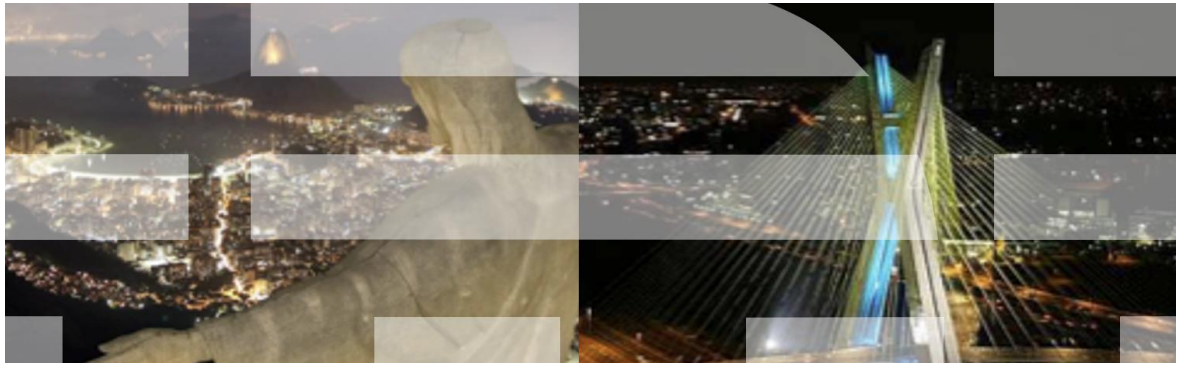
Language resources for natural language ~~processing~~ understanding



Alexandre Rademaker

IBM Research

Brazil

Different levels of understanding

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

→ Bram Stoker

What is understanding?

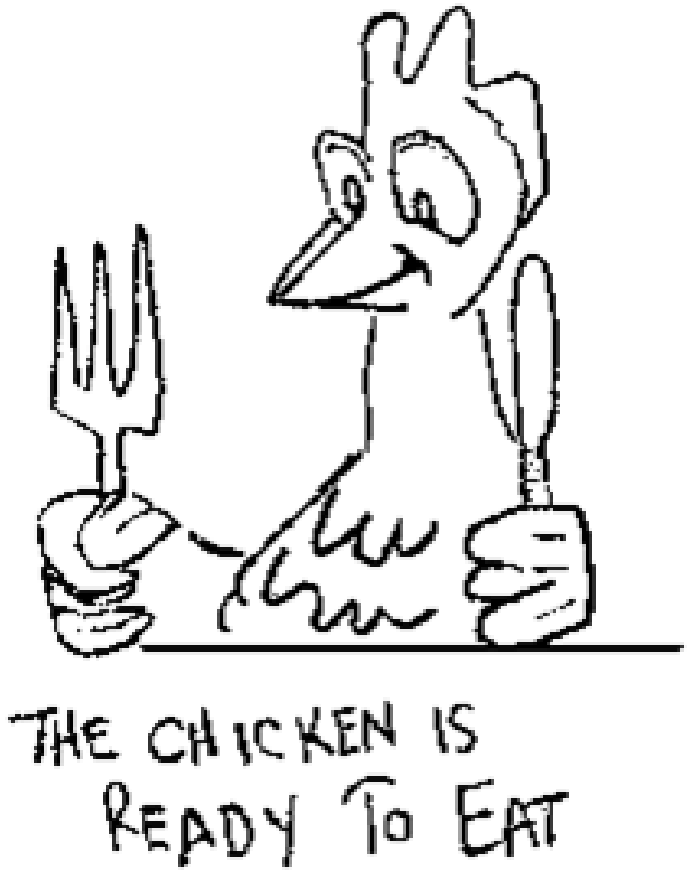
- There is no one universal measure
- one evaluation metric: the detection of entailment and contradiction relations between (portions of) texts.
- RTE is not a *sufficient* criterion for LU. But RTE is a minimal, *necessary* criterion.
- If you understand sentences (1) and (2), then you can recognize that they are contradictory. If you fail to recognize the contradiction, then you cannot have understood.

(1) No civilians were killed in the Najaf suicide bombing.

(2) Two civilians died in the Najaf suicide bombing.

common sense, knowledge

- Domain knowledge
  - “…produced from the Muddy sands of Cretaceous age…”
  - “…produced from the Muddy sands from 135 millions year ago…”
- I shot a…
  - “I shot a picture”
  - “I shot a person”
- “I saw a man with a telescope”
  - “I saw a man with a telescope in the picture.”
  - “I saw a man with a telescope in the terrace.”



7,102

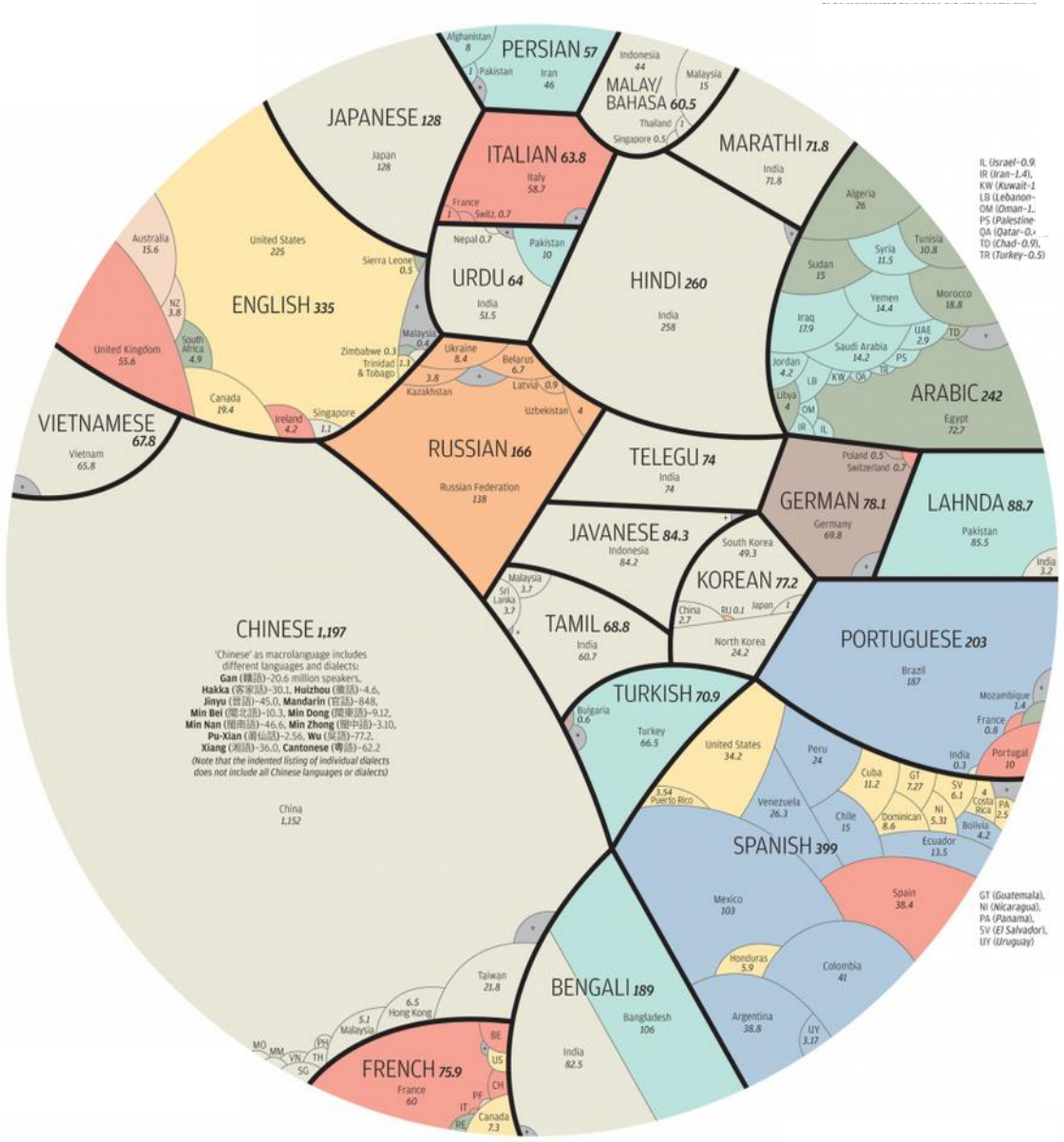
known languages

23

most spoken language

4.1+ Billion

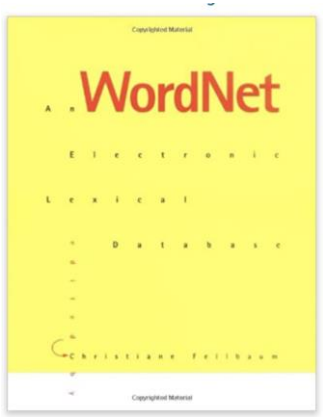
people



https://www.iflscience.com/environment/worlds-most-spoken-languages-and-where-they-are-spoken/

lexical semantics: word-sense disambiguation

- WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.
- The Global Wordnet Association



