Building Resources for Human and Computational Language Processing of Portuguese

Sílvio Cordeiro, Carlos Ramisch, Marco Idiart, Rodrigo Wilkens, Leonardo Zilio, Jorge Wagner, Aline Villavicencio

Federal University of Rio Grande do Sul (Brazil)

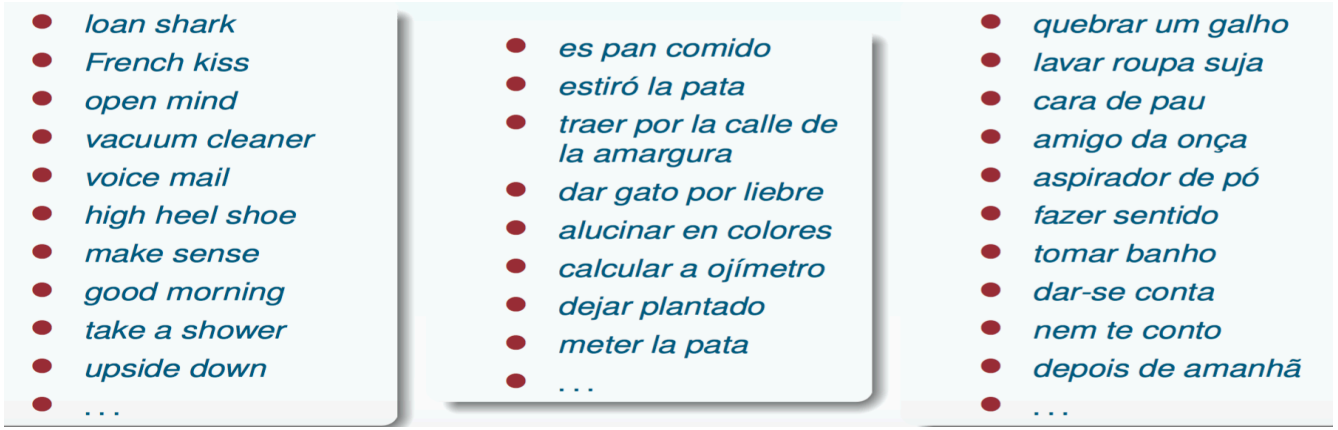
Aix Marseille Université, CNRS, LIF UMR 7279 (France)

University of Essex (UK)

MWE-aware processing with the mwetoolkit

Multiword Expressions in a Nutshell

- A combination of words that must be treated as a unit at some level of linguistic processing (Calzolari et al., 2002)
  - Compound Nouns
  - Verb-particle constructions
  - Light-verb constructions
  - Idioms



Multiword Expressions in a Nutshell

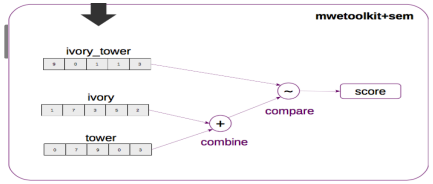
- Lexical, syntactic, semantic, pragmatic, statistical idiosyncrasies
  - *Ad hoc, wine and dine (Kim and Baldwin 2010)*
- Arbitrariness and Institutionalisation
  - *salt and pepper*, *?pepper and salt* (Smadja, 1993)
- Frequency
  - Same order of magnitude as words in mental lexicon (Jackendoff, 1997)
- Limited lexical, syntactic and semantic variability
  - *kick the bucket/?pail/?container* (Sag et al., 2002)

MWEs and NLP

- Real understanding requires MWE-aware corpus processing
  1. Corpus processing
  2. MWE discovery from corpus (to build MWE lexica)
  3. MWE representation (in lexicon and grammar)
  4. MWE token identification in corpus (to annotate MWEs)
  5. MWE semantic processing
  6. MWE integration in applications

