Daniel Crawford
12 December 2023
LING2773 – Morphology
Prof. M Kanwit

# Lexical Access Prediction from Semantic Embedding

## I. Introduction

The study of morphology allows us to understand the process by which words are formed. Implicit in that pursuit is the desire for a psychologically reflective method of explaining word formation: we seek to understand not only the word-formation processes but the cognitive patterns that underpin them. As we consider this, we learn about the nature of accessing words from our lexicon.

## II. Literature Review and Explanation of Morphological Pattern

### A. The Lexicon

The lexicon can be considered an 'internal dictionary.' That is, the lexicon is a mental list of words that can be accessed by a speaker to express ideas, usually in sentences. In the attempt to have a psychologically reflective model of morphology, we want to understand what elements exist in the lexicon: There is a continuum of representations of the lexicon.

One may consider a lexicon comprising entirely of morphemes. Some of these morphemes will be free – they do not need concatenation and can be expressed along, such as the verb 'do.' A native English speaker is comfortable with this word being used alone. Other morphemes are bound: they cannot be found (in standard context) on their own. An example would be the English 'un- 'prefix, which must be attached to another constituent. In this framework, a morpheme-based lexicon model, the lexicon consists of morphemes assembled into words – there is no preexisting construction. While this would be a convenient model for constructing a hierarchal view of language (phonemes create morphemes, which create words, which create sentences), it demands a high level of saliency of morphemic boundaries, which are not guaranteed to be present in analytical languages as much as in synthetic/agglutinative ones.

Contrastively, it could be that the lexicon is composed of very preconstructed words that the morpheme-based model does not permit. This would be a word-based model of the lexicon. Adopting this model suggests that speakers have a list of words that hold no morphemic boundaries – save for semantic processes – and access the words wholly. This would require a speaker's memory to be highly adept at recalling words as a whole, and does not explain the phenomena of novel constructions – there is no account for parsing new words, which speakers can clearly do.

Haspelmath & Sims (2010) offer a mediating position between these two models: moderate word-based lexicon. This offers those processes relating to both lexicon models at work. Some words are accessed directly, while others are formed and accessed as a whole. This allows for a more flexible approach that is still able to account for experimental results and make theoretical predictions. Researchers in this area have termed the process of accessing the word wholly as the direct route (from the word-based model) and composite route (reflective of the morpheme-based modes.) (Haspelmath & Sims, 2010; Jennifer Hay, 2001; Taft & Forster, 1975)

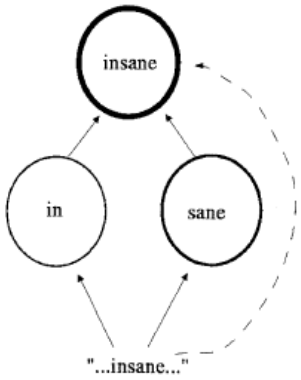



*Figure 1. A schematic of the two methods of lexical access.(Jennifer Hay, 2001)*

B. Lexical Access Routes

The conditions that affect the election of these two routes are of interest because they aid morphology in being psychologically reflective of speakers' thinking. Consider Example 1; Hay (2001) suggests that it would be natural for the morpheme *in-* to exist independently of any base in the lexicon. Thus, when the speaker wishes to produce the word '*insane*,' they simply need to combine it with the base '*sane*'. However, assuming this is the case may not always be accurate. In more synthetic words (and indeed languages), these words may exist as complete, preconstructed units. Consider Example 2: Taft and Forster (1975) suggest that the word '*unremittingly*' may need to be accessed as a complete word due to its complexity and the fact that '*mit*' (in the core meaning used here) would not have much meaning, and probably not be stored independently.

| | |
|---|---|
| *in-sane*<br>NEG-sane<br>'insane' | *un-re-mit-ing-ly*<br>NEG-REP-mit-PRS-ADV<br>'unremittingly' |

*Example 1 & 2. Examples of readily decomposed words that may be accessed in multiple ways. (Jennifer Hay, 2001; Taft & Forster, 1975)*

C. Lexical Access Route Influences

Segmentability and allomorphy influence the access route (Haspelmath & Sims, 2010). Naturally, the more familiar the speaker is with morphology, and the more salient the morphological constituents are, the more segmentable the word is likely to be. This would lead to an increase in composition route access. If the composition of the word contains allomorphic variances, the saliency of the constituents would decrease, suggesting a direct access route is being implemented. It has also been posited that there is a saliency of bases, and speakers more readily identify real bases than fake ones, suggesting that the base morphemes are stored individually (Taft & Forster, 1975).

