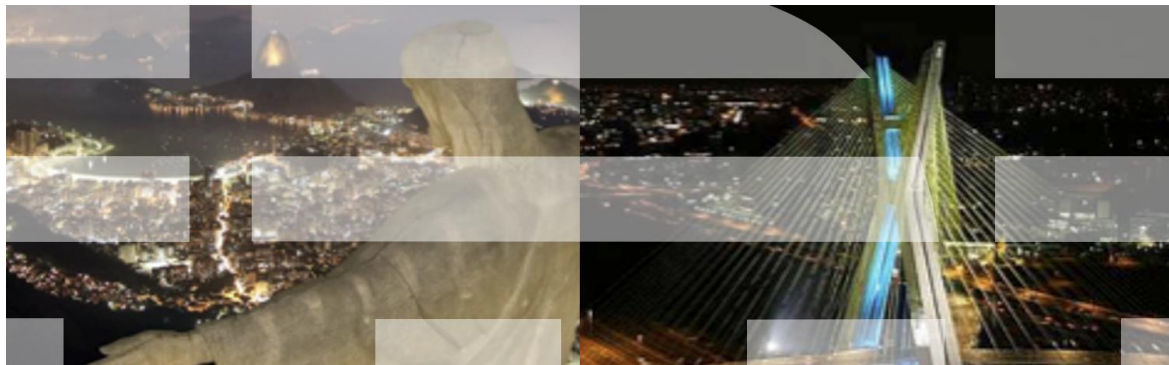


Language resources for natural language ~~processing~~ understanding



Alexandre Rademaker
IBM Research
Brazil

Different levels of understanding



Information Extraction

Subject: curriculum meeting

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg

Date: Jan-16-2012

Start: 10:00am

End: 11:30am

Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in

-Chris

[Create new Calendar entry](#)

- Won Jeopardy on February 16, 2011!

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



Bram Stoker

What is understanding?

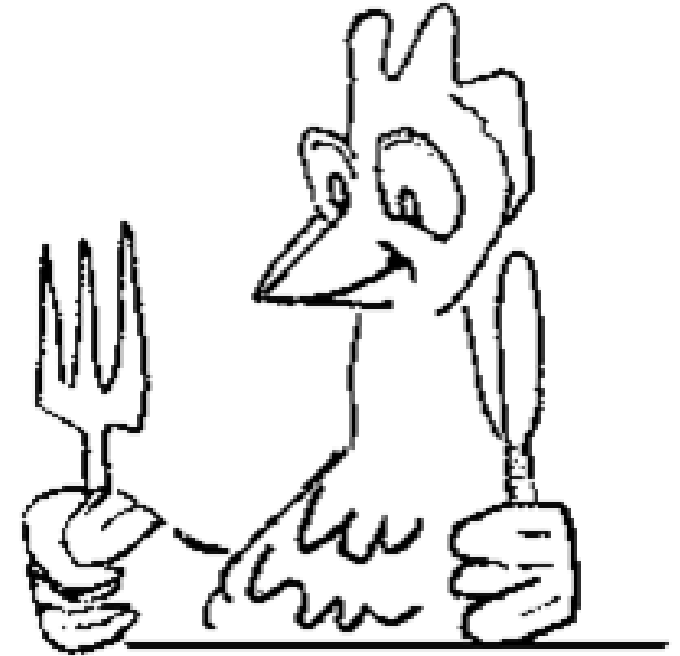
- There is no one universal measure
- one evaluation metric: the detection of entailment and contradiction relations between (portions of) texts.
- RTE is not a *sufficient* criterion for LU. But RTE is a minimal, *necessary* criterion.
- If you understand sentences (1) and (2), then you can recognize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



