The impact of inflection on word vectors

Dániel Lévai, András Kornai

Institute for Computer Science and Control, Hungarian Academy of Sciences
Research Group for Human Language Technologies
{levai, kornai}@ilab.sztaki.hu

Abstract. We present a method to evaluate the similarity of word vector clusters, and use it to determine the coherence (self-similarity) and relatedness of morphologically defined clusters
Keywords: word vector, cluster similarity, morphology, skip-gram

1 Introduction

Word vectors encode not just semantic relations [1], but also morphological ones, as in $\overrightarrow{\text{goes}} - \overrightarrow{\text{go}} + \overrightarrow{\text{see}} = \overrightarrow{\text{sees}}$. In agglutinative languages it is common to treat inflectionally related tokens as separate types (form-based, rather than stem-based modeling). Our main aim is to show that the tokens considered unrelated by the form-based model are indeed related on a morphological level. Furthermore, the more specific case endings (delative, translative, ...) dominate the word vector as opposed to the less specific case endings (nominative, accusative, ...) where the word vectors contain richer semantic relations.

2 Methods

The idea of neural networks dates back to the 1940s [2], when McCulloch created a computational network. The main idea is that our brain is composed of neurons (nodes) and synapses (edges). We learn and memorize by creating and strengthening synapses, and an artificial neural network – by analogy – should learn by strengthening and weakening weights on the edges based on the sample it receives. Constructing and training a neural network is a difficult task, because we do not have a strong idea how to interpret the weights of the edges or the nodes themselves – a neural network is a black box, and we do not always know how the architecture of the network should look like, or how we should train a network. The architecture in a skip-gram model [3] [4] consists only of a single hidden layer and an output layer with hierarchical softmax classifier. The task of the model is, for every word in the vocabulary, to learn the probabilities of every other word being in the context of the vocabulary word.

The input is a one-hot vector representing the word, and the output is a probability vector. As we can see in Fig. 1, there are separate weights for each coordinate, and the number of nodes in the hidden layer defines the number of

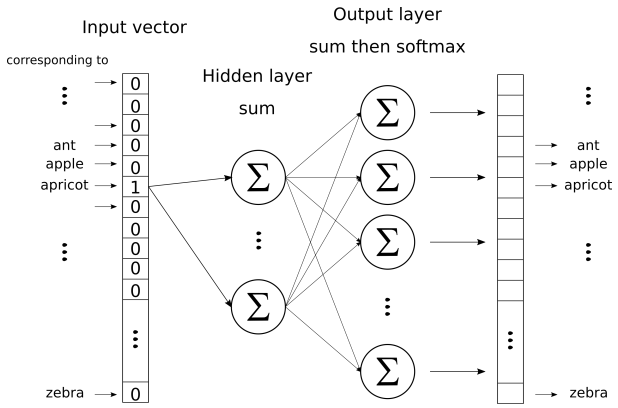

Fig. 1: Network architecture

weights. If the hidden layer counts 200 neurons, then we will have 200 weights for each coordinate, thus for every word. To summarize, we create a model to predict contexts only to learn the input weights to be used as vectors. The same way, the model also learns the output weights, and the classification problem reduces to a matrix dot product and to a softmax classification problem. For adjusting the weights, the model uses backpropagation. The main differences from the traditional neural networks are the subsampling, negative sampling and the use of skip-grams with negative sampling [5].

We use the most noise-reduced tier of the Hungarian Webcorpus[1] [6,7] that has the duplicates, foreign language pages, and script-generated text (such as dates, headlines, tables of content) removed, leaving 710m word and punctuation tokens. For morphological analysis, we used the `emMorph module`[2] [8] of `e-magyar` [9]. To establish the clusters we trimmed the analyses until the last stem, since the derivation does not concern us in this paper, and we used the `<>` sign concatenating the analyses returned by `emMorph`. The following tables show some sample lines before and after the normalization.

elméleti	`elmélet[/N]i[_Adjz:i/Adj][Nom]`
elméleti	`elméleti[/Adj][Nom]`
számítógépes	`számít[/V]ó[_ImpfPtcp/Adj]gép[/N]es[_Nz:s/N][Nom]`
számítógépes	`számítógép[/N]es[_Adjz:s/Adj][Nom]`
számítógépes	`számítógép[/N]es[_Nz:s/N][Nom]`

These two words become *elméleti* – `[/Adj][Nom]` 'theoretical', *számítógépes* – `[/Adj][Nom]<>[/N][Nom]` 'computational, computer-related' after the normalization.