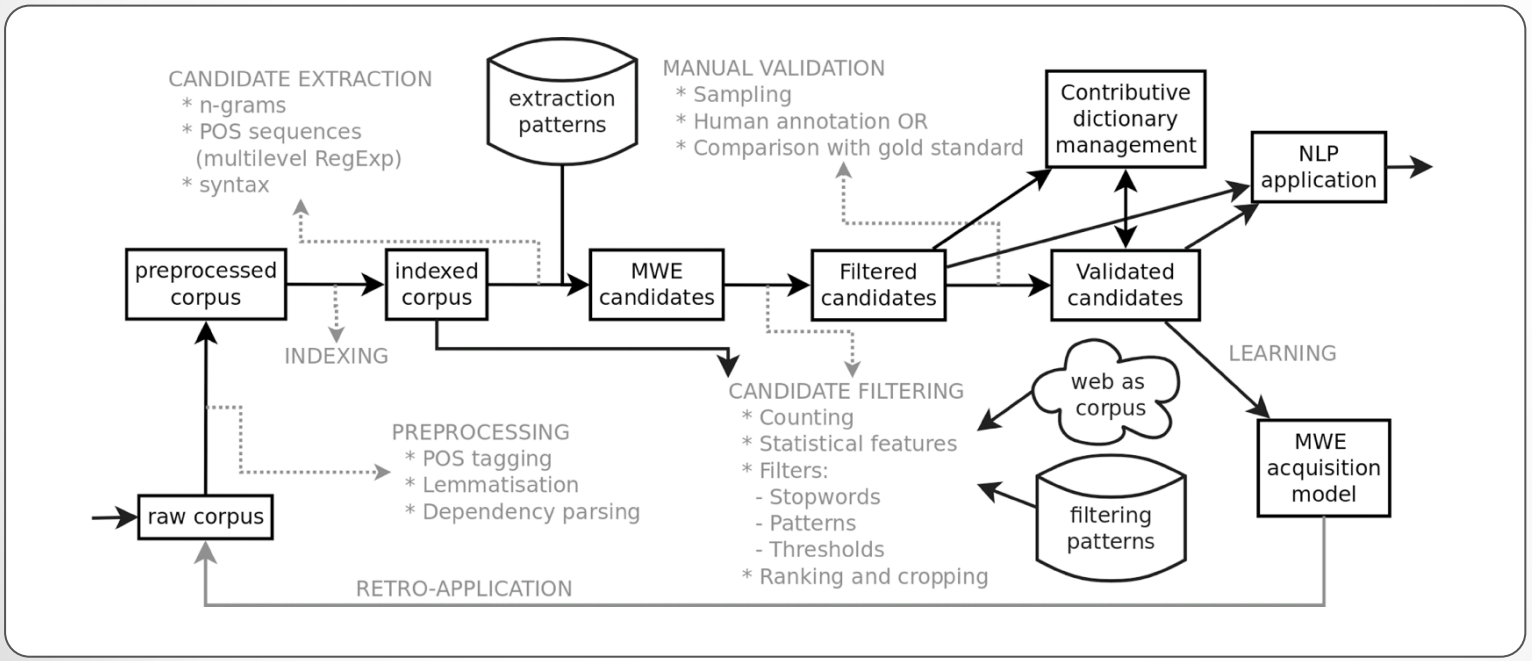
mwetoolkit

- Language independent framework for MWE processing
- Extracts MWE from corpora
- Annotates corpora with MWEs
- Calculates AMs
- Pre-processes MWEs in corpora for DSM construction
- Imports DSMs (word2vec, glove, PI
- Provides functions for vector combinations
- Calculates compositionality
- Evaluates against gold standard



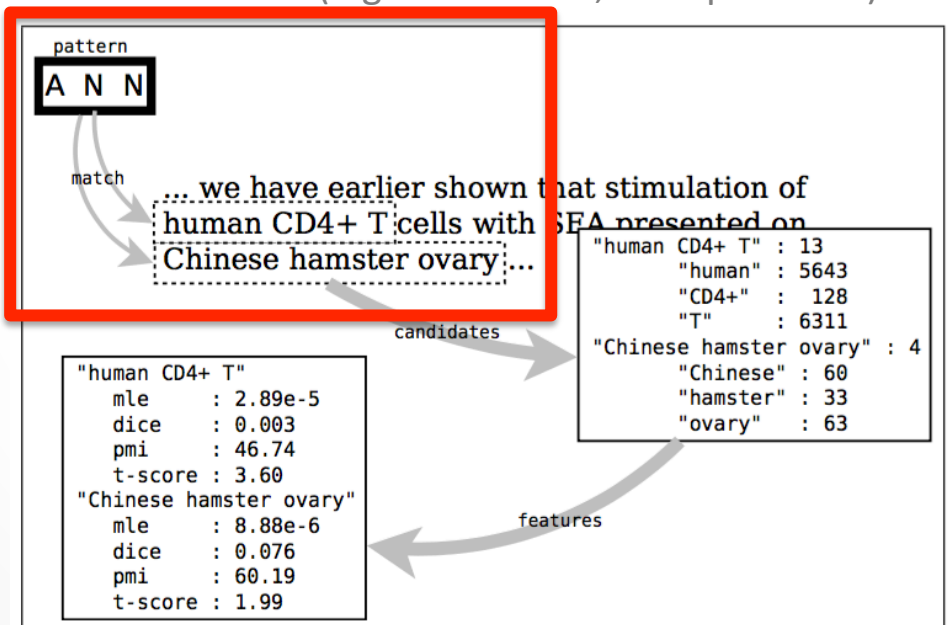
Project CAPES-COFECUB (France-Brazil)

