One Size Does Not Fit All: Finding the Optimal Subword Sizes for FastText Models across Languages*

Vít Novotný
Faculty of Informatics
Masaryk University
witiko@mail.muni.cz

Eniafe Festus Ayetiran
Faculty of Informatics
Masaryk University
ayetiran@mail.muni.cz

Dalibor Bačovský
Faculty of Informatics
Masaryk University
456662@mail.muni.cz

Dávid Lupták
Faculty of Informatics
Masaryk University
dluptak@mail.muni.cz

Michal Štefánik
Faculty of Informatics
Masaryk University
stefanik.m@mail.muni.cz

Petr Sojka
Faculty of Informatics
Masaryk University
sojka@fi.muni.cz

Abstract

Unsupervised representation learning of words from large multilingual corpora is useful for downstream tasks such as word sense disambiguation, semantic text similarity, and information retrieval. The representation precision of log-bilinear fastText models is mostly due to their use of subword information.

In previous work, the optimization of fastText's subword sizes has not been fully explored, and non-English fastText models were trained using subword sizes optimized for English and German word analogy tasks.

In our work, we find the optimal subword sizes on the English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian word analogy tasks. We then propose a simple n-gram coverage model and we show that it predicts better-than-default subword sizes on the Spanish, French, Hindi, Turkish, and Russian word analogy tasks.

We show that the optimization of fastText's subword sizes matters and results in a 14% improvement on the Czech word analogy task. We also show that expensive parameter optimization can be replaced by a simple n-gram coverage model that consistently improves the accuracy of fastText models on the word analogy tasks by up to 3% compared to the default subword sizes, and that it is within 1% accuracy of the optimal subword sizes.

1 Introduction

Bojanowski et al. (2017) have shown that taking word morphology into account is important for accurate continuous representations of words. However, they only show the optimal n-gram sizes on the German and English word analogy tasks (Bojanowski et al., 2017, Section 5.5). We continue their experiment by finding the optimal parameters on the Czech, Italian, Spanish, French, Hindi, Turkish, and Russian word analogy tasks and we show an up to 14% improvement in accuracy compared to the default subword sizes.

Furthermore, we propose a cheap and simple n-gram coverage model that can suggest near-optimal subword sizes for under-resourced languages, where the optimal subword sizes are unknown. We train our n-gram coverage model on the English, German, Czech, and Italian word analogy tasks, and we show that it suggests subword sizes that improve the accuracy by up to 3% on the Spanish, French, Hindi, Turkish, and Russian word analogy tasks and are within 1% accuracy of the optimal subword sizes on average. To make it easy for others to reproduce and build upon our work, we have publicly released a reference implementation of our n-gram coverage model.[1]

The rest of the paper is structured as follows: In Section 2, we discuss the related work. In Section 3, we discuss our methods and we propose our n-gram coverage model. In Section 4, we show and discuss our results. We conclude in Section 5 by summarizing our contribution. We outline the future work in Section 6.

2 Related work

Mikolov et al. (2013) described the *Word2vec language model*, which uses a shallow neural network to learn continuous representations of words: *word embeddings*. They also produced the *English word analogy task*, which tests how well word embeddings represent language regularities such as analogical relations (man is to woman what a king is to a queen), and evaluated Word2vec on their task.

*First author's work was graciously funded by the South Moravian Centre for International Mobility as a part of the Brno Ph.D. talent project. Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

[1] See https://github.com/MIR-MU/fasttext-optimizer.

Proceedings of Recent Advances in Natural Language Processing, pages 1068–1074
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_120

$n = 1$:	H, e, l, o, w, r, d
$n = 2$:	He, el, ll, lo, wo, or, rl, ld
$n = 3$:	Hel, ell, llo, wor, orl, rld
$n = 4$:	Hell, ello, worl, orld
$n = 5$:	Hello, world

(a) 27 unique subwords in a corpus of two words: *Hello* and *world*

$n = 1$:	$^7/_{27} = 25.93\%$
$n = 2$:	$^8/_{27} = 29.63\%$
$n = 3$:	$^6/_{27} = 22.22\%$
$n = 4$:	$^4/_{27} = 14.81\%$
$n = 5$:	$^2/_{27} = 7.41\%$