More recently, Hay has argued that the lexical access route has a social component. It seems that the speaker is making the election based on the social domain of the discourse, as well

as the gauged status of the speaker (J. Hay, Walker, Sanchez, & Thompson, 2019). They argue that the lexicon contains social elements or notes that will influence access routes.

The influence most relevant to this project is the relative frequency of the base morpheme and the complex word. It has been suggested that, with other factors equal, if the complex word is of higher frequency than its simpler base, it will be accessed directly (Jennifer Hay, 2001). This project is an extension of this research.

D. Semantic Transparency

In the 2001 investigation, Hay appeals to semantic transparency: it is suggested that the more semantically transparent a word is, the more likely it is to be accessed via the composite route. The less the complex word has undergone semantic drift, the more likely it is to have a directly relational meaning to the base word. Because of this, speakers would conceive of the base word and then concatenate the affix to achieve the desired output – which is the composite route. In contrast, words that do (*now*) differ greatly in their meaning, even if they are related morphologically to one another, are more likely to be accessed directly. This is because, without the aspect of semantic transparency, there is a lower chance for the speaker to associate those two words.

To determine the level of semantic transparency, Hay (2001) investigated the presence of the base word of the complex (comprising of multiple morphemes) word in the dictionary entry of the complex word itself and conducted a binary classification (the base was present or not present in the definition of the complex word). This provides a useful metric for determining semantic transparency: if the complex word is a simpler extension of the base (even if it is a negation, the complex is still highly related to the base), then it would be reasonable to assume that the base would be present in the definition of the complex word. This approach is subject to fluctuations in particular dictionary entries. Further, the binary classification may coarse some details in the rest of the study.

The research presented here is an extension of this idea: a computational approach to understanding the effects of semantic transparency lexical access routes. The assertion is that by using novel computational tools, there will be a higher degree of nuance in the usage of semantic transparency, which would be a more rigorous metric of semantic transparency. Because it has been established that there is a relation in relative frequency to the lexical access route, we are seeking a computational model that is able to predict the relative frequency of a base to its complex form. This, in turn, will provide a lexical access prediction.

E. Computational Models of Semantic Representation

With the dawn of several computational methods that are applicable to linguistics, such as word2vec, neural embeddings, and semantic analysis, there are novel ways to investigate semantic transparency. The word2vec algorithm seeks to translate a word into some multidimensional vector (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). This family of algorithms utilizes words that co-occur with the word in

question to determine its position in a high-dimensional vector space. This is thought to be a semantic space for the vector, in which meaning can be extrapolated from relative position.

When this is done for thousands of words, it is suggested that there will be relationships formed between these words that reflect their semantic meanings. The input to a word2vec model is a corpus. The corpora can hold billions of words. By moving a sliding window over these words, and performing statistical transformations to accounts for type-frequency and document frequency, a multi-dimensional array is the output. A shallow neural network is implemented through which the corpus passes. The resulting vectors, or embeddings, provide linguistics with a computational avenue for analyzing and predicting linguistic phenomena, based off of a model for semantics. More details on the corpora and metrics are discussed in the Methods section.

Research has been conducted in the area of segmenting words into constituent morphemes. In the past, these have been based on probabilistic frameworks. (Creutz & Lagus, 2007; Galinsky, Kovalenko, Yakovleva, & Filchenkov, 2018). The strategy has been to create (literally) a morpheme-based lexicon. These probabilistic models have been extended to more sophisticated approaches (Sorokin, 2022). This work centers around leveraging the (relatively) novel transformer (Vaswani et al., 2017) – based BERT architecture (Devlin, Chang, Lee, & Toutanova, 2019) to create a morpheme embedding. While this project is not directly related to morpheme segmentation, it is important to understand that the work is being done.

## III. Research Questions

This research investigates methods of predicting the lexical access route by associating relative frequency semantic similarity based on semantic embedding. This will extend the knowledge already present in the field (Libben & Jarema, 2002) by providing researchers with a reliable and replicable computational methodology for predicting, based on the semantics of a word (as provided by word2vec), whether speakers will access that word by the direct route or the composite route. It is also a direct extension of Hay's (2001) notion of semantic transparency.

