Morpho-syntactic Regularities in Continuous Word Representations: A Multilingual Study

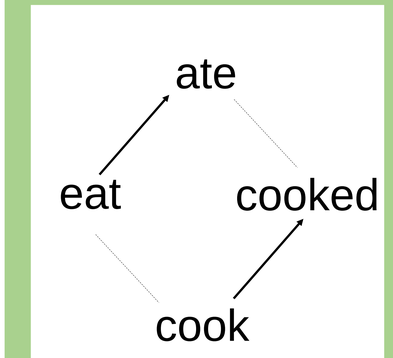
Garrett Nicolai[1] Colin Cherry[2] Greg Kondrak[1]

[1]Department of Computing Science
University of Alberta

[2]National Research Council
Ottawa, Ontario, Canada

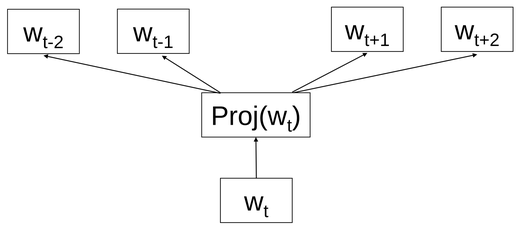
- We replicate the experiments of Mikolov et al. (2013) that find syntactic regularities in English, using *Word2Vec*.
- We then extend the experiments to four languages with more complex morphology.
- Finally, we investigate the role of window size and training corpus size on analogy accuracy.

Syntactic Analogy



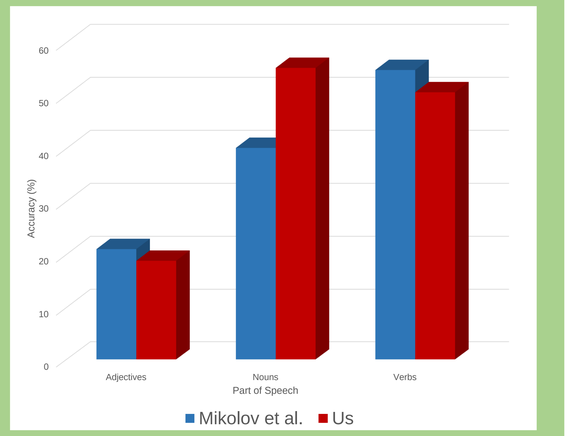
- Syntactic analogies are of the form:
  a is to *b* as *c* is to *?*
- Analogies are created by tagging our raw corpus, and selecting the 100 most frequent base forms, which are paired with other base forms from the same list.
- Morphological information is obtained from CELEX.

Continuous Skip-gram Representation



- For replication, we train 640-dimensional vectors using the skip-gram model with a hierarchical softmax, a context window of 10, sub-sampling of 1e-3, and a minimum threshold of 10.
- We use the same test set as Mikolov et al. (2013), but modify the nominal set to exclude possessive forms.

Replication

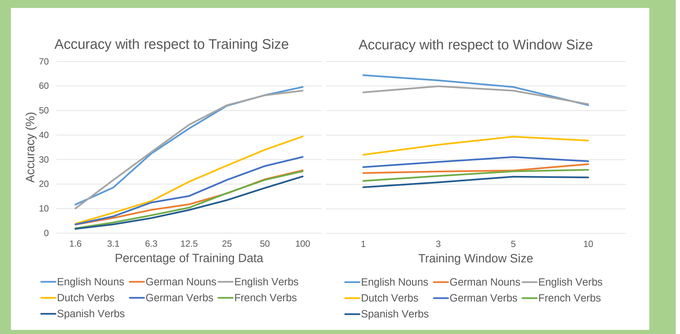


We replicate the results of Mikolov et al. (2013).
Differences can be attributed to:

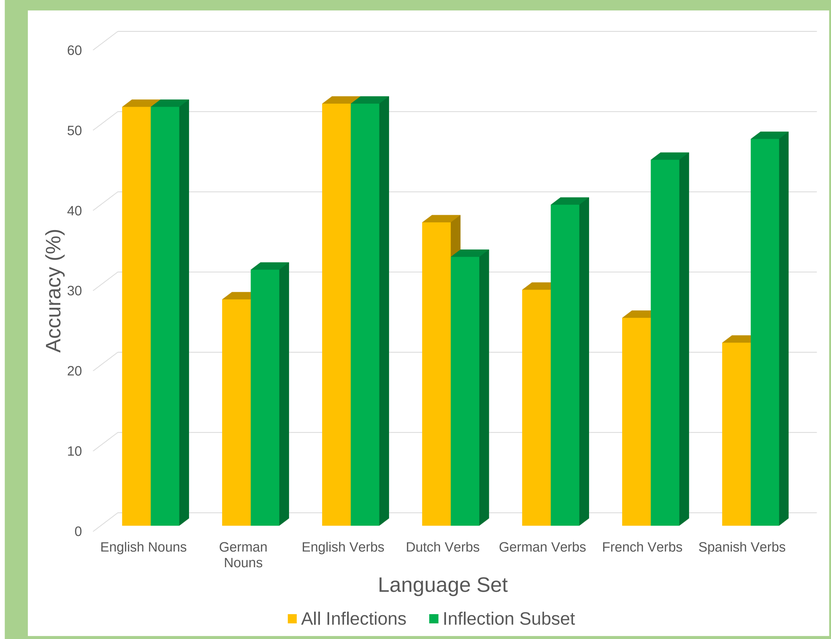
- Different training corpora
- Approximation of training parameters

Multi-lingual Experiments

Set		Inflections	Example
Verbs		5	go:gone see:?
		9	gaan:gegaan zien:?
		27	gehen:gegangen sehen:?
		48	aller:allé voir:?
		57	ir:ido ver:?
Nouns		2	bear:bears lion:?
		8	Bär:Bären Löwe:?



- We see that while English is starting to converge, the other languages continue to benefit from more training data.
- While accuracy for English is higher when the window size is small, larger window sizes benefit other languages.



- As the number of inflections in a paradigm increases, the accuracy of the system decreases (yellow).
- For a subset of common inflections, this trend reverses (green).
- We conjecture that a larger number of inflections may make individual forms easier to disambiguate.