# Word embeddings and morphology Do language models learn what linguists are describing?

Yulia Zinova

Computational linguistics, Heinrich Heine University of Düsseldorf

June 29, 2023

# Cast your vote!

https://directpoll.com/r?
XDbzPBd3ixYqg8XGqfWvc6TLbjcxPk92i8Z0RyQ

# Verbal Aspect (English)

John was crossing the street when he was hit by a motorcycle. Did he cross the street?

# Verbal Aspect (English)

John was crossing the street when he was hit by a motorcycle. Did he cross the street?

Yes, based on the information provided, it is stated that John was crossing the street when he was hit by a motorcycle. Therefore, it can be inferred that John did indeed cross the street before the accident occurred.

# Some theory: Types of languages

- Morphology is not equally prominent in all languages.
- What one language expresses morphologically may be expressed by different means in another language.
- English: Aspect is expressed by certain syntactic structures:

(1) a. John wrote / has written a letter.
b. John was writing a letter.

- Russian: Aspect is marked mostly by prefixes:

(2) a. Maša napisala pis'mo.
Masha NA.write.PST.SG.F letter.SG.ACC
Masha wrote a letter.
b. Maša pisala pis'mo.
Masha write.PST.SG.F letter.SG.ACC
Masha was writing a letter.

# Some theory: Types of languages

- Morphology is not equally prominent in all languages.
- What one language expresses morphologically may be expressed by different means in another language.
- English: Aspect is expressed by certain syntactic structures:

(1) a. John wrote / has written a letter.
    b. John was writing a letter.

- Russian: Aspect is marked mostly by prefixes:

(2) a. Maša napisala pis'mo.
    Masha NA.write.PST.SG.F letter.SG.ACC
    Masha wrote a letter.
  b. Maša pisala pis'mo.
    Masha write.PST.SG.F letter.SG.ACC
    Masha was writing a letter.

# Some theory: Types of languages

- Morphology is not equally prominent in all languages.
- What one language expresses morphologically may be expressed by different means in another language.
- English: Aspect is expressed by certain syntactic structures:

(1) a. John wrote / has written a letter.
    b. John was writing a letter.

- Russian: Aspect is marked mostly by prefixes:

(2) a. Maša napisala pis'mo.
    Masha NA.write.PST.SG.F letter.SG.ACC
    Masha wrote a letter.
  b. Maša pisala pis'mo.
    Masha write.PST.SG.F letter.SG.ACC
    Masha was writing a letter.

# Types of languages: analytic and synthetic

- Two basic morphological types of language structure: analytic and synthetic
- Analytic languages have only free (occurring on their own) morphemes, sentences are sequences of single-morpheme words.
- Synthetic languages have both free and bound (occuring only with affixes) morphemes.

# Subtypes of synthetic languages

- Agglutinating languages: each morpheme has a single function, it is easy to separate them.
- Fusional languages: like agglutinating, but affixes tend to "fuse together", one affix has more than one function.
- Polysynthetic languages: extremely complex, many roots and affixes combine together, often one word corresponds to a whole sentence in other languages.

# Types of languages: continuum

- The distinction between analytic and (poly)synthetic languages is a continuum, ranging from the most radically isolating to the most highly polysynthetic languages.
- Degree of synthesis (Haspelmath, 2002)

| Language | Morphemes per word |
|---|---|
| Greenlandic Eskimo | 3.72 |
| Sanskrit | 2.59 |
| Swahili | 2.55 |
| Old English | 2.12 |
| Lezgian | 1.93 |
| German | 1.92 |
| Modern English | 1.68 |
| Vietnamese | 1.06 |

# Verbal Morphology (Prefixation, Russian)

- Imperfective aspect:
  - *čitat'* 'to read'
- Perfective aspect:
  - *pročitat'* 'to read completely'
  - *počitat'* 'to read for some time'
  - *dočitat'* 'to finish reading'
  - *perečitat'* 'to read again'
- Much more in Zinova (2021)

# Chat GPT and Russian Verbal Morphology

- **Scenario description:** When Alexandra reads, she always reads for 30 minutes. Alexandra **started reading a book a month ago,** but then she abandoned it. She returned to the book yesterday and finished it today.
- How long did it take Alexandra to finish reading (*dočitat'*) the book?
- How long did it take Alexandra to read (*pročitat'*) the book?