We used `gensim` [10] skip-gram with negative sampling with the default hyperparameters in the creation of our models: the dimension of the word embed-

[1] http://mokk.bme.hu/resources/webcorpus/
[2] https://github.com/dlt-rilmta/emMorph

ding is 200, the window used is 5 words in both directions, 5 training epochs. Negative sampling is set to a factor of 5, and minimal sample size is 5. The model was generated using the surface forms only and morphological analyses were assigned to the words subsequently. In the resulting embedding we can observe (Fig. 2) a strong correlation ($\rho = 0.939$) between the log-frequency of the words and the length of their vectors posited in [11]. Knowing this, we can project the vectors to the surface of the unit sphere without great loss of information. Later in this paper we will use a set of 200000 uniformly distributed random vectors as a baseline for comparing to the actual word vectors projected onto the surface of a unit-radius 200-ball.

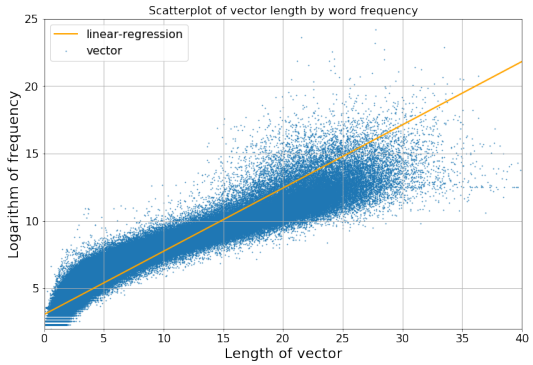

Fig. 2: Length versus $\log_2$(frequency)

Another characteristic of the skip-gram model is that it prefers placing the words in a specific part of n-dimensional space [12]. One technique we used to measure the spatial preference of the model is to count the relative frequency of each coordinate being positive, then plotting these numbers in an ascending order. Fig. 3 shows that some coordinates are highly likely to be positive and others negative, whereas for a random set of points the line would be flat since every coordinate would have 0.5 probability to be positive or negative.

3 The statistics of grammatically defined clusters

We hypothesize that there is a coherent structure in the embedding and each vector encodes a certain meaning and grammatical structure. The clustering methodology we will be using here is a viable approach to classify word vectors to the extent that we can analyze these clusters in a way that helps understanding them. The model used has no a priori knowledge of these grammatical categories, yet we will see that the clusters are indeed coherent.

Visualizing high-dimensional data is a difficult exercise [13]. We can use principal component analysis to maximize the information retained in the first few

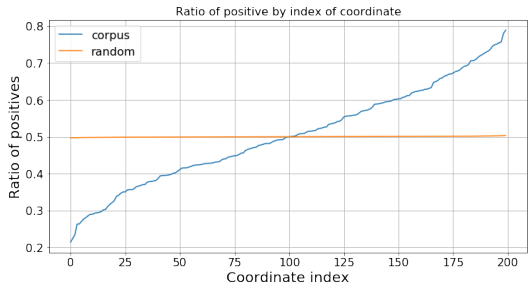

Fig. 3: Probability of a certain coordinate being positive

dimensions. Plotting a sample of 1000 vectors from the spherical projection of the first 3 principal components of some clusters of word vectors yielded Fig. 4, which makes clear we have 3 clusters each restricted to a dominant orthant. What we need to verify is that this phenomenon persists in the whole 200-dimensional space. One way of doing this is by comparing the standard deviations and the entropy of the clusters. If a cluster's standard deviation is high, it indicates low density, the lack of a core, and incoherent structure. If the standard deviation is lower, it indicates a higher density, a more characteristic core. Number of occurrences and entropy (y axis) are plotted against the standard deviation (x axis) in Fig. 5.