(b) Frequencies of unique subwords of size n

	1	2	3	4	5
1	25.93	55.55	77.77	92.59	100.00
2		29.63	51.85	66.66	74.07
3			22.22	37.03	44.44
4				14.81	22.22
5					7.41

(c) N-gram coverages for the subword sizes i–j, where $1 \leq i \leq j \leq 5$

Table 1: An example of n-gram coverages. We start in Subtable (a) by producing all unique subwords of size less than 10 from the corpus. In Subtable (b), we compute the frequencies of unique subwords of different sizes. In Subtable (c), we compute the n-gram coverages for various subword sizes i–j.

Berardi et al. (2015); Köper et al. (2015); Svoboda and Brychcín (2016); Cardellino (2019); Güngör and Yıldız (2017); Korogodina et al. (2020) produced the *Italian, German, Czech, Spanish, Turkish, and Russian word analogy tasks* for evaluating the performance of non-English word embeddings. Their findings revealed that, despite the morphological complexity of the languages, Word2vec language models can generate semantically and syntactically meaningful word embeddings.

In order to take word morphology into account, Bojanowski et al. (2017) developed the *fastText language model* based on Word2vec. Their improvements consisted of representing each word as a sequence of *subwords* with their own embeddings. They evaluated their models on the English, German, Czech, and Italian word analogy tasks. They also showed the optimal subword sizes of fastText on the English and German word analogy tasks. However, they did not optimize the subword sizes of fastText on the Czech and Italian word analogy tasks and used subwords of size 3–6[2], which they described as "an arbitrary choice" (Bojanowski et al., 2017, Section 5.5).

Grave et al. (2018) produced the *French and Hindi word analogy tasks*. Furthermore, they also trained and publicly released fastText language models for 157 languages. Like Bojanowski et al., they also neglected to optimize the subword sizes. Unlike Bojanowski et al., they used subwords of size 5–5 for all languages, noting that "using character n-grams of size 5, instead of using the default range of 3–6, does not significantly decrease the accuracy (except for Czech)." (Grave et al., 2018, Section 4.3)

[2]For subword sizes, we adopt the notation of Bojanowski et al. (2017) and (Grave et al., 2018). For example, subwords of size 3–6 are all subwords whose size is 3, 4, 5, or 6.

3 Methods

In this section, we describe our methods and propose our n-gram coverage model, which can be used to suggest subword sizes for a fastText model without expensive parameter optimization.

3.1 Optimal subword sizes

In the first part of our experiment, we train fastText language models on the English (22 GiB), German (8.3 GiB), Czech (1.2 GiB), Italian (4.2 GiB), Spanish (5.2 GiB), French (7.4 GiB), Hindi (0.57 GiB), Turkish (0.72 GiB), and Russian (9.9 GiB) Wikipedia corpora. We use subword sizes i–j for all i, j, where $1 \leq i \leq j \leq 10$, and we report the accuracies and the optimal subword sizes i–j on the English (Mikolov et al., 2013), German (Köper et al., 2015), Czech (Svoboda and Brychcín, 2016), Italian (Berardi et al., 2015), Spanish (Cardellino, 2019), French and Hindi (Grave et al., 2018), Turkish (Güngör and Yıldız, 2017), and Russian[3] word analogy tasks.

3.2 *N*-gram coverage

In the second part of our experiment, we compute and report the ratio between the frequencies of unique subwords of size i–j and the frequencies of all unique subwords of size less than 10 on the English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian Wikipedia corpora. In the following text, we call this ratio the *n-gram coverage*. Table 1 shows how the n-gram coverage is computed by example.

3.3 Suggested subword sizes

In the third part of our experiment, we show that the n-gram coverage can be used to suggest subword sizes that are close to the optimal subword sizes on word analogy tasks.

[3]See https://rusvectores.org/static/testsets/ru_analogy.txt.