# Chat GPT and Russian Verbal Morphology

- **Scenario description:** When Alexandra reads, she always reads for 30 minutes. Alexandra **started reading a book a month ago,** but then she abandoned it. She returned to the book yesterday and finished it today.
- How long did it take Alexandra to finish reading (*dočitat'*) the book?
- How long did it take Alexandra to read (*pročitat'*) the book?

# Chat GPT and Russian Verbal Morphology

- **Scenario description:** When Alexandra reads, she always reads for 30 minutes. Alexandra **started reading a book a week ago,** but then she abandoned it. She returned to the book yesterday and finished it today.
- How long did it take Alexandra to finish reading (*dočitat'*) the book?
- How long did it take Alexandra to read (*pročitat'*) the book?

# Chat GPT and Russian Verbal Morphology

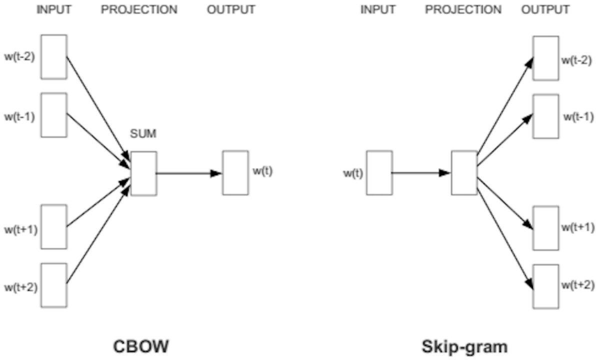
- **Scenario description:** When Alexandra reads, she always reads for 30 minutes. Alexandra **started reading a book a week ago,** but then she abandoned it. She returned to the book yesterday and finished it today.
- How long did it take Alexandra to finish reading (*dočitat'*) the book?
- How long did it take Alexandra to read (*pročitat'*) the book?

# Outline

# Table of Contents

# Word embeddings: word2vec

A way to represent words as dense multidimensional vectors (Mikolov et al., 2013)



# Word embeddings: FastText

How FastText (Bojanowski et al., 2017) differs from word2vec:

- learning character n-gram representations;
- word embeddings are sums of the embeddings of all their n-grams;
- embdeddings for character n-grams allow to represent out-of-vocabulary (oov) words;
- overall, FastText embeddings allow to better capture morphology.

# Table of Contents

# Morphosyntactic analogies

- FastText's n-grams are able to answer morphosyntactic analogy questions
- $a_1 :: a_2 = x :: b_2$, where x has to be guessed from the entire lexicon
- For English singular/plural pairs this predicts x accurately in 91.8% of the cases. For present/past verb forms 76.5%.

# Fasttext-based analogies for various languages

| Category | sl | en | ru |
|---|---|---|---|
| Capitals and countries | 28.13 | 95.23 | 81.26 |
| Family | 38.77 | 92.03 | 58.64 |
| City in country | 45.44 | 89.92 | 95.26 |
| Animals | 1.13 | 11.72 | 14.90 |
| City with river | 5.92 | 44.81 | 11.34 |
| Adjective to adverb | 36.62 | 27.32 | 29.31 |
| Opposite adjective | 30.42 | 50.00 | 0.00 |
| Comparative adjective | 31.38 | 96.88 | 37.55 |
| Superlative adjective | 19.28 | 97.31 | 23.08 |
| Verb to verbal noun | 65.33 | 82.37 | 19.05 |
| Country to nationality | 31.43 | 56.56 | 67.71 |
| Singular to plural | 32.68 | 91.78 | 57.35 |
| Genitive to dative | 26.68 | N/A | 33.19 |
| Present to past | 51.63 | 76.50 | 77.00 |
| Present to other tense | 54.17 | 32.55 | 78.50 |

From Ulčar et al. 2020, Multilingual Culture-Independent Word Analogy Datasets

# Analogy test results: Nouns, inflection

| Form | MultiLing | Random |
|---|---|---|
| Sg to pl | 57.35% | 43% |
| Pl to sg | – | 37% |
| Gen to dat | 33.19% | 43% |
| Dat to gen | – | 49,3% |
| Nom to gen | – | 40,3% |
| Gen to nom | – | 43% |

# Table of Contents

# Pipeline Outline

Pipeline described in Wiemerslage et al. (2022). Steps:

- Cluster word forms into paradigms on the basis of their orthographic similarity;
- Assess which orthographic changes of the word forms express the same inflectional information;
- Use information about word embeddings to assess the distribution of such inflections;
- Assign labels to word forms;
- Train a morphological learner with the assigned labels.

