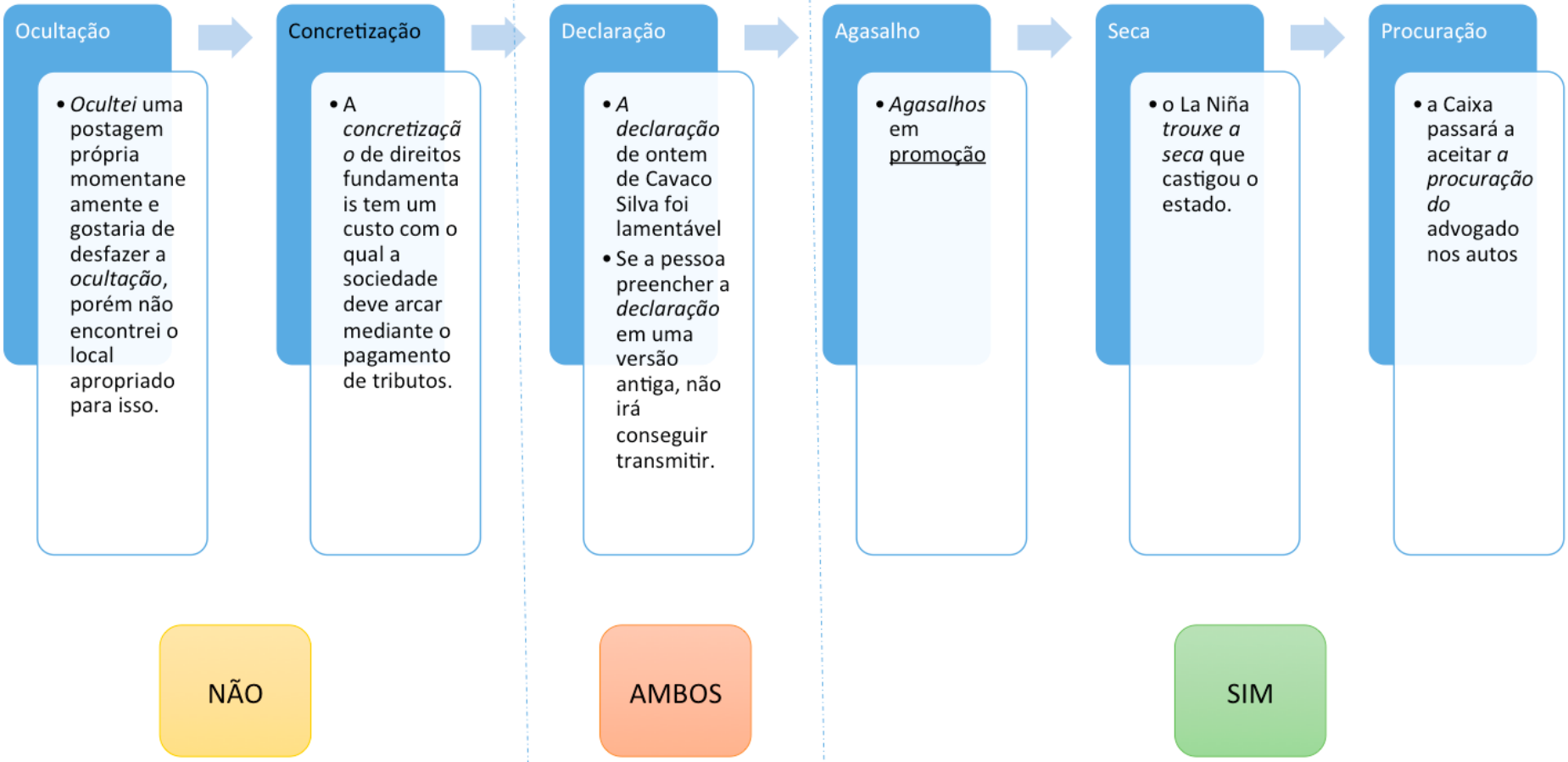
- **Agentive**: the paraphrase one that Xs/ o que Xs (where X is the verb) is possible.
- **Animacy**: A (animate), I (inanimate), or both (underspecified). For instance, **pintor** (**painter**) is animate,**utilização** (**utilization**) is inanimate. the label 'both' applies to nouns that can be used as either, e.g. **recolhedor** (**collector**).