Results here are interesting in the fields of linguistics, morphology, psychology, cognitive science, and computational linguistics because they will provide both insights into the elements of the lexicon, and inform computational representations of the lexicon for applications across the (computational) linguistics spectrum. Further, results will increase the parallels between formal descriptions of morphological processes and the psychological procedures they are describing.

Therefore, the central research goal is to determine if the word2vec family of semantic embedding algorithms can be combined with a measure of semantic distance to accurately predict lexical access routes. This will require the answering of the following research questions:

- What is an appropriate corpus to utilize for predicting lexical access routes?
- What metric of distance embedding allows for the detection lexical access routes?
- What are the predictions made by combining the preferred corpora + metric pairing?

Results of experiments found Hay (2001) will be utilized to compare the results found in this investigation. (Details on the experiment were provided for this investigation so that they can be recreated in this computational counterpart.)

It is predicted that there will be a suitable metric and corpus pairing that allows for the accurate prediction of lexical access routes. Indeed, the goal of semantic embeddings is to accurately reflect the meaning of a word *relative* to other words. With the increasing number of corpora, finding suitable data is not expected to be an issue. Further, there has been research into improving various word2vec algorithm, and the distance metrics are very accessible. With these in mind and regarding predictions, therefore, when controlling for the meaning of a morpheme addition, it is expected that there is a way to capture the severity of this change in meaning, allowing for predictability in accordance with Hay (2001): the greater the semantic distance, the lower the semantic transparency, the greater the relative frequency of the base to the complex, the greater probability of being accessed directly.

## IV. Methods

This section outlines the methods for creating the list of words to investigate, the creation of the computational models, and the metrics utilized to determine similarity.

### A. Dataset

Hay (2001) provides the list of words that were utilized in the experiment. These words were tagged for relevant parts of speech. Note, in the case that the word could be multiple parts of speech, the one that was most common was used. Eg: *list* NOUN *listless* ADJ. Also, the meaning of the word that was captured in *both* entries was utilized. Eg: *perfect* ADJ *impefect* ADJ, even though *perfect* may be used as a verb. This is indeed a reference to semantic transparency, however, allowing for the creative and uncommon uses of the word to determine part of speech would not reflect in a model, which is of course an approximation. Also, it was the case that the more natural part of speech classification was more common.

### B. Corpora

Five corpora were utilized for this study. Three were facilitated by the Python programming library spaCy (M. Honnibal, 2020). The smallest of the three `en_core_web_sm` contained a large corpus from various genres of text: news, conversational telephone speech, weblog, newsgroups, broadcasts, and talk shows (Ralph Weischedel, 2013). About 625,000 English words were included total. Also used was `en_core_web_md`. This contains the same elements as the previous, and has the addition data from Wikipedia (Tiedemann, 2016; Wikimedia.org, 2016), OpenSubtitles (Tiedemann, 2016; Tom Kocmi, 2022), WMT Newscrawl (Tom Kocmi, 2022), and OSCAR 21.09 (Pedro Ortiz Suarez, 2019). This contains over 8 million English tokens. The largest of the three corpora accessed through spaCy is `en_core_web_lg`, which includes over ten times as much data, so can be thought of as an expansion. These corpora will be referred to as spaCy Small, spaCy Medium, and spacy Large.

To diversify the corpora included, the gensim library in python was also utilized (R Rehuvrek, 2010). Two corpora were harnessed with this library: a corpus of google news articles, the Google News dataset of about 100 billion words (Google, 2013), and a Twitter corpus of 2 billion tweets, with 27 billion tokens (J Pennington, 2014). These will be referenced as `google` and `twitter` respectively. Note that the corpus trained the models which were then downloaded for this investigation.

C. Distance Metrics

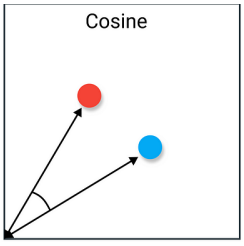
Three options were investigated for distance metrics. When working with semantic embeddings, the similarity metric provides a way to compare two vectorized words. That is, the closer in semantic meaning the words are, the closer they will be in the space. This follows from the discussion of semantic embeddings. Therefore, a metric that is accurately able to assess similarity would be a powerful tool. It is standard to compare human intuition to the similarity metric to determine this accuracy.