Overview of the mwetoolkit



MWE Type Discovery

- Candidate extraction
  - Pattern-based heuristics (e.g. noun-noun, verb-particle…)



MWE-aware corpus processing

- MWE token identification
  - Pattern-based heuristics
  - Contiguous/gappy identification
  - Shortest/longest/all match distances
  - Projecting extracted MWE types back in source corpus

```
<pat>  <!-- Finding past forms of "wake (up)" -->
  <w lemma="wake" pos="Verb">
     <neg surface="wake"/> <neg surface="wakes"/>
  </w>
  <pat repeat="?">
     <w surface="up"/>
  </pat>
</pat>
```

Pattern **Verb (Word*) Particle**

- Shortest: I have **picked** it **up** and **put** it **down**. ✓
- Longest: I have **picked** it up and put it **down**. ✗

Pattern **Noun Noun$^+$**

- Shortest: The **science fiction** writers are on strike.
- Longest: The **science fiction writers** are on strike. ✓

Annotation Options

- Corpus but no MWE List:
  - Generate MWE list from corpus and projecting them back
    - Corpus → MWE list → Annotated Corpus
- Corpus and MWE List
  - Annotation based on external lists of MWEs
    - Corpus + MWE list → Annotated Corpus

MWE semantic processing

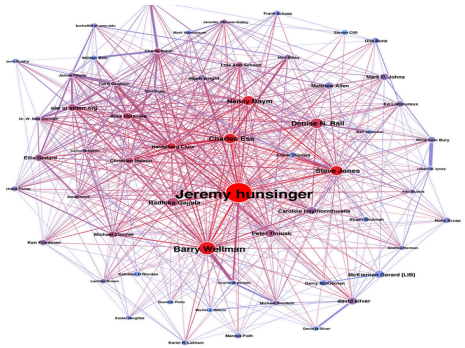
- Meaning of MWE may not be understood from meaning of individual words
  - *brick wall* is a wall *made of* bricks,
  - *cheese knife* is not a *knife made of* cheese → *knife for cutting* cheese (Girju et al., 2005).
  - *Loan shark* is not a shark for loan but a person who offers **loans** at extremely high interest rates

How to detect compositionality?

Idiomaticity ⟷ Compositionality

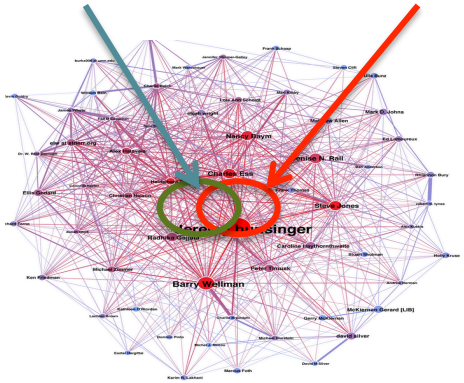
Cloud nine | Grandfather clock | Access road

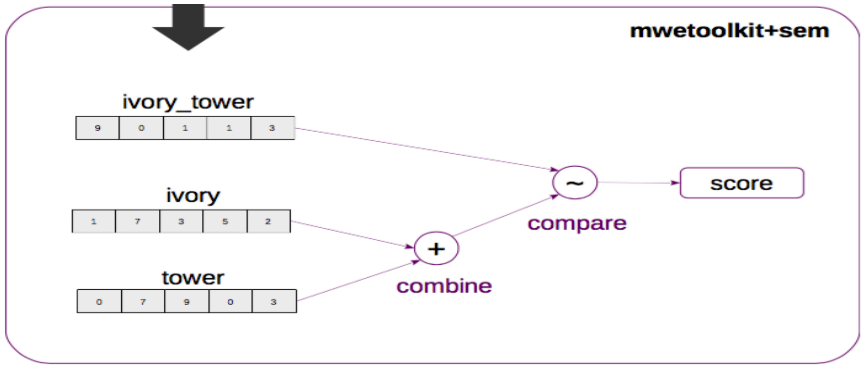
- Distributional Semantic Models (DSMs)
  - Position words in multidimensional semantic space
    - Each word/MWE represented as a vector in the semantic space
  - Proximity in space indicates semantic relatedness



How to detect compositionality?