A limpeza do córrego está sendo realizada com barreiras para evitar a disseminação do óleo, com equipamentos de sucção e com *recolhedores* de óleo. **The cleaning of the creek is done with barriers to avoid dissemination of the oil, with suction equipment and with collectors of oil.** (inanimate)
Primeiro, trabalhou como *recolhedor* das apostas e há três anos teria ganho alguns pontos de jogo na zona norte. **First (he) worked as a bet collector and three years ago (he) was supposed to have won some game spots in the North Zone.** (animate)

Lexicalization

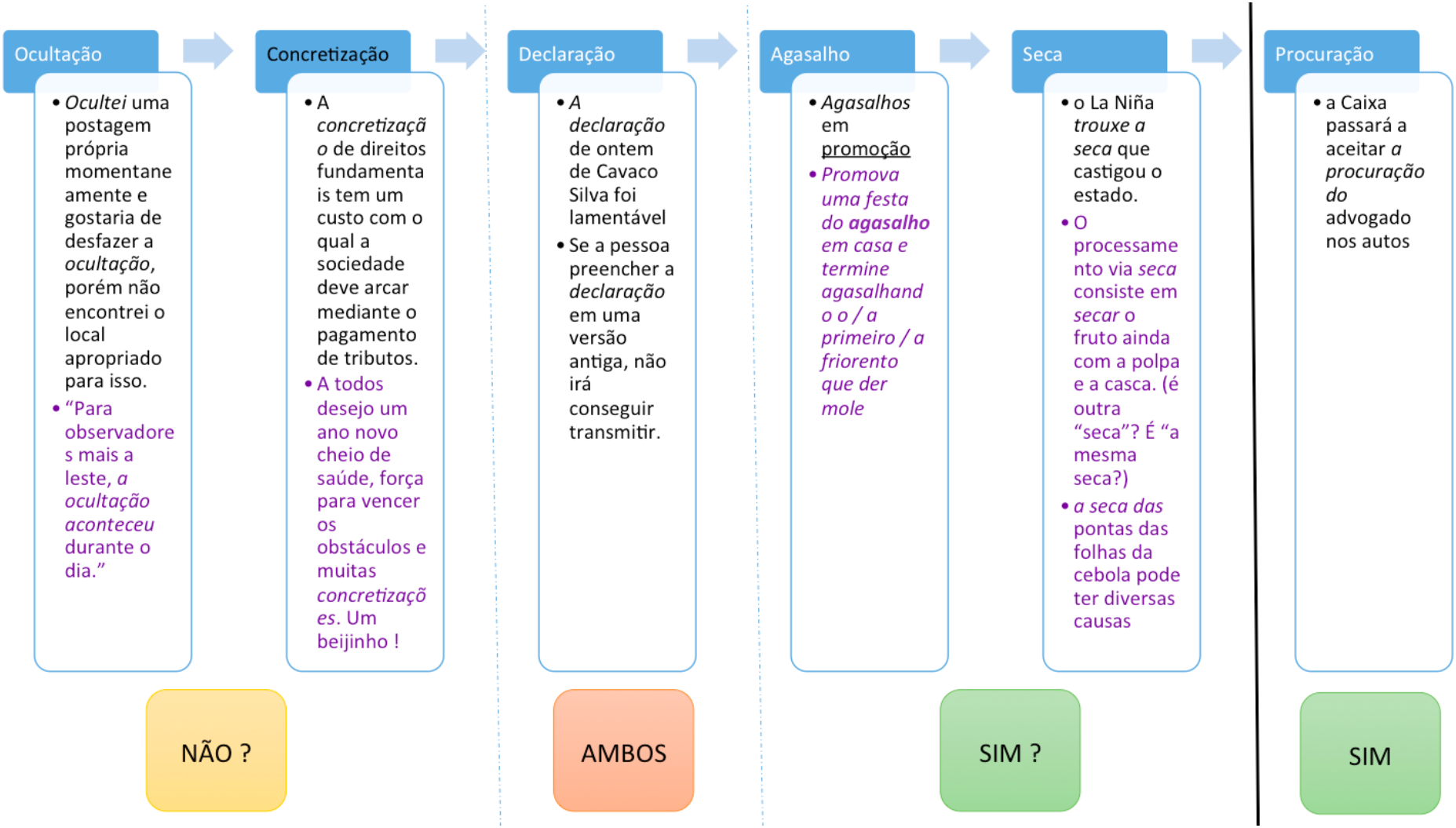
- Harder to mark
- Origin of most controversy
- Lexicalized nominals are those in which the actual meaning of the word no longer corresponds to the meaning resulting from the morphological process.
- Example: **procuração** ('power of attorney' in English) does not mean the act of **procurar** (**to search**).
- It is not always easy to decide whether we are facing a lexicalized nominal or not. The lexicalization process would be better described as a continuum, and some words can be more lexicalized than others.