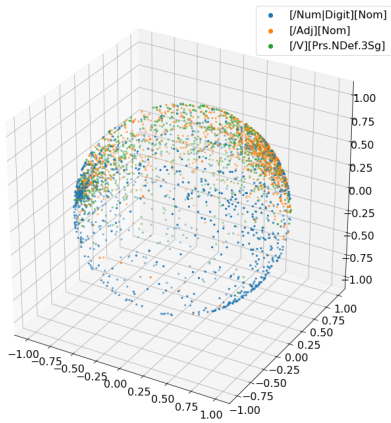

Fig. 4: Clusters on the unit sphere

On the left panel we can see a square-like shape, showing weak correlation between the frequency and the standard deviation. The scatter plot of the entropy-standard deviation shows that higher entropy generally means higher standard

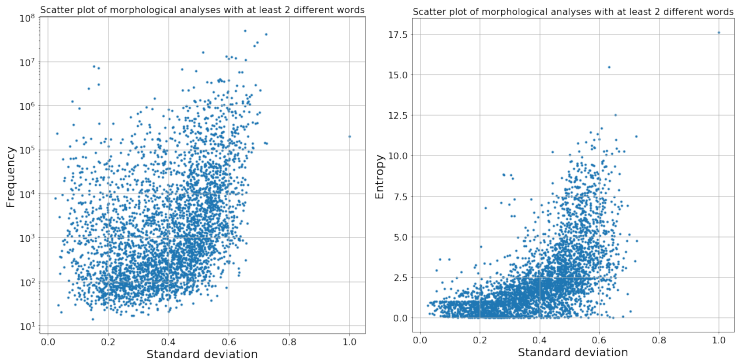

Fig. 5: Scatter plots of clusters

deviation. After filtering out morphological analyses with low number of words, first 5, then 50, Fig. 6 the correlations weaken. The Spearman correlation coefficients for the 4 figures are: 0.512, 0.957, 0.384, 0.057.

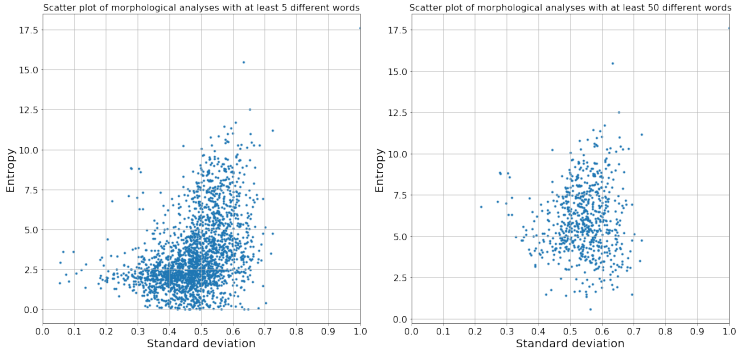

Fig. 6: Scatter plots of clusters

4 Quantifying similarity

Here we define and explain the intuition behind a similarity measure between sets of vectors on the n-sphere. Since it is hard to have intuition in 200-dimensional space, we begin with the definition of a *cap* (vectors at a small angle to an axis):

Definition 1. *n-cap*
Let $\boldsymbol{m} \in S^n$, let $\alpha \in [-\pi, \pi]$. The cap defined by $\boldsymbol{m}$, α is

$$cap_\alpha(\boldsymbol{m}) = \{\boldsymbol{x} | \boldsymbol{x} \in S^n \wedge \langle \boldsymbol{m}, \boldsymbol{x} \rangle \geq \cos(\alpha)\}$$

which is equivalent to

$$cap_\alpha(\boldsymbol{m}) = \{\boldsymbol{x} | \boldsymbol{x} \in S^n \wedge \mathrm{sim}_{cos}(\boldsymbol{m}, \boldsymbol{x}) \geq \cos(\alpha)\}$$

and a theorem about the surface of the n-sphere [14]:

Theorem 1. *Let $I_x(a, b)$ be the regularized incomplete beta function, and $A_n = 2\pi^{n/2}/\Gamma(\frac{n}{2})$ be the surface area of the r-radius n-sphere. Then the area of the spherical cap characterized by its h height is:*

$$A = \frac{1}{2} A_n r^{n-1} I_{(2rh-h^2)/r^2}\left(\frac{n-1}{2}, \frac{1}{2}\right) \tag{1}$$