For training, we compute *the mean n-gram coverage for the optimal subword sizes* on the English, German, Czech, and Italian word analogy tasks. For testing, we suggest subword sizes for the Spanish, French, Hindi, Turkish, and Russian word analogy tasks, so that the n-gram coverages for the suggested subword sizes on the testing word analogy tasks are *the closest to the mean n-gram coverage for the optimal subword sizes* on the training word analogy tasks. Notice that the suggested subword sizes are not based on the optimal subword sizes for the testing word analogy tasks.

After the performance estimation, we fold the training and testing word analogy tasks and we compute the mean n-gram coverage for the optimal subword sizes on all word analogy tasks (English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian). This means n-gram coverage can be used in applications of fastText to suggest the optimal subword sizes without expensive parameter optimization.

3.4 Language distances

In the final part of our experiment, we interpret suggested subword sizes as *two-dimensional vectors* and use *the Euclidean distance* to measure distances between languages. To see if our language distance measure represents interpretable linguistic phenomena, we compare it to the typological, geographical, and phylogenetic language distance measures of Littell et al. (2017):

Typological Littell et al. define three typological language distance measures: *syntactic*, *phonological*, and *inventory*. Each distance measure is defined as the cosine distances between different feature vectors:

- The *syntactic* features describe the sentence structure of a language and have been adapted from the World Atlas of Language Structures (WALS), Syntactic Structures of World Languages, and Ethnologue databases.
- The *phonological* features describe the structure of the sound and sign systems of a language and have been adapted from the WALS and Ethnologue databases.
- The *inventory* features describe the presence or absence of distinctive speech sounds in the sound system of a language and have been adapted from the PHOIBLE database.

Geographical The *geographical* language distance measure is defined as the cosine distance between feature vectors, where the features have been adapted from declarations of language location in the Glottolog, WALS, and SSWL databases.

Phylogenetic The *phylogenetic* language distance measure is defined as the cosine distance between feature vectors, where the features correspond to the shared membership in language families, according to the world language family tree in the Glottolog database.

To compare our language distance measure with the language distance measures of Littell et al., we compute and report the Pearson's correlation coefficient (r) between the distance measures.

3.5 Implementation details

We reproduce the experimental setup of Bojanowski et al. (2017, Section 4): skip-gram architecture, hash table bucket size $2 \cdot 10^6$, 300 vector dimensions, negative sampling loss with 5 negative samples, initial learning rate 0.05 with a linear decay to zero, sampling threshold 10^{-4}, window size 5, and 5 epochs.

Like Bojanowski et al. (2017), we use a reduced vocabulary of the $2 \cdot 10^5$ most frequent words to solve word analogies. We use the implementation of word analogies in Gensim (Řehůřek and Sojka, 2010), which uses Unicode upper-casing in the `en_US.UTF-8` locale for caseless matching.

To compute Pearson's r between two language distance measures, we use the Representational Similarity Analysis (RSA) framework of Kriegeskorte et al. (2008); Chrupała and Alishahi (2019): we produce two matrices of all pairwise distances between 282 Wikipedia languages[4] and we compute Pearson's r between the upper-triangulars, excluding the diagonals.

To make it easy for others to reproduce and build upon our work, we have published a reference implementation of our n-gram coverage model, which suggests subword sizes for fastText models.[5] The reference implementation contains pre-computed subword frequencies for 288 Wikipedia languages, which makes the suggestions instantaneous.

[4] For six out of the 288 Wikipedia languages, Littell et al. did not provide feature vectors: Bhojpuri (`bh`), Emilian-Romagnol (`eml`), Western Armenian (`hyw`), Nahuatl (`nah`), Simple English (`simple`), and Sakizaya (`szy`).

[5] See https://github.com/MIR-MU/fasttext-optimizer.

