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

**Garrett Nicolai**[†] **Colin Cherry**[‡] **Grzegorz Kondrak**[†]

[†]Department of Computing Science
University of Alberta
Edmonton, AB, T6G 2E8, Canada
{nicolai,gkondrak}@ualberta.ca

[‡]National Research Council Canada
1200 Montreal Road
Ottawa, ON, K1A 0R6, Canada
Colin.Cherry@nrc-cnrc.gc.ca

## Abstract

We replicate the syntactic experiments of Mikolov et al. (2013b) on English, and expand them to include morphologically complex languages. We learn vector representations for Dutch, French, German, and Spanish with the WORD2VEC tool, and investigate to what extent inflectional information is preserved across vectors. We observe that the accuracy of vectors on a set of syntactic analogies is inversely correlated with the morphological complexity of the language.

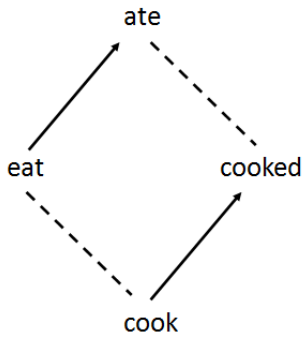


Figure 1: An example of vector offsets.

## 1 Introduction

Mikolov et al. (2013b) demonstrate that vector representations of words obtained from a neural network language model provide a way of capturing both semantic and syntactic regularities in language. They observe that by manipulating vector offsets between pairs of words, it is possible to derive an approximation of vectors representing other words, such as *queen* $\approx$ *king* $-$ *man* $+$ *woman*. Similarly, an abstract relationship between the present and past tense may be computed by subtracting the base form *eat* from the past form *ate*; the result of composing such an offset with the base form *cook* may turn out to be similar to the vector for *cooked* (Figure 1). They report state-of-the-art results on a set of analogy questions of the form "$a$ is to $b$ as $c$ is to _", where the variables represent various English word forms.

Our work is motivated by two observations regarding Mikolov et al.'s experiments: first, the syntactic analogies that they test correspond to morphological inflections, and second, the tests only evaluate English, a language with little morphological complexity. In this paper, we replicate their syntactic experiments on four languages that are more morphologically complex than English: Dutch, French, German, and Spanish.

## 2 Replication Experiments

In order to to validate our methodology, we first replicate the results of Mikolov et al. (2013b) on English syntactic analogies.

### 2.1 Training Corpus for Word Vectors

The vectors of Mikolov et al. (2013b) were trained on 320M tokens of broadcast news data, as described by Mikolov et al. (2011). Since we have no access to this data, we instead train English vectors on a corpus from the Polyglot project (Al-Rfou et al., 2013), which contains tokenized Wikipedia dumps intended for the training of vector-space models. For comparison with the results of Mikolov et al. (2013b), we limit the data to the first 320M lowercased tokens of the corpus.



*Proceedings of NAACL-HLT 2015*, pages 129–134,
Denver, Colorado, May 31 – June 5, 2015. 

Mikolov et al. (2013b) obtain their best results with vectors of size 1600 that combine several models, but do not elaborate how this composite model was constructed. Instead, we take as a point of reference their second-best model, which employs 640-dimensional vectors produced by a single recursive neural network (RNN) language model.[1]

Rather than use an RNN model to learn our own vectors, we employ the far simpler skip-gram model. Mikolov et al. (2013a) show that higher accuracy can be obtained using vectors derived using this model, which is also far less expensive to train. The skip-gram model eschews a language modeling objective in favor of a logistic regression classifier that predicts surrounding words. The WORD2VEC package includes code for learning skip-gram models from very large corpora.[2] 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 frequency threshold of 10.

