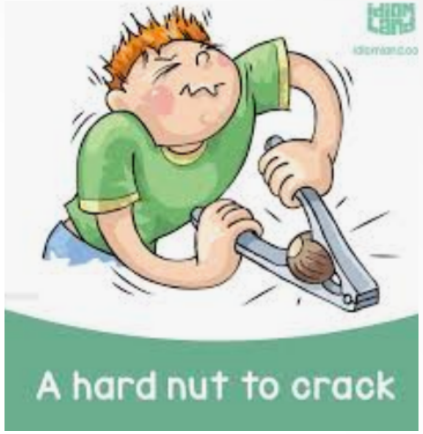


Multiword Expressions and Idiomaticity:

How Much of the Sailing Has Been Plain?

Are MWEs still a hard nut to crack?

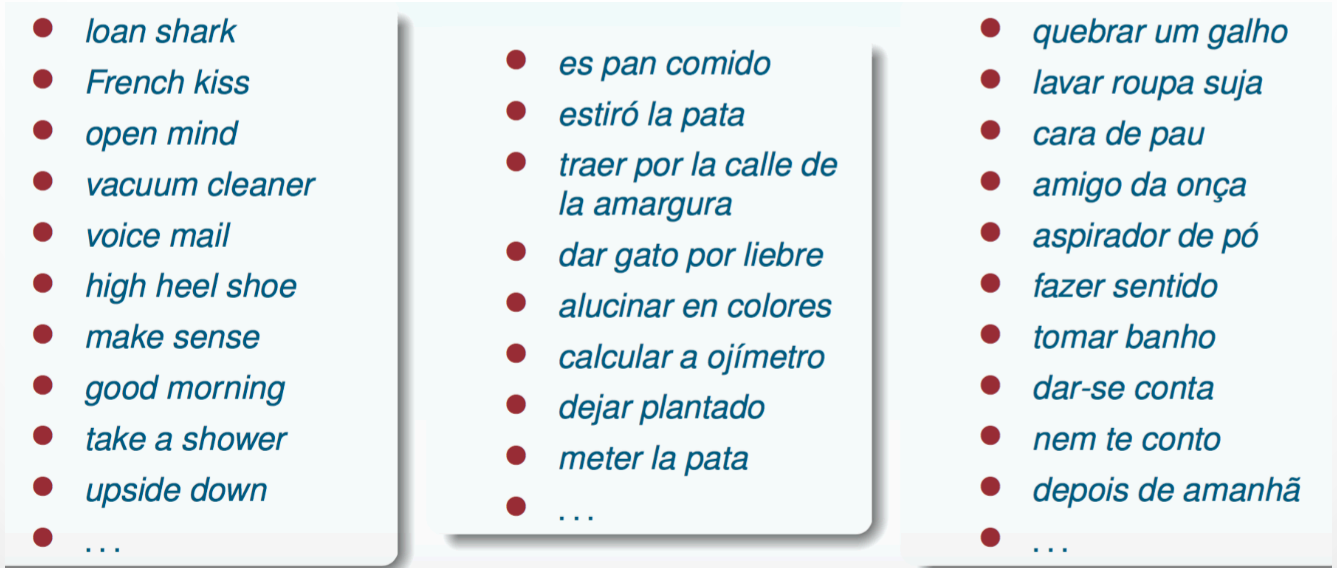
Aline Villavicencio

University of Sheffield (UK)

Federal University of Rio Grande do Sul (Brazil)

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



- *loan shark*
- *French kiss*
- *open mind*
- *vacuum cleaner*
- *voice mail*
- *high heel shoe*
- *make sense*
- *good morning*
- *take a shower*
- *upside down*
- . . .

- *es pan comido*
- *estiró la pata*
- *traer por la calle de la amargura*
- *dar gato por liebre*
- *alucinar en colores*
- *calcular a ojímetro*
- *dejar plantado*
- *meter la pata*
- . . .

- *quebrar um galho*
- *lavar roupa suja*
- *cara de pau*
- *amigo da onça*
- *aspirador de pó*
- *fazer sentido*
- *tomar banho*
- *dar-se conta*
- *nem te conto*
- *depois de amanhã*
- . . .

Multiword Expressions in a Nutshell

- Lexical, syntactic, semantic, pragmatic, statistical idiosyncrasies
    - o *Ad hoc, wine and dine (Kim and Baldwin 2010)*
- Arbitrariness and Institutionalisation
    - o *salt and pepper, ?pepper and salt* (Smadja, 1993)
- Limited lexical, syntactic and semantic variability
    - o *kick the bucket/?pail/?container* (Sag et al., 2002)

MWEs are all around

- 4 MWEs produced per minute of discourse (Glucksberg 1989)
- Same order of magnitude as single words in mental lexicon of native speakers (Jackendoff 1997)
- Large proportion of technical language (Biber et al. 1999)
- Faster processing times compared to non-MWEs (Cacciari and Tabossi 1988; Arnon and Snider 2010; Siyanova-Chanturia 2013)

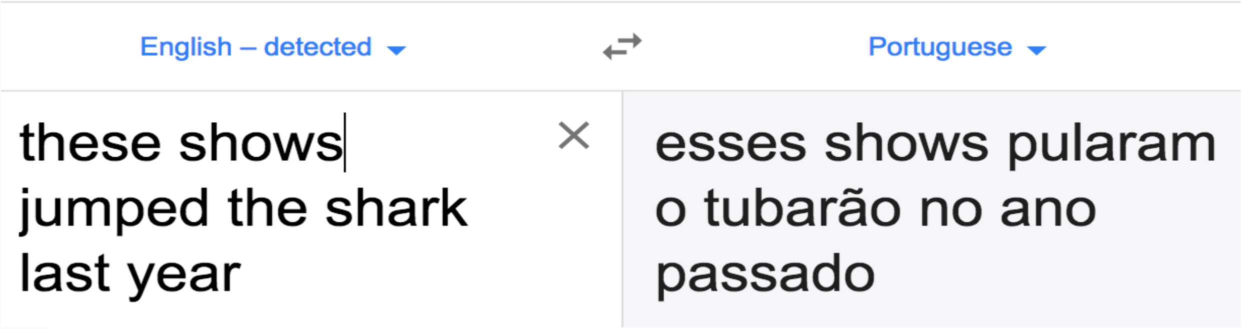
What happens if we ignore them?

11 TV Shows That Jumped The Shark

- Refers to the specific moment when a TV show goes downhill. Originally from *Happy Days*

MWEs and NLP

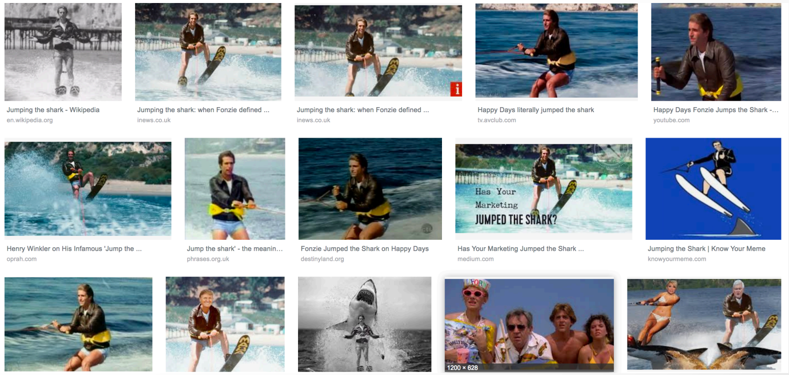
- Machine Translation



these shows jumped the shark last year

esses shows pularam o tubarão no ano passado

- Text Simplification
  - They moved over the fish
- Information Retrieval


Jumping the shark - Wikipedia
en.wikipedia.org
Jumping the shark: when Fonzie defined ...
inews.co.uk
Jumping the shark: when Fonzie defined ...
inews.co.uk
Happy Days literally jumped the shark
tv.avclub.com
Happy Days Fonzie Jumps the Shark -...
youtube.com
Henry Winkler on His Infamous 'Jump the ...
oprah.com
Jump the shark' - the meanin...
phrases.org.uk
Fonzie Jumped the Shark on Happy Days
destinyland.org
Has Your Marketing JUMPED THE SHARK?
Has Your Marketing Jumped the Shark ...
medium.com
Jumping the Shark | Know Your Meme
knowyourmeme.com
1200 × 628

Processing MWEs

- For NLP, given a combination of words determine if
  - It is a MWE
    - *Rocket science* vs. *small boy*
  - How syntactically flexible it is
    - *Kick the bucket, ?the bucket has been kicked*
  - If it is idiomatic
    - *Rocket science* vs. *olive oil*
  - Decide if it can be processed accurately using compositional approaches