THE CHICKEN IS
READY TO EAT

7,102

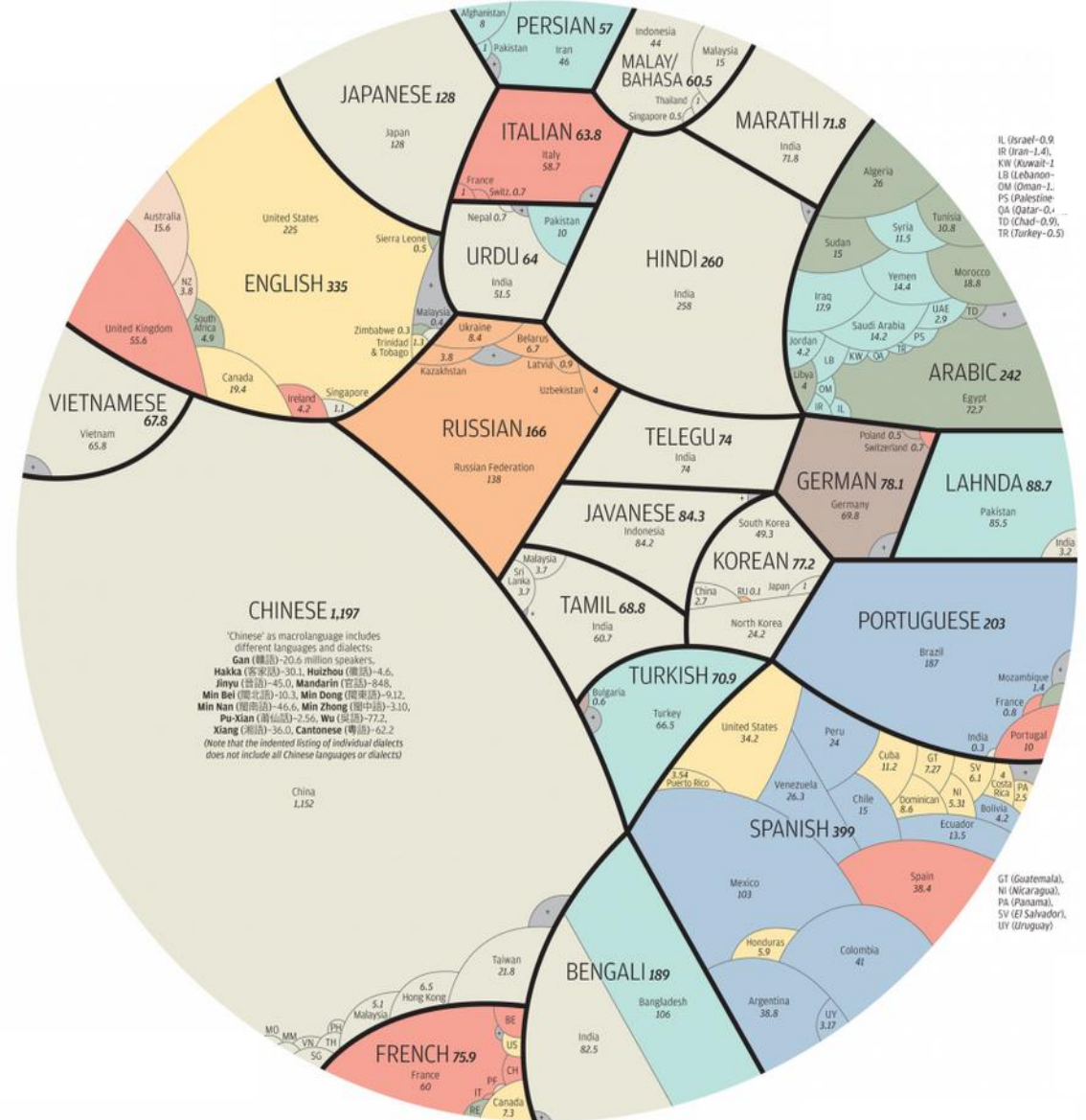
known languages

23

most spoken language

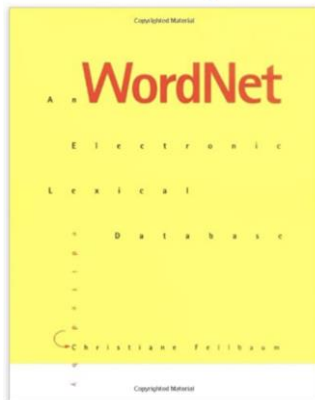
4.1+ Billion

people



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



They can fish



WordNet 

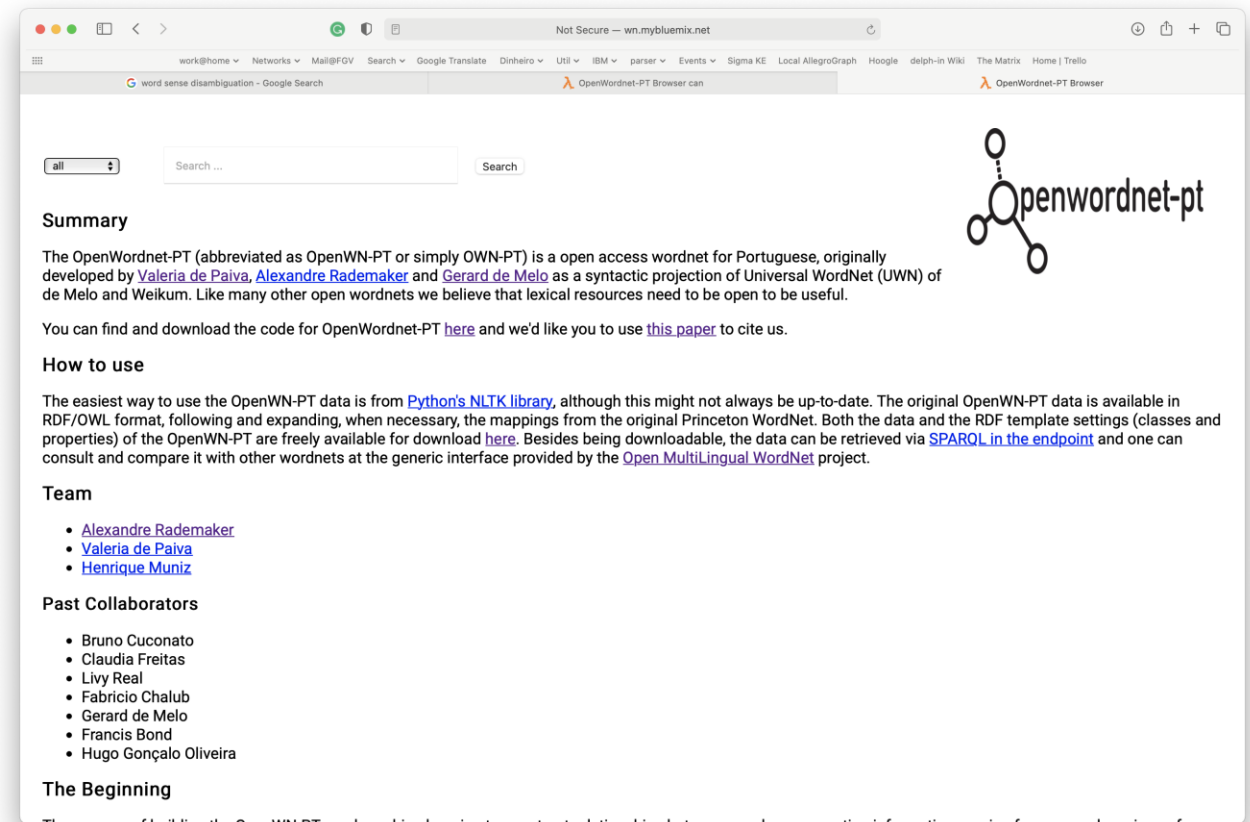
- ▶ S: (n) **mouse**
 - ▷ S: (n) rodent, gnawer
- ▶ S: (n) shiner, black eye, **mouse**
 - ▷ S: (n) bruise, contusion
- ▶ S: (n) **mouse**
 - ▷ S: (n) person, individual, someone, somebody, mortal, soul
- ▶ S: (n) **mouse**, computer mouse
 - ▷ S: (n) electronic device

*"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

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>.