### 2.2 Test Set

The test set of Mikolov et al. (2013b) is publicly available[3]. They extract their gold standard inflections, as well as frequency counts, from tagged newspaper text. Their test set was constructed as follows: after tagging 267M words, the 100 most frequent plural nouns, possessive nouns, comparative adjectives, and verbal infinitives were selected. Each was paired with 5 randomly-selected words of the same part-of-speech, and analogy questions were constructed for each word pair. For example, for the pair *people* and *city*, two questions are created: *people*:*person* :: *cities*:*city*, and its mirror: *person*:*people* :: *city*:*cities*.

To solve the analogies in this test set, we apply the word-analogy tool that is included with WORD2VEC. For each analogy $a : b :: c :?$, the tool searches the entire vocabulary for the vector $d$ that is most similar to the vector estimated by performing a linear analogy on the query triplet $a, b, c$:

$$d = \text{argmax}_{d'} = \cos(d', c + b - a) \qquad (1)$$

We calculate accuracy as the percentage of analogies whose answers are correctly predicted, according to an exact match.

| Test Set | M13 | Ours |
|---|---|---|
| Adjectives | 21.0 | 18.8 |
| Nouns | 40.1 | 55.2 |
| Verbs | 54.8 | 50.6 |

Table 1: The results of replicating the experiments of Mikolov et al. (2013b) on English.

The analogies involve nouns, adjectives, and verbs. Nominal analogies consist of comparisons between singular and plural forms, and possessive and nominative forms. Due to the tokenization method used in our training corpus, we are unable to build vectors for English possessives. We therefore modify the nominal test set to only include questions that contain the singular vs. plural distinction. We make no changes to the adjectival and verbal analogy sets. The adjectival set contains analogies between the comparative and the superlative, the comparative and the base, and the superlative and the base. The verbal set includes comparisons between the preterite, the infinitive, and the $3rd$ person singular present, but not the past and present participles.

### 2.3 Results

In Table 1, we report two numbers for each part of speech. The first, labeled as M13, is the result of applying the vectors of Mikolov et al. (2013b) to their test set. The results match the results reported in their paper, except for the nominal results, which reflect our modifications described in Section 2.2. The removal of the possessives improves the accuracy from 25.2% reported in the original paper to 40.1%. The second column, labeled as Ours, reports the results for our vectors, which were trained using WORD2VEC on the English data described in Section 2.1.

Our verbal and adjectival vectors obtain slightly lower accuracies than the RNN trained vectors of Mikolov et al. (2013b), but they are not far off. For nouns, however, we obtain higher accuracy than Mikolov et al. The tokenization method that removes possessives from consideration may produce better vectors for singular and plural forms, as it increases the frequency of these types.

[1] The vectors are available at *http://rnnlm.org*.

[2] *https://code.google.com/p/word2vec*.

[3] *http://research.microsoft.com/en-us/projects/rnn*

## 3 Multilingual Experiments

Our second set of experiments examine to what extent the syntactic regularities are captured by word vectors in four other languages: Dutch, French, German, and Spanish.

### 3.1 Training Corpora for Word Vectors

As in the previous experiment, our training corpora are from the Polyglot project. We limit each corpus to the first 320M lowercased tokens, except for the Dutch corpus, which has only 180M tokens. Since the WORD2VEC tool cannot handle Unicode, we map all non-ASCII characters to unused ASCII characters. We run WORD2VEC with exactly the same hyper-parameters as in Section 2.1. The English experiments in this section use the same training data and vectors as in Section 2, but we construct a new test set to match our methodology for the other languages.

### 3.2 Test Sets

In order to make results between multiple languages comparable, we made several changes to the construction of syntactic analogy questions. We follow the methodology of Mikolov et al. (2013b) in limiting analogy questions to the 100 most frequent verbs or nouns. The frequencies are obtained from corpora tagged by TREETAGGER (Schmid, 1994).