- Cosine similarity between the MWE vector and the sum of the vectors of the component words
  - The closer vectors are the more compositional they are (Reddy et al. 2011)
  - **cos($w_1w_2$vector, $w_1$vector+$w_2$vector)**



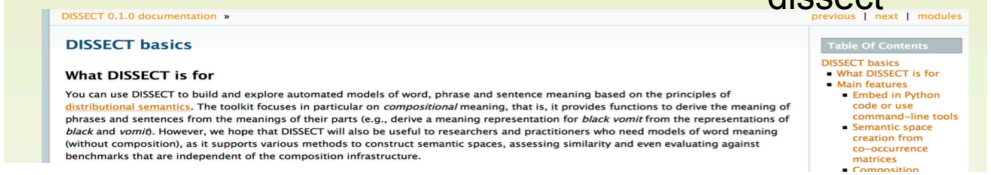


Distributional Semantic Models

- Techniques and tools for constructing DSMs
  - Dissect (Dinu et al., 2013), Minimantics (Ramisch et al. 2013), word2vec (Mikolov et al., 2013) and Glove4 (Pennington et al., 2014).

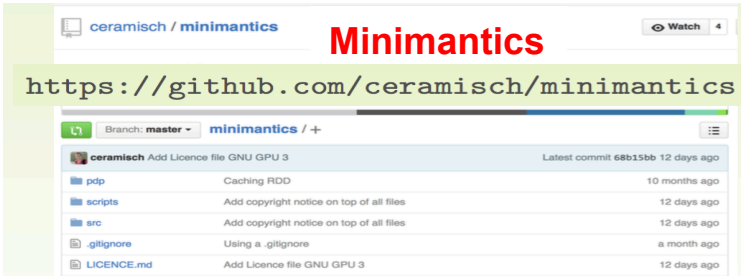
Baroni et al. http://clic.cimec.unitn.it/composes/toolkit/introduction.html

dissect



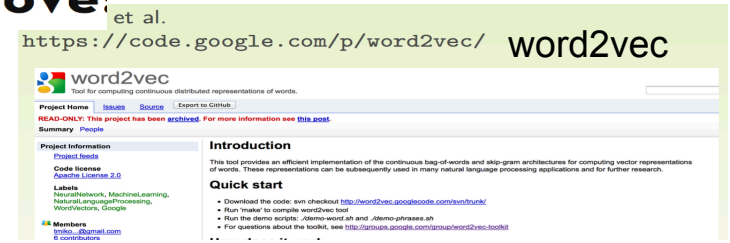
Minimantics

https://github.com/ceramisch/minimantics



GloVe et al.

https://code.google.com/p/word2vec/ word2vec



LexVec (Lexical Vectors)

Matrix Factorization using Window Sampling and Negative Sampling for Improved Word Representations

Alexandre Salle[1] **Marco Idiart**[2] **Aline Villavicencio**[1]
[1] Institute of Informatics
[2] Physics Department
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
{alexandre.salle,avillavicencio}@inf.ufrgs.br, idiart@if.ufrgs.br

Abstract

In this paper, we present Lexical Vectors (LexVec), a method for factorizing PPMI matri-

Gold Standards for Evaluation

- Roller et al. (2013) 244 German compounds
  - around 30 judgments by crowdsourcing scale from 1 to 7
- Farahmand et al. (2015) 1,042 English compounds
  - 4 experts judges binary scale for non-compositionality and conventionality
- Reddy et al. (2011) 90 English compounds
  - around 30 judgments by crowdsourcing scale from 0 to 5

DSMs and Compositionality

- Dataset of nominal compounds with human judgments about literality/compositionality
  - 180 compounds for English, French and Portuguese
  - Resource freely available
    - http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds&lang=en

Compositionality of Nominal Compounds - Datasets

- Authors: Silvio Cordeiro, Carlos Ramisch, Aline Villavicencio, Leonardo Zilio, Marc
- Version 1.0 - August 2, 2016
- Download the data set

Description

This package contains numerical judgements by human native speakers about 180 nominal

Judgements were obtained using Amazon Mechanical Turk (EN and FR) and a web interface (fully compositonal) and are averaged over several annotators (around 10 to 20 depending

The datasets are described in detail and used in the experiments of papers below. Please c

- How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Com
- Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings
- Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgment

Our methodology is inspired from Reddy, McCarthy and Manandhar (2011). We include thei full EN dataset.