<http://openwordnet-pt.org>

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

MorphoBr
Resources for morphological analysis of Portuguese

Python ☆ 15 🍴 3

aelius
Python/NLTK-based package for shallow parsing of Brazilian Portuguese

Python ☆ 3

tools
Tools for checking the compatibility between a lexical resource and a treebank

Python ☆ 1

BrGram
Computational grammar fragment of Brazilian Portuguese in the LFG formalism implemented in XLE

Prolog ☆ 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

delphin-docker
A docker container for running all DELPH-IN tools in a Linux box.

Dockerfile

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

Repositories

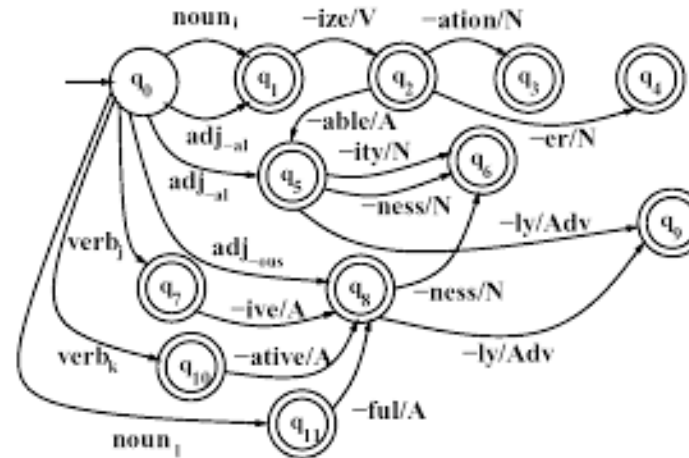
Find a repository... Type Language Sort New

MorphoBr
Resources for morphological analysis of Portuguese

Python ☆ 15 Apache-2.0 🍴 3 🕒 46 (1 issue needs help) 🛠️ 0 Updated 7 days ago

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

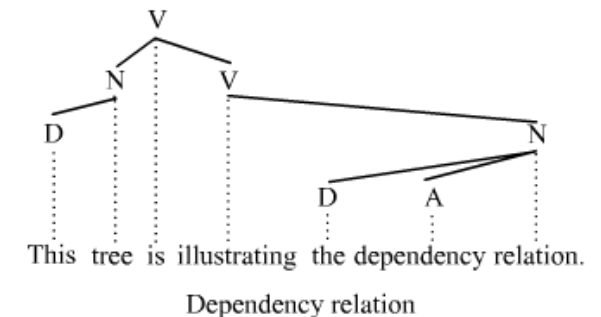
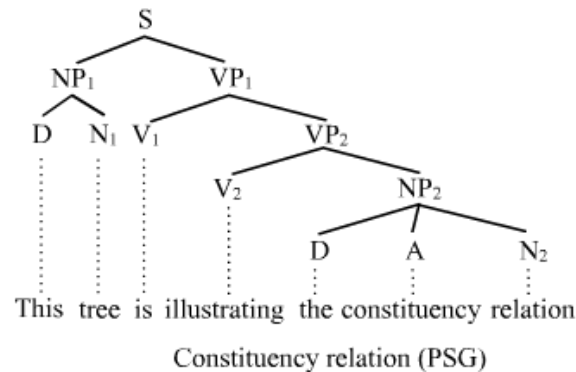
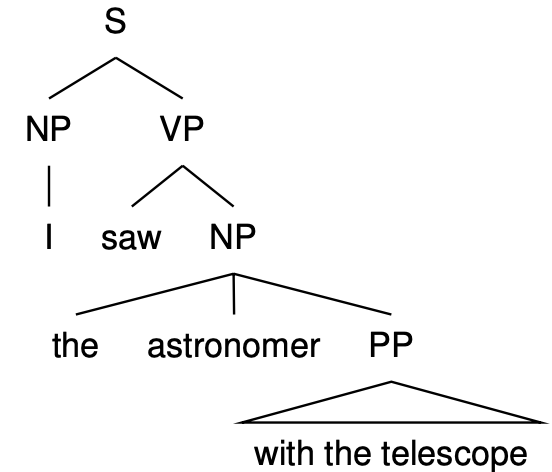
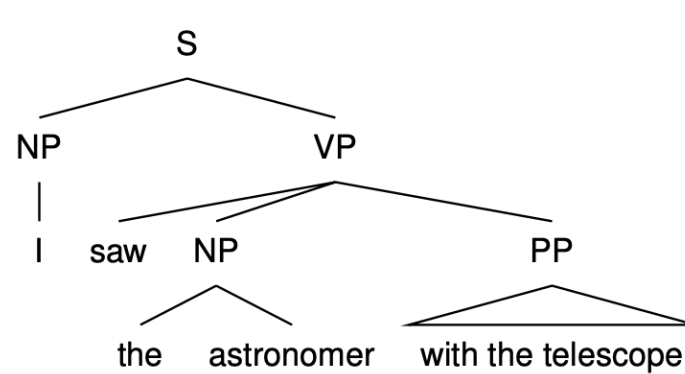


simples	simples+A+F+PL
simples	simples+A+F+SG
simples	simples+A+M+PL
simples	simples+A+M+SG
simplicíssimas	simples+A+SUPER+F+PL
simples	simples+A+SUPER+M+PL
simples	simples+A+SUPER+F+SG
trabalhadreira	trabalhador+A+F+SG
trabalhadreizinhas	trabalhador+A+DIM+F+PL
trabalhadreizita	trabalhador+A+DIM+F+SG
trabalhadreirinha	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
felicitíssimas	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?
- **Syntactic** parsing: assigning syntactic structure
- **Semantic** parsing: assigning semantic structure