We identify inflections using manually constructed inflection tables from several sources. Spanish and German verbal inflections, as well as German nominal inflections, are from a Wiktionary data set introduced by Durrett and DeNero (2013).[4] Dutch verbal inflections and English verbal and nominal inflections are from the CELEX database (Baayen et al., 1995). French verbal inflections are from Verbiste, an online French conjugation dictionary.[5]

Whereas Mikolov et al. create analogies from various inflectional forms, we require the analogies to always include the base dictionary form: the infinitive for verbs, and the nominative singular for nouns. In other words, all analogies are limited to comparisons between the base form and an inflected form. This is to prevent a combinatorial explosion of the number of analogies in languages that contain dozens of different inflection forms. We also create new English test sets using this methodology, in order to ensure a fair cross-lingual comparison. Table 2 shows the number of analogy questions for each language set. Note that the languages are ordered according to increasing morphological complexity.

| Set | I | Q | Example |
|---|---|---|---|
| EN-V | 5 | 3096 | go:gone see:? |
| NL-V | 9 | 5136 | gaan:gegaan zien:? |
| DE-V | 27 | 6514 | gehen:gegangen sehen:? |
| FR-V | 48 | 15573 | aller:allé voir:? |
| ES-V | 57 | 22579 | ir:ido ver:? |
| EN-N | 2 | 876 | bear:bears lion:? |
| DE-N | 8 | 1804 | Bär:Bären Löwe:? |

Table 2: The number of inflectional slots (I) and analogy questions (Q) for each language set.

Following Mikolov et al., we ensure that all analogies contain at least one pair of non-syncretic forms. It would make little sense to include analogies such as "*set* is to *set* as *put* is to *?*" because both verbs in question have the same present and past tense form. However, we do allow analogies which involve syncretic forms for one half of the analogy. For example, either *taken* or *took* is a correct answer to "*play* is to *played* as *take* is to *?*". These types of questions account for an average of 2.8% of analogies, ranging from 0% for English nouns to 8.9% for German verbs.

The number of questions for each language is a function of the number of inflectional forms, but it is not a simple linear relationship. If each English verb had five different inflections, each with sufficient frequency in the training corpus, we would expect 4000 questions for 100 verbs. This is because each verb should ideally be compared to five other verbs, with the base form paired with the other four inflectional forms, in both directions. The actual number of questions is smaller because some forms are identical, while other forms are observed less frequently than our minimum threshold of 10.

[4] We exclude Finnish because of its high morphological complexity and the small size of the corresponding Polyglot corpus.

[5] *http://perso.b2b2c.ca/sarrazip/dev/verbiste.html*

| Set | All Inflections | | Inflection Subset | |
|---|---|---|---|---|
| EN-V | 52.6 | (21.3k) | 52.6 | (21.3k) |
| NL-V | 37.8 | (4.5k) | 33.5 | (7.0k) |
| DE-V | 29.4 | (5.0k) | 40.0 | (8.9k) |
| FR-V | 25.9 | (0.5k) | 45.6 | (8.6k) |
| ES-V | 22.8 | (0.5k) | 48.2 | (10.6k) |
| EN-N | 52.2 | (46.9k) | 52.2 | (46.9k) |
| DE-N | 28.2 | (18.0k) | 31.9 | (35.6k) |

Table 3: Accuracy on analogy questions. The median frequencies of the types involved are provided in brackets.

### 3.3 Results

We conduct two experiments to quantify the extent that the syntactic regularities observed in English hold in the other languages. In the first experiment, which is referred to as *All Inflections*, we measure the accuracy of vectors on all inflected forms. In the second experiment, named *Inflection Subset*, we attempt to factor out the variation in the number of inflectional forms across languages by considering only the forms that are observed in English (five forms for verbs, and two forms for nouns).

The results of the experiments are in Table 3. In the *All Inflections* column, we see that the overall accuracy decreases as the morphological complexity increases. However, the *Inflection Subset* column reveals an opposite trend: the accuracy is increasing towards the bottom of the table, (although English stands out as a clear exception). Looking across the rows, the accuracy on the inflection subset is higher than on all inflections, except on Dutch. Noun analogies are only tested on two languages, but they seem to follow the same trends as verbs.