	2	3	4	**5**	6
2	73	73	73	74	74
3		74	75	76	75
4			76	**76**	76
5				76	76
6					75

(a) English

	2	3	4	5	**6**
2	51	52	54	56	57
3		55	56	57	58
4			57	58	59
5				59	60
6					**61**

(b) German

	2	3	**4**	5	6
1	44	58	**60**	60	58
2	41	57	59	58	57
3		53	56	54	55
4			49	52	49
5				46	46

(c) Czech

	2	3	4	**5**	6
1	44	50	53	54	53
2	46	51	53	**54**	52
3		51	53	53	53
4			53	53	52
5				53	52

(d) Italian

	2	3	4	**5**	6
2	51	53	55	55	55
3		55	57	57	56
4			57	57	57
5				**57**	57
6					57

(e) Spanish

	2	3	4	5	**6**
2	63	63	65	67	67
3		66	66	67	68
4			68	68	69
5				69	69
6					**70**

(f) French

	2	3	4	5	6
1	15	15	14	14	14
2	**17**	15	14	13	14
3		16	13	13	12
4			13	12	12
5				12	11

(g) Hindi

	2	**3**	4	5	6
1	32	37	38	38	37
2	34	39	39	39	38
3		**40**	39	39	38
4			37	38	37
5				36	35

(h) Turkish

	2	3	4	5	**6**
2	46	43	46	50	51
3		46	47	50	52
4			51	51	52
5				52	**53**
6					53

(i) Russian

Table 2: Accuracies on English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian word analogy tasks. Optimal subword sizes for the different word analogy tasks are **bold**: 4–5 for English, 6–6 for German, 1–4 for Czech, 2–5 for Italian, 5–5 for Spanish, 6–6 for French, 2–2 for Hindi, 3–3 for Turkish, and 5–6 for Russian. Our training and testing word analogy tasks are shown on separate lines.

	2	3	4	**5**	6
2	0.26	0.75	1.72	4.51	10.50
3		0.49	1.45	4.25	10.24
4			0.97	**3.76**	9.75
5				2.79	8.78
6					5.99

(a) English

	2	3	4	5	**6**
2	0.08	0.25	0.85	2.68	6.87
3		0.17	0.77	2.60	6.79
4			0.60	2.43	6.62
5				1.83	6.02
6					**4.19**

(b) German

	2	3	**4**	5	6
1	0.21	0.82	**3.28**	10.40	23.23
2	0.18	0.80	3.25	10.37	23.20
3		0.61	3.07	10.19	23.02
4			2.46	9.58	22.40
5				7.12	19.95

(c) Czech

	2	3	4	**5**	6
2	0.14	0.44	1.34	**3.81**	8.92
3		0.30	1.20	3.67	8.78
4			0.90	3.37	8.48
5				2.47	7.58
6					5.11

(d) Italian

	2	3	4	**5**	6
2	0.25	0.78	2.35	6.67	15.44
3		0.53	2.10	6.42	15.19
4			1.57	5.89	14.66
5				**4.32**	13.09
6					8.77

(e) Spanish

	2	3	4	5	**6**
2	0.28	0.85	2.51	7.00	16.21
3		0.57	2.23	6.73	15.93
4			1.66	6.16	15.36
5				4.50	13.70
6					**9.20**

(f) French

	2	3	4	5	6
1	0.70	2.79	8.15	18.01	30.59
2	**0.57**	2.66	8.03	17.89	30.46
3		2.09	7.46	17.32	29.89
4			5.36	15.22	27.80
5				9.86	22.44

(g) Hindi

	2	**3**	4	5	6
1	0.31	1.08	3.72	10.66	22.49
2	0.27	1.04	3.68	10.63	22.45
3		**0.77**	3.41	10.35	22.18
4			2.64	9.59	21.41
5				6.95	18.77

(h) Turkish

	2	3	4	5	**6**
2	0.17	0.61	2.25	7.13	16.57
3		0.43	2.08	6.96	16.40
4			1.64	6.52	15.96
5				4.88	**14.32**
6					9.44

(i) Russian

Table 3: The n-gram coverages for English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian. The n-gram coverages for the optimal subword sizes on the different word analogy tasks are **bold**: 3.76% for English, 4.19% for German, 3.28% for Czech, 3.81% for Italian, 4.32% for Spanish, 9.20% for French, 0.57% for Hindi, 0.77% for Turkish, and 14.32% for Russian.