The theorem and the definition together show the ratio of the surface of the cap to the n-sphere. For example, putting $n = 200, h_1 = \cos\left(\frac{11\pi}{24}\right), h_2 = \cos\left(\frac{10\pi}{24}\right)$ into the theorem above we get 0.0327 and 10^{-4} respectively. We can verify this upper bound by placing uniformly random points on the surface of the n-sphere, counting the points inside cap_α (we performed simulations with 200k random vectors) and compare the ratio given by theorem 1 to the random sample. To measure the compactness of clusters, we use an increasing cap around the cluster centroid, and plot the ratio of word vectors lying in the cap as a function of the minimal similarity of words to the cluster centroid. As we

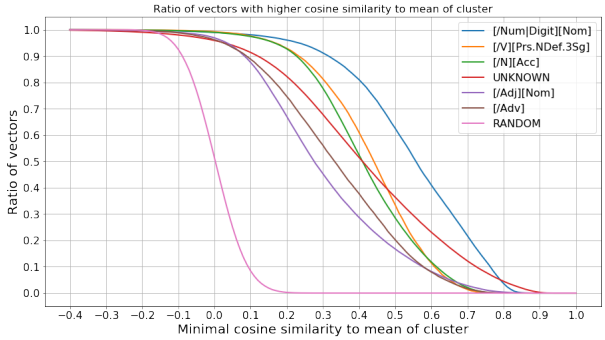

Fig. 7: Ratio of points in a $\mathrm{cap}_\alpha(\mathrm{mean}_{\mathrm{cluster}})$

can see on Fig. 7, the `RANDOM` cap vanishes around $\cos(\alpha) = 0.2$ (for this α, theorem 1 limits the relative surface of the cap to 0.0023), while the other clusters, most notably the `[/Num|Digit][Nom]` (digit in nominative case) shows the strongest coherence, which seems intuitive, as the numbers mostly indicate quantity and amount (counterexamples are dates, or symbolical numbers like 7, 3, 24/7). The `UNKNOWN` cluster shows high coherence, as it is dominated by nouns. The `[/V][Prs.NDef.3Sg]` cluster (third person singular verbs) show the same coherence as the `[/N][Acc]` cluster (accusative nouns), while the `[/Adj][Nom]`

cluster (adjectives in noun case) shows lower coherence than any of the clusters other than the `RANDOM` presented on the figure.

Since we want to filter out noise, and our ultimate goal is to measure similarity, we can use the ratio of the words in a cap_α with fixed α to measure self-similarity, and we can also calculate the ratio of some words in other clusters' cap. That way, we obtain an asymmetrical similarity measure. Obtaining the fixed α is based on filtering out the most noise. We use `RANDOM` as a base of comparison: on Fig. 8, we show the ratios with that corresponding to `RANDOM` subtracted. The plot shows the maximal difference to be around $\cos(\alpha) = 0.13$, so we have chosen $\alpha = \frac{11\pi}{24}$ (82.5°) (for which $\cos(\alpha) \approx 0.1305$). Thus the formal

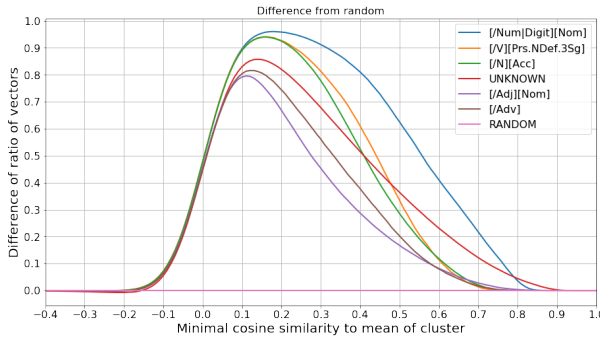

Fig. 8: Difference of the ratios from `RANDOM`

definition of the cluster similarity defined above:

Definition 2. *Cluster similarity*
Let C_1, C_2 be two sets of points on the n-sphere, $\boldsymbol{m}$ be the mean vector of C_1. The (signed) similarity of C_1 and C_2 is:

$$\text{sim}_{cl}(C_1, C_2) = \frac{\left|\left\{\boldsymbol{x} \in C_2 | \text{sim}_{\cos}(\boldsymbol{m}, \boldsymbol{x}) \geq \cos(\frac{11\pi}{24})\right\}\right|}{|C_2|}$$

where $|.|$ is the cardinality of a set.