The results in Table 3 are not easy to interpret. It appears the lower frequencies of multiple inflected word forms make the task more difficult, which is reflected in the *All Inflections* results. The median frequencies of individual verb forms in French and Spanish are approximately one-tenth of the corresponding numbers in Dutch and German, which in turn are about one-fourth of the English median. However, these ratios are not neatly correlated with the accuracy results in Table 3.

Regarding the contrasting results in the *Inflection Subset* column, we conjecture that a larger number of inflections may make *individual forms* easier to disambiguate. This in turn allows WORD2VEC to learn more precise vectors for each word type. The median frequencies of the forms in the inflection subset tend to be higher than the corresponding values computed for all inflections, but there is a substantial variation between different languages. Dutch, in particular, sees a similar increase in median frequency to German, but while German accuracy increases, Dutch decreases. We conclude that although frequency is an important factor when performing syntactic analogies with vectors, there must be other factors contributing to these results.

It is perhaps unsurprising that English is the winner on its own inflection set. However, another reason that English does not follow the trend in the *Inflection Subset* column may be related to the frequencies of its small set of wordforms, which are uniformly higher than in other languages. The experiments that we describe in the next section provide additional insights into these results.

## 4 Hyper-Parameter Experiments

In this section, we describe experiments that quantify how the quality of the vectors is affected by the window size and the amount of training data.

### 4.1 Window size

First, we investigate the role that the window size has on the accuracy of learned vectors. We expect that larger window sizes may create more topic-oriented vectors, while small windows result in vectors that capture syntactic information (Turney, 2012). While all experiments in Section 3 used a window size of 10, the languages have different syntactic and morphological patterns, and some of the results observed in Section 3 may simply be a side effect of better or worse window sizes for particular languages. We run an experiment that tests window sizes of 1, 3, 5 and 10, calculating the analogy accuracy for each language and each window size.

Figure 2 shows the results for varying window sizes. While no single window size is best for all languages, we observe that the morphologically complex languages perform better with larger windows. One benefit that larger window sizes may provide is access to more information during vector training,

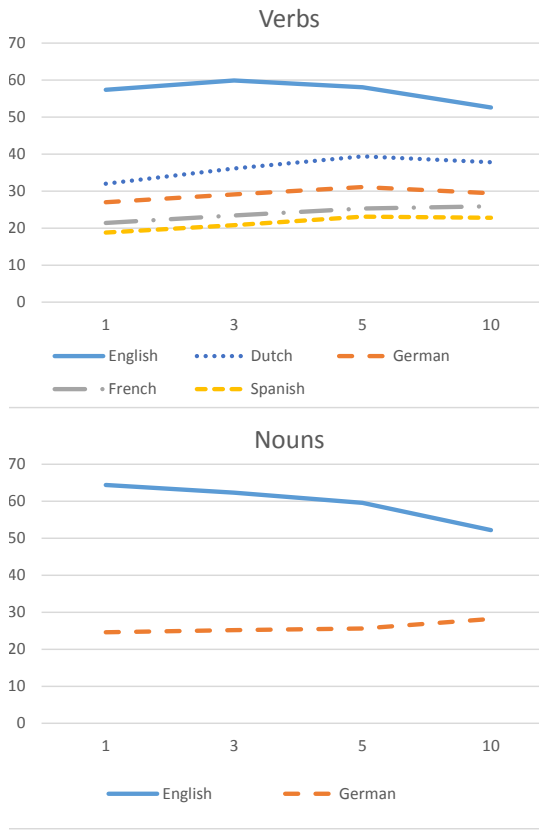


Figure 2: Accuracy for different context windows.