*Equation 1. Cosine Similarity*

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

In this investigation, the first metric considered is cosine similarity. This is one of the most common metrics utilized in NLP, Machine Learning, as well as any field that utilized multidimensional spaces. This is because it is attuned to small fluctuations, and is computationally efficient to compute. The equation for cosine similarity is given in Equation 1: the dot-product of two vectors, divided by the product of their magnitude. This is highly useful because it captures the difference in angle between the two vectors in question, as shown in Figure 2. The greater the cosine similarity, the closer the vectors are, and therefore closer the words are thought to be.

*Figure 2. Costine Similarity in 2D space.*




*Equation 2. Minkowski Distance, D*

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

The second metric used is Minkowski Distance. This is a generalization of the Euclidean distance. With two dimensions (n = 2) and p = 2, this is the Pythagorean Theorem. In the case here, we will see that it extends the notion of Euclidean (flat plane) distance to the multi-dimensional space that the semantic embedding exists in.

*Figure 4. Minkowkski Distance (Representing Multiple Dimensions)*

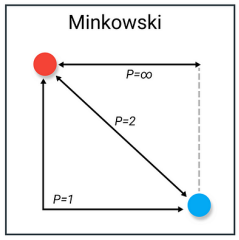
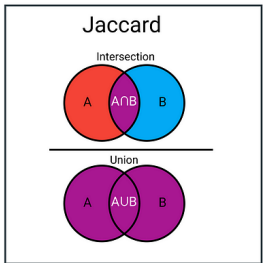



*Figure 3. Jaccard Index*




Finally, the Jaccard Index (Jaccard Similarity, Jaccard Distance) is a metric for determining the distance for two sets. While it is indeed being used here to determine similarity between two vectors, the same notion will apply: it determines the ratio of overlapping in the multidimensional space to union of space occupied by both vectors. This is a natural metric for studying semantic distances because it reflects natural overlap between two vectors.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

*Equation 3. Jaccard Index (Set Notation)*

The cosine similarity we implemented with the native architecture of spaCy and genism. The Minkowski Distance and Jaccard index were implemented with scikit learn. (Pedregosa, 2011). Cosine similarity and the Jaccard Index range from 0 to 1, with greater values being more similar, and the Minkowski distance ranges from 0 (most similar, because of least distance) to infinity. It is good to include these two kinds of ranges because it diversifies the metrics.

## V. Results

Recall that the goal of this investigation was do determine an appropriate/useful corpus-metric pairing of semantic similarity for predicting lexical access routes, via relative frequency. This is an extension of Hay' (2001) notion of semantic transparency. Initially the results did not provide much of a pattern: there was no strong signal for determining if a base word had a higher frequency than the complex word from the semantic similarity.

To investigate deeper, a delineation was made and two categories are proposed: cases where the complex word is a different word-class (part of speech) from the base word. When this separation is made, there does appear to be patterns. A series of two-way ANOVA were conducted to determine if there is significant difference semantic similarity between the base word and the complex word, under the influence of the two variables: if the Base Frequency is greater than the Complex Frequency, and if the Complex wordform is of a different word class. Table 1 provides the p-values of each of the distance metrics, and the corresponding factors, for each corpus used.

*Table 1. P-values of Corpus + Metric Pairing, reflecting ability to differentiate features.*

| | Cosine Similarity | | Minkowski Distance | | Jaccard Index | |
|---|---|---|---|---|---|---|
| | Ch. WC | B > C | Ch. WC | B > C | Ch. WC | B >C |
| Spacy Small | **<0.0001** | **<0.01** | **<0.0001** | 0.13 | **0.038** | 0.361 |
| Spacy Medium | 0.13 | **0.016** | **0.043** | 0.172 | 0.057 | **0.037** |
| Spacy Large | **0.025** | 0.359 | **0.031** | 0.162 | 0.333 | 0.395 |
| Google | 0.745 | 0.591 | 0.239 | 0.312 | 0.309 | 0.875 |
| Twitter | 0.871 | 0.639 | 0.950 | 0.824 | 0.038 | 0.362 |

*Figure 5.*

While there are a few cases of one of the factors appearing to make a difference in semantic similarity, it appears that using the cosine similarity metric on embedding form the Spacy Small corpus yields low p-values, suggesting that this pairing would be a valuable candidate for trying to

predict relative frequency. This suggestion is based on the indication that there is a difference similarity across the two factors.