Global WordNet Association

They can fish

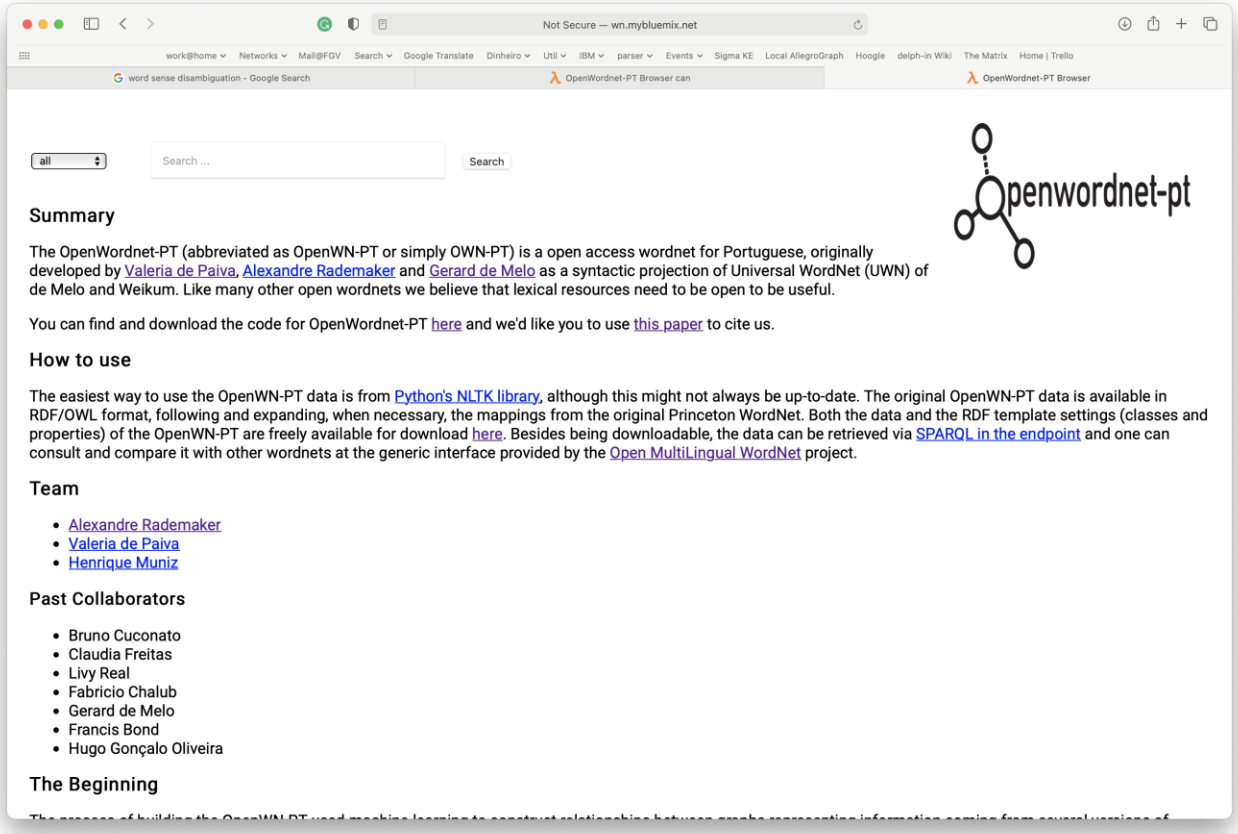
WordNet

- S: (n) **mouse**
  - S: (n) rodent, gnawer
- S: (n) shiner, black eye, **mouse**
  - S: (n) bruise, contusion
- S: (n) **mouse**
  - S: (n) person, individual, someone, somebody, mortal, soul
- S: (n) **mouse**, computer mouse
  - S: (n) electronic device

*"a **mouse** takes much more room than a trackball"*

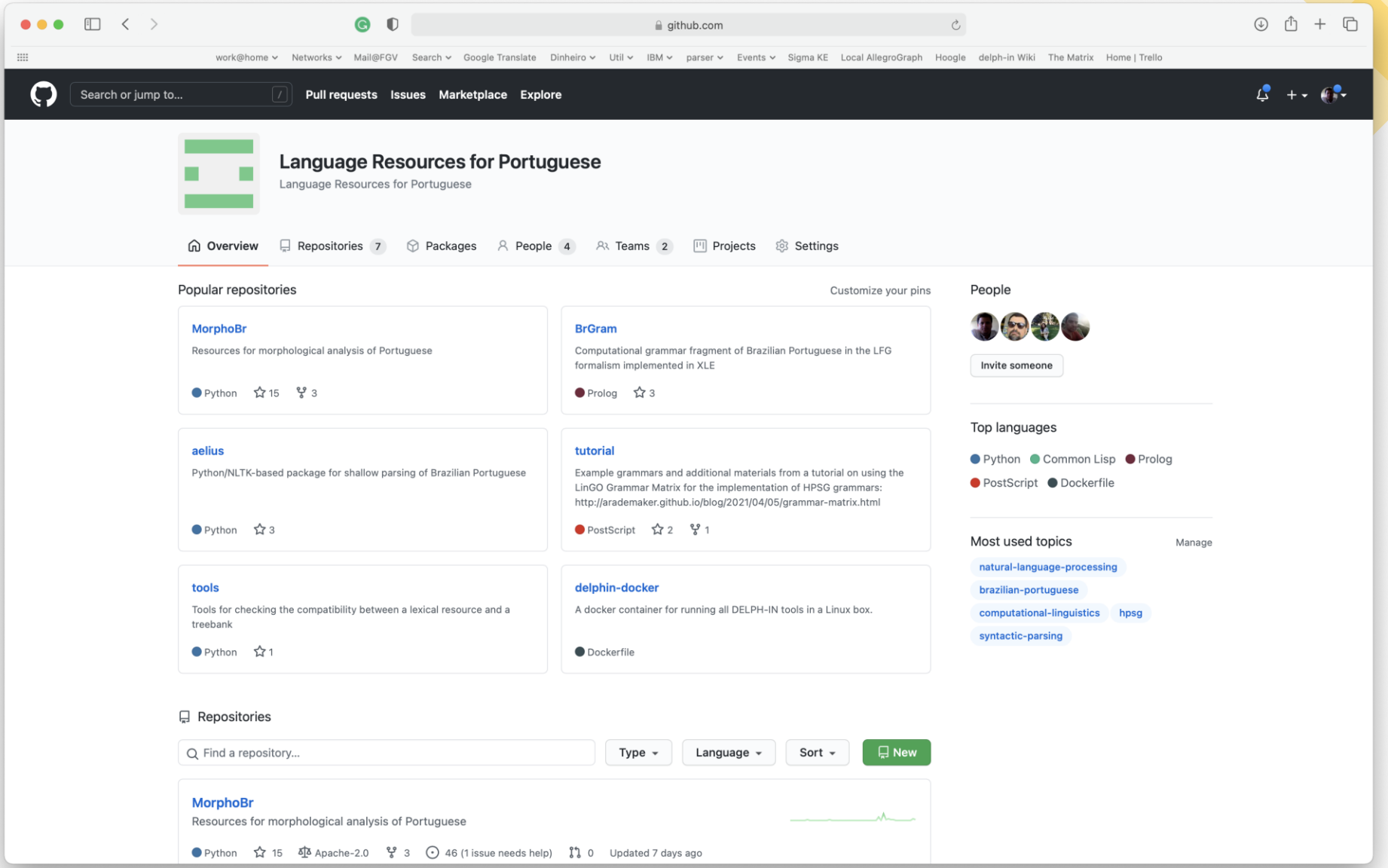
The OpenWordnet for Portuguese (OWN-PT)