Quick start

If you only want to use our datasets to evaluate your compositionality prediction models, you're probably interested in the scores present in the column named `compositionality` of files:

How Naked is the Naked Truth?
A Multilingual Lexicon of Nominal Compound Compositionality

Carlos Ramisch[1], **Silvio Cordeiro**[1,2], **Leonardo Zilio**[2]
Marco Idiart[3], **Aline Villavicencio**[2], **Rodrigo Wilkens**[2]
[1] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr lzilio@inf.ufrgs.br
marco.idiart@gmail.com avillavicencio@inf.ufrgs.br rswilkens@inf.ufrgs.br

Abstract

We introduce a new multilingual resource containing judgments about nominal compound compositionality in English, French and Portuguese. It covers 3 × 180 noun-noun and adjective-noun compounds for which we provide numerical compositionality scores for the head word, for the modifier and for the compound as a whole, along with possible paraphrases. This resource was constructed by native speakers via crowdsourcing. It can serve

Eliciting quantitative judgments about compositionality from non-linguists may be too abstract, even with accompanying guidelines and training. We propose a more constrained way of obtaining these judgments, with the participation of non-experts through crowdsourcing. We first focus the participants' attention on compound interpretation in context, by requesting paraphrases in example sentences. Then, we inquire about the degree to which the meaning of a given compound arises from each of its elements. The assumption is that if the interpretation of the compound comes from both nouns (e.g. *access road*), then it is fully compo-

not inc

DSMs and Compositionality

- Dataset of Lexical Substitution of Nominal Compounds in Portuguese (LexSubNC)
  - 180 compounds for Portuguese
  - Resource freely available
    - http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds&lang=en

LexSubNC - Lexical Substitution of Nominal Compounds in Portuguese

- *Rodrigo Wilkens, Leonardo Zilio, Silvio Cordeiro, Felipe S. F. Paula, Carlos Ramisch, Marco Idiart, Aline Villavicencio*
- *Version 1.0 - September 20, 2017*
- Download the data set

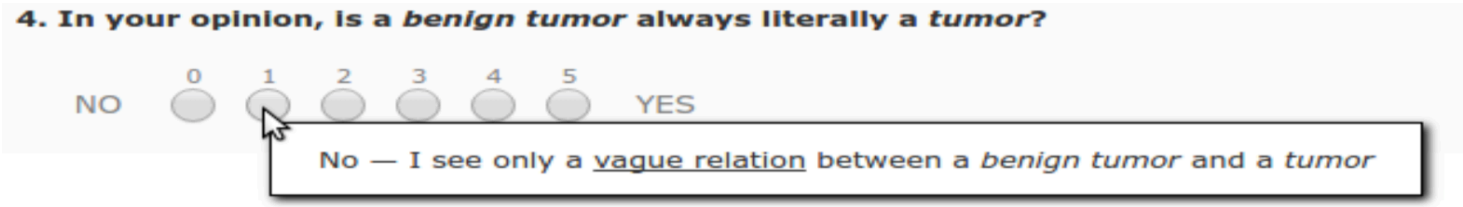
Description

This package is an extension of the original compositionality datasets and includes more detailed annotation for Portuguese lexical substitution candidates in the original dataset. 180 compounds in Portuguese as the compositionality dataset. It additionally contains frequency and PMI from a large Brazilian Portuguese corpos (around 1.2 billion words), as well a annot the following categories:

- Invalid: the substitution candidate is not fit for substitution, either for being too specific for a given context or for simply not being valid for the target MWE.
- Syn-SW: the substitution candidate is a single-word matching synonym in relation to the target MWE.
- NearSyn-SW: the substitution candidate is a single-word quasi-synonym in relation to the target MWE.
- Syn-MWE: the substitution candidate is a multiword matching synonym in relation to the target MWE.
- NearSyn-MWE: the substitution candidate is a multiword quasi-synonym in relation to the target MWE.
- Paraphrase: the substitution candidate is a paraphrasis of the target MWE.
- Definition: the substitution candidate is a definition of the target MWE.
- Head