NomLex-PT: lexicalized or not

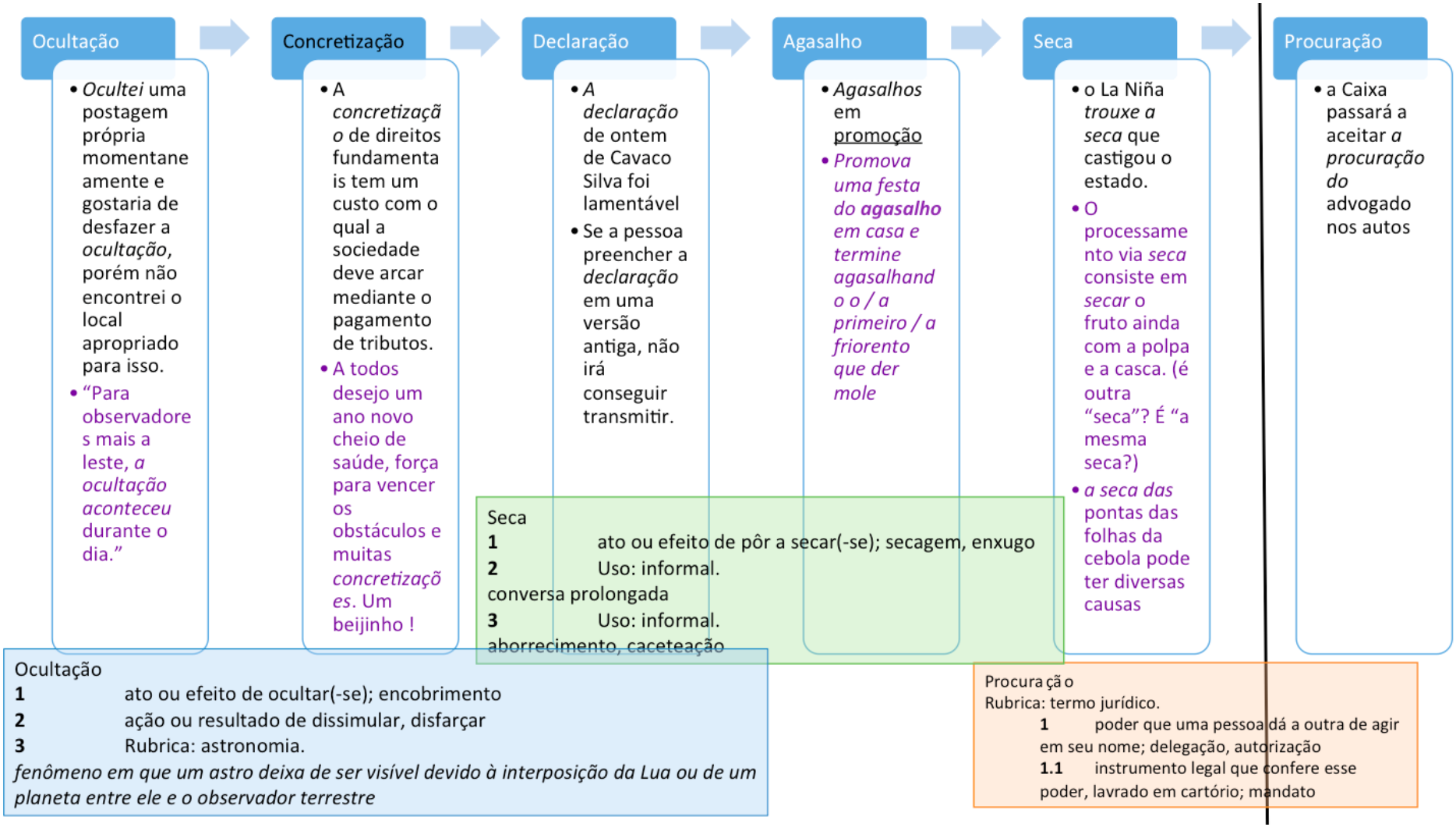


concealment (the act of hiding); the act of making something concrete; declaration; coat (coating); dry; power of attorney)