Processing MWEs

- Clues from:
  - Collocational Preferences
    - Recurrent word combinations
  - Contextual Preferences
    - (Dis)similarities between contexts of MWE and of its components
  - Canonical Form Preferences
    - Limited number of variant forms
  - Multilingual Preferences
    - (A)symmetries for MWE in different languages

Processing MWEs

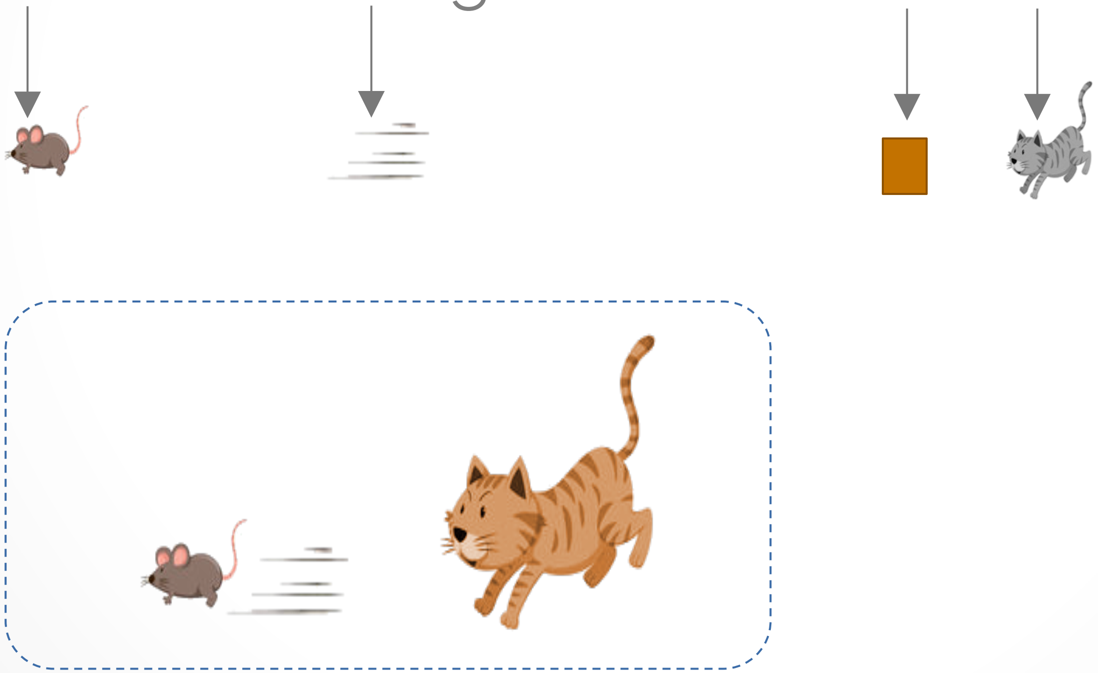
- Clues from:
  - Collocational Preferences
    - Recurrent word combinations
  - **Contextual Preferences**
    - **(Dis)similarities between contexts of MWE and of its components**
  - Canonical Form Preferences
    - Limited number of variant forms
  - Multilingual Preferences
    - (A)symmetries for MWE in different languages

Outline

- Multiword Expressions (MWEs)
- Idiomaticity Detection
- Distributional Semantic Models (DSMs)
- Gold Standards for Compositionality
- Multilingual Evaluation
- Conclusions and Future Work

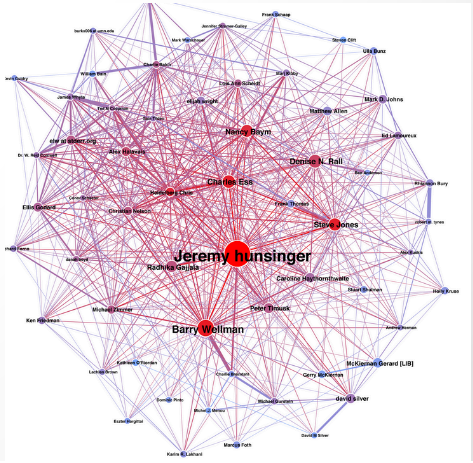
NLP and the Principle of Compositionality

- The meaning of the **whole** comes from the meaning of the **parts**.
- *“The mouse is running from the brown cat”*



NLP and the Principle of Compositionality

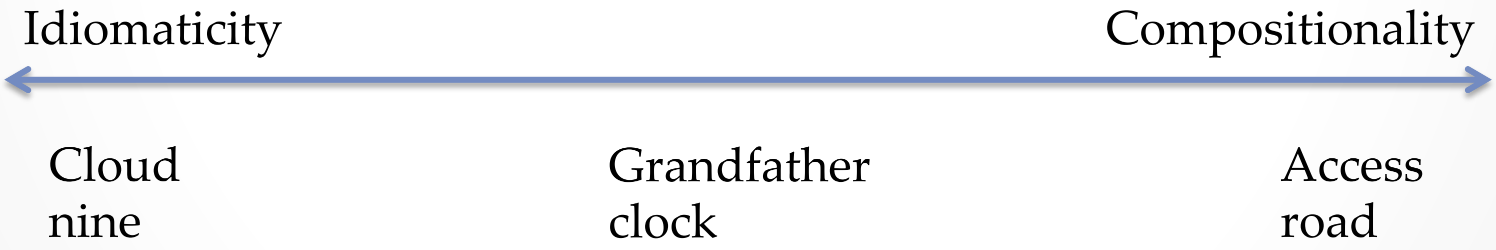
- Distributional Semantic Models (DSMs)
  - o You shall know a word by the company it keeps (Firth 1957)
    - *Famous author* ***writes*** *book under a pseudonym*
  - o Words that occur in similar contexts have similar meanings (Turney and Pantel 2010)
    - Author ***writes/rewrites/composes/creates/prepares*** book
  - o Position words in multidimensional semantic space
    - Each word represented as a vector
      - o coordinates in the semantic space
    - Proximity in space indicates semantic relatedness



-

Compositionality vs. Idiomaticity

- 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



Idiomaticity

Compositionality

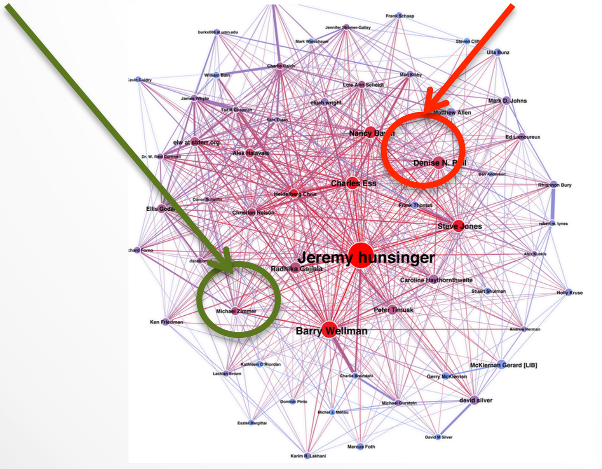
Cloud nine

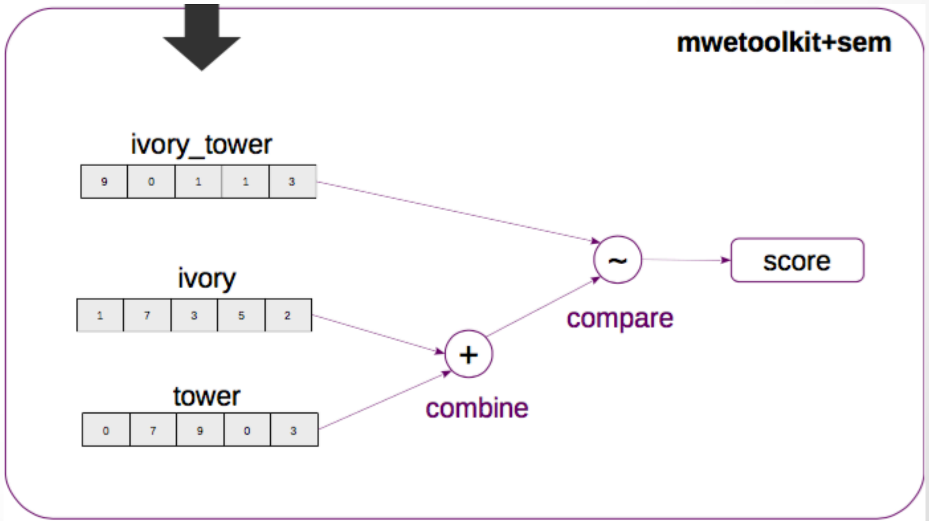
Grandfather clock

Access road

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
  - Additive operation (Mitchell and Lapata, 2010, Reddy et al. 2011, Cordeiro et al. 2019)
  - Other operations (Socher et al. 2011, Salehi et al. 2015, Zhao et al. 2015, Qi et al. 2019)
  - **cos($w_1w_2$vector, $w_1$vector+$w_2$vector)**