- The more comprehensive wordnet for Portuguese
- Freely available for browsing and download (RDF)
- Used by many projects and tools (Freeling, Google Translate, etc)
- Incorporate in many other resources (BabelNet, Open Multilingual Wordnet, etc)
- Under development since 2010



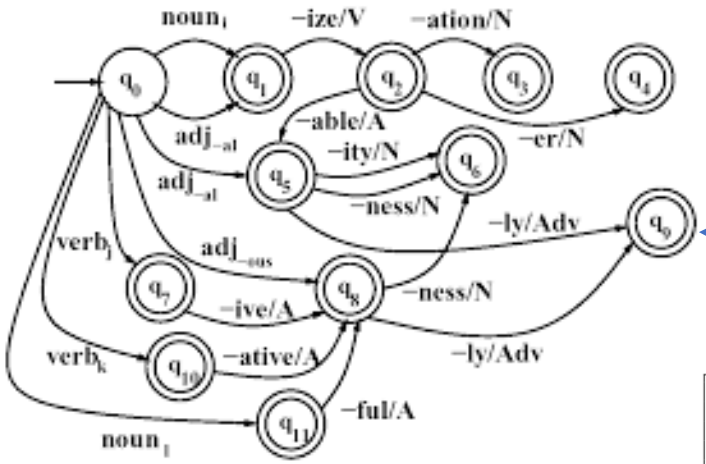
http://openwordnet-pt.org

Paiva, Valeria de, Alexandre Rademaker, and Gerard de Melo. 2012. "OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning." In *Proceedings of COLING 2012: Demonstration Papers*, 353–60. Mumbai, India: The COLING 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-3044.


github.com
work@home
Networks
Mail@FGV
Search
Google Translate
Dinheiro
Util
IBM
parser
Events
Sigma KE
Local AllegroGraph
Hoogle
delph-in Wiki
The Matrix
Home | Trello
Search or jump to...
Pull requests
Issues
Marketplace
Explore
Language Resources for Portuguese
Language Resources for Portuguese
Overview
Repositories 7
Packages
People 4
Teams 2
Projects
Settings
Popular repositories
Customize your pins
MorphoBr
Resources for morphological analysis of Portuguese
Python 15 3
BrGram
Computational grammar fragment of Brazilian Portuguese in the LFG formalism implemented in XLE
Prolog 3
aelius
Python/NLTK-based package for shallow parsing of Brazilian Portuguese
Python 3
tutorial
Example grammars and additional materials from a tutorial on using the LinGO Grammar Matrix for the implementation of HPSG grammars: http://arademaker.github.io/blog/2021/04/05/grammar-matrix.html
PostScript 2 1
tools
Tools for checking the compatibility between a lexical resource and a treebank
Python 1
delphin-docker
A docker container for running all DELPH-IN tools in a Linux box.
Dockerfile
Repositories
Find a repository...
Type
Language
Sort
New
MorphoBr
Resources for morphological analysis of Portuguese
Python 15 Apache-2.0 3 46 (1 issue needs help) 0 Updated 7 days ago
People
Invite someone
Top languages
Python Common Lisp Prolog
PostScript Dockerfile
Most used topics
Manage
natural-language-processing
brazilian-portuguese
computational-linguistics
hpsg
syntactic-parsing

MorphoBr - morphological analysis of Portuguese

- finite state morphology
- coverage
  - 538,081 nouns
  - 2,373,600 verbs
  - 543,694 adjectives
  - 21,182 adverbs
- consistency among resources
  - OWN-PT
  - UD Corpus
  - HPSG Grammar for Portuguese
- Consolidation of PT-PT and PT-BR

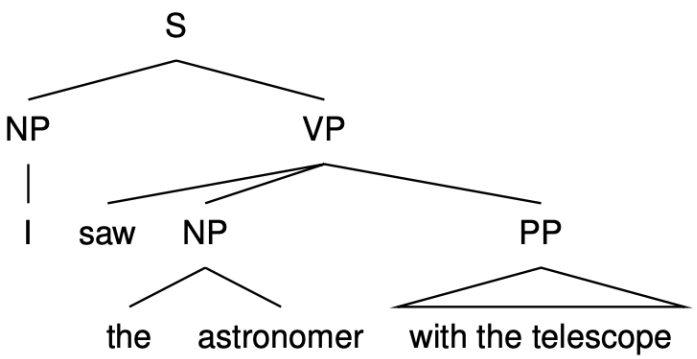


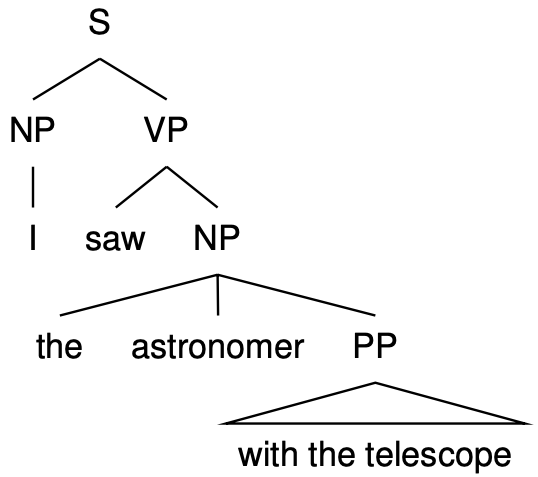

```
simples  simples+A+F+PL
simples  simples+A+F+SG
simples  simples+A+M+PL
simples  simples+A+M+SG
simplicíssimas   simples+A+SUPER+F+PL
simplérrimos     simples+A+SUPER+M+PL
simplíssima      simples+A+SUPER+F+SG

trabalhadeira    trabalhador+A+F+SG
trabalhadeirazinhas      trabalhador+A+DIM+F+PL
trabalhadeirazita        trabalhador+A+DIM+F+SG
trabalhadeirinha         trabalhador+A+DIM+F+SG
trabalhadora     trabalhador+A+F+SG
trabalhadorinha  trabalhador+A+DIM+F+SG
trabalhadorazinhas       trabalhador+A+DIM+F+PL
trabalhadorazita         trabalhador+A+DIM+F+SG
trabalhadores    trabalhador+A+M+PL
trabalhadorzinho         trabalhador+A+DIM+M+SG
trabalhadorezinhos       trabalhador+A+DIM+M+PL
trabalhadorzinhos        trabalhador+A+DIM+M+PL
trabalhadorzitos         trabalhador+A+DIM+M+PL

feliz    feliz+A+F+SG
feliz    feliz+A+M+SG
felizes  feliz+A+F+PL
felizes  feliz+A+M+PL
felicíssimas     feliz+A+SUPER+F+PL
felizão  feliz+A+AUG+M+SG
felizona         feliz+A+AUG+F+SG
felizãozinho     feliz+A+AUG+DIM+M+SG
felizonazinha    feliz+A+AUG+DIM+F+SG
felizoninha      feliz+A+AUG+DIM+F+SG
felizezinhas     feliz+A+DIM+F+PL
felizezinhos     feliz+A+DIM+M+PL
felizinha        feliz+A+DIM+F+SG
felizinhas       feliz+A+DIM+F+PL
felizinho        feliz+A+DIM+M+SG
felizinhos       feliz+A+DIM+M+PL
```

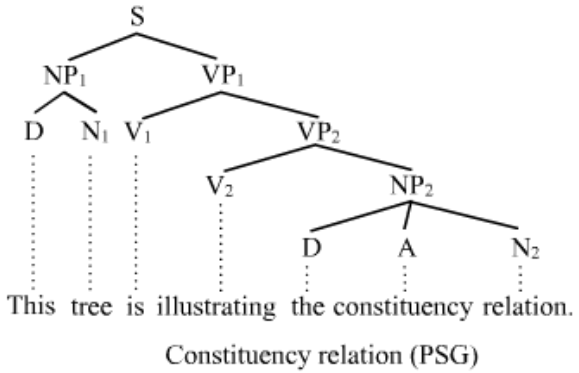

Alencar, Leonel Figueiredo de, Bruno Cuconato, and Alexandre Rademaker. 2018. “MorphoBR: An Open Source Large-Coverage Full-Form Lexicon for Morphological Analysis of Portuguese.” *Texto Livre: Linguagem e Tecnologia* 11 (3): 1–25. https://doi.org/https://doi.org/10.17851/1983-3652.11.3.1-25.