# Pipeline Outline

Pipeline described in Wiemerslage et al. (2022). Steps:

- Cluster word forms into paradigms on the basis of their orthographic similarity;
- Assess which orthographic changes of the word forms express the same inflectional information;
- Use information about word embeddings to assess the distribution of such inflections;
- Assign labels to word forms;
- Train a morphological learner with the assigned labels.

# Pipeline Outline

Pipeline described in Wiemerslage et al. (2022). Steps:

- Cluster word forms into paradigms on the basis of their orthographic similarity;
- Assess which orthographic changes of the word forms express the same inflectional information;
- Use information about word embeddings to assess the distribution of such inflections;
- Assign labels to word forms;
- Train a morphological learner with the assigned labels.

# Pipeline Outline

Pipeline described in Wiemerslage et al. (2022). Steps:

- Cluster word forms into paradigms on the basis of their orthographic similarity;
- Assess which orthographic changes of the word forms express the same inflectional information;
- Use information about word embeddings to assess the distribution of such inflections;
- Assign labels to word forms;
- Train a morphological learner with the assigned labels.

# Pipeline Outline

Pipeline described in Wiemerslage et al. (2022). Steps:

- Cluster word forms into paradigms on the basis of their orthographic similarity;
- Assess which orthographic changes of the word forms express the same inflectional information;
- Use information about word embeddings to assess the distribution of such inflections;
- Assign labels to word forms;
- Train a morphological learner with the assigned labels.

# Unsupervised Learning of Morphology: Evaluation

- Model trained on digitized children's books and the Bible.
- Languages of training: German, Greek, Icelandic, and Russian.
- Evaluation: correct paradigm reconstructions with paradigm slots aligned between different lemmas but in random order; the best possible correspondence to true labels is selected for the evaluation.
- Best result: 27% correctly generated word forms (Russian digitized children's books).
- Worst result: $< 10\%$ (for the Bible translation of Greek).

# Unsupervised Learning of Morphology: Evaluation

- Model trained on digitized children's books and the Bible.
- Languages of training: German, Greek, Icelandic, and Russian.
- Evaluation: correct paradigm reconstructions with paradigm slots aligned between different lemmas but in random order; the best possible correspondence to true labels is selected for the evaluation.
- Best result: 27% correctly generated word forms (Russian digitized children's books).
- Worst result: $< 10\%$ (for the Bible translation of Greek).

# Unsupervised Learning of Morphology: Evaluation

- Model trained on digitized children's books and the Bible.
- Languages of training: German, Greek, Icelandic, and Russian.
- Evaluation: correct paradigm reconstructions with paradigm slots aligned between different lemmas but in random order; the best possible correspondence to true labels is selected for the evaluation.
- Best result: 27% correctly generated word forms (Russian digitized children's books).
- Worst result: $< 10\%$ (for the Bible translation of Greek).

# Unsupervised Learning of Morphology: Evaluation

- Model trained on digitized children's books and the Bible.
- Languages of training: German, Greek, Icelandic, and Russian.
- Evaluation: correct paradigm reconstructions with paradigm slots aligned between different lemmas but in random order; the best possible correspondence to true labels is selected for the evaluation.
- Best result: 27% correctly generated word forms (Russian digitized children's books).
- Worst result: $< 10\%$ (for the Bible translation of Greek).

# Unsupervised Learning of Morphology: Evaluation

- Model trained on digitized children's books and the Bible.
- Languages of training: German, Greek, Icelandic, and Russian.
- Evaluation: correct paradigm reconstructions with paradigm slots aligned between different lemmas but in random order; the best possible correspondence to true labels is selected for the evaluation.
- Best result: 27% correctly generated word forms (Russian digitized children's books).
- Worst result: $< 10\%$ (for the Bible translation of Greek).

