# Determining the Semantic Compositionality of Croatian Multiword Expressions

Petra Almić and Jan Šnajder

University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab

Ninth Language Technologies Conference
Information Society
Jožef Stefan Institute
Ljubljana, October 9–10, 2014

# The problem

- MWEs require special attention in NLP

**Semantic compositionality**

Degree to which the features of the parts of an MWE combine to predict the features of the whole [Baldwin, 2006].
Compositional MWEs: *world war, yellow tape*
Non-compositional MWEs: *cold war, red tape*

- In reality, MWEs populate a continuum between two extremes [Bannard et al., 2003]
- Determining compositionality useful for many NLP tasks (machine translation, information retrieval, word sense disambiguation...)

# Our approach

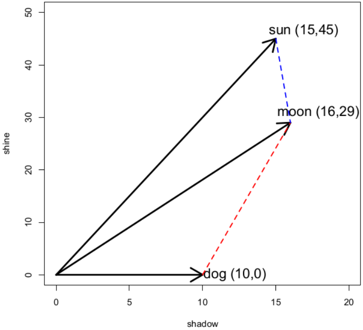
- We follow up on the works of Katz and Giesbrecht [2006] and Biemann and Giesbrecht [2011]
- Idea: compare the meaning of an MWE against the meaning of the composition of its parts
  $\rightarrow$ world $\oplus$ war $=$ world war ?
- To model the meanings of words, we use *distributional semantics*
- Our contribution:
  - we build a small dataset of Croatian MWEs annotated with semantic compositionality scores
  - we build and evaulate a semantic compositionality model based on **Latent Semantic Analysis** [Landauer et al., 1998]
  - results comparable to relevant RW

# Distributional semantics

- Representation of word meaning based on distributional hypothesis [Harris, 1954]:
  - correlation between similarity of words' contexts and words' semantic similarity
- Words represented as vectors of context features obtained from corpus
- Semantic similarity predicted via vector similarity
- Distributional semantic models used in many applications [Turney and Pantel, 2010]

# Distributional semantic models

| | planet | night | full | shadow | shine | crescent |
|---|---|---|---|---|---|---|
| moon | 10 | 22 | 43 | 16 | 29 | 12 |
| sun | 14 | 10 | 4 | 15 | 45 | 0 |
| dog | 0 | 4 | 2 | 10 | 0 | 0 |




(Marco Baroni's EACL 2012 tutorial: Compositionality in Distributional Semantics)

# Dataset

- Corpus: fHrWaC [Šnajder et al., 2013], filtered version of hrWaC [Ljubešić and Erjavec, 2011]
- Three MWE types:
  1. **AN**: *žuti karton (yellow card)*
  2. **SV**: *podatak govori (data says)*
  3. **VO**: *popiti kavu (drink coffee)*
- We extracted the most frequent MWEs and pre-annotated each as compositional (C) or non-compositional (NC)
- Final dataset was balanced to include roughly equal number of C and NC MWEs

# Annotation

- Setup: 200 MWEs, 24 annotators
- Score aggregation: median

| MWE | Score |
|---|---|
| *maslinovo ulje (olive oil)* | 5 |
| *telefonska linije (telephone line)* | 4 |
| *pružiti pomoć (to offer help)* | 4 |
| *kućni ljubimac (a pet)* | 3.5 |
| *crno tržište (black market)* | 3 |
| *voditi brigu (to worry)* | 3 |
| *ostaviti dojam (to leave an impression)* | 2.5 |
| *zeleno svjetlo (green light)* | 1 |
| *hladni rat (cold war)* | 1 |
| ⋮ | ⋮ |

- Average Spearman's correlation coefficient: 0.77
- Dataset split in development (100 MWEs) and test set (100 MWEs)

# Compositionality model

**Step 1:** model the meaning of constituent words and MWEs

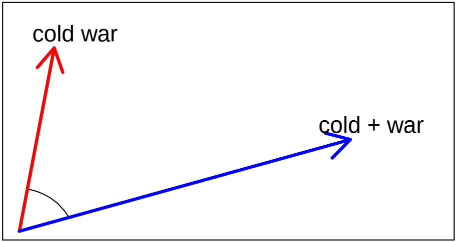
- Latent Semantic Analysis
- $\pm 5$ words context window, 10K most freq. words (excl. stopwords)

**Step 2:** model the composed meaning from constituents

- six compositional models

**Step 3:** compare composed meaning against MWE meaning

- cosine similarity between word vectors



# Distributional semantic composition

($\vec{z}$ – composed vector; $\vec{x}$, $\vec{y}$ – constituents' vectors)

- **multiplicative**: $\vec{z} = \vec{x} \odot \vec{y}$
- **simple additive**: $\vec{z} = \vec{x} + \vec{y}$
- **weighted additive**: $\vec{z} = \alpha\vec{x} + \beta\vec{y}$
  - opt: weights optimized globally on the train set
  - dyn: constituent more similar to MWE more important (*gray economy*)

$$\alpha = \frac{\cos(\overrightarrow{xy}, \vec{x})}{\cos(\overrightarrow{xy}, \vec{x}) + \cos(\overrightarrow{xy}, \vec{y})}, \quad \beta = 1 - \alpha$$