Parsing

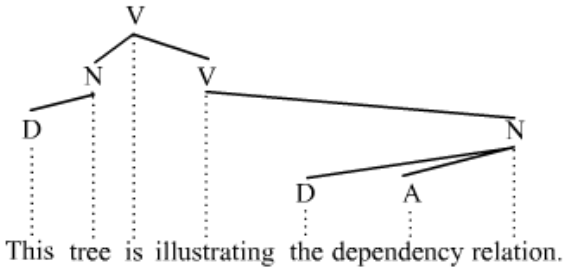
- Recognizing string as input and assigning structure to it
- Parsing: Making explicit structure that is inherent (implicit) in natural language strings
  - What is that structure?
  - Why would we need it?







Constituency relation (PSG)



Dependency relation

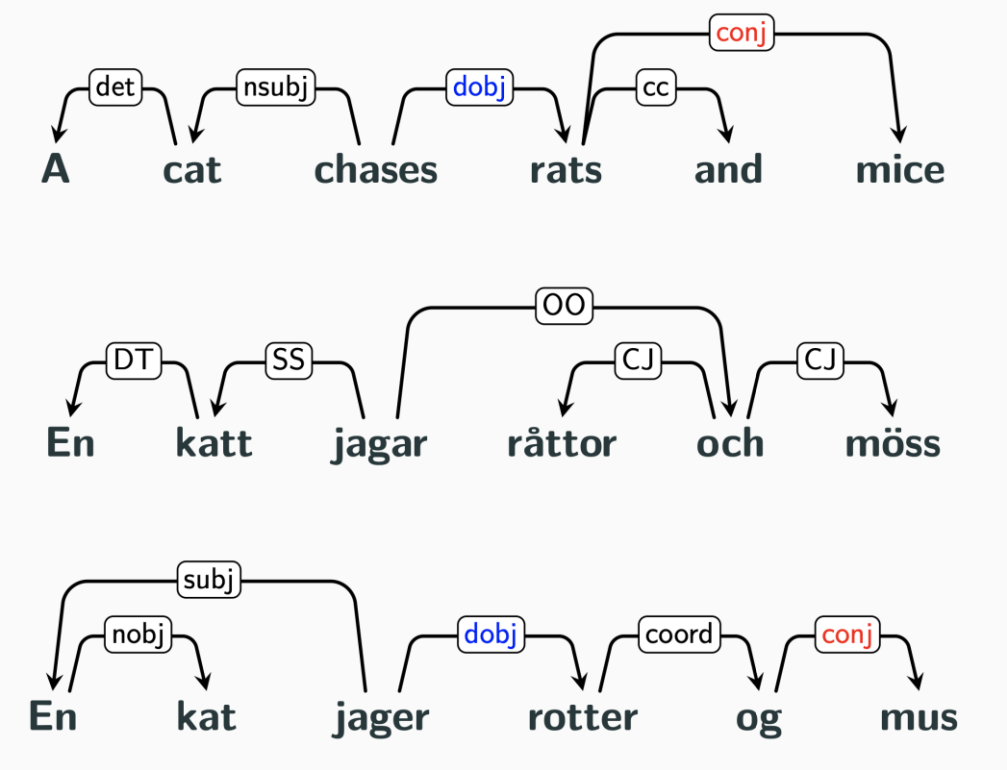
- **Syntactic** parsing: assigning syntactic structure
- **Semantic** parsing: assigning semantic structure

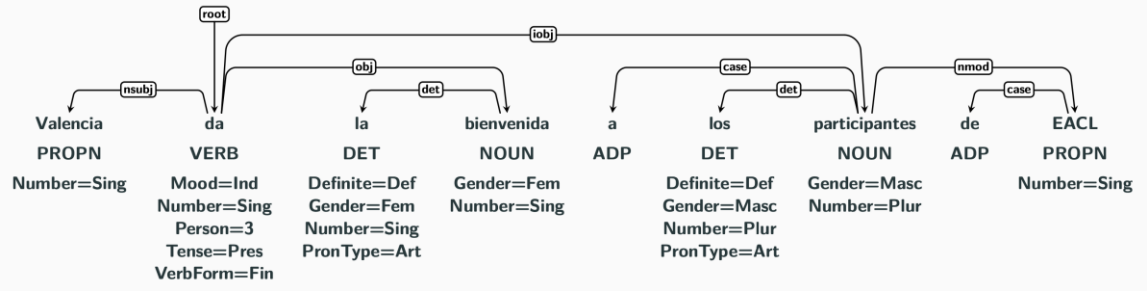
"A grammar is better, but in practice people use language models."
D. Jurafsky

"To produce a statistically based simulation of ... a [bee] dance without attempting to understand why the bee behaved that way... is ...a notion of [scientific] success that's very novel. I don't know of anything like it in the history of science." Chomsky

Universal Dependencies

- Increasing interest in multilingual NLP
  - Multilingual evaluation campaigns to test generality
  - Cross-lingual learning to support low-resource languages
- Increasing awareness of methodological problems
  - Current NLP relies heavily on annotation
  - Annotation schemes vary across languages
- UD Design principles





http://universaldependencies.org

202 treebanks, 114 languages, released May 15, 2021

UD Design Principles

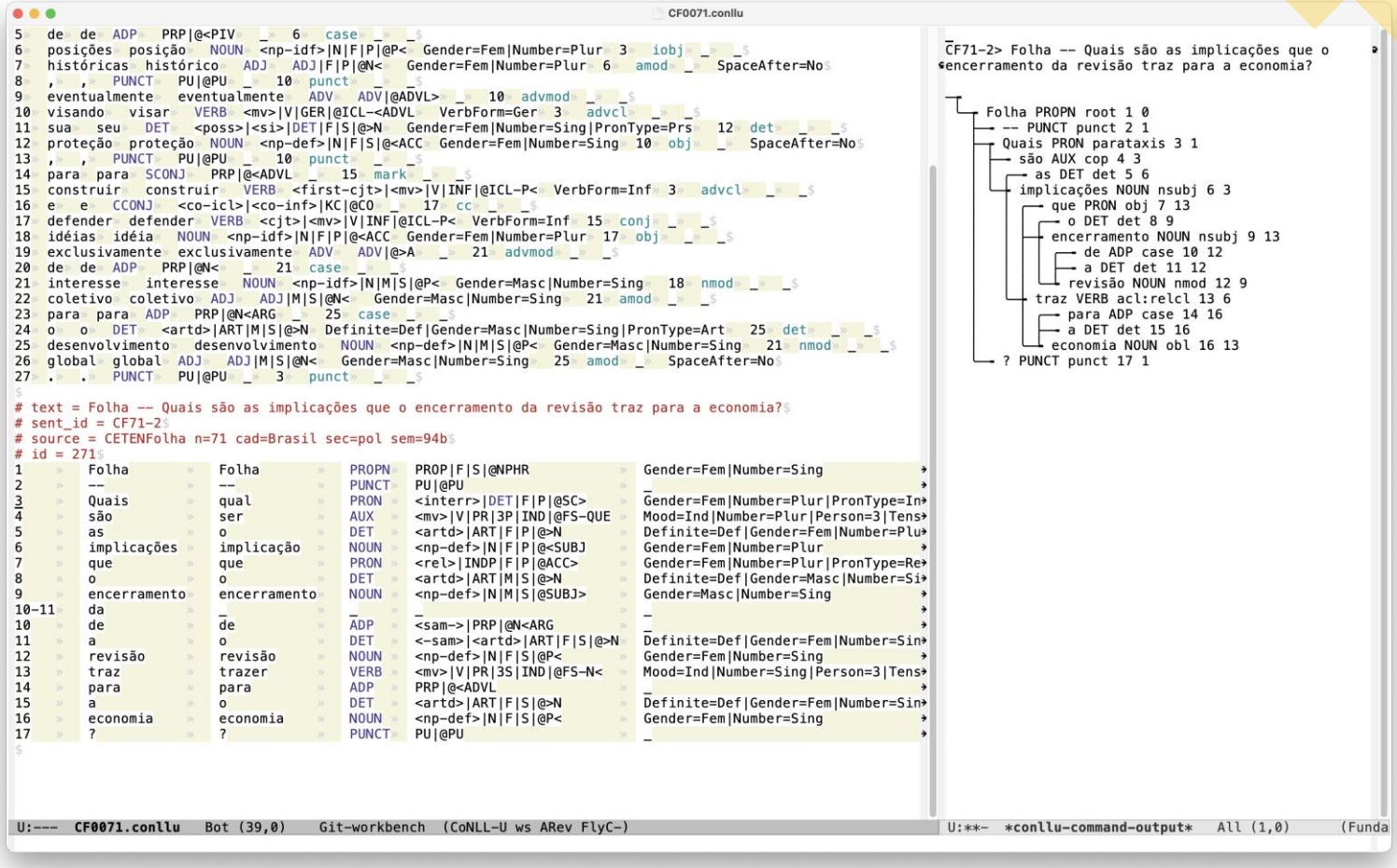
- UD needs to be reasonably satisfactory on linguistic analysis grounds for individual languages—a journeyman's universal grammar.
- UD needs to be good for linguistic typology: It should bring out crosslinguistic parallelism across languages and language families.
- UD must be suitable for rapid, consistent annotation by a human annotator.
- UD must be easily comprehended and used by non-linguist users with prosaic needs.
- UD must be suitable for computer parsing with high accuracy.
- UD must support well downstream language understanding tasks, such as relation extraction, reading comprehension, machine translation, and so on.