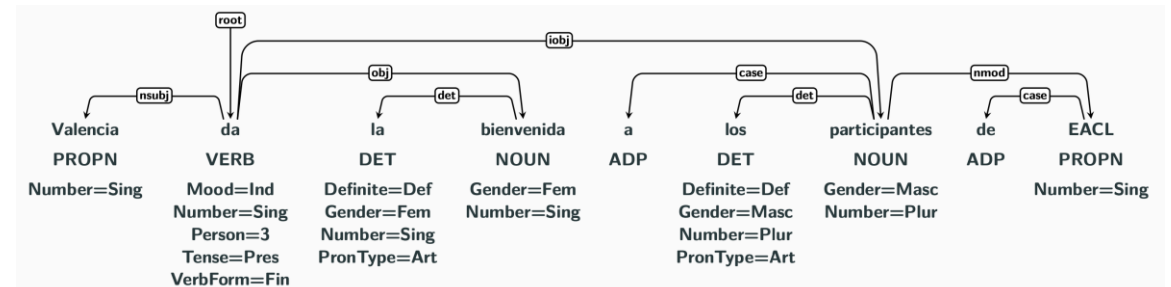
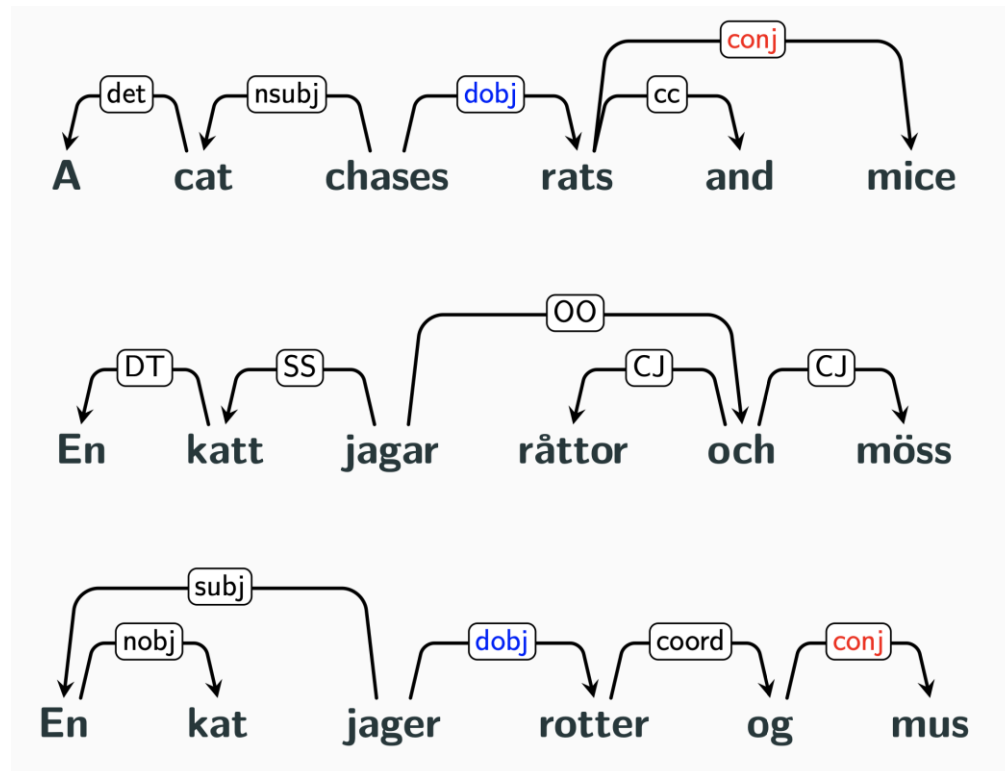


“A grammar is better, but in practice people use language models.”
D. Jurafsky

“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

CF0071.conllu

```

5 de de ADP PRP|@<PIV _ 6 case _
6 posições posição NOUN <np-idf>|N|F|P|@P< Gender=Fem|Number=Plur 3 iobj _
7 históricas histórico ADJ ADJ|F|P|@N< Gender=Fem|Number=Plur 6 amod _ SpaceAfter=No
8 , PUNCT PU|@PU _ 10 punct _
9 eventualmente eventualmente ADV ADV|@ADVL _ 10 advmod _
10 visando visar VERB <mv>|V|GER|@ICL-<ADVL VerbForm=Ger 3 advcl _
11 sua seu DET <poss>|<si>|DET|F|S|@>N Gender=Fem|Number=Sing|PronType=Prs 12 det _
12 proteção proteção NOUN <np-def>|N|F|S|@<ACC Gender=Fem|Number=Sing 10 obj _ SpaceAfter=No
13 , PUNCT PU|@PU _ 10 punct _
14 para para SCONJ PRP|@<ADVL _ 15 mark _
15 construir construir VERB <first-cjt>|<mv>|V|INF|@ICL-P< VerbForm=Inf 3 advcl _
16 e e CCONJ <co-icl>|<co-inf>|KC|@CO _ 17 cc _
17 defender defender VERB <cjt>|<mv>|V|INF|@ICL-P< VerbForm=Inf 15 conj _
18 idéias idéia NOUN <np-idf>|N|F|P|@<ACC Gender=Fem|Number=Plur 17 obj _
19 exclusivamente exclusivamente ADV ADV|@>A _ 21 advmod _
20 de de ADP PRP|@N< _ 21 case _
21 interesse interesse NOUN <np-idf>|N|M|S|@P< Gender=Masc|Number=Sing 18 nmod _
22 coletivo coletivo ADJ ADJ|M|S|@N< Gender=Masc|Number=Sing 21 amod _
23 para para ADP PRP|@N<ARG _ 25 case _
24 o o DET <artd>|ART|M|S|@>N Definite=Def|Gender=Masc|Number=Sing|PronType=Art 25 det _
25 desenvolvimento desenvolvimento NOUN <np-def>|N|M|S|@P< Gender=Masc|Number=Sing 21 nmod _
26 global global ADJ ADJ|M|S|@N< Gender=Masc|Number=Sing 25 amod _ SpaceAfter=No
27 . PUNCT PU|@PU _ 3 punct _

```

text = Folha -- Quais são as implicações que o encerramento da revisão traz para a economia?
sent_id = CF71-2
source = CETENFolha n=71 cad=Brasil sec=pol sem=94b
id = 271