Collecting Human Judgments

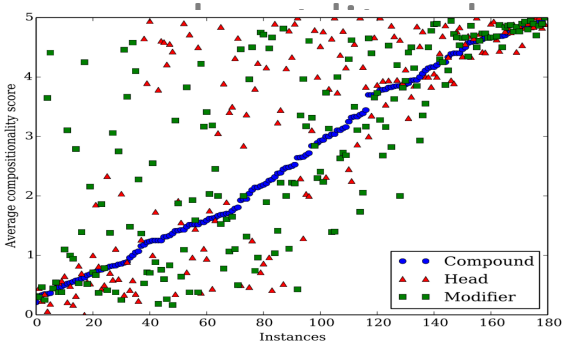
- Judgments with likert scale (0 to 5)
  - For compound
  - For w_1 and w_2 separately



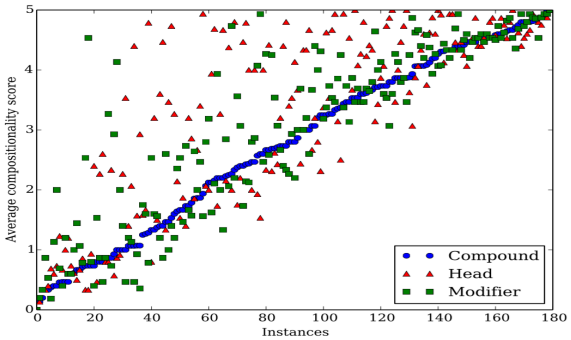
- Agreement for Portuguese
  - For subset of annotators
    - $\alpha = .52$ for head,
    - $\alpha = .36$ for modifier
    - $\alpha = .42$ for compound
  - Same annotator after 1 month: 0.59 for compound

Collecting Human Judgments - Agreement

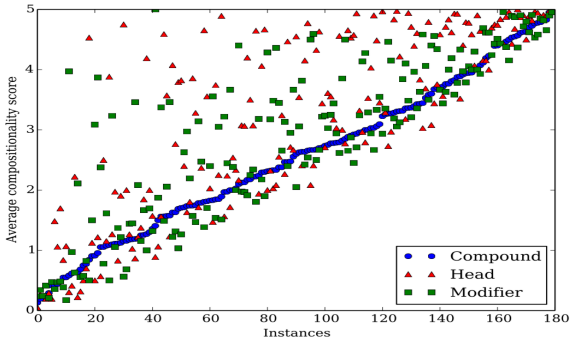
- Greater agreement between score for compound and head (or modifier) for extremes
  - totally idiomatic and fully compositional
- For PT and FR compound score determined by score of the



(a) English



(b) French



(c) Portuguese

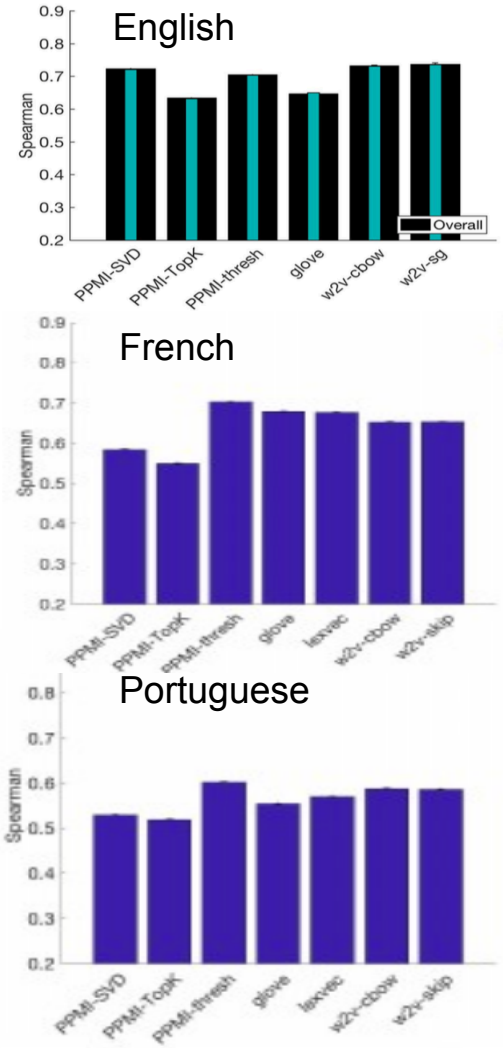
Agreement

- Most/least variation in scores (average±σ score)

	compound	head	mod	comp
English	brass ring	3.9 ±2.0	3.7 ±1.9	3.7 ±1.8
	fish story	4.8 ±0.4	1.5 ±1.8	1.7 ±1.8
	tennis elbow	4.3 ±1.3	2.2 ±1.8	2.5 ±1.8
	brick wall	3.5 ±1.9	3.2 ±2.2	3.8 ±1.7
	dirty word	4.1 ±1.4	2.0 ±1.4	2.5 ±1.7
	prison guard	4.8 ±0.4	4.9 ±0.3	4.9 ±0.3
	graduate student	5.0 ±0.0	4.7 ±0.5	4.9 ±0.3
	engine room	5.0 ±0.0	4.9 ±0.3	4.9 ±0.3
	climate change	4.8 ±0.4	4.9 ±0.3	5.0 ±0.2
	insurance company	4.9 ±0.5	5.0 ±0.0	5.0 ±0.0