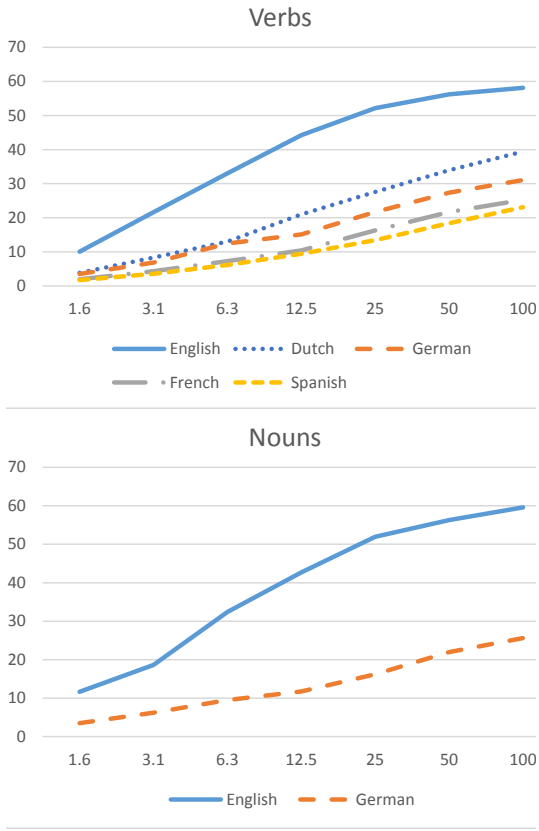


Figure 3: Learning curves.

which may be important when each type is observed less frequently. Our next experiment directly investigates the impact of the training data size.

### 4.2 Learning curves

In this section, we investigate how varying the size of the vector training data affects the vector accuracy. We progressively subsample the training data: starting with the complete training set, we construct a 50% subsample by selecting each sentence for inclusion with probability 0.5. We then iterate this process, each time sampling roughly 50% of the sentences from the previously created subsample, until we have a subsample that is only 1.6% of the original training data. This gives us training sets with approximately 1.6, 3.1, 6.3, 12.5, 25, 50, and 100% of the full corpora. We set the window size to 5 for this experiment; the other hyper-parameters are the same as those in Section 2.1.

The learning curves for verbs and nouns are shown in Figure 3. We see that the trends observed in Section 3 hold regardless of the amount of data that is used for training: namely, the accuracy of the vectors is inversely correlated with the number of inflection slots in a given language set. Secondly, while the English curves are beginning to level off, the curves for the other languages continue to rise, even as we reach 100% of our data. This suggests that there would be little gain in adding more English data, but a potential gain to be seen by adding more data to the other languages. This seems to support our hypothesis that the sparsity of the data is at least partially responsible for the lower accuracies on the morphologically complex languages.

## 5 Conclusion

The results of our experiments show that it is possible to learn vectors that preserve morphological information even for languages with complex inflectional systems. The accuracy of vectors on a set of syntactic analogies in four tested languages is lower than in English, and it appears to be in-

versely proportional to morphological complexity, as measured by the number of inflections in the language. When we limit our test set to the small set of inflections common across languages, we see improvements in the accuracy, which positively correlate with the complexity of the language. This suggests that for frequently observed phenomena, morphological complexity may be an advantage, making each type distinct and easier to model. Additional experiments suggest that the accuracy on more complex languages may further improve if more training data is provided.

These results suggest two possible avenues for future work. The first is to build morphologically-aware vectors, such as those of Botha and Blunsom (2014), so that the more morphologically complex languages can make better use of limited training data. The second is to investigate methods that can distinguish syncretic forms in context. For example, it could be possible to modify the joint word-sense and vector induction algorithm of Neelakantan et al. (2014) to focus on syntactic parts-of-speech instead of topical senses.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and the Alberta Innovates – Technology Futures.

## References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Jan A. Botha and Phil Blunsom. 2014. Compositional Morphology for Word Representations and Language Modelling. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, Beijing, China, jun. *Award for best application paper*.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *HLT-NAACL*, pages 1185–1195.

Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan Cernocky. 2011. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR, 2013*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient nonparametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Peter D Turney. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585.