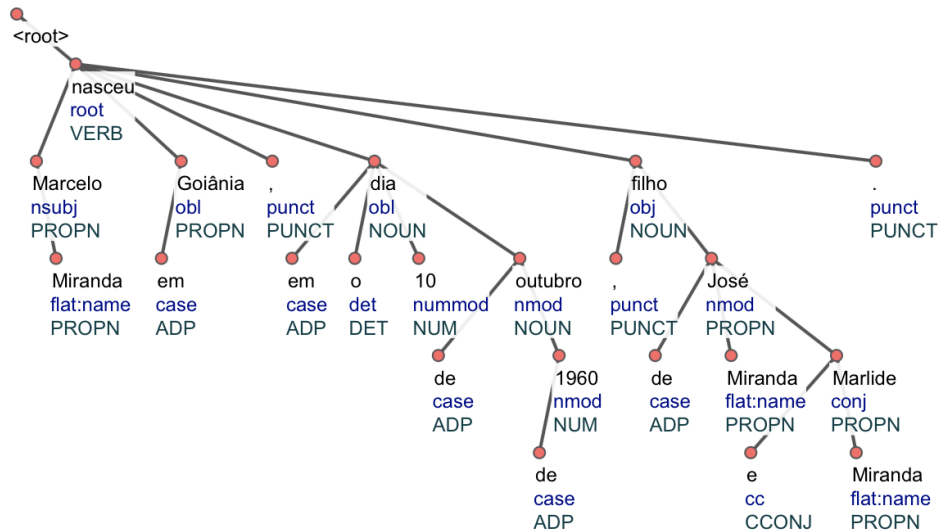
1	Folha	Folha	PROPN	PROP F S @NPHR	Gender=Fem Number=Sing
2	--	--	PUNCT	PU @PU	--
3	Quais	qual	PRON	<interr> DET F P @SC>	Gender=Fem Number=Plur PronType=In
4	são	ser	AUX	<mv> V PR 3P IND @FS-QUE	Mood=Ind Number=Plur Person=3 Tens
5	as	o	DET	<artd> ART F P @>N	Definite=Def Gender=Fem Number=Plu
6	implicações	implicação	NOUN	<np-def> N F P @<SUBJ	Gender=Fem Number=Plur
7	que	que	PRON	<rel> INDP F P @<ACC>	Gender=Fem Number=Plur PronType=Re
8	o	o	DET	<artd> ART M S @>N	Definite=Def Gender=Masc Number=Si
9	encerramento	encerramento	NOUN	<np-def> N M S @SUBJ>	Gender=Masc Number=Sing
10-11	de	de	ADP	<sam-> PRP @N<ARG	--
11	a	o	DET	<-sam> <artd> ART F S @>N	Definite=Def Gender=Fem Number=Sin
12	revisão	revisão	NOUN	<np-def> N F S @P<	Gender=Fem Number=Sing
13	traz	trazer	VERB	<mv> V PR 3S IND @FS-N<	Mood=Ind Number=Sing Person=3 Tens
14	para	para	ADP	PRP @<ADVL	--
15	a	o	DET	<artd> ART F S @>N	Definite=Def Gender=Fem Number=Sin
16	economia	economia	NOUN	<np-def> N F S @P<	Gender=Fem Number=Sing
17	?	?	PUNCT	PU @PU	--

U:---- CF0071.conllu Bot (39,0) Git-workbench (CoNLL-U ws AReV FlyC-) U:*** *conllu-command-output* All (1,0) (Fund

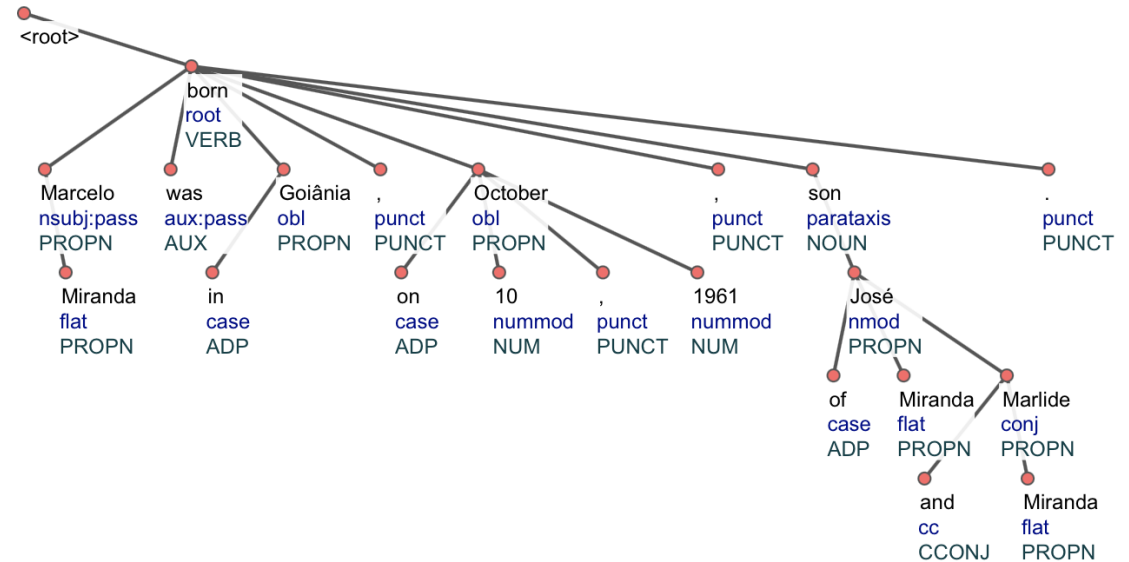
Rademaker, Alexandre, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal Dependencies for Portuguese. 2017. "Universal Dependencies for Portuguese." In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, 197–206. Pisa, Italy. <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.