Figure 5 is a plot of the similarity and relative frequency. Note that each marker is a base-complex word pairing. There appears to be a general positive linear trend: as the similarity increases is seems that the relative frequency increases: the base becomes more frequent that the complex. The trend lines indicate the linear regression model that was determined from these data. These models are summarized below in Table 2.

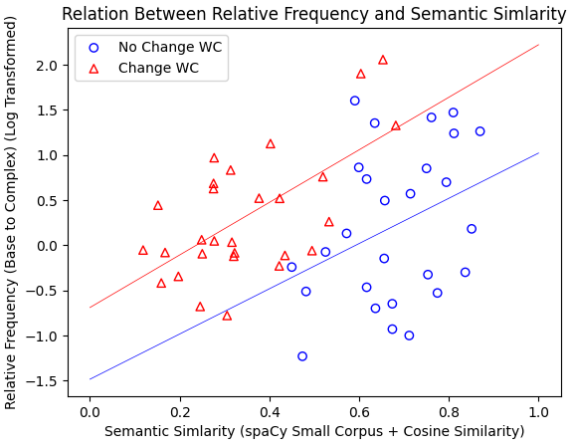



*Table 2. Summary of Linear Models*

| MODEL | N | COEF. | 95% CI | ST. ERR. | F | P | $R^2$ |
|---|---|---|---|---|---|---|---|
| **CHANGE WC** | 28 | 1.2383 | [0.613, 1.864] | 0.305 | 16.50 | 0.0003 | 0.379 |
| **NO CHANGE WV** | 27 | 0.3782 | [-0.105, 0.861] | 0.235 | 2.589 | 0.120 | 0.091 |

## VI. Discussion

The results of this investigation do support a model of predictive relative frequency from semantic similarity. As the base-complex word pair becomes closer in semantic similarity, we see an increase in the base frequency relative to the complex frequency. This is congruent with the expectations. As we are utilizing semantic similarity metrics in embeddings as a model for semantic transparency, we see that that more semantically similar (transparent) the base-complex word pairs are, the more frequency the base is relative to the complex. This in turn predicts lexical access via the composition route, just as Hay (2001) suggests.

The model is reliable for the cases complex word form is of another word class. Indeed, there does not appear to be a significant relationship to infer a relationship in the cases where word class was not altered. However, regarding the change in word class, note that Figure 5 does appear to capture the intuition that preserving word class in creating the complex word form does correlate with an increase in semantic similarity. Indeed, the average similarity (spaCy Small Corpus + Cosine Similarity) for cases with changing word-class is 0.34 compared to 0.67 when there is not this change.

An important note about the results is regarding the similarity metrics. When we consider the high dimensionality of the embedding space, the overall distance that a complex word may differ from the base word is small, compared to the hundreds of other dimensions. The resulted in very high similarity metrics (almost 1.0 in many cases) for the Minkowski and Jaccard metrics. While there was occasionally an ability to distinguish certain feature (if the base frequency is greater than complex frequency and if the complex word is in a different word class), because of the low variance, it is reasonable to conclude the some of the results were truncated by precision limitations of the libraries and computer used for the investigation.

It is interesting that the spaCy small corpus is the only corpus to show significant delineation across both features. This is especially interesting because spaCy medium is an extension of spaCy small (it contains all of spaCy small) and further, spaCy large is an extension of both. It is possible that the larger corpora are more susceptible to noise, similar to how the Minkowski and Jaccard metrics had the signals obfuscated by the high dimensionality of the embedding space. But this does not necessarily align with the intuition that with a more robust corpus, we would see a convergence to a particular conclusion.

Recall the research questions posed:

- What is an appropriate corpus to utilize for predicting lexical access routes?
- What metric of distance embedding allows for the detection lexical access routes?
- What are the predictions made by combining the preferred corpora + metric pairing?

It seems that the word2vec algorithm conducted on the spaCy Small corpus results in a useful embedding for this investigation. It may not be the case that it is the only corpus to do this, and as discussed, invites questions as to why it is the case. The cosine similarity is, predictable, adept at measuring semantic transparency as it relates to relative frequency. This is because it is highly fined tuned to the fluctuations in position.