ivory_tower

9	0	1	1	3

ivory

1	7	3	5	2

tower

0	7	9	0	3

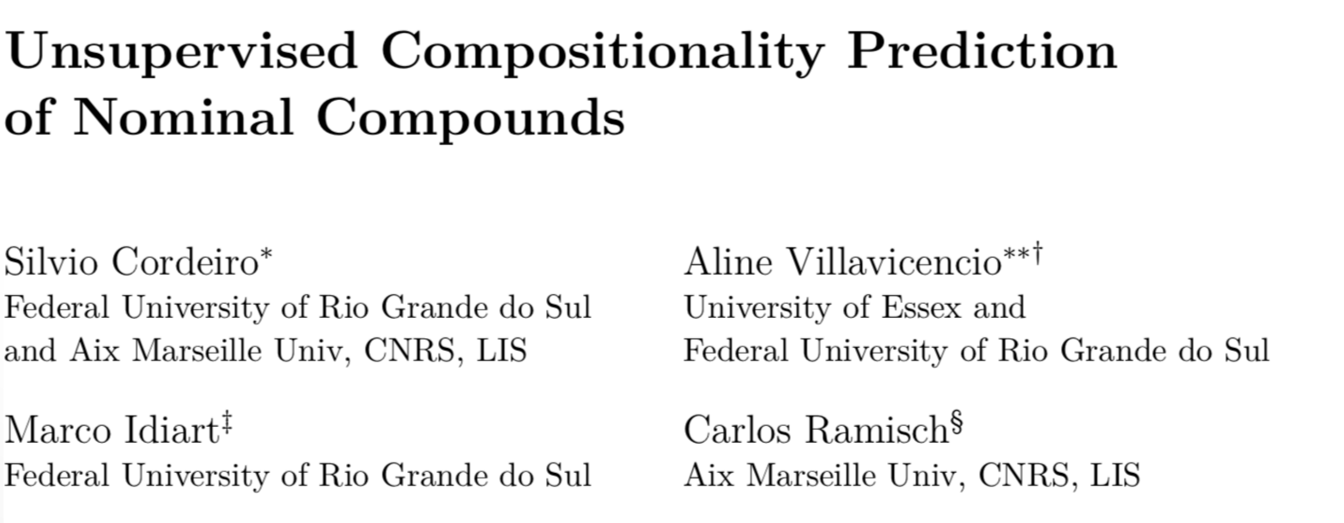
How to detect compositionality?

- To what extent the meaning of MWE can be computed from the meanings of component words using DSMs
  - Is accuracy in prediction dependent on
    - characteristics of the DSMs ?
    - the language/corpora ?

How to detect compositionality?

- Over 9,000 analyses and 680 DSMs detailed in

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, Carlos Ramisch, "**Unsupervised Compositionality Prediction of Nominal Compounds**", *Computational Linguistics*, 45(1):1--57, 2019, MIT Press.



Unsupervised Compositionality Prediction of Nominal Compounds

Silvio Cordeiro*
Federal University of Rio Grande do Sul
and Aix Marseille Univ, CNRS, LIS

Aline Villavicencio**†
University of Essex and
Federal University of Rio Grande do Sul

Marco Idiart‡
Federal University of Rio Grande do Sul

Carlos Ramisch§
Aix Marseille Univ, CNRS, LIS

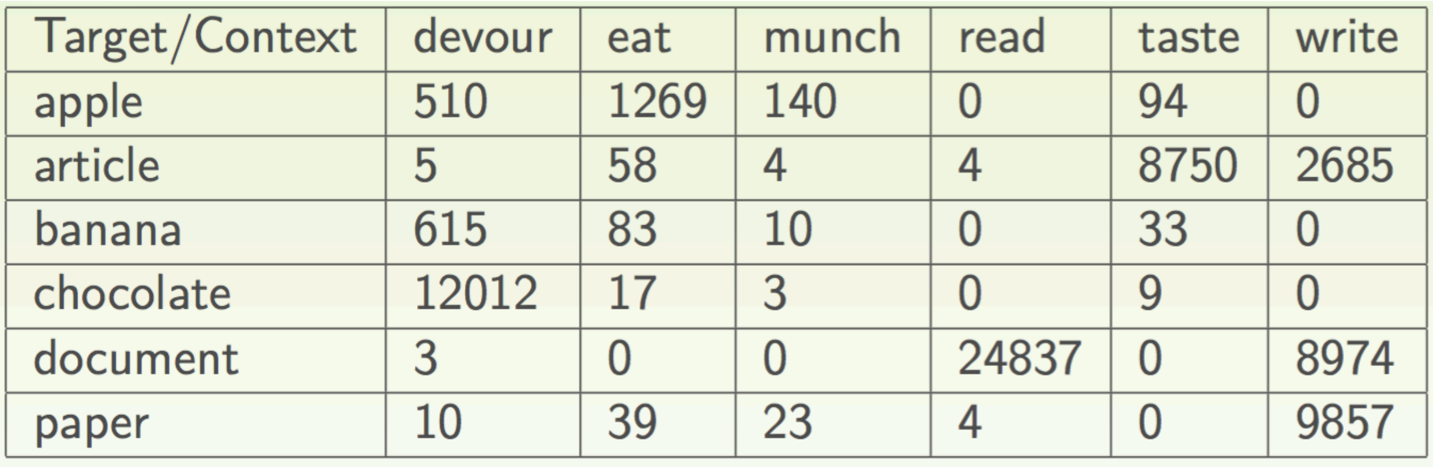
-

Outline

- Multiword Expressions (MWEs)
- Idiomaticity Detection
- Distributional Semantic Models (DSMs)
- Gold Standards for Compositionality
- Multilingual Evaluation
- Conclusions and Future Work

Distributional Semantic Models

- Distributional Hypothesis
  1. Count Targets and Contexts in corpus
     - The man ate chocolate → (eat,man), (eat,chocolate)
  2. Compute association strength between targets and contexts
  3. Compute similarity between targets



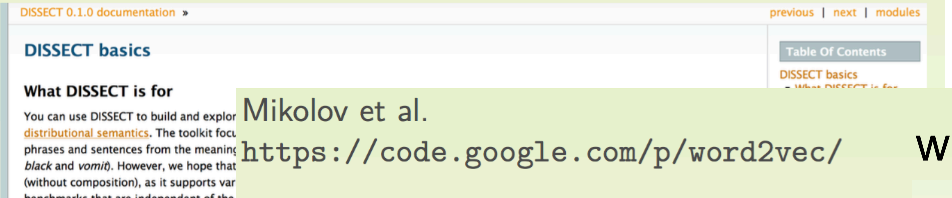
Target/Context	devour	eat	munch	read	taste	write
apple	510	1269	140	0	94	0
article	5	58	4	4	8750	2685
banana	615	83	10	0	33	0
chocolate	12012	17	3	0	9	0
document	3	0	0	24837	0	8974
paper	10	39	23	4	0	9857

Distributional Semantic Models

- Constructing DSMs
  - Dissect (Dinu et al., 2013), Minimantics (Ramisch et al. 2013), word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014).

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

dissect



DISSECT 0.1.0 documentation » previous | next | modules

DISSECT basics

What DISSECT is for

You can use DISSECT to build and explor distributional semantics. The toolkit focu phrases and sentences from the meaning *black* and *vomit*). However, we hope that (without composition), as it supports var benchmarks that are independent of the

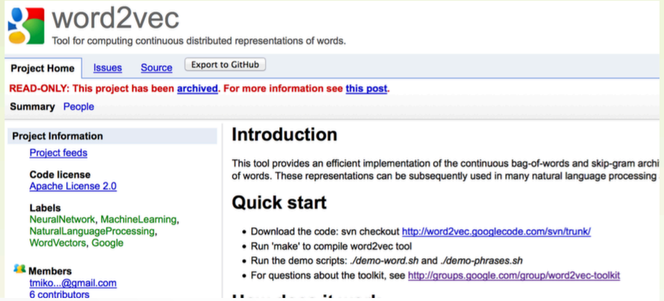
Table Of Contents

DISSECT basics

Mikolov et al.

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

word2vec



word2vec

Tool for computing continuous distributed representations of words.

READ-ONLY: This project has been archived. For more information see this post.