CoNLL-U Format

```
# sent_id = 1
# text = They buy and sell books.
1   They     they    PRON    PRP    Case=Nom|Number=Plur                2   nsubj   2:nsubj|4:nsubj   _
2   buy      buy     VERB    VBP    Number=Plur|Person=3|Tense=Pres     0   root    0:root            _
3   and      and     CONJ    CC     _                                   4   cc      4:cc              _
4   sell     sell    VERB    VBP    Number=Plur|Person=3|Tense=Pres     2   conj    0:root|2:conj     _
5   books    book    NOUN    NNS    Number=Plur                         2   obj     2:obj|4:obj       SpaceAfter=No
6   .        .       PUNCT   .      _                                   2   punct   2:punct           _

# sent_id = 2
# text = I have no clue.
1   I       I       PRON    PRP   Case=Nom|Number=Sing|Person=1     2   nsubj   _   _
2   have    have    VERB    VBP   Number=Sing|Person=1|Tense=Pres   0   root    _   _
3   no      no      DET     DT    PronType=Neg                      4   det     _   _
4   clue    clue    NOUN    NN    Number=Sing                       2   obj     _   SpaceAfter=No
5   .       .       PUNCT   .     _                                 2   punct   _   _
```

Universal Dependencies for Portuguese

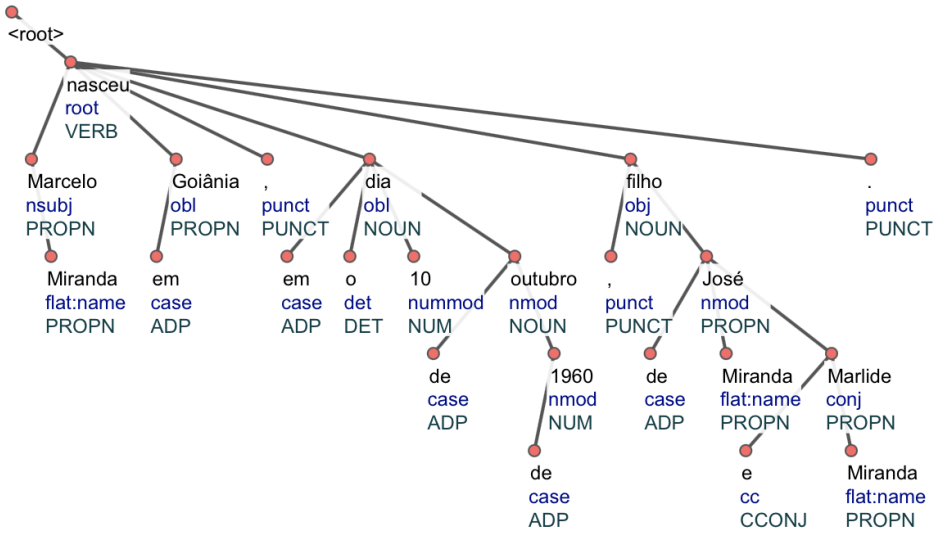
- Portuguese Corpora
  - Bosque (9K sentences vs English EWT 16K sentences)
  - GSD
  - PUD
  - DHBB (TBA, 300K sents)
  - 5 indigenous languages
- GitHub Issues (143 open for Bosque)
- Tools: editor, search/browsing and visualization, software library for batch processing etc.
- We need consistent annotations
- Long Term project
- Efficient collaboration



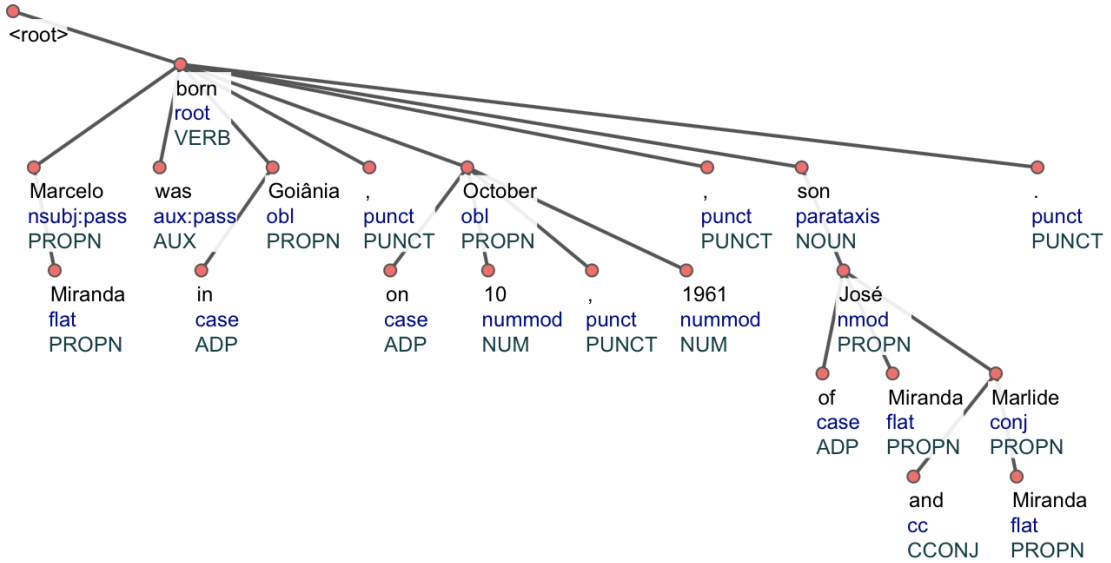
Rademaker, Alexandre, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva Universal Dependencies for Portuguese. 2017. “Universal Dependencies for Portuguese.” In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, 197–206. Pisa, Italy. https://www.aclweb.org/anthology/W17-6523/.

Universal Dependencies

Marcelo Miranda nasceu em Goiânia, no dia 10 de outubro de 1961, filho de José Edmar Miranda e Marlide Miranda.



Marcelo Miranda was born in Goiânia, on October 10, 1961, son of José Miranda and Marlide Miranda.