5 The role of affix frequency

We begin by examining the clusters based on their case endings to see whether some specific case endings contribute significantly more to cluster coherence. Table 1. summarizes the clusters and their respective self-similarities. We can see that the more specific case endings like `[/Adj][Transl]` and `[/Adj][Temp]` (translative and temporal case for adjectives) show higher self-similarity, while the more general ones like `[/Adj][Nom]` and `[/Adj][Supe]` (nominative and superessive) show lower self-similarity. This tendency continues with the cases

affixed to nouns, where [/N][All] and [/N][Transl] (allative and translative) are among the highest self-similarity cases and [/N][Nom] has one of the lowest self-similarity from the paradigm. Let us now consider clusters that are more

Cluster	Sim	Cluster	Sim	Cluster	Sim
[/Adj][Nom]	0.822	[/N][EssFor:képp]	0.889	[/Num][Nom]	0.908
[/Adj][Supe]	0.910	[/N][Nom]	0.922	[/Num][Del]	0.955
[/Adj][Subl]	0.924	[/N][Ess]	0.926	[/Num][Dat]	0.957
[/Adj][Ine]	0.929	[/N][EssFor:ként]	0.936	[/Num][Ter]	0.960
[/Adj][Ela]	0.936	[/N][Ine]	0.937	[/Num][Cau]	0.971
[/Adj][Acc]	0.941	[/N][EssFor:képpen]	0.941	[/Num][Ill]	0.977
[/Adj][Ade]	0.945	[/N][Cau]	0.946	[/Num][All]	0.978
[/Adj][Ins]	0.951	[/N][Ade]	0.949	[/Num][Ine]	0.980
[/Adj][Abl]	0.959	[/N][Hyph:Hyph]	0.957	[/Num][Acc]	0.983
[/Adj][Ill]	0.960	[/N][Ter]	0.962	[/Num][Subl]	0.984
[/Adj][Cau]	0.961	[/N][Supe]	0.962	[/Num][Ela]	0.985
[/Adj][Del]	0.961	[/N][Abl]	0.964	[/Num][Ade]	0.988
[/Adj][Ter]	0.963	[/N][Acc]	0.966	[/Num][Ins]	0.992
[/Adj][Dat]	0.967	[/N][Temp]	0.966	[/Num][Abl]	1.000
[/Adj][All]	0.978	[/N][Ela]	0.968	[/Num][Transl]	1.000
[/Adj][Transl]	0.994	[/N][Del]	0.969	[/Num][Temp]	1.000
		[/N][Ill]	0.969	[/Num][Supe]	1.000
		[/N][Dat]	0.969		
		[/N][Subl]	0.969		
		[/N][Ins]	0.972		
		[/N][Transl]	0.979		
		[/N][All]	0.979		

Table 1: Clustering by case ending

frequent or have higher entropy. We partitioned the clusters into 20 equal bins, each 0.05 wide, by their respective standard deviation, then calculated the mean and the standard deviation (σ) of the vectors of each cluster, see Fig. 9.

Most clusters lay in the 1σ stripe, and the 2σ stripe is also rather populated. Each stripe is monotonically increasing. The interesting clusters are the ones above the 2σ stripe, because compared to their high entropy their variance is smaller than expected. Summarizing on Table 2 the biggest non-ambiguous clusters (counting more than 5000 words) outside the 2σ line on Fig. 9, it shows us that nouns, be they plural or singular, form highly coherent clusters. The presence of infinitive and plural third person verbs among these most coherent clusters is very interesting, because verbs in general did not show strong coherence.