Project Information

Project feeds

Code license
Apache License 2.0

Labels
NeuralNetwork, MachineLearning, NaturalLanguageProcessing, WordVectors, Google

Members
tmiko...@gmail.com
6 contributors

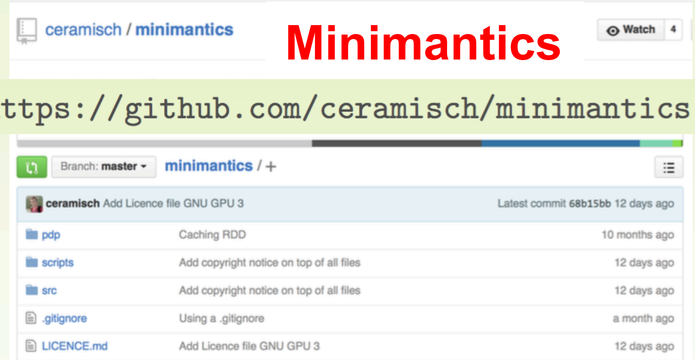
Introduction

This tool provides an efficient implementation of the continuous bag-of-words and skip-gram archi of words. These representations can be subsequently used in many natural language processing

Quick start

- Download the code: svn checkout http://word2vec.googlecode.com/svn/trunk/
- Run 'make' to compile word2vec tool
- Run the demo scripts: *./demo-word.sh* and *./demo-phrases.sh*
- For questions about the toolkit, see http://groups.google.com/group/word2vec-toolkit

Minimantics



ceramisch / minimantics

`https://github.com/ceramisch/minimantics`

ceramisch Add Licence file GNU GPU 3 — Latest commit 68b15bb 12 days ago

pdp	Caching RDD	10 months ago
scripts	Add copyright notice on top of all files	12 days ago
src	Add copyright notice on top of all files	12 days ago
.gitignore	Using a .gitignore	a month ago
LICENCE.md	Add Licence file GNU GPU 3	12 days ago

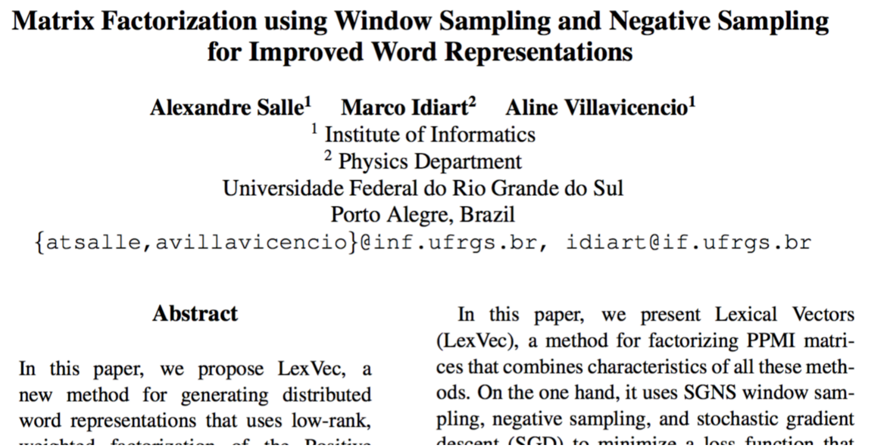
GloVe

ELMo

Distributional Semantic Models

- LexVec (Lexical Vectors)
  - Alternative that outperforms word2vec and GloVe in word similarity tasks
    - Freely available

https://github.com/alexandres/lexvec



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
{atsalle,avillavicencio}@inf.ufrgs.br, idiart@if.ufrgs.br

Abstract

In this paper, we propose LexVec, a new method for generating distributed word representations that uses low-rank, weighted factorization of the Positive

In this paper, we present Lexical Vectors (LexVec), a method for factorizing PPMI matrices that combines characteristics of all these methods. On the one hand, it uses SGNS window sampling, negative sampling, and stochastic gradient descent (SGD) to minimize a loss function that

Project SAMSUNG

The models

- DSMs
  - PPMI models – positive PMI (Minimantics)
  - GloVe (Pennington et al. 2014)
  - Word2vec (Mikolov et al 2013) Skipgram, CBOW
  - LexVec (Salle et al. 2016, 2018)
- 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

Outline

- Multiword Expressions (MWEs)
- Idiomaticity Detection
- Distributional Semantic Models (DSMs)
- Gold Standards for Compositionality
- Multilingual Evaluation
- Conclusions and Future Work

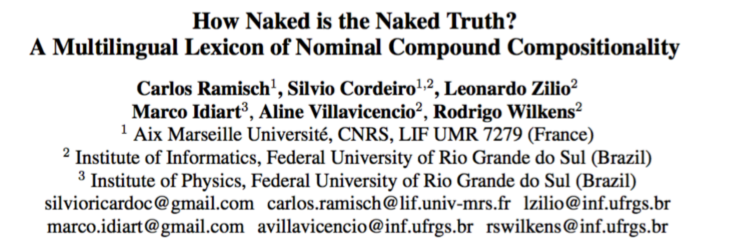
Gold Standards

- 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
- We used Reddy's protocol as basis to add 180 compounds and expand to other languages

Collecting Human Judgments

- Multilingual dataset
  - 270 English compounds: N_1 N_2, and A_1 N_2
    - *olive oil*
    - extends Reddy et al. 2011 with 180 compounds
  - 180 French compounds: N_2 A_1
    - *mort cellulaire (cell death)*
  - 180 Portuguese compounds: N_2 A_1
    - *morte celular (cell death)*
- Balanced for compositionality
  - 60 idiomatic, 60 partially compositional and 60 compositional

ACL 2016



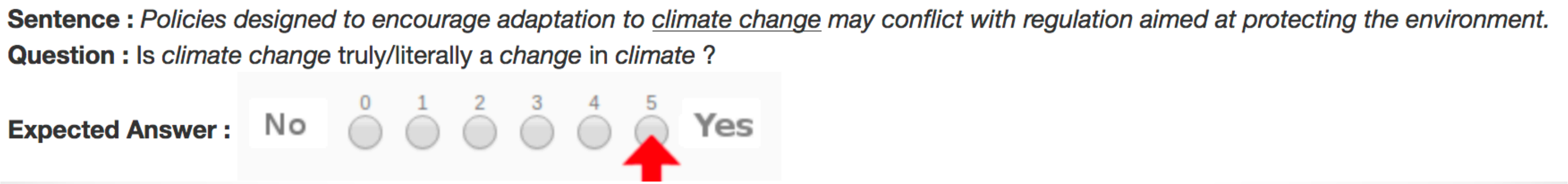
How Naked is the Naked Truth?
A Multilingual Lexicon of Nominal Compound Compositionality

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

Project FAPERGS-CNRS-INRIA (France Brazil)

Collecting Human Judgments

- Following Reddy et al. (2011) use literality to approximate compositionality
- Judgments with likert scale (0 to 5)
  - For **compound**
  - For w_1 and
  - For w_2 separately



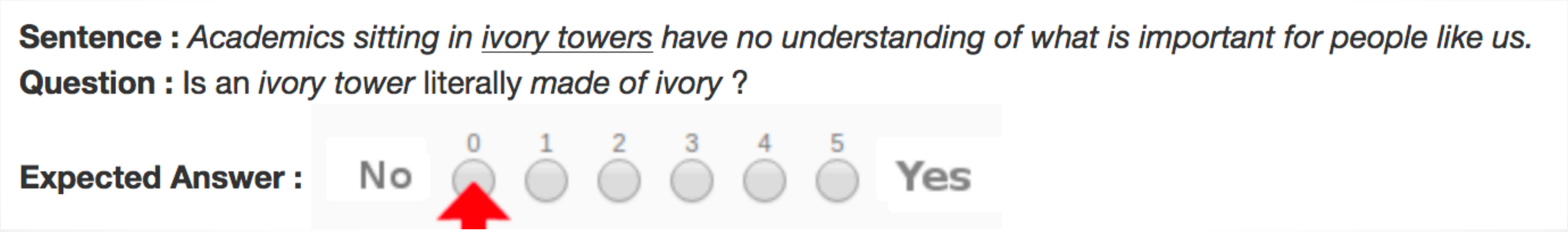
Sentence : *Policies designed to encourage adaptation to climate change may conflict with regulation aimed at protecting the environment.*

Question : Is *climate change* truly/literally a *change* in *climate* ?

Expected Answer : No 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ [✓ 5 ○] Yes

Collecting Human Judgments

- Following Reddy et al. (2011) use literality to approximate compositionality
- Judgments with likert scale (0 to 5)
  - For compound
  - For $\mathbf{w}_1$ and for w_2 separately



Sentence : *Academics sitting in ivory towers have no understanding of what is important for people like us.*

Question : Is an *ivory tower* literally *made of ivory* ?