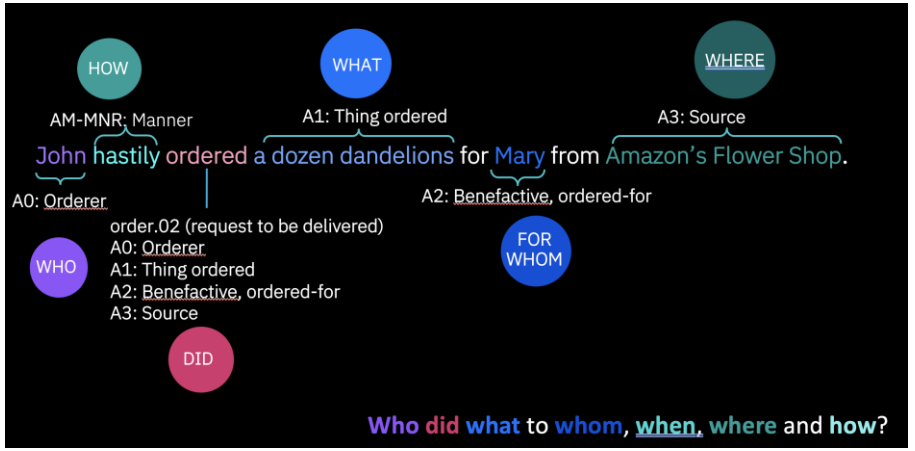


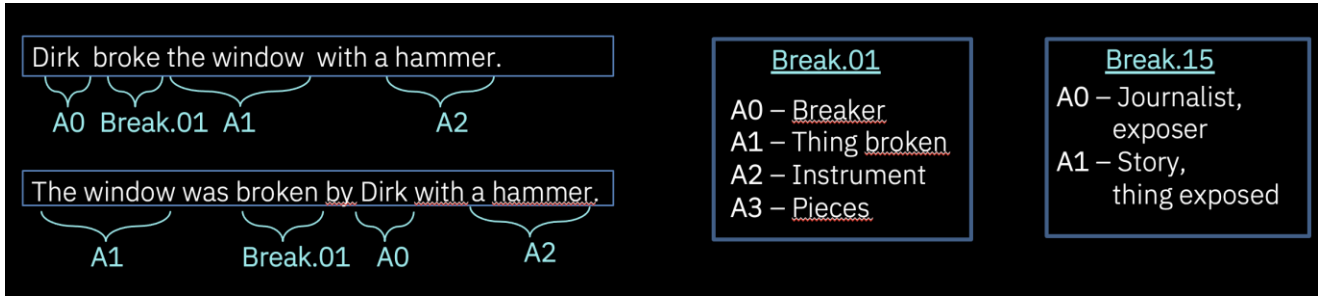
semantic representations: semantic role labelling

http://propbank.github.io

Universal Proposition Banks: UD + SRL

https://github.com/System-T/UniversalPropositions

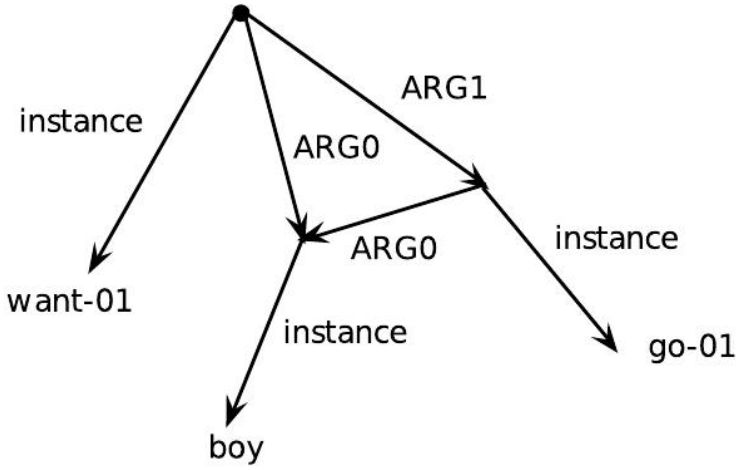




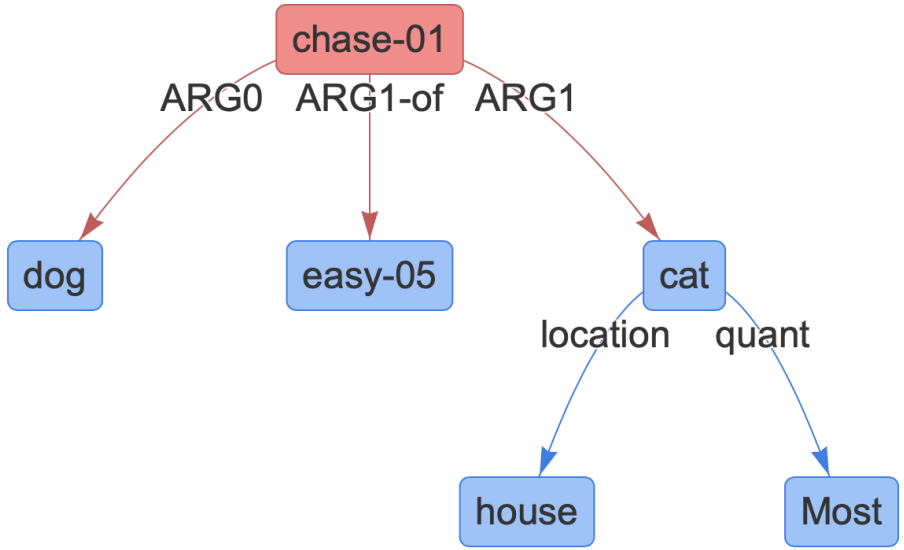
semantic representation: abstract meaning representation (AMR)

The boy wants to go

```
(w / want-01
   :ARG0 (b / boy)
   :ARG1 (g / go-01
            :ARG0 b))
```



Most house cats are easy for dogs to chase.



"we observe a tendency which we view as ill-advised, to conflate sentence meaning and speaker meaning into a single mapping, whether done by annotators or by a parser."

Bender, E. M., Flickinger, D., Oepen, S., Packard, W., & Copestake, A. (2015, April). Layers of interpretation: On grammar and compositionality. In Proceedings of the 11th international conference on Computational Semantics (pp. 239-249).

‘deep’ linguistic processing of human language

- DELPH-IN Consortium is a collaboration among computational linguists
- highly lexicalized, constraint-based grammar, Head-Driven Phrase Structure Grammar (HPSG) and
- Minimal Recursion Semantics (MRS)
- grammar engineering
  - the grammar MATRIX
  - lexicon acquisition
- language per se vs information encoded in language

rain-spattered window with a scene outside (patterns the raindrops make in the window vs. scene outside)

https://youtu.be/ax6Kal18Ki4

Natural Language Processing with Language in Focus, Emily M. Bender

The HPSG English Grammar

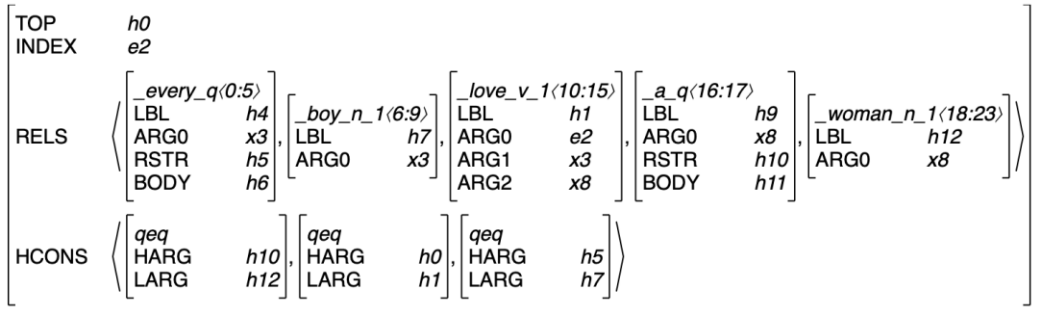
http://delph-in.github.io/delphin-viz/demo/

Most house cats are easy for dogs to chase.