NomLex-PT: lexicalized or not



NomLex-PT: lexicalized or not



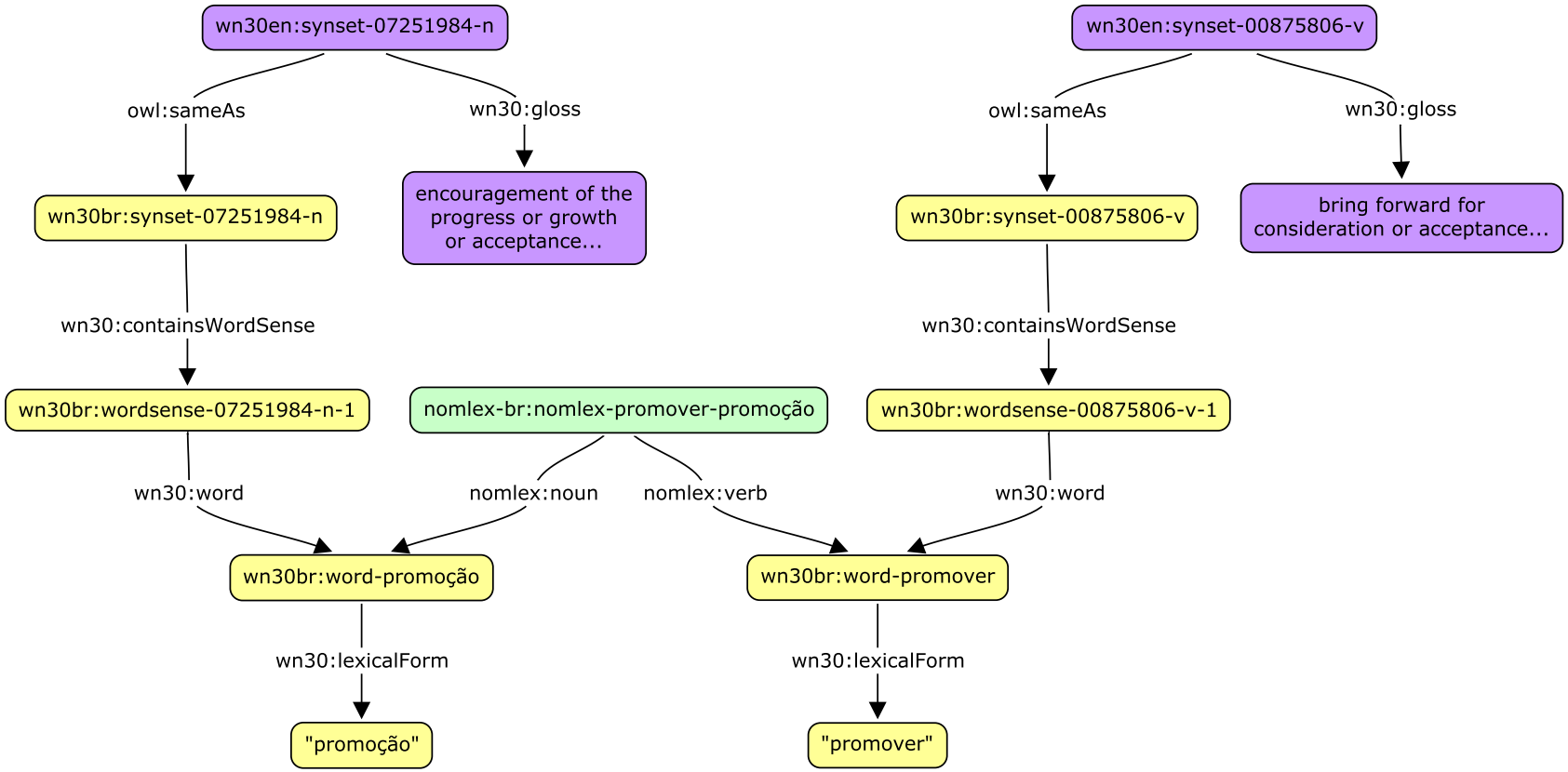
RDF Representation

- Interoperability between lexical resources. Linked Data and Semantic Web standards such as RDF and OWL.
- The emergence of Linked Data projects for lexical and reasoning resources make OpenWN-PT/NomLeX-PT encoded and distributed in RDF/OWL.
- Standards allow both data model and data in the same format. Tools including databases (triple stores) with SQL-like query interfaces (SPARQL). Schema Free.
- Standard W3C encoding of WordNet in RDF since 2006 [1]. OpenWN-PT is modelled after and fully interoperable with Princeton WordNet. Our own lisp parser [2].
- Part of a large ecosystem of compatible resources.