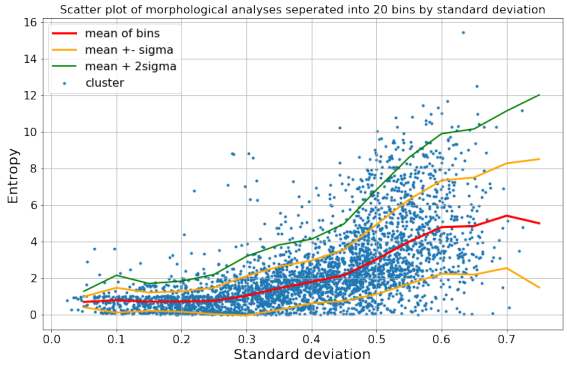

Fig. 9: Binning clusters by standard deviation

6 Asymmetrical similarity

In Section 4. we have already given the idea to compare one cluster's mean to another cluster's elements. When comparing not round-shaped clusters, this way of measuring similarity introduces asymmetry. Plotting a histogram of the differences $\mathrm{sim}_{\mathrm{cl}}(C_1, C_2) - \mathrm{sim}_{\mathrm{cl}}(C_2, C_1)$ shows a distribution quite close to normal.

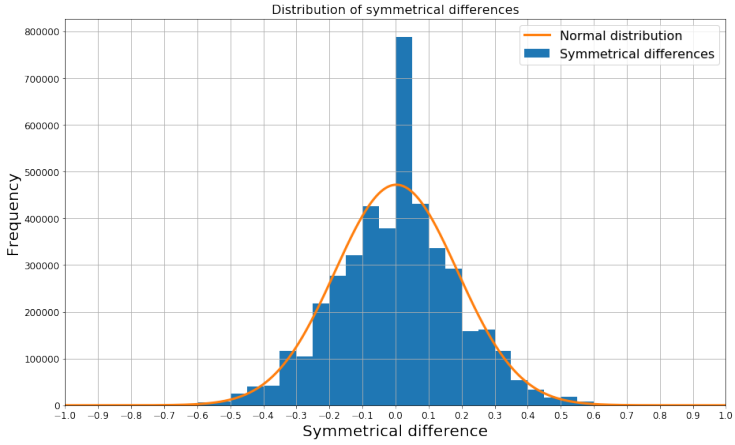

Fig. 10: Distribution of symmetrical differences

Most of these differences are around 0, showing that most of the clusters are round shaped. The two tails of the distribution are the important parts, because they show us pairs of clusters whose pairwise similarity in one direction is 1, while in the other direction this similarity is 0. One example to this phenomenon is the pair of `[/N|Pro][Subl][1Sg]`, `[/N|Pro][3Pl][Dat]<`

Bin	σ_{diff}	**Cluster**	$\textbf{Sim}_{\text{self}}$	**#Words**	**Frequency**
0.45	5.75	`[/V][Inf]`	0.988	14071	6677547
0.50	2.88	`[/Num|Digit][Nom]`	0.977	26666	6000563
0.55	2.25	`[/V][Prs.Def.3Pl]`	0.990	5230	1312497
0.55	2.56	`[/N][Pl][Subl]`	0.980	6795	582972
0.55	2.20	`[/N][Pl][Supe]`	0.979	5325	519220
0.55	2.72	`[/N][Pl][Ins]`	0.982	10135	955157
0.55	2.17	`[/V][Prs.NDef.3Pl]`	0.989	10486	3296388
0.55	2.94	`[/N][All]`	0.979	13858	1073443
0.60	2.60	`[/N][Ins]`	0.972	32886	3868455
0.60	2.42	`[/N][Subl]`	0.969	25687	3469518
0.60	2.21	`[/N][Pl][Acc]`	0.974	20702	3601070
0.60	2.25	`[/N][Abl]`	0.964	10270	649706
0.60	2.22	`[/N][Ela]`	0.968	12717	1028608
0.65	2.04	`[/N][Ade]`	0.949	7324	363706
0.65	3.99	`UNKNOWN`	0.892	199475	5643460
0.65	2.58	`[/N][Acc]`	0.966	61671	12617934
0.65	2.06	`[/N][Poss.3Sg][Acc]`	0.962	14164	2823258
0.70	2.48	`[/N][Nom]`	0.922	144945	50298170