Language	Default subword sizes		Suggested subword sizes	Optimal subword sizes
	3–6	5–5		
Spanish	*57.00*	**57.60**	**57.60 (5–5)**	**57.60 (5–5)**
French	68.38	*69.33*	*69.33 (5–5)*	**69.60 (6–6)**
Hindi	12.87	12.10	*15.03 (1–3)*	**16.95 (2–2)**
Turkish	38.04	36.10	*38.34 (1–4)*	**39.51 (3–3)**
Russian	51.89	*52.51*	*52.51 (5–5)*	**52.75 (5–6)**

Table 4: Accuracies on the Spanish, French, Hindi, Turkish, and Russian word analogy tasks using the default subword sizes of Bojanowski et al. (3–6) and Grave et al. (5–5), the subword sizes suggested by n-gram coverage, and the optimal subword sizes. Best accuracies for each language are **bold**, second best are in *italics*.

4 Results

In this section, we show and discuss the optimal subword sizes, accuracies, and n-gram coverages on the English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian word analogy tasks. We also show that the n-gram coverage can be used to suggest subword sizes that are close to the optimal subword sizes.

4.1 Optimal subword sizes

In Table 2 on the previous page, we show the accuracies and the optimal subword sizes on the English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian word analogy tasks. The optimal subword sizes 4–5 for English and 6–6 for German reproduce and confirm the results of Bojanowski et al. (2017, Section 5.2).

The optimal subword sizes for English (4–5), Italian (2–5), Spanish (5–5), French (6–6), and Russian (5–6) word analogy tasks are equal or within 1% accuracy of the default subword sizes suggested by Bojanowski et al. (3–6) and Grave et al. (5–5). In contrast, we see an improvement of up to 14% for Czech (1–4), 5% for Hindi (2–2), 4% for Turkish (3–3), and 3% for German (6–6).

To understand these differences, we look to the linguistic typology of languages: Czech, Hindi, and Turkish are synthetic languages and benefit from short subwords that represent morphemes. German and Russian are also synthetic, but the long compound nouns in German and the use of separate characters for yers (ъ and ь) in Russian make both languages benefit from longer subwords.

4.2 *N*-gram coverage

In Table 3 on the preceding page, we show the n-gram coverages for English, German, Czech, Italian, Spanish, French, Hindi, Turkish, and Russian.

The mean n-gram coverage for the optimal subword sizes on the training word analogy tasks (English, German, Czech, and Italian), which we use to suggest subword sizes for the testing word analogy tasks (Spanish, French, Hindi, Turkish, and Russian), is 3.76%. The mean n-gram coverage for the optimal subword sizes on all word analogy tasks, which can be used in applications of fastText to suggest the optimal subword sizes, is 4.91%.

4.3 Suggested subword sizes

In Table 4, we compare the accuracies on the testing word analogy tasks (Spanish, French, Hindi, Turkish, and Russian) using the default subword sizes of Bojanowski et al. (3–6) and Grave et al. (5–5), the subword sizes suggested by the n-gram coverage, and the optimal subword sizes.

Using the suggested subword sizes is never worse than using the default subword sizes. For Hindi and Turkish, the suggested subword sizes always improve the accuracy: by 2.58% on average compared to the weaker default subword sizes and by 1.23% on average compared to the stronger default subword sizes. For Spanish, French, and Russian, the suggested subword sizes equal the default subword sizes of Grave et al. and they improve the accuracy by 0.72% on average compared to the default subword sizes of Bojanowski et al.

For Spanish, the optimal subword sizes equal the suggested subword sizes. For French, Hindi, Turkish, and Russian, the optimal subword sizes improve the accuracy by only 0.90% on average compared to the suggested subword sizes, whereas they improve the accuracy by 2.59% on average compared to the weaker default subword sizes and by 1.52% on average compared to the stronger default subword sizes.