- **first constituent**: $\vec{z} = \vec{x}$
- **second constituent**: $\vec{z} = \vec{y}$
- **linear combination**:

$$\lambda = a_0 + a_1 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x + y}) + a_2 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x \odot y}) + a_3 \cdot \cos(\overrightarrow{xy}, \vec{x}) + a_4 \cdot \cos(\overrightarrow{xy}, \vec{y})$$

# Results – Predicting compositionality scores

| Model | AN+SV+VO | AN | SV+VO |
|---|---|---|---|
| Multiplicative | $-0.19$ | $-0.20$ | $-0.18$ |
| Simple additive | 0.45 | 0.54 | **0.35** |
| Weighted additive (Opt) | 0.46 | 0.56 | 0.28 |
| Weighted additive (Dyn) | 0.46 | **0.57** | 0.26 |
| First constituent | 0.41 | 0.50 | 0.19 |
| Second constituent | 0.28 | 0.31 | 0.31 |
| Linear combination ($\lambda$) | **0.48** | 0.56 | 0.34 |
| Annotators | 0.77 | 0.77 | 0.74 |

- Combining multiple models beneficial
- AN compositionality easier to predict (AN easier to model?)

# Results – Compositionality classification

- Dataset: score $\leq 3 \Rightarrow$ MWE is non-compositional
- Linear combination model
- The threshold optimized on the train set by optimizing the F1-score

| | AN+SV+VO | AN | SV+VO |
|---|---|---|---|
| Precision | $0.58$ | $0.74$ | $0.43$ |
| Recall | $0.73$ | $0.65$ | $0.77$ |
| Accuracy | $0.65$ | $0.72$ | $0.54$ |
| F1-score | $0.65$ | $0.69$ | $0.56$ |

# Conclusion

- A composition-based model for determining semantic compositionality of Croatian MWEs
- The best-performing model combines the additive and the multiplicative compositional models and the representations of the two individual words
- Annotated dataset available from takelab.fer.hr/cromwesc
- Future work wishlist:
  - enlarge the dataset
  - consider using an unbalanced dataset
  - error analysis
  - supervised compositionality classification
  - experiment with neural word embeddings
  - token based semantic compositionality detection

# References I


Timothy Baldwin. Compositionality and multiword expressions: Six of one, half a dozen of the other. In *Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 2006.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 65–72. ACL, 2003. doi: 10.3115/1119282.1119291. URL http://dx.doi.org/10.3115/1119282.1119291.

Chris Biemann and Eugenie Giesbrecht. Distributional semantics and compositionality 2011: Shared task description and results. In *Proc. of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. ACL, 2011. URL http://dl.acm.org/citation.cfm?id=2043121.2043125.

Zelig S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

# References II


Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proc. of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. ACL, 2006.

T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998. URL http://lsa.colorado.edu/papers/dp1.LSAintro.pdf.

Nikola Ljubešić and Tomaž Erjavec. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, pages 395–402. Springer, 2011.

Jan Šnajder, Sebastian Padó, and Željko Agić. Building and evaluating a distributional memory for Croatian. In *In Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789. ACL, 2013.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

# Annotation (1)

Annotation setup:

- 200 MWEs randomly split in 4 groups (A, B, C, D)
- 24 annotators $\Rightarrow$ each MWE annotated by 6 annotators
- 10% overlap
- question: how literal an MWE is on the scale from 1 (non-compositional) to 5 (compositional)?
- one context sentence provided for each MWE
- final score: median

# Annotation (2)

Inter-annotator agreement (Krippendorff's $\alpha$):

| Sample | AN+SV+VO | AN | SV+VO |
|---|---|---|---|
| Group A | 0.587 | 0.620 | 0.535 |
| Group B | 0.506 | 0.510 | 0.478 |
| Group C | 0.490 | 0.544 | 0.337 |
| Group D | 0.586 | 0.505 | 0.648 |
| Overlap (10%) | 0.456 | 0.452 | 0.439 |

# Levels of compositionality

- MWEs come in different "flavors of compositionality"
- In an attempt to identify different levels of non-compositionality, we developed the following typology:
  - **NC3**: completely non-compositional
    $\rightarrow$ *žuti karton (yellow card)*
  - **NC2**: partially compositional
    $\rightarrow$ *siva ekonomija (gray economy)*
  - **NC1**: non-compositional considering the dominant senses
    $\rightarrow$ *planinski lanac (mountain chain)*

# Results analysis

- Moderate level of correlation
- Comparable to Biemann and Giesbrecht [2011] and Katz and Giesbrecht [2006]
- Possible causes of error:
  - low quality of vector representations for some words
  - polysemy