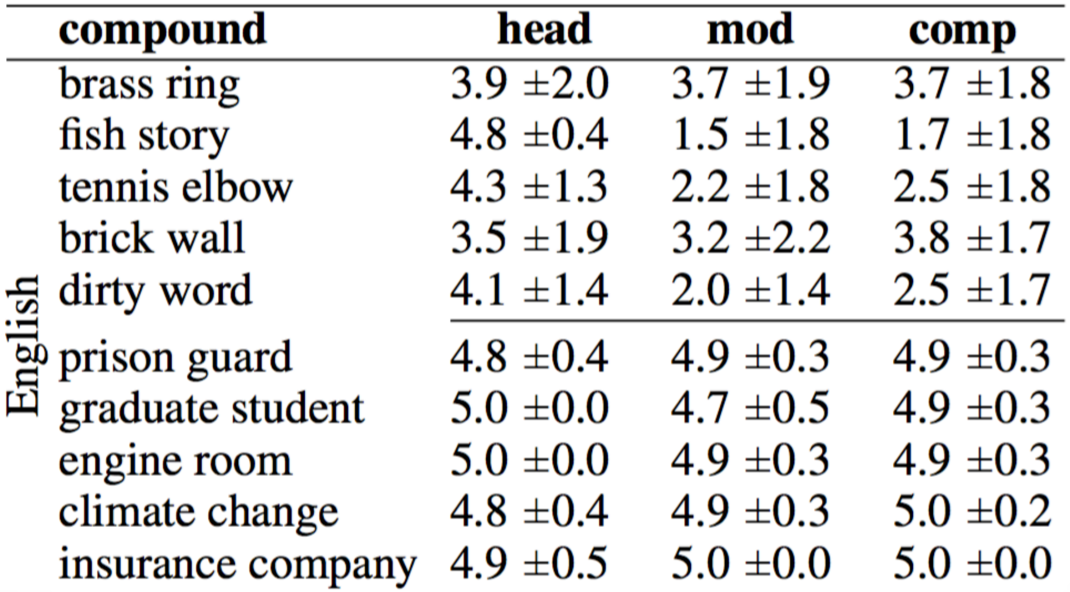
Expected Answer : No [✓ ○ 0] ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Yes

Collecting Human Judgments - Agreement

- Context: 3 sentences per compound
  - Compound has same meaning in all sentences
- Participants: linguists, CS students, AMT workers
  - Non-expert participants
- For Portuguese
  - For subset of annotators
    - α = .52 for head,
    - α = .36 for modifier
    - α = .42 for compound
  - Same annotator after 1 month:
    - α = .59 for compound
    - ρ = .77 for compound
      - qualitative upper bound for compositionality prediction on *PT-comp*.

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

Outline

- Multiword Expressions (MWEs)
- Idiomaticity Detection
- Distributional Semantic Models (DSMs)
- Gold Standards for Compositionality
- Multilingual Evaluation
- Conclusions and Future Work

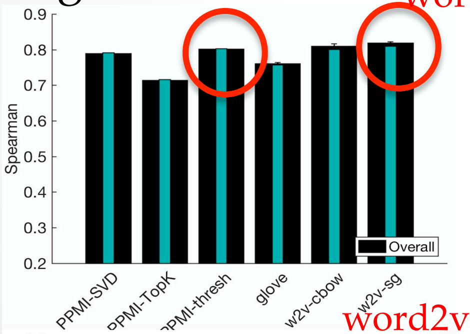
Evaluation

- Comparing model predictions with average human judgment
  - English Reddy: word2vec, Spearman ρ =0.82
  - English Reddy++: word2vec, Spearman ρ =0.73
  - French: PPMI global context, Spearman ρ =0.70
  - Portuguese: PPMI global context, Spearman ρ =0.60

English

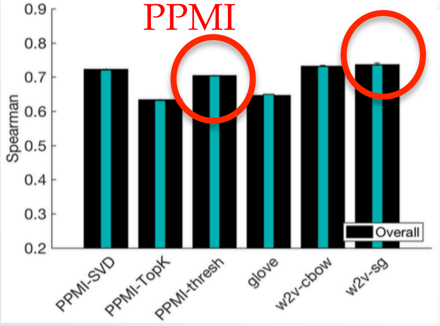
PPMI

word2vec



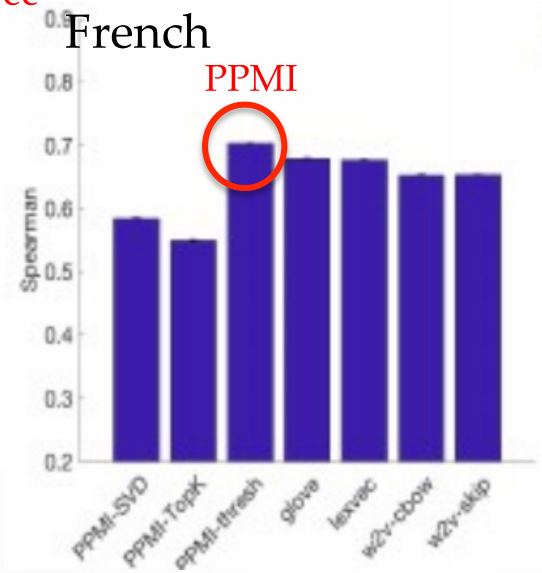
PPMI

word2vec



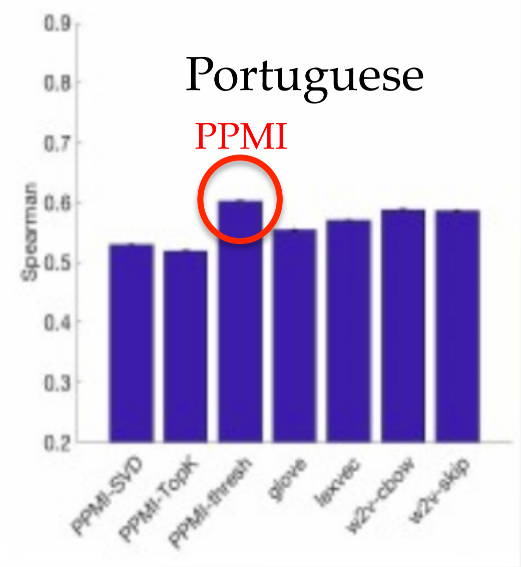
French

PPMI



Portuguese

PPMI

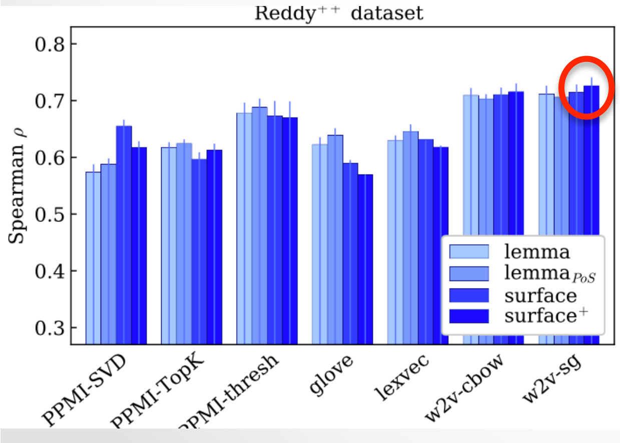


Evaluation – Type of Preprocessing

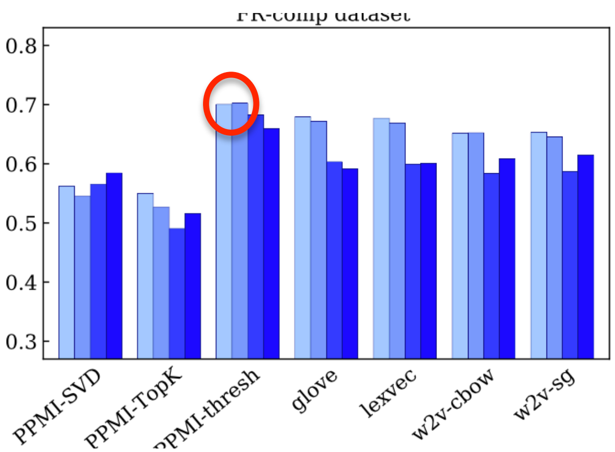
- Do less sparse representations lead to better results?
  - Not for English: preprocessing makes no differences for best model
  - Yes for French and Portuguese: lemma-based models considerably better for best models