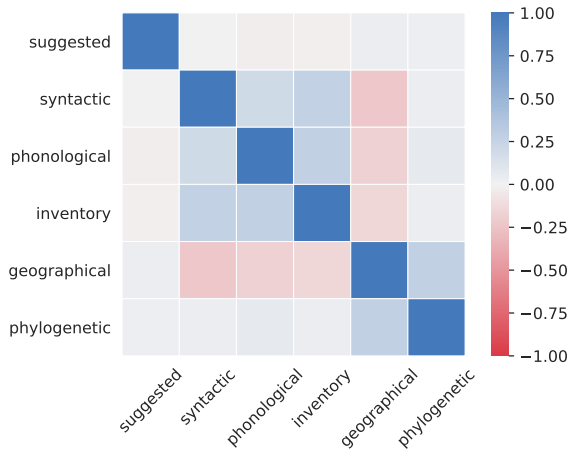

Figure 1: Pearson's correlation coefficients (r) between our language distance measure (*suggested*) as well as the typological (*syntactic*, *phonological*, and *inventory*), geographical, and phylogenetic language distance measures of Littell et al. (2017). Best viewed in color.

4.4 Language distances

In Figure 1, we show Pearson's r between pairs of different language distance measures: our language distance measure, which is based on the Euclidean distance between our suggested subword sizes, as well as the typological, geographical, and phylogenetic language distance measures of Littell et al.

Pearson's r between our language distance measure and the language distance measures of Littell et al. range between -0.03 (*phonological*) and 0.03 (*geographical*). Since the absolute values of Pearson's r are consistently smaller than random, our language distance measure does not either correlate or anti-correlate with the other language distance measures. This is because our suggested subword sizes are based on latent data-driven features of text, which complement the hand-crafted linguistic features of Littell et al.

5 Conclusion

Subword sizes have a profound impact on the accuracy of fastText language models and their word embeddings. However, they are expensive to optimize on large corpora.

In this work, we showed the optimal subword sizes for Czech, Italian, Spanish, French, Hindi, Turkish, and Russian fastText language models, we confirmed prior optimal subword sizes reported for English and German, and we showed that the optimization of subword sizes improves the accuracy of fastText on word analogy tasks by up to 14% compared to the default subword sizes. Our optimal subword sizes can be used in applications of fastText as the new default.

Furthermore, we proposed a cheap and simple n-gram coverage model that consistently improves the accuracy of fastText models on the word analogy tasks by up to 3% compared to the default subword sizes, and that it is within 1% accuracy of the optimal subword sizes on average. Subword sizes suggested by our n-gram coverage model can be used in applications of fastText as the new default for under-resourced languages, where the optimal subword sizes are unknown.

6 Future work

Although the word analogy intrinsic task is a convenient proxy for the usefulness of fastText word embeddings, Ghannay et al. (2016); Chiu et al. (2016); Rogers et al. (2018) show that it is no substitute for actual extrinsic end tasks. In future work, we will evaluate our n-gram coverage model on extrinsic tasks.

In recent machine translation models (Vaswani et al., 2017), text is tokenized into words and subwords using word-piece (Wu et al., 2016) and byte-pair (Sennrich et al., 2016) models. Our experiments suggest that we can remove the subword size parameter from fastText models and draw subwords from byte-pair models with little adverse effect on the word analogy accuracy. In future work, we will evaluate the use of word-piece and byte-pair models for subword selection in fastText models both on the intrinsic word analogy task and on other extrinsic tasks.

References

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word Embeddings Go to Italy: A Comparison of Models and Training Datasets. In *CEUR Workshop Proceedings of 6th Italian Information Retrieval Workshop, IIR 2015*, volume 1404, page 8.

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

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings. Visited on [2021-08-18].

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, Berlin, Germany. ACL.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating Neural and Symbolic Representations of Language. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 2952–2962, Florence, Italy. ACL.

Sahar Ghannay, Benoit Favre, Yannick Estève, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 300–305, Portorož, Slovenia. European Language Resources Association (ELRA).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Onur Güngör and Eray Yıldız. 2017. Linguistic features in Turkish word representations. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Olga Korogodina, Olesya Karpik, and Edward Klyshinsky. 2020. Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings. *Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020)*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(4):1–28.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th international conference on Computational Semantics*, pages 40–45. ACL.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. ACL.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA. ACL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Lukáš Svoboda and Tomáš Brychcín. 2016. New Word Analogy Corpus for Exploring Embeddings of Czech Words. In *19th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing, 2016, Konya, Turkey, April 3–9, 2016*, pages 103–114. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144v2*.