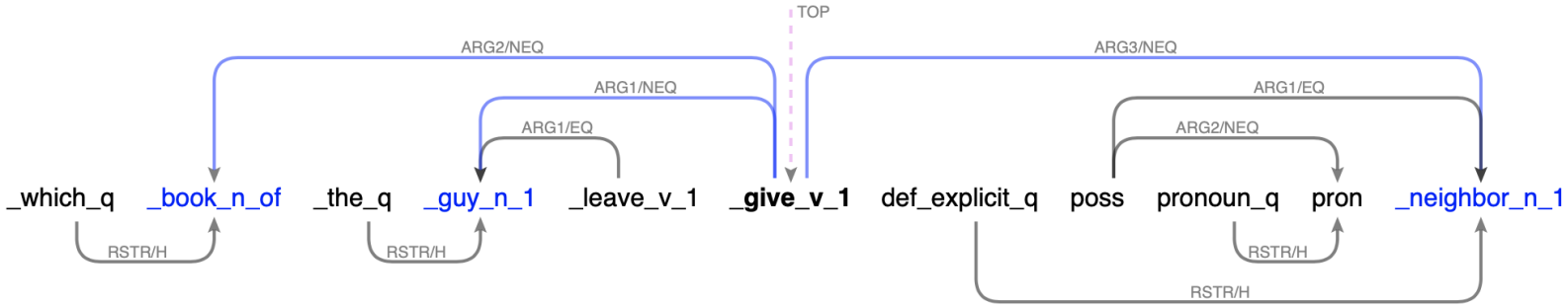
$$\langle h_1, e_3,$$
$$\left| \begin{array}{l} h_4{:}\_most\_q(x_5, h_6, h_7), \\ h_8{:}compound(e_{10}, x_5, x_9), \\ h_{11}{:}udef\_q(x_9, h_{12}, h_{13}), \\ h_{14}{:}\_house\_n\_of(x_9, i_{15}), \\ h_8{:}\_cat\_n\_1(x_5), \\ h_2{:}\_easy\_a\_for(e_3, h_{16}, x_{17}), \\ h_{18}{:}udef\_q(x_{17}, h_{19}, h_{20}), \\ h_{21}{:}\_dog\_n\_1(x_{17}), \\ h_{22}{:}\_chase\_v\_1(e_{23}, x_{17}, x_5) \end{array} \right|$$
$$\{ h_1 =_q h_2, h_6 =_q h_8, h_{12} =_q h_{14}, h_{16} =_q h_{22}, h_{19} =_q h_{21} \} \rangle$$

MRS

Every boy loves a woman



Which book did the guy who left give to his neighbor?



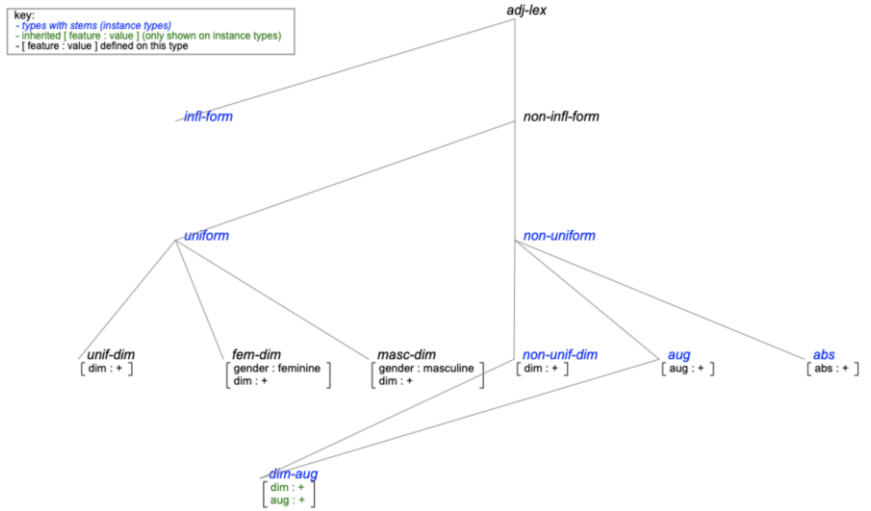
DMRS

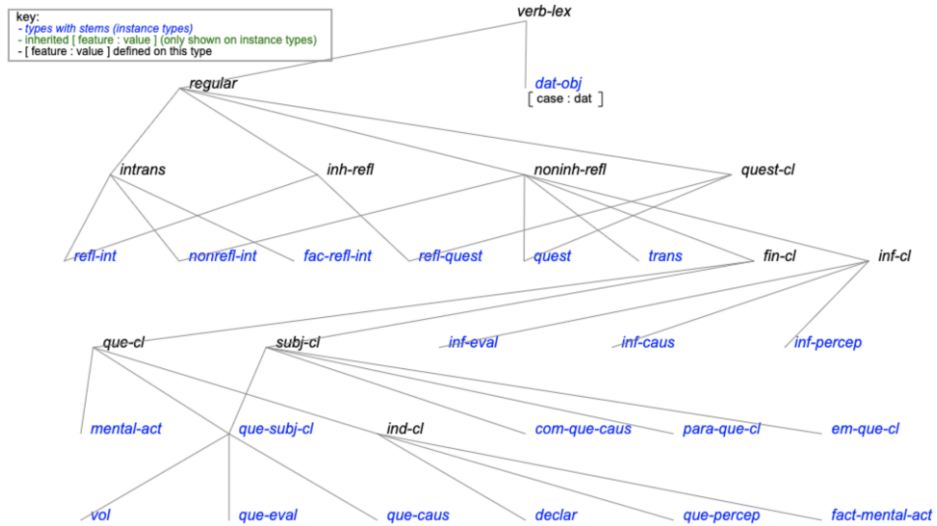
"… there is no existing logic which is adequate for all the phenomena of natural language.." Ann Copestake

The Portuguese HPSG Grammar

- Long-term (3 years): parsing of unrestricted texts in standard language
- Medium-term (10 months): syntactic and semantic annotation of part of the Brazilian Historical-Biographic Dictionary (DHBB)
- Short-term (6 months): parsing the MRS and HP Test Suite (CSLI profile) test sets

- Using UD corpora and MorphoBr based on the MATRIX customization system





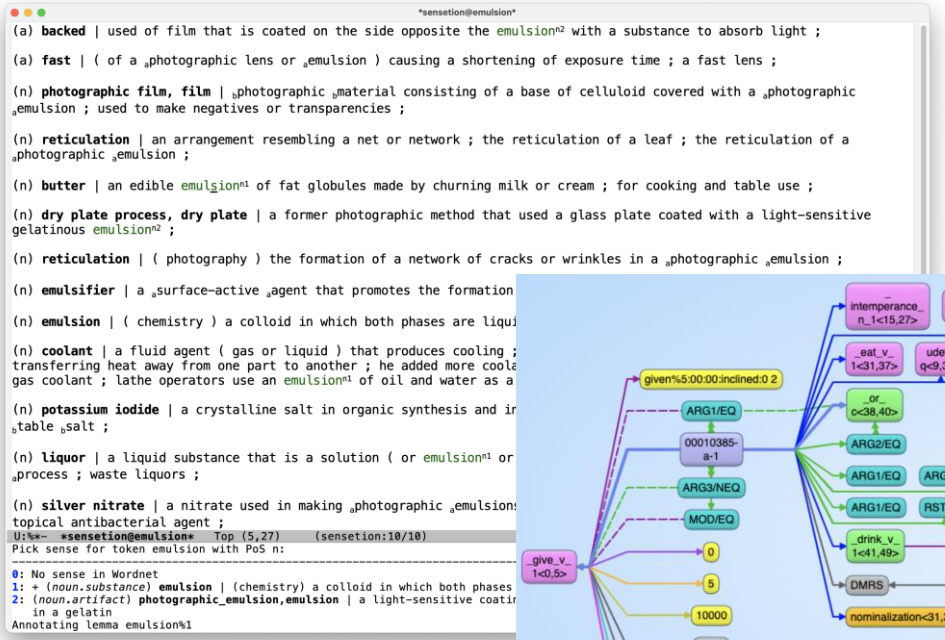
semantically tagged PWN glosses

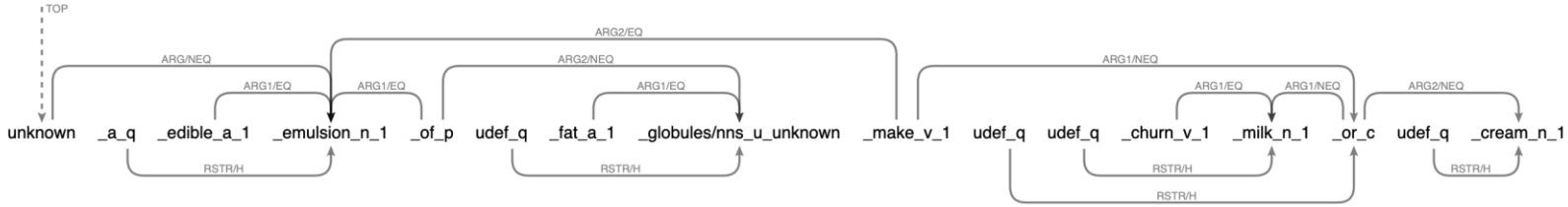
- scalable knowledge graph from PWN definitions and examples
- word-sense disambiguated
- semantic representation

What is butter?