▨ lemma
▨ lemma$_{PoS}$
▨ surface
▨ surface^{+}

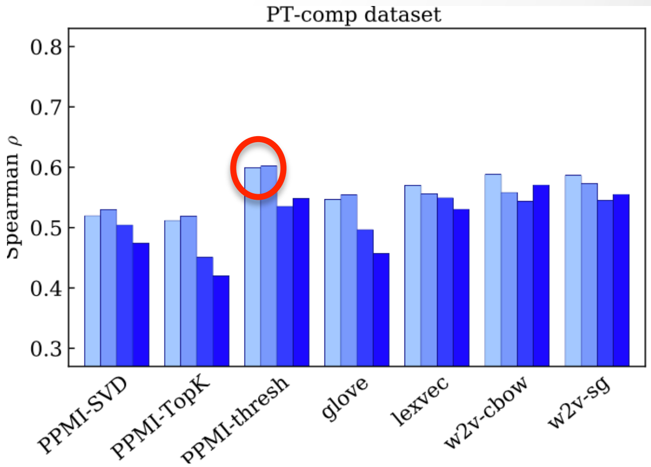
English



French

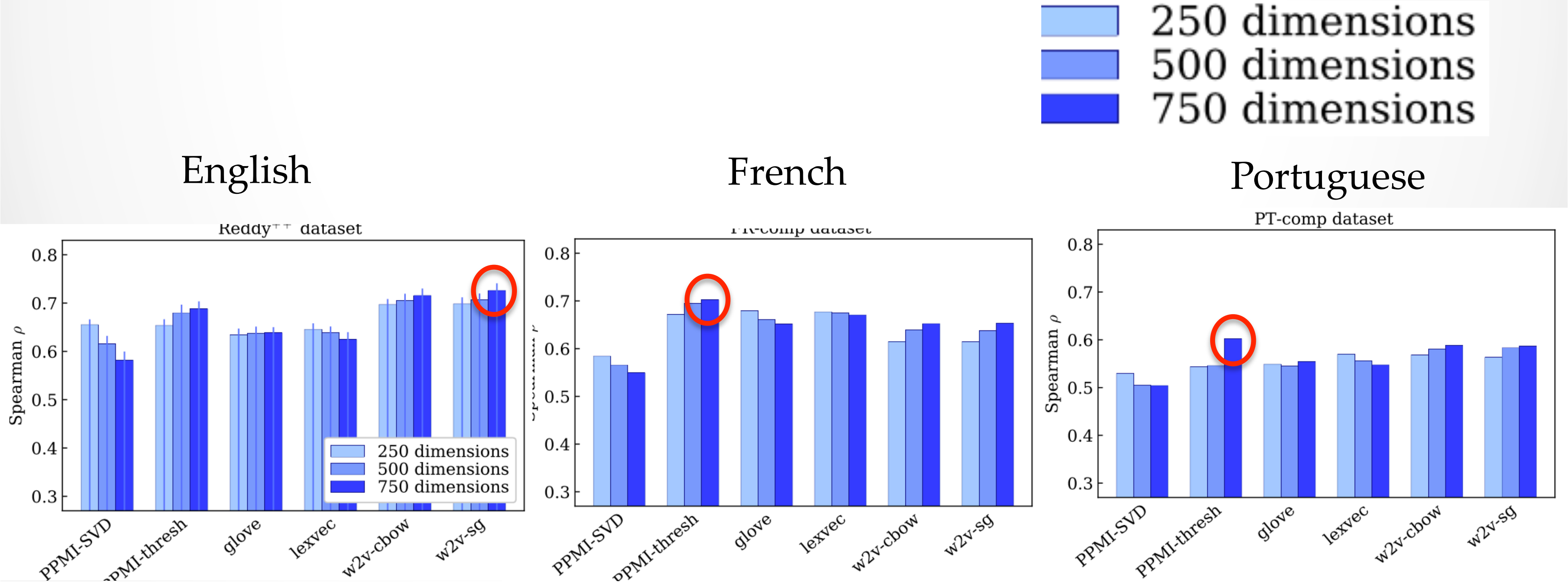


Portuguese



Evaluation – Number of Dimensions

- Do larger dimensions lead to more accurate models/better results?
  - Yes for English, French and Portuguese: more dimensions lead to better results


English

French

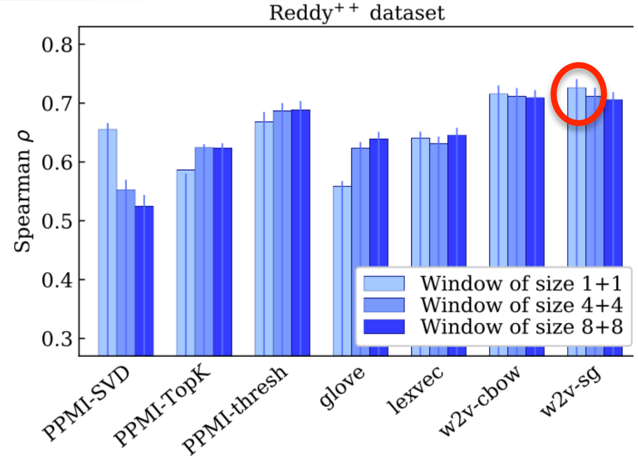
Portuguese

Evaluation – Size of Context Window

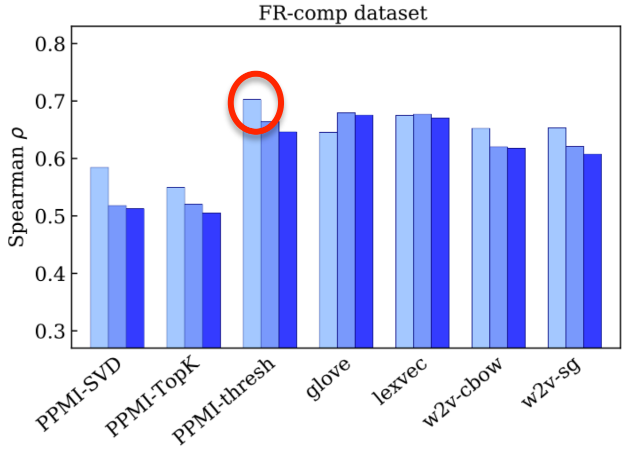
- Do larger window sizes lead to better results?
  - Not for English, French and Portuguese: trend for smaller windows in best models

▭ Window of size 1+1
▭ Window of size 4+4
■ Window of size 8+8

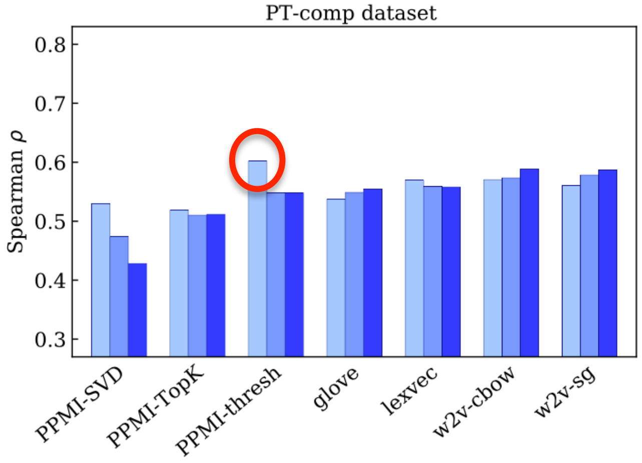
English



French

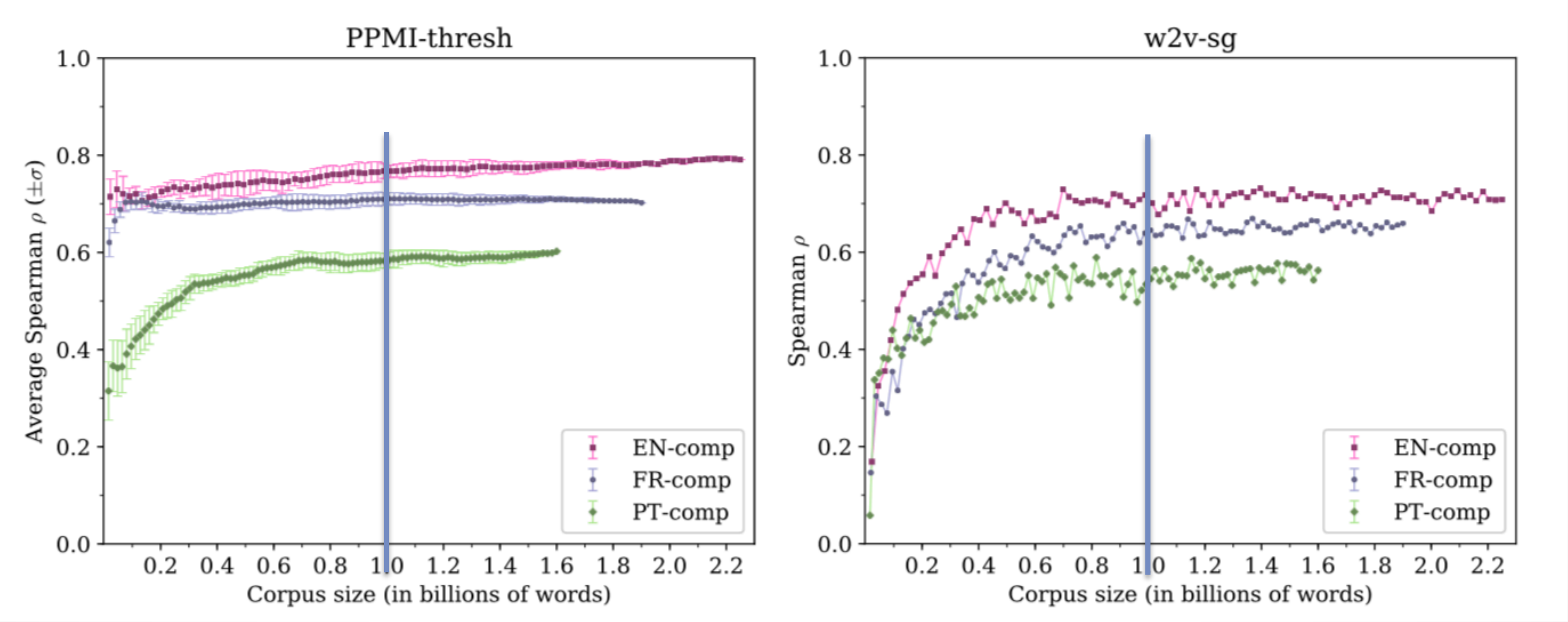


Portuguese



Evaluation – Corpus Size

- Are better results for English due to larger corpus size?