The models

- WaCky Corpora (Baroni et al., 2009):
  - ukWaC for English (~2 billion tokens)
  - frWaC (~1.6 billion tokens) for French
  - brWaC (~2.3 billion tokens) for Portuguese (Wagner Filho et al. 2016)
  - Pre-processing
    - *surface+*: the original corpus
    - *surface*: with stopword removal.
    - *lemma*: stopword removal and lemmatization;
    - *lemmaPOS*: stopword removal, lemmatization and POS-tagging
  - Context Window size: 1,4 and 8
  - Dimension size: 250, 500, 750
- DSMs
  - PPMI models – positive PMI (Minimantics)
  - GloVe (Pennington et al. 2014)
  - Word2vec (Mikolov et al 2013) Skipgram, CBOW
  - LexVec



Resources for Text Simplification for Portuguese

Rodrigo Wilkens, Leonardo Zilio, Marco Idiart, Jorge Wagner Filho, Eduardo Ferreira, Luis Mollmann, Bianca Pasqualini, Aline Villavicencio

Federal University of Rio Grande do Sul (Brazil)

Text Simplification (TS)

- TS quality dependent on resources
  - English:
    - Corpora:
      - Simple English Wikipedia parallel corpus, with alignments between the Simple and the Standard English Wikipedia
      - Penn Treebank, British National Corpus, ukWaC
    - Resources for Lexical Substitutions: WordNet, Roget, Moby
    - Gold Standard Substitution Lists: SemEval Lexical Substitution Task
  - Portuguese:
    - Corpora
      - Por Simples (Aluísio et al)
    - Thesauri
      - WordNet.Pt, OntoPT,

Text Simplification

- Two main tasks (Shardlow, 2014):
  - lexical simplification (LS),
    - replacing complex expressions with simpler synonyms,
  - syntactic simplification (SS)
    - change the structure of a sentence by using simpler syntactic constructions (Siddharthan, 2002).
  - TS with MT techniques for monolingual translation
    - learning alignments between simple and standard sentences

General Corpora

- WaCky (Baroni et al. 2009)
  - ukWaC (Baroni et al. 2009)
  - brWaC (Boos et al. 2014, Wagner Filho et al. 2018)
    - Crawling, from medium frequency content words as seeds
      - Linguateca Corpora Frequency List
    - Cleaning
      - HTML and boilerplate stripping, using density metrics and shallow text features
    - Near-duplicate detection and removal
      - pairwise comparison of all documents

Simple Corpora

- For English,
  - Simple English Wikipedia aligned with English Wikipedia (Coster and Kauchak , 2011)

- For Portuguese
  - Coleção É Só o Começo (Wilkens et al. 2014)
    - 5 books manually simplified by linguists.
  - Caseli et al. 2009
    - manually annotated corpus of syntactic and lexical simplifications
  - WikiJunior
    - illustrated books for children up to 12 years old.
  - Projeto PorPopular (Finatto et al. 2012)
    - Tabloids for low literacy readers

Simple Corpora

- Wikilivros Readability Corpus (WRC)
  - Book library from Wikilivros
    - L1: 33 books from 1st to 9th grades
    - L2: 65 books from 10th to 12th grades
    - L3: 21 books for college education
- Readability Assessed WaC (RAW)
  - readability assessment module (Wagner Filho et al. 2016)
    - intermediate module of readability assessment
    - several readability features used as features for classifier
  - 129,000 sentences from L1,
    - 13.5 words per sentence
  - 236,000 sentences from L2
    - 15.2 words per sentence
  - 96,000 sentences from L3,
    - 17.4 words per sentence

Multiword Expressions (MWEs)