Table 2: Clusters with unexpectedly high coherence

`>[/N|Pro][Poss.3Pl][Dat]` clusters. Both of the clusters contain 4 vectors, but the words of the first cluster have 42228 occurrences in the corpus, and the words of the second cluster count 545 occurrences. The words are pronouns in both cases, the first clusters' words are *énrám, reám, rám, énreám,* ($s = 0.1$) meaning 'onto me' with variable spelling, the difference is only stylistic, the words of the second cluster are *némelyiküknek, valamelyiküknek, mindegyiküknek, bármelyiküknek*, ($s = 0.382$) the first one meaning 'to some of them' and 'of some of them', while the rest meaning the same, but changing 'some' to 'specific one', 'all', 'any'. One reason for this strange phenomenon is that the *énrám, reám, rám, énreám* have identical meanings, the standard deviation of their cluster is very low, while the other cluster of 4 words have significant difference in their meanings.

In the following sections, the asymmetry is of less importance. As shown on Fig. 10, most of the pairwise similarities have difference below 0.1, thus we do not lose much by symmetrizing the similarity measure by taking the mean of the similarities, $(\text{sim}_{\text{cl}}(C_1, C_2) + \text{sim}_{\text{cl}}(C_2, C_1))/2$.

6.1 Subcategories

E-magyar creates multiple subcategories for adjectives, nouns and numbers, and we can measure the pairwise similarity of their paradigms. If some subcategories show high similarity, we can say that it is not worth preserving as separate categories. Comparison of the subcategories to the `[/Adj]` categories yields interesting results.

Cluster$_1$	**Cluster$_2$**	**similarity**	**cases**
`[/Adj][.]`	`[/Adj][.]`	0.954	22
`[/Adj][.]`	`[/Adj|col][.]`	0.921	14
`[/Adj][.]`	`[/Adj|nat][.]`	0.900	16
`[/Adj][.]`	`[/Adj|Attr][.]`	0.865	7
`[/Adj][.]`	`[/Adj|Pro][.]`	0.843	18
`[/Adj][.]`	`[/Adj|Pro|Rel][.]`	0.549	7
`[/Adj][.]`	`[/Adj][Pl][.]`	0.884	17
`[/Adj][Pl][.]`	`[/Adj][Pl][.]`	0.956	17
`[/Adj][Pl][.]`	`[/Adj|col][Pl][.]`	0.949	9
`[/Adj][Pl][.]`	`[/Adj|nat][Pl][.]`	0.943	16
`[/Adj][Pl][.]`	`[/Adj][Poss.1Sg][.]`	0.855	10

Table 3: Adjectival subcategories

`[.]` marks the pairwise comparison of single morphemes, so in the first few examples, we compare singular forms to singular form (because singular forms are not marked, thus a single morpheme after the word root must mean singular), and in the cases after, the plural forms. We can see a declining similarity when comparing more and more specific clusters, with the `[/Adj|col][.]` (adjectives describing colors) and `[/Adj|nat][.]` (adjectives describing nationality) being relatively similar to `[/Adj][.]`, while `[/Adj|Pro][.]` (pronominal adjectives) and especially the `[/Adj|Pro|Rel]` (relative pronouns like *amilyen* or *amekkora*, 'such as', 'as large as, as much as') show significantly less similarity. As we noted at the beginning, more specific case endings may dominate the word vectors' similarity clusterwise, which is indeed the case in the last examples. Comparing plural adjectives, the similarities are significantly higher than their singular counterparts' similarities, while comparing singular to plural yields very low similarity.

6.2 Paradigm self-similarities

In the previous section, we have already used the `[.]` to indicate the comparison of paradigms. While the nominative forms may have lower similarities, the paradigm comparisons are dominated by the abundance of cases and case endings, producing very high self similarities. `[.]` denotes only a single morpheme, so this table aggregates only the 2-morpheme-long morphological analyses.