From the pairing of these two, it is suggested that there can be predictions made regarding lexical access route from similarity metrics made with semantic embeddings. This is an extension of the role of semantic transparency, which also acts as predictor for lexical access. The specific prediction is that as similarity increases, there is an increase in base frequency relative to complex frequency, again support the selection to access the complex word via the composition route.

## VII. Future Research

As this investigation was a direct extension of the Hay's (2001) study, many decisions were made or predetermined. These decisions were primarily focused on the word list. Hay (2001) utilized 58 words (3 here were removed because in three cases, one of the members of the base-complex pair was not included in the corpus). So, a natural extension would be to investigate more base-complex word pairs. It would be reasonable to expect different behaviors from different kinds of affixation processed, so grouping into classes like prefixation, suffixation, negation, would serve to refine the investigation helpfully.

Perhaps the most significant aspect is the reliance on the token frequency of the members of the word list. The relative frequency was calculated from Hay's (2001) frequency, which was founded on the CELEX corpus. (Baayen, Gulikers, Piepenbrock, Centre for Lexical, & Max Planck Instituut voor, 1995). It could be that the corpora used in this investigation do not reflect the same or similar frequency, which may alter the results. Indeed, comparing corpora across linguistic domains is known to create variation in results. (Kanwit, 2021) Therefore, a very positive next step would be to investigate this with intra-corpus frequency, and further, comparing the frequencies would be important as well.

Looking at the resulting plot in Figure 5 it may seem that there is more than just a linear pattern. The linear regression modeled utilized to report the findings, could readily be extended. There are specialized clustering models and more sophisticated grouping mechanism that may be able to be

leveraged to refine the scope of predictability. It may even be possible to have neural network layers added to the embedding architecture that can be used to specifically predict lexical access route directly, rather than just demonstrate prediction of relative frequency.

With respect to the specific predictions made by semantic transparency, an important next step would be to conduct experiments to ascertain lexical access predictions. The Hay (2001) study that this project extends does this. However, those data were not explicitly given, but relative frequency was, so it was chosen to be the variable to serve as a target to predict. But there are certainly predictions made here that could be tested with a routine test of providing native speakers with words and determining the parsing rate or reading speed of words and determining if that positive correlation of semantic similarity and access via the composition route continues.

## VIII. Conclusions

The primary morphological phenomenon under investigation here is lexical access route selection. As Hay (2001) puts forth this competing model, the appeal semantic transparency is offered as a means to augment relative frequency. In this current investigation, the notion of semantic transparency is captured in semantic similarity, found by semantic embeddings. The predictions from the previous study are supported.

The limitations of the study are primarily related to word choice and reliance on frequency data form Hay (2001). There fact that the frequency data was not intertwined with the semantic embedding data could affect the results here. The goal with corpus linguistics and semantic embeddings would be to have it be the case that there is no dependence on an individual corpus, I am certainly an import limitation of the work here. Further, there is a strong case for interpreting the resulting without the linear model, and removing the idea of predictions. Rather we may consider semantic similarity as a simple augmentation to relative frequency in determining lexical access routes. This is brough most to light by the fact that the experimental data were not utilized here.

A useful interpretation of this morphological phenomena is that speakers utilize the meaning of a base-complex word pairing when accessing their lexicon. If the complex counterpart has a more transparent meaning, there is thought to be an increased selection of the composition route. This investigation has provided a computational way of corroborating this theory. The model put forth related semantic similarity to relative frequency, and uses that to make a prediction of lexical access routes. Because the model is predicated on semantic dimensions, it could be a valuable tool because we could intuit a certain ‘dimension’ that is traversed by creating the complex word form the base word. This then suggests a validity to the model: what we are capturing is traversal of this dimension,

## References

Baayen, R. H., Gulikers, L., Piepenbrock, R., Centre for Lexical, I., & Max Planck Instituut voor, P. s. (1995). The Celex lexical database. [Philadelphia, Pa.]: Linguistic Data Consortium [Philadelphia, Pa.].

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process., 4*(1), Article 3. doi:10.1145/1187415.1187418

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Paper presented at the North American Chapter of the Association for Computational Linguistics.

Galinsky, R., Kovalenko, T., Yakovleva, J., & Filchenkov, A. (2018, 2018//). *Morpheme Level Word Embedding.* Paper presented at the Artificial Intelligence and Natural Language, Cham.