[1] `http://www.w3.org/TR/wordnet-rdf/`

[2] `https://github.com/arademaker/wordnet2rdf`

RDF Representation: NomLex-PT and OWN-PT



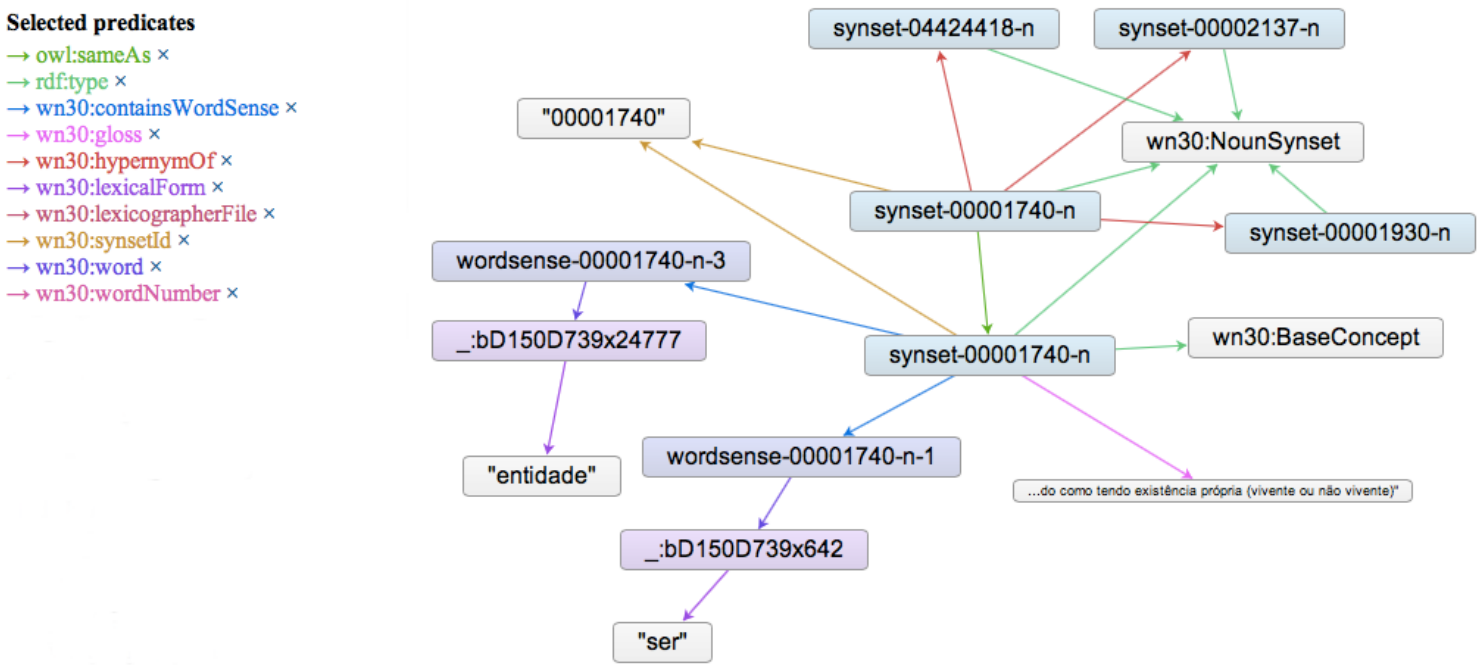
But **nomlex:noun** and **nomlex:verb** should point to **wn30:WordSense** not **wn30:Word**! Future work!

Some queries

- By provenance `http://bit.ly/Mohmni`
- By suffix `http://bit.ly/LmAXn4`; `http://bit.ly/1fKEnKr` and `http://bit.ly/1fyia3a`.

RDF Representation (cont.)

One can easily find Portuguese equivalents for specific English word senses and vice versa. See `http://bit.ly/1aPxd7J`.



URIs for name resources

- `http://arademaker.github.com/nomlex/schema/`
- `http://arademaker.github.com/wn30/schema/`
- `http://arademaker.github.com/own-pt/instances/`
- `http://arademaker.github.com/nomlex-pt/instances/`

We are still thinking in better and stable URIs! Persistent Uniform Resource Locators (PURL)? Other domain?

Participating in the Ontology-Lexicon W3C Group.

URIs are important! They allow interoperability, identify provenance and ownership.

Conclusion

- We discussed the construction and improvement of NOMLEX-PT and its connection to OpenWordNet-PT, an open WordNet for Portuguese.
- Recent improvements include better coverage (via the AC/DC corpora) and nominalization links connecting nouns and verbs.
- The resource has been used in developing a high-throughput commercial system as well as in a cultural heritage project, and we anticipate that numerous further applications will follow.
- Freely available from `http://github.com/arademaker/openWordnet-PT` and `http://github.com/arademaker/nomlex-pt` and a SPARQL Endpoint at `http://logics.emap.fgv.br:10035`.
- Browsing via Open Multilingual Wordnet is fun.

Next steps

- We are developing our own web interface for browsing and collaborative editing. Most important pending issue!
- First finish translating the "core" synsets in the Princeton WordNet to Portuguese.
- Finish to embed Nomlex-PT into OpenWN-PT (anchor floating words, `http://bit.ly/1aQdpkr`).
- Since we have a first target corpus, DHBB, we can also calculate word frequency to prioritize expansion of the OpenWN-PT and go back to the ontology building.
- Use and test the accuracy of the resource! **IBM plans to bring Watson to Brazil**! (`http://goo.gl/9V7Z5q`)
- OpenVerbNet-PT?
- Possibly glosses and automatic suggestions of extensions with Hugo Gonçalo Oliveira (from Onto.PT).

Thanks! Obrigado!