7 Conclusions and further research

Clustering word vectors by their morphological analysis has proven a good way to examine the impact of inflection on word vectors. Because of the high dimension, naive statistical testing of the distances from the mean does not produce easily interpretable results. In contrast, the 'cap similarity' introduced here,

$\textbf{Cluster}_1$	$\textbf{Sim}_{\text{self}}$	**cases**	$\textbf{Cluster}_1$	$\textbf{Sim}_{\text{self}}$	**cases**
`[/Adj|Pro|Rel][.]`	1.000	7	`[/N|Unit][.]`	0.992	14
`[/Num|Pro][.]`	1.000	9	`[/Adj|nat][.]`	0.995	16
`[/Num|Roman][.]`	1.000	6	`[/V][.]`	0.989	54
`[/N|Acronx][.]`	1.000	13	`[/Post][.]`	0.987	8
`[/N|Pro|Rel][.]`	1.000	15	`[/N|Unit|Abbr][.]`	0.984	14
`[/Adj|col][.]`	1.000	14	`[/Num][.]`	0.979	18
`[/N|mat][.]`	0.998	17	`[/Num|Digit][.]`	0.975	14
`[/N|Ltr][.]`	0.997	13	`[/N|Acron][.]`	0.974	14
`[/N|Abbr][.]`	0.996	13	`[/N][.]`	0.958	24
`[/N|Pro][.]`	0.996	16	`[/Adj|Pro][.]`	0.958	20
`[/Adj|Attr][.]`	0.995	7	`[/Adj][.]`	0.955	22

Table 4: High self-similarity

while asymmetrical, has produced acceptable results, showed high coherence and similarity where expected, and showed lower similarity where difference was expected, thus justifying the selection of clusters for most cases. There are exceptions however, such as treating `[/Adj|Pro|Rel]` as a subcategory of `[/Adj]`, which our method shows to be mistake due to their low similarity. Other future work may also include using disambiguated text corpus to have bigger clusters thus more data to perform the same analysis.

There are supervised methods for creating meaningful ultradense subspaces for polarity, concreteness, frequency and part-of-speech (POS) [15,16], supporting operations like 'give me a neutral word for *greasy*'. We plan on analyzing the POS subspace, comparing the similarities of the clusters projected onto the subspace with the similarities obtained without projection.

Acknowledgment

This research is partially supported by National Research, Development and Innovation Office NKFIH grant #120145 'Deep Learning of Morphological Structure' and by 2018-1.2.1-NKP-00008 'Exploring the Mathematical Foundations of Artificial Intelligence'.

References

1. Siklósi Borbála, N.A.: Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra. In Tanács, A., Varga, V., Vincze, V., eds.: Proc. MSZNY 2016. Szegedi Tudományegyetem (2016) 3–14
2. McCulloch, W., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biophysics **5** (1943) 115–133
3. Mikolov, T., Yih, W.t., Zweig, G.: 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 (NAACL-HLT 2013), Atlanta, Georgia, Association for Computational Linguistics (2013) 746–751
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. (2013) International Conference on Learning Representations (ICLR 2013).
5. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., eds.: Advances in Neural Information Processing Systems 26. Curran Associates, Inc. (2013) 3111–3119
6. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), ELRA (2004) 203–210
7. Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In Kilgariff, A., Baroni, M., eds.: Proc. 2nd Web as Corpus Workshop (EACL 2006 WS01). (2006) 1–8
8. Novák, A., Siklósi, B., Oravecz, C.: A new integrated open-source morphological analyzer for Hungarian. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
9. Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: `e-magyar`: digitális nyelvfeldolgozó rendszer. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), Szeged (2017)
10. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA (2010) 45–50
11. Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A.: Rand-walk: A latent variable model approach to word embeddings. Transactions of the Association for Computational Linguistics (TACL) **4** (2016) 385–399
12. Mimno, D., Thompson, L.: The strange geometry of skip-gram with negative sampling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2017) 2873–2878
13. Grinstein, G., Trutschl, M., Cvek, U.: High-dimensional visualizations. In: Proceedings of the Visual Data Mining Workshop, KDD 2, 120. (2002)
14. Li, S.: Concise formulas for the area and volume of a hyperspherical cap. Asian Journal of Mathematics & Statistics **4** (2011) 66–70
15. Rothe, S., Ebert, S., Schütze, H.: Ultradense word embeddings by orthogonal transformation. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, Association for Computational Linguistics (2016) 767–777
16. Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Rojas-Barahona, L.M., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Counter-fitting word vectors to linguistic constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, Association for Computational Linguistics (2016) 142–148