- Not for English, French and Portuguese:
  - stable performance after ~1 billion words
    - all compounds may be frequent enough for accurate representations

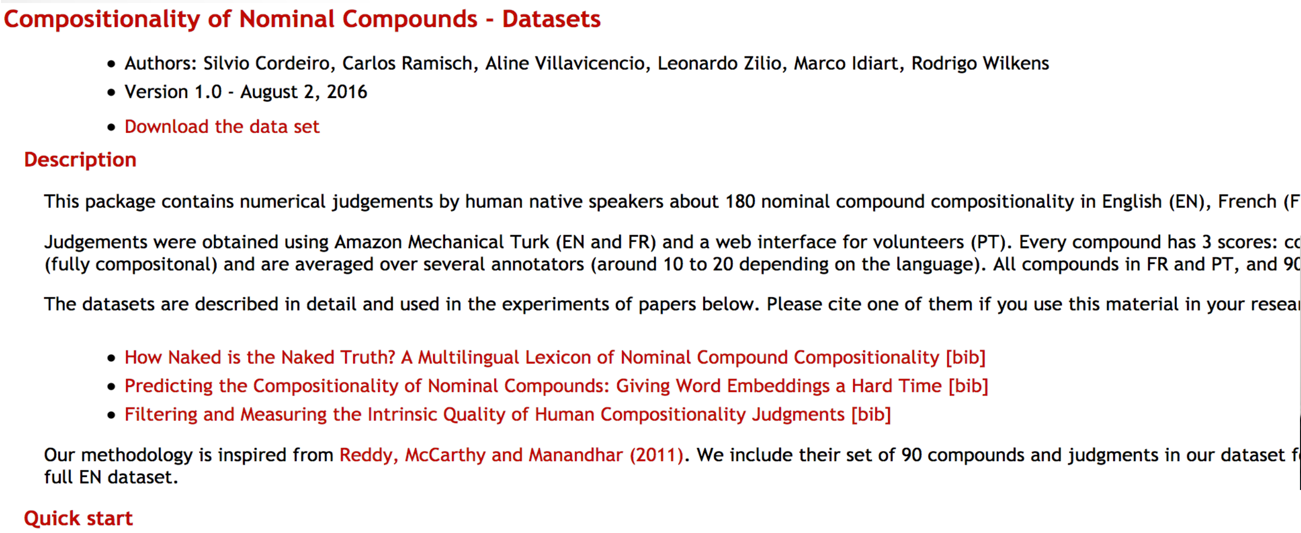
Conclusions

How to detect compositionality?

- To what extent the meaning of MWE can be computed from the meanings of component words
  - Compared to human judgments how accurate DSMs are for MWEs of various levels of compositionality?
  - Is accuracy in prediction dependent on characteristics of the DSMs ?
  - Is accuracy in prediction dependent on the language/ corpora ?

DSMs and Compositionality

- Dataset of nominal compounds with human judgments about literality/compositionality
  - o 270 compounds for English, 180 compounds for French and Portuguese
  - o 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, Marco Idiart, Rodrigo Wilkens
- Version 1.0 - August 2, 2016
- Download the data set

Description

This package contains numerical judgements by human native speakers about 180 nominal compound compositionality in English (EN), French (FR) and Brazilian Portugu

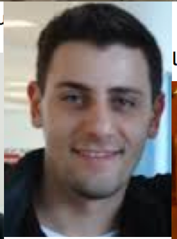
Judgements were obtained using Amazon Mechanical Turk (EN and FR) and a web interface for volunteers (PT). Every compound has 3 scores: co (fully compositonal) and are averaged over several annotators (around 10 to 20 depending on the language). All compounds in FR and PT, and 90

The datasets are described in detail and used in the experiments of papers below. Please cite one of them if you use this material in your resear

- How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality [bib]
- Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time [bib]
- Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments [bib]

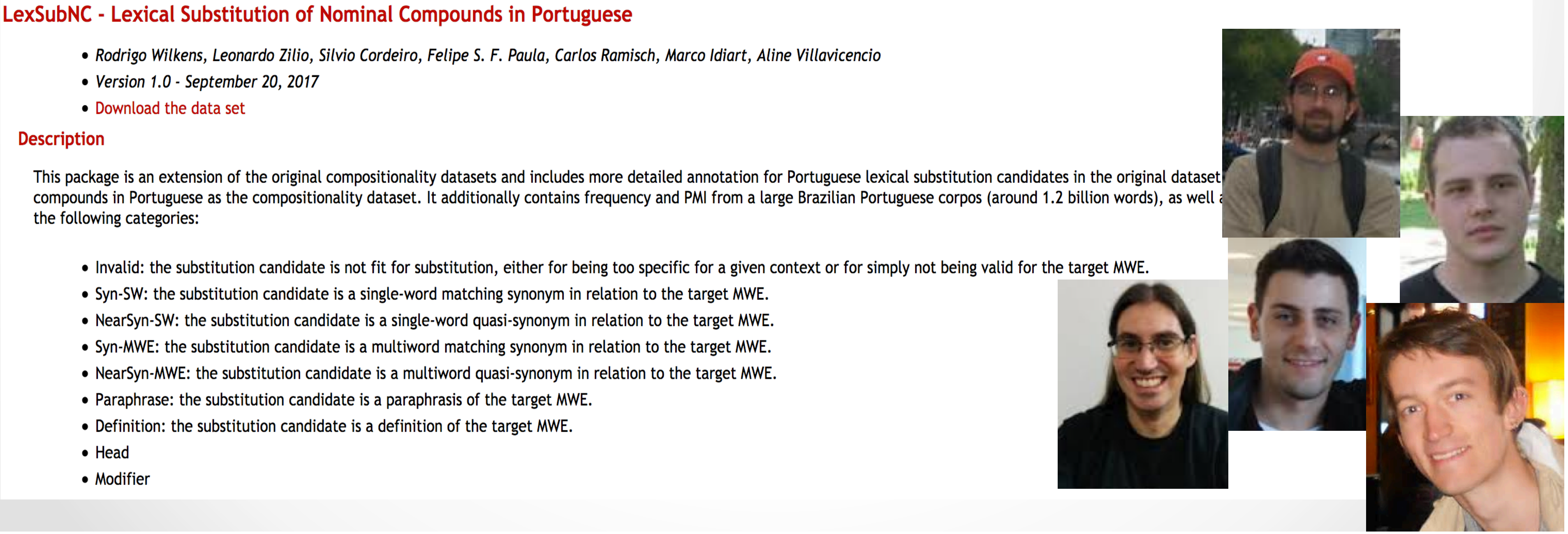
Our methodology is inspired from Reddy, McCarthy and Manandhar (2011). We include their set of 90 compounds and judgments in our dataset f apers. We do not in full EN dataset.

Quick start



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 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 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
- Modifier

DSMs and Compositionality

- Large-scale multilingual analysis of DSMs for compound compositionality prediction
  - in English, French and Portuguese
  - Over 600 DSMs and
  - Almost 9000 evaluations
  - 3 families of models: word2vec, GloVe, and PPMI-based models.

mwetoolkit.sf.net

mwetoolkit: Home

Home

> Installing the mwetoolkit
> Quick Start
> Defining Patterns
> Preprocessing a corpus using TreeTagger
> Preprocessing a corpus using Rasp
> Developers

Welcome to the mwetoolkit

News

- 09 Mar 2012: Release 0.5 is now available on sourceforge! Lots of new features and up-to-date documentation are here! Enjoy :-)
- 17 May 2011: The mwetoolkit release 0.4 is now available on sourceforge!
- 17 May 2011: We created a Quick Start guide for you to understand how the toolkit works

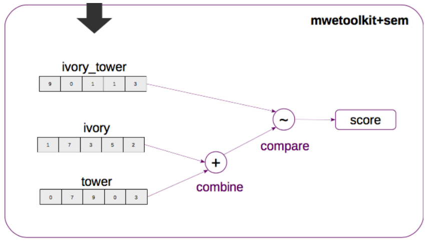
The Multiword Expressions toolkit is a research tool that aids in the automatic identification of multiword units such as idiomatic expressions (kick the bucket) and phrasal verbs (take off, give up) in large text bases, independently of the language.