# Table of Contents

# Russian Paradigms

| Case | Num | ‘table’ | ‘mother’ | ‘elephant’ |
|---|---|---|---|---|
| Nom | Sg | stol | mama | slon |
| Gen | Sg | stola | mamy | slona |
| Dat | Sg | stolu | mame | slonu |
| Acc | Sg | stol | mamu | slona |
| Ablt | Sg | stolom | mamoj | slonom |
| Loc | Sg | stole | mame | slone |
| Nom | Pl | stoly | mamy | slony |
| Gen | Pl | stolov | mam | slonov |
| Dat | Pl | stolam | mamam | slonam |
| Acc | Pl | stoly | mam | slonov |
| Ablt | Pl | stolami | mamami | slonami |
| Loc | Pl | stolax | mamax | slonax |

# Russian nominal inflection: Same suffixes for different paradigm sells

| Case | Num | 'table' | 'mother' | 'elephant' |
|---|---|---|---|---|
| Nom | Sg | stol | mama | slon |
| Gen | Sg | stola | mam**y** | slona |
| Dat | Sg | stolu | mame | slonu |
| Acc | Sg | stol | mamu | slona |
| Ablt | Sg | stolom | mamoj | slonom |
| Loc | Sg | stole | mame | slone |
| Nom | Pl | stol**y** | mam**y** | slon**y** |
| Gen | Pl | stolov | mam | slonov |
| Dat | Pl | stolam | mamam | slonam |
| Acc | Pl | stol**y** | mam | slonov |
| Ablt | Pl | stolami | mamami | slonami |
| Loc | Pl | stolax | mamax | slonax |

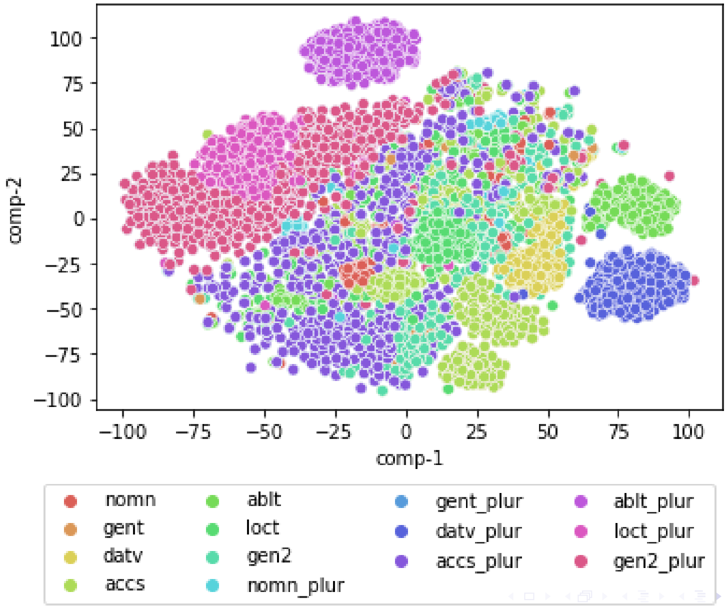
# Russian nominal inflection: Syncretism

| Case | Num | 'table' | 'mother' | 'elephant' |
|---|---|---|---|---|
| Nom | Sg | **stol** | mama | slon |
| Gen | Sg | stola | mamy | **slona** |
| Dat | Sg | stolu | **mame** | slonu |
| Acc | Sg | **stol** | mamu | **slona** |
| Ablt | Sg | stolom | mamoj | slonom |
| Loc | Sg | stole | **mame** | slone |
| Nom | Pl | stoly | mamy | slony |
| Gen | Pl | stolov | mam | slonov |
| Dat | Pl | stolam | mamam | slonam |
| Acc | Pl | stoly | mam | slonov |
| Ablt | Pl | stolami | mamami | slonami |
| Loc | Pl | stolax | mamax | slonax |