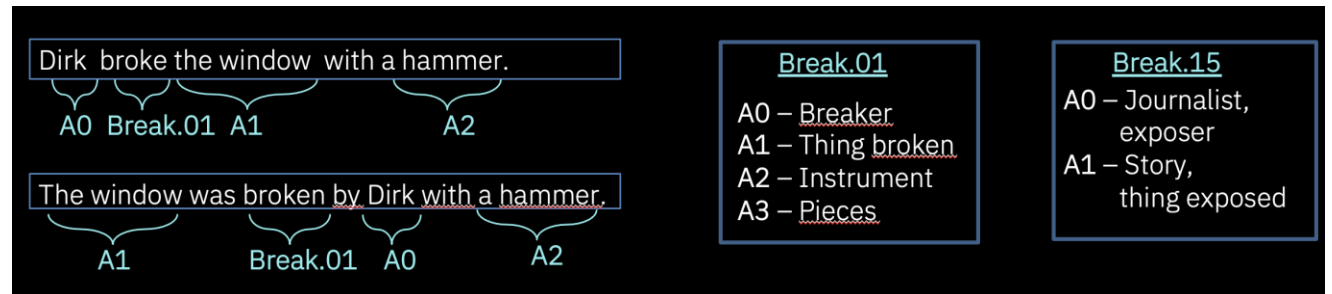
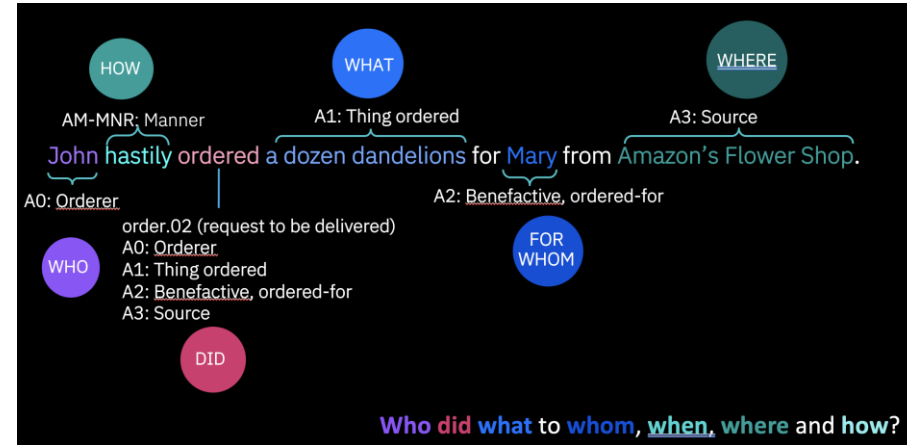


semantic representations: semantic role labelling

<http://proppbank.github.io>

Universal Proposition Banks: UD + SRL

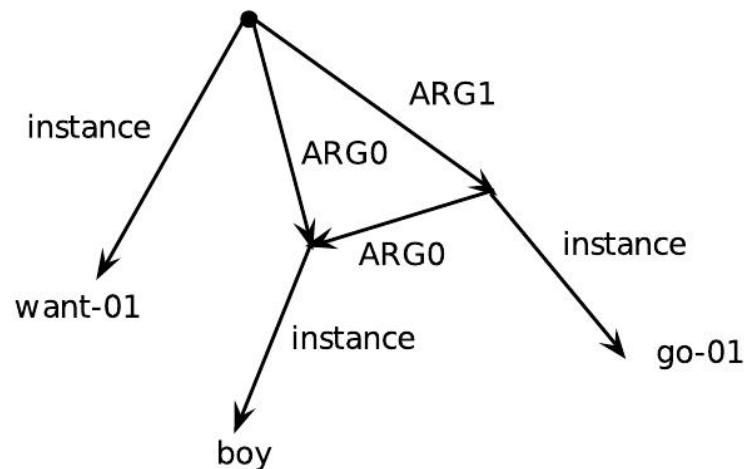
<https://github.com/System-T/UniversalPropositions>



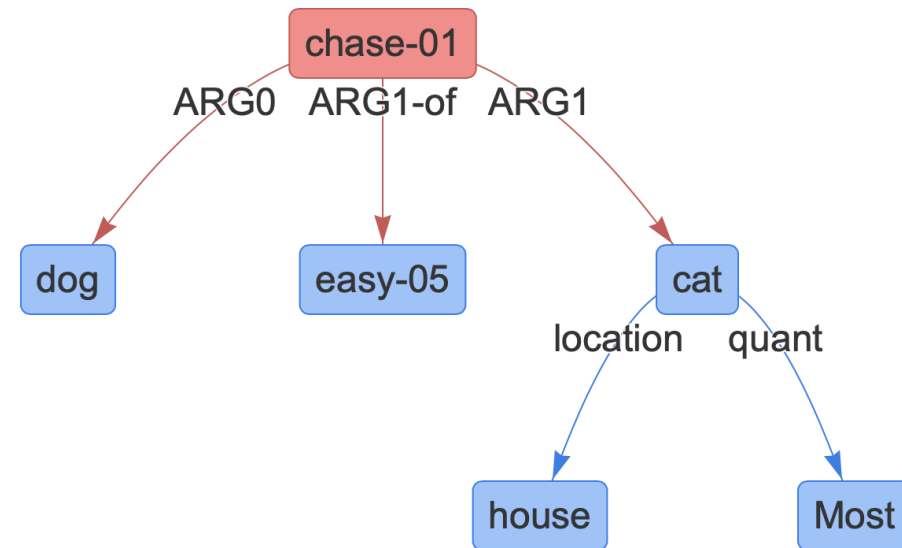
semantic representation: abstract meaning representation (AMR)

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/ax6Ka18Ki4>

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

The HPSG English Grammar

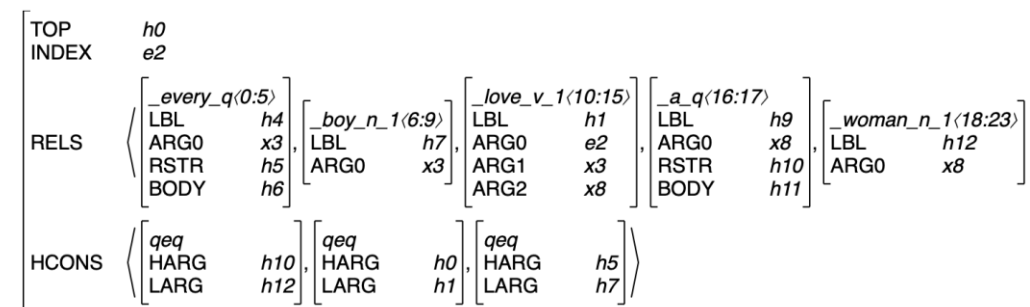
<http://delph-in.github.io/delphin-viz/demo/>