- For English
  - NomLex, WordNet
  - Verb-Particle Constructions (Baldwin 2005)
  - Compound Nouns (Nakov 2010, Reddy et al 2011, Yazdani et al 2015, Ramisch et al 2016)
- For Portuguese
  - Light Verb Constructions
    - Duran et al. 2011
  - Compound Nouns
    - Noun Preposition Nouns from Europarl (Zilio et al. 2016)
      - Parsing-based (FIPS) combined with Statistical (PMI)
    - Noun Adjective (Cordeiro et al. 2016)

Simple Words Lists

- Manually created lists
  - For English
    - Oxford 3000
  - For Portuguese
    - 3,853 words from Oxford 3000 translation complemented with most frequent words in corpora (Finatto et al. 2013)

Lexical Substitution

- Manual resources
  - For English
    - WordNet (Fellbaum 1998), Roget, Moby
  - For Portuguese
    - Onto.PT (Oliveira et al. 2010),
    - OpenWN-PT (Paiva et al. 2012),
    - MultiWordnet (Branco et al.),
    - WordNet.PT(Marrafa, 2002),
    - WordNet.Br (Dias da Silva et al., 2008)

Lexical Substitution

- Distributional Semantic Models
  - GloVe, word2vec, Minimantics, Dissect, LexVec
- Quality Evaluation
  - For English
    - WordNet-Based Synonymy Test (WBST) (Freitag et al. 2005)
    - Word similarity and analogy tasks
  - For Portuguese
    - BabelNet-Based Semantic Gold Standard (B^2SG) (Wilkens et al. 2016)
      - synonymy, antonymy and hypernymy for nouns and verbs

Semantic Role Labeling (SRL)

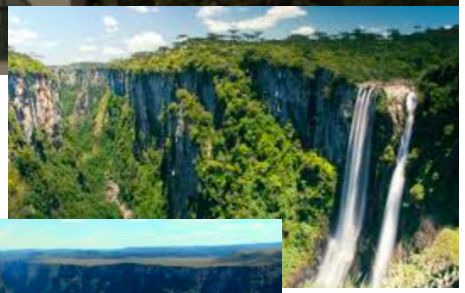
- For word substitution in context
  - For English
    - FrameNet (Baker et al. 1998), PropBank (Kingsbury et al. 2002)
  - For Portuguese
    - PropBank.Br (Duran and Aluísio 2012), VerbNet.Br (Scarton 2013), and FrameNet Brasil (Salomao, 2009).
    - VerbLexPor (Zilio et al. 2016)
      - Cardiology papers vs. Newspaper articles
        - 15,281 annotated arguments (4,192 in CARD and 11,089 in DG)

Resources in Numbers

Resources	Size in English	Size in Portuguese
WaC	>2 billion	3 billion
Lists of simple words	3,000	1,024
WordNet-like	155,287	150,000
Semantic Gold Standard	23,570	2,875
MWE lists	71,888	3,204

Conclusion and Future Work

- English and Portuguese
  - Difference in magnitude and availability of manually constructed resources
- Alternative: language independent methods
  - Extrapolate from manually created resources
  - Corpora
    - brWaC (Wagner Filho et al. 2016)
    - Readability Assessed Wac (Wagner Filho et al. 2016, Wagner Filho et al. 2018)
  - Distributional Semantic Models
    - LexVec (Salle et al. 2016)
  - Gold standards
    - NC Compositionality Dataset (Cordeiro et al. 2016)
    - NC LexSub (Cordeiro et al. 2017)
    - BabelNet-Based Semantic Gold Standard (B^2SG) (Wilkens et al. 2016)



Main conference

PROPOR2018 is the 13th edition of the PROPOR conferences.

PROPOR – International Conference on the Computational Processing of Portugu ng in the area of language and speech technologies for the Portuguese language and on t ssues related to this language. **The event is supported by the PROPOR steering committe**

Acknowledgments

- This work has been funded by the
  - Brazilian Agency CNPq (482520/2012-4 and 312114/2015-0)
  - Projects "Simplificação Textual de Expressões Complexas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91, and
  - French Agence Nationale pour la Recherche through projects PARSEME-FR (ANR-14-CERA-0001) and ORFEO (ANR-12-CORP-0005), and by
  - French-Brazilian cooperation projects CAMELEON (CAPES-COFECUB 707/11) and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7).

Building Resources for Human and Computational Language Processing of Portuguese

Sílvio Cordeiro, Carlos Ramisch, Marco Idiart, Rodrigo Wilkens, Leonardo Zilio, Jorge Wagner, Aline Villavicencio

Federal University of Rio Grande do Sul (Brazil)

Aix Marseille Université, CNRS, LIF UMR 7279 (France)

University of Essex (UK)