# Table of Contents

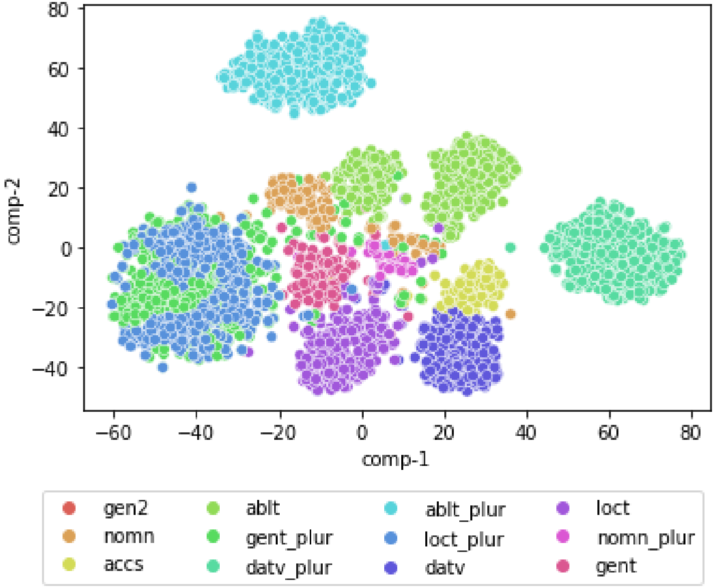
# Visualising word embeddings

All nouns, original word vectors



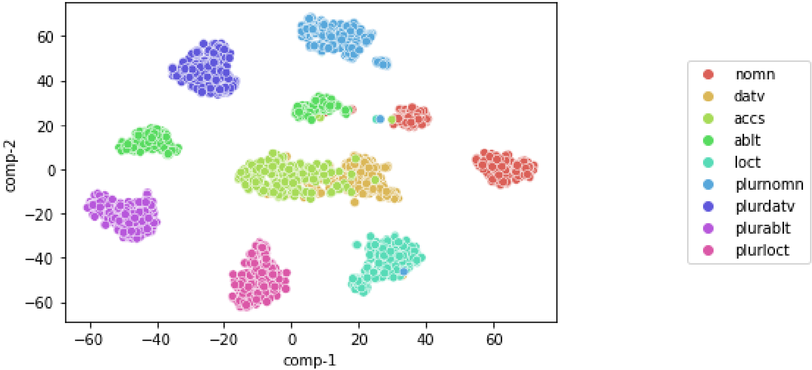
# Visualising word embeddings

No syncretism, original word vectors



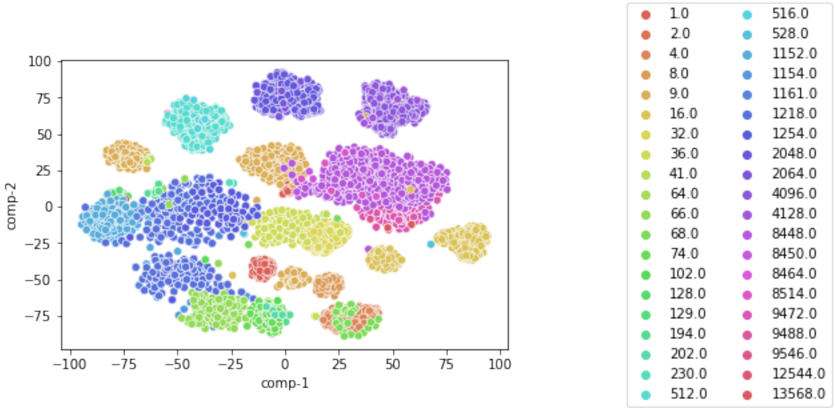
# Visualising word embeddings

No syncretism, difference vectors, average as base form



# Visualising word embeddings

Syncretism, difference vectors, average as base form



# Table of Contents

# Summary and Outlook

- Even lots of data does not solve the problems.
- For many languages, there is not so much data and a lot of morphology.
- Linguistic insights can be used to improve machine learning of morphology.
- Insights from analysing embeddings can be used to evaluate morphological theory ('Evaluation of Russian Noun Word Embeddings For Cases and a Number' tomorrow at 11:30).
- Rapidly developing area with lots of potential.

# Thank You!

Questions?

# References I


Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Martin Haspelmath. 2002. *Understanding Morphology*. Arnold Publishers.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D. McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next.

# References II

Yulia Zinova. 2021. *Russian verbal prefixation*. Number 7 in Empirically Oriented Theoretical Morphology and Syntax. Language Science Press, Berlin.