Most house cats are easy for dogs to chase.

```

⟨ h1, e3,
  h4:_most_q(x5, h6, h7),
  h8:compound(e10, x5, x9),
  h11:undef_q(x9, h12, h13),
  h14:_house_n_of(x9, i15),
  h8:_cat_n_1(x5),
  h2:_easy_a_for(e3, h16, x17),
  h18:undef_q(x17, h19, h20),
  h21:_dog_n_1(x17),
  h22:_chase_v_1(e23, x17, x5)
  { h1 =q h2, h6 =q h8, h12 =q h14, h16 =q h22, h19 =q h21 } ⟩
  
```

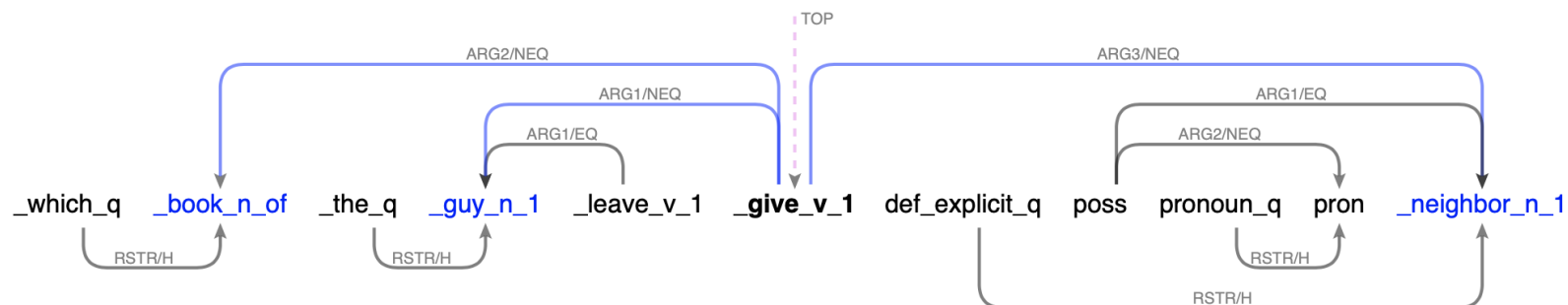
Every boy loves a woman



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

MRS

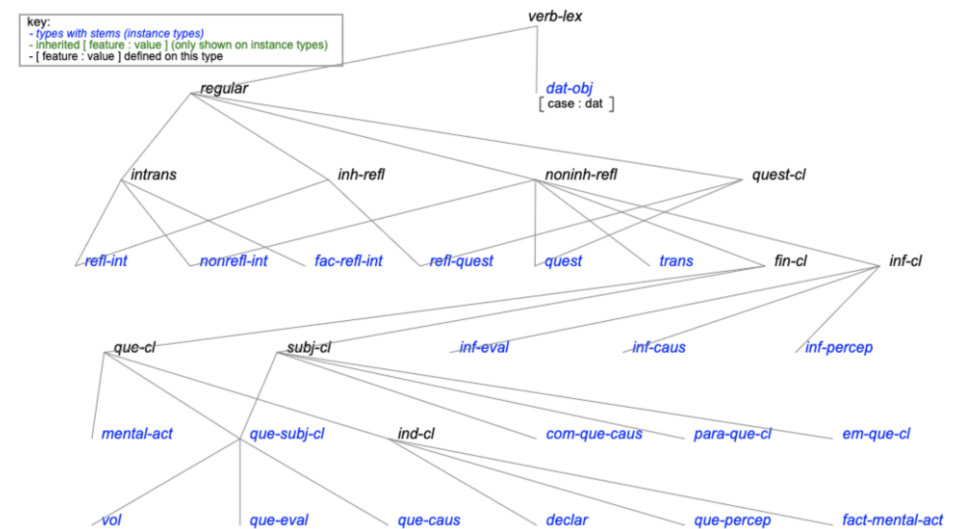
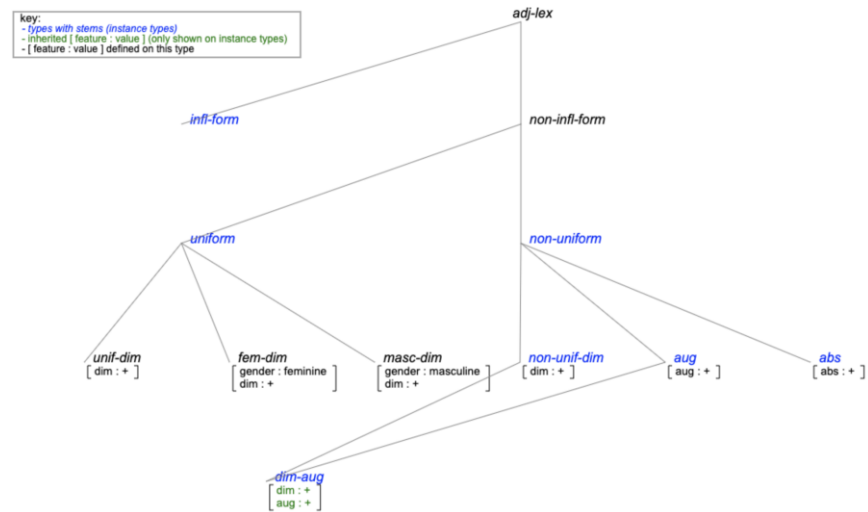
“... there is no existing logic which is adequate for all the phenomena of natural language..” Ann Copestake