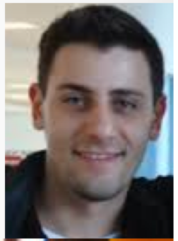
Links

- Sourceforge project
- Multiword Expressions community website
- Mweoolkit Developer Wiki

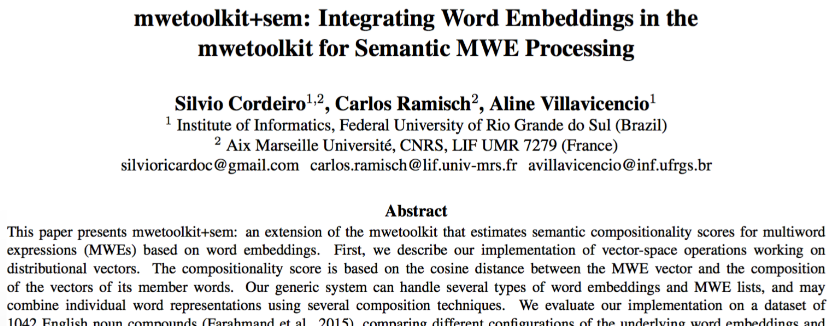
mwetoolkit

- 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, PPMI)
- Provides functions for vector combinations
- Calculates compositionality
- Evaluates against gold standard





LREC 2016



mwetoolkit+sem: Integrating Word Embeddings in the mwetoolkit for Semantic MWE Processing

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

Abstract

This paper presents mwetoolkit+sem: an extension of the mwetoolkit that estimates semantic compositionality scores for multiword expressions (MWEs) based on word embeddings. First, we describe our implementation of vector-space operations working on distributional vectors. The compositionality score is based on the cosine distance between the MWE vector and the composition of the vectors of its member words. Our generic system can handle several types of word embeddings and MWE lists, and may combine individual word representations using several composition techniques. We evaluate our implementation on a dataset of 1042 English noun compounds (Farahmand et al., 2015), comparing different configurations of the underlying word embeddings and

Project CAPES-COFECUB (France-Brazil)

Future Work

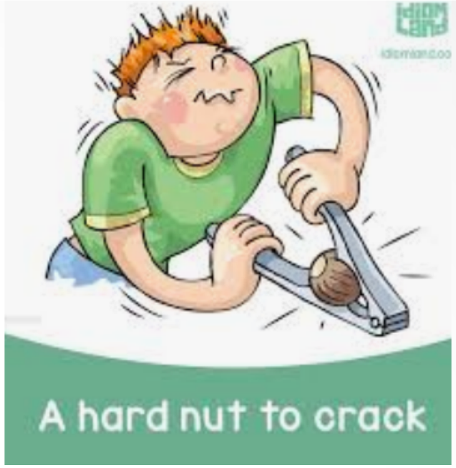
- More accurate MWE representation
  - ACL 2019: Jana et al. 2019, Qi et al. 2019
  - MWE 2019
- Token idiomaticity identification
  - Gharbieh et al. 2017, Taslimipoor et a.l 2017, King and Cook 2018
  - Fixedness detection as indication of idiomaticity
    - Limited degree of variation for idiomatic MWEs (Ramisch et al. 2008, Geeraert et al. 2017)
    - Preference for canonical form for idiomatic MWEs (Fazly et al. 2009, Taslimipoor et al. 2017, King and Cook 2018)
    - Less similarity with variants for idiomatic MWEs in DSMs (Senaldi et al. 2019)

This research was done in collaboration with Carlos Ramisch, Marco Idiart, Silvio Cordeiro, Rodrigo Wilkens and Leonardo Zilio

This work was partly supported by the Brazilian Research Council (CNPq 423843/2016-8).

Thank you

• • •



Multiword Expressions and Idiomaticity:

How Much of the Sailing Has Been Plain?

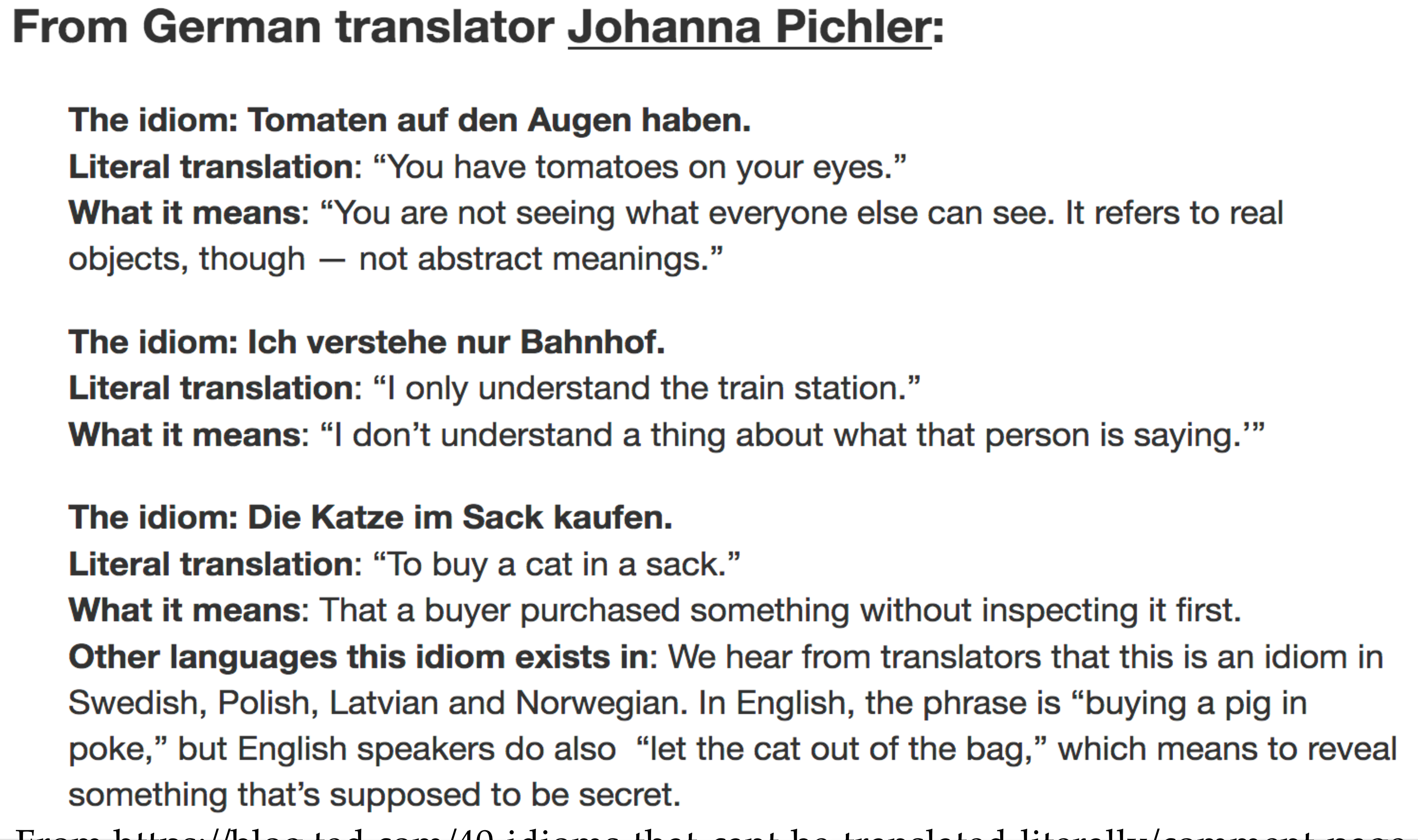
Are MWEs still a hard nut to crack?

Aline Villavicencio

University of Sheffield (UK)

Federal University of Rio Grande do Sul (Brazil)

Idioms that can’t be translated literally ?



From German translator Johanna Pichler:

The idiom: Tomaten auf den Augen haben.
Literal translation: “You have tomatoes on your eyes.”
What it means: “You are not seeing what everyone else can see. It refers to real objects, though — not abstract meanings.”

The idiom: Ich verstehe nur Bahnhof.
Literal translation: “I only understand the train station.”
What it means: “I don’t understand a thing about what that person is saying.’”

The idiom: Die Katze im Sack kaufen.
Literal translation: “To buy a cat in a sack.”
What it means: That a buyer purchased something without inspecting it first.
Other languages this idiom exists in: We hear from translators that this is an idiom in Swedish, Polish, Latvian and Norwegian. In English, the phrase is “buying a pig in poke,” but English speakers do also “let the cat out of the bag,” which means to reveal something that’s supposed to be secret.

From https://blog.ted.com/40-idioms-that-cant-be-translated-literally/comment-page-10/