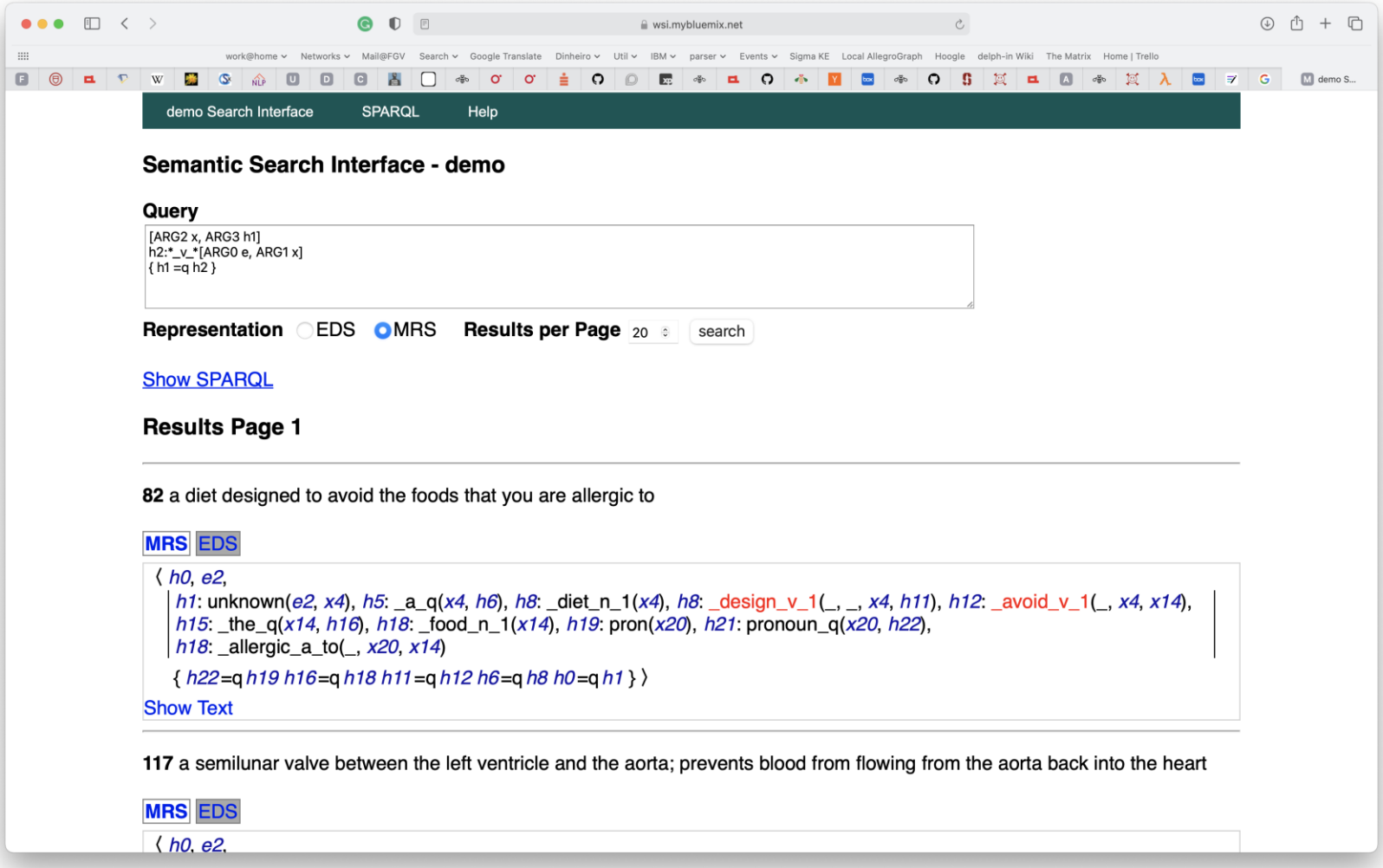
“an edible emulsion of fat globules made by churning milk or cream; for cooking and table use”

- Hypernym of:
  [lemon_butter, drawn_butter, stick, yak_butter, beurre_noisette]
- Hyponym of: [solid_food, dairy_product]





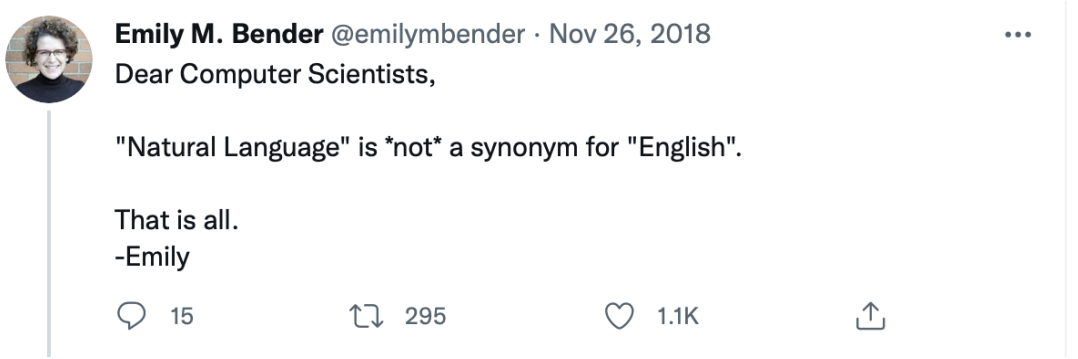
Lightweight scalable ontology and KR

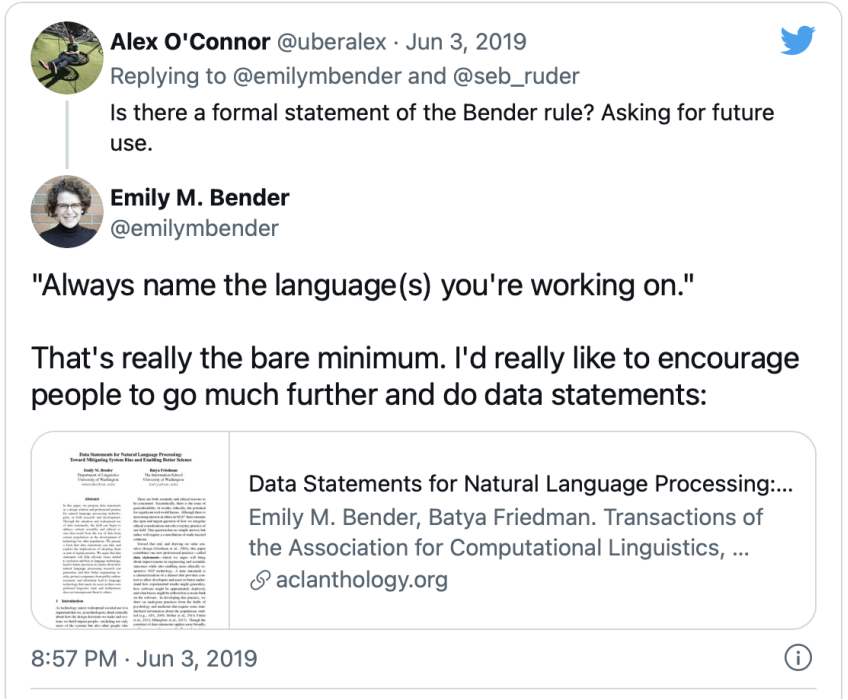


fingerprint search example: ‘Object’ Control

Conclusions

- Linguistic resources are very easy to start, hard to improve, and extremely difficult to maintain.
- Size of linguistic resources are easy to compare, quality is hard.
- Interoperability is complex but improves the quality
- Resources have many shapes: dictionaries, corpora, grammar, annotated data, datasets (QA, TE etc)
- **The #BenderRule and data statements**





Thank You