DMRS

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

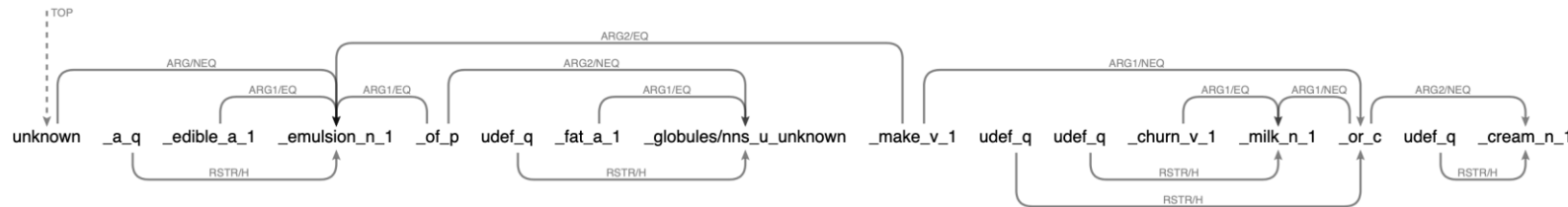
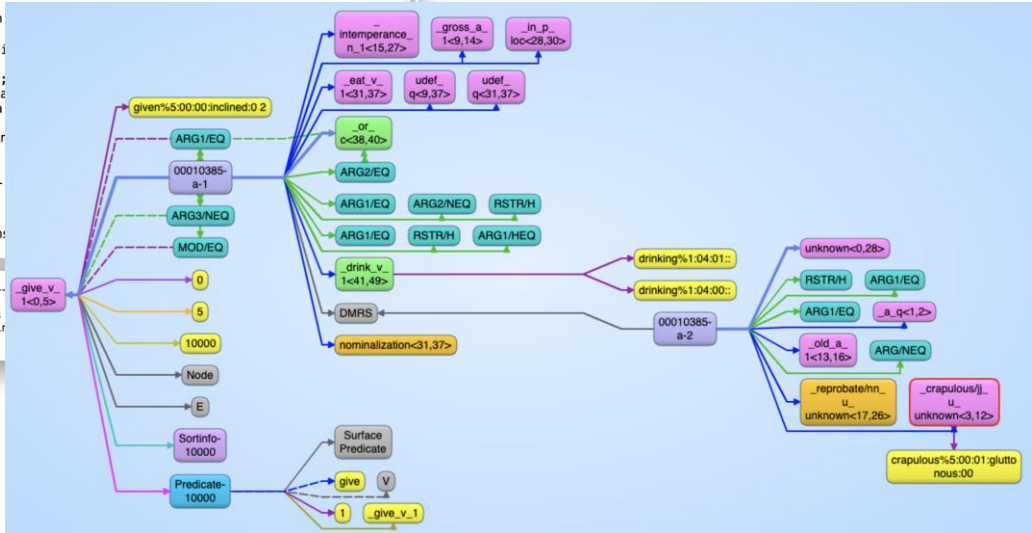
What is butter?

“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](#)]

```

(a) backed | used of film that is coated on the side opposite the emulsionn2 with a substance to absorb light ;
(a) fast | ( of a ,photographic lens or ,emulsion ) causing a shortening of exposure time ; a fast lens ;
(n) photographic film, film | ,photographic ,material consisting of a base of celluloid covered with a ,photographic ,emulsion ; used to make negatives or transparencies ;
(n) reticulation | an arrangement resembling a net or network ; the reticulation of a leaf ; the reticulation of a ,photographic ,emulsion ;
(n) butter | an edible emulsionn1 of fat globules made by churning milk or cream ; for cooking and table use ;
(n) dry plate process, dry plate | a former photographic method that used a glass plate coated with a light-sensitive gelatinous emulsionn2 ;
(n) reticulation | ( photography ) the formation of a network of cracks or wrinkles in a ,photographic ,emulsion ;
(n) emulsifier | a ,surface-active ,agent that promotes the formation
(n) emulsion | ( chemistry ) a colloid in which both phases are liquid
(n) coolant | a fluid agent ( gas or liquid ) that produces cooling ; transferring heat away from one part to another ; he added more coolant ; lathe operators use an emulsionn1 of oil and water as a
(n) potassium iodide | a crystalline salt in organic synthesis and in
table ,salt ;
(n) liquor | a liquid substance that is a solution ( or emulsionn1 or ,process ; waste liquors ;
(n) silver nitrate | a nitrate used in making ,photographic ,emulsion: topical antibacterial agent ;
U: %s- *sensation@emulsion* Top (5,27) (sensation:10/10)
Pick sense for token emulsion with PoS n:
0: No sense in Wordnet
1: + (noun.substance) emulsion | (chemistry) a colloid in which both phases
2: (noun.artifact) photographic_emulsion,emulsion | a light-sensitive coating in a gelatin
Annotating lemma emulsion%1
    
```



Lightweight scalable ontology and KR

The screenshot shows a web browser window with the URL `wsr.mybluemix.net`. The page title is "demo Search Interface" and it includes navigation links for "SPARQL" and "Help".

Semantic Search Interface - demo

Query

```
[ARG2 x, ARG3 h1]
h2:*_v_*[ARG0 e, ARG1 x]
{ h1=q h2 }
```

Representation EDS MRS **Results per Page** 20

[Show SPARQL](#)

Results Page 1

82 a diet designed to avoid the foods that you are allergic to

```
< h0, e2,
  h1: unknown(e2, x4), h5: _a_q(x4, h6), h8: _diet_n_1(x4), h8: _design_v_1(_, _, x4, h11), h12: _avoid_v_1(_, x4, x14),
  h15: _the_q(x14, h16), h18: _food_n_1(x14), h19: pron(x20), h21: pronoun_q(x20, h22),
  h18: _allergic_a_to(_, x20, x14)
  { h22=q h19 h16=q h18 h11=q h12 h6=q h8 h0=q h1 } >
```

[Show Text](#)

117 a semilunar valve between the left ventricle and the aorta; prevents blood from flowing from the aorta back into the heart

```
< h0, e2,
```

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



Emily M. Bender @emilymbender · Nov 26, 2018

Dear Computer Scientists,

"Natural Language" is **not** a synonym for "English".

That is all.
-Emily



15



295



1.1K



Alex O'Connor @uberalex · Jun 3, 2019

Replying to @emilymbender and @seb_ruder

Is there a formal statement of the Bender rule? Asking for future use.



Emily M. Bender
@emilymbender

"Always name the language(s) you're working on."

That's really the bare minimum. I'd really like to encourage people to go much further and do data statements:



Data Statements for Natural Language Processing:...

Emily M. Bender, Batya Friedman. Transactions of the Association for Computational Linguistics, ...
aclanthology.org

8:57 PM · Jun 3, 2019