Google. (2013). word2vec. Retrieved from https://code.google.com/archive/p/word2vec/

Haspelmath, M., & Sims, A. D. (2010). Understanding Morphology.

Hay, J. (2001). Lexical frequency in morphology: is everything relative? *Linguistics, 39(6)*.

Hay, J., Walker, A., Sanchez, K., & Thompson, K. (2019). Abstract social categories facilitate access to socially skewed words. *PLoS One, 14*(2), e0210793. doi:10.1371/journal.pone.0210793

J Pennington, R. S., C D Manning. (2014). GloVe: Global Vectors for Word Representation.

Kanwit, M., & Berríos, J. (2021). No se sabía de que eso iba a pasar: Do lexical frequency and structural priming condition dequeísmo? *The Routledge handbook of variationist approaches to Spanish, M. Díaz-Campos (Ed.)*, 453-467. doi:DOI: 10.4324/9780429200267-41

Libben, G., & Jarema, G. (2002). Mental Lexicon Research in the New Millennium. *Brain and Language, 81*(1), 2-11. doi:https://doi.org/10.1006/brln.2002.2654

M. Honnibal, I. M., S. Van Landeghem, A. Boyd. (2020). spaCy: Industrial-strength Natural Language Processing in Python. doi:10.5281/zenodo.1212303

Mikolov, T., Chen, K., Corrado, G. s., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR, 2013*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. s., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems, 26*.

Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P.and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research, 12*, 2825--2830.

Pedro Ortiz Suarez, B. S., Laurent Romary. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, 9-16. doi:10.14618/ids-pub-9021

R Rehuvrek, P. S. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45-50.

Ralph Weischedel, M. P., Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, Ann Houston. (2013). OntoNotes Release 5.0. *Linguistic Data Consortium*. doi:https://doi.org/10.35111/xmhb-2b84

Sorokin, A. (2022, 2022//). *Improving Morpheme Segmentation Using BERT Embeddings.* Paper presented at the Analysis of Images, Social Networks and Texts, Cham.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 638-647. doi:https://doi.org/10.1016/S0022-5371(75)80051-X

Tiedemann, P. L. a. J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora

from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation*. Retrieved from http://www.opensubtitles.org/

Tom Kocmi, R. B., Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović. (2022). Findings of the 2022 Conference on Machine Translation (WMT22). *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention is all you need*. Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.

Wikimedia.org. (2016). Wikimedia Dumps.

Appendix.

Below are resulting data from this investigation, with the corpus + metric pairing as the column header:

| Base Words | Complex Words | Base POS | Complex POS | Base Freq | Complex Freq | Rel Freq | log(Rel F) | spaCySM_COS | spaCyMd_COS | spaCyLg_COS | twitter_COS | google_COS | spaCySM_MIN | spaCyMd_MIN | spaCyLg_MIN | twitter_MIN | google_MIN | spaCySM_JAC | spaCyMd_JAC | spaCyLg_JAC | twitter_JAC | google_JAC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| couth | uncouth | ADJ | ADJ | 2 | 34 | 0.058823529 | -1.230448921 | 0.473711983 | 0.015359628 | 0.626523159 | 0.1237852 | 0.4953191 | 7.818833828 | 72.87664795 | 18.88362122 | 5.870859146 | 2.779042721 | 1 | 1 | 1 | 1 | 1 |
| mutable | immutable | ADJ | ADJ | 4 | 40 | 0.1 | -1 | 0.711687182 | 0.999999984 | 0.873206973 | 0.4635941 | 0.5291927 | 6.502598763 | 0 | 20.5240078 | 5.396452904 | 3.206757069 | 1 | 0 | 1 | 1 | 1 |
| animate | inanimate | ADJ | ADJ | 4 | 34 | 0.117647059 | -0.929418926 | 0.674140102 | 1.000000071 | 0.850202871 | 0.2201768 | 0.49537197 | 6.783192635 | 0 | 19.64858246 | 6.328373909 | 3.187576532 | 1 | 0 | 1 | 1 | 1 |
| scruff | scruffy | NOUN | ADJ | 7 | 42 | 0.166666667 | -0.77815125 | 0.305651089 | 0.571636989 | 0.825750182 | 0.7165928 | 0.34034416 | 9.306271553 | 35.36585617 | 11.04666328 | 3.963967323 | 4.297562599 | 1 | 1 | 1 | 1 | 0.996666667 |
| mobile | immobile | ADJ | ADJ | 11 | 55 | 0.2 | -0.698970004 | 0.636452974 | 0.999999942 | 0.649277032 | 0.15965669 | 0.14074576 | 6.693105698 | 0 | 34.47269821 | 6.924239159 | 3.916531563 | 1 | 0 | 1 | 1 | 0.996666667 |
| exact | exactly | VERB | ADV | 532 | 2535 | 0.209861933 | -0.678066331 | 0.245672666 | 0.725937286 | 0.725937286 | 0.75309664 | 0.596412 | 10.01514912 | 31.32122803 | 31.32122803 | 3.865584373 | 2.314585209 | 1 | 1 | 1 | 1 | 0.996666667 |
| canny | uncanny | ADJ | ADJ | 20 | 89 | 0.224719101 | -0.648360011 | 0.673959261 | 0.553700354 | 0.544599635 | 0.35243234 | 0.37957215 | 5.992546082 | 30.35396004 | 25.88967133 | 6.754140377 | 3.411348104 | 1 | 1 | 1 | 1 | 0.996666667 |
| leash | unleash | VERB | VERB | 16 | 54 | 0.296296296 | -0.528273777 | 0.774598144 | 0.075888741 | 0.224440298 | 0.24398857 | 0.13694629 | 5.686510563 | 55.68616104 | 44.32323837 | 6.523563385 | 4.480111122 | 1 | 1 | 1 | 1 | 1 |
| vamp | revamp | NOUN | NOUN | 4 | 13 | 0.307692308 | -0.511883361 | 0.481641842 | -0.070010258 | 0.145334598 | 0.06322495 | 0.029695999 | 7.739679337 | 49.5295639 | 32.80180359 | 6.714096546 | 4.317526817 | 1 | 1 | 1 | 1 | 1 |
| audible | inaudible | ADJ | ADJ | 100 | 292 | 0.342465753 | -0.465382851 | 0.616800179 | 0.438022093 | 0.850545097 | 0.4934836 | 0.40217108 | 7.515321732 | 51.62599564 | 22.08394051 | 5.111028671 | 3.371095657 | 1 | 1 | 1 | 1 | 1 |
| frequent | frequently | ADJ | ADV | 396 | 1036 | 0.382239382 | -0.417664569 | 0.159030689 | 0.789902855 | 0.789902855 | 0.5976107 | 0.6497709 | 11.69923401 | 23.36991119 | 23.36991119 | 4.763568401 | 2.152412891 | 1 | 1 | 1 | 1 | 1 |
| list | listless | NOUN | ADJ | 19 | 42 | 0.452380952 | -0.344495689 | 0.197095204 | 0.203777461 | 0.437809555 | -0.19776641 | -0.054619763 | 9.689188004 | 70.72164154 | 63.72740173 | 7.533518791 | 4.507861137 | 1 | 1 | 1 | 1 | 1 |
| mortal | immortal | ADJ | ADJ | 53 | 112 | 0.473214286 | -0.324942153 | 0.753360269 | 0.760026151 | 0.760026151 | 0.39211228 | 0.5565232 | 5.8262887 | 21.91371346 | 21.91371346 | 5.971897602 | 2.779873371 | 1 | 1 | 1 | 1 | 1 |
| patient | impatient | ADJ | ADJ | 114 | 227 | 0.502202643 | -0.299121006 | 0.836591291 | 1.000000045 | 0.745050651 | 0.4766648 | 0.14989424 | 4.366622448 | 0 | 32.25460434 | 6.019310474 | 3.794145584 | 1 | 0 | 1 | 1 | 1 |
| slime | slimy | NOUN | ADJ | 35 | 61 | 0.573770492 | -0.241261791 | 0.44990498 | 0.398588555 | 0.757844852 | 0.3384686 | 0.6539536 | 7.530131817 | 33.51640701 | 20.94417381 | 5.86784029 | 2.654242992 | 1 | 1 | 1 | 1 | 1 |
| hap | hapless | NOUN | ADJ | 13 | 22 | 0.590909091 | -0.228479329 | 0.421934881 | 0.025083708 | 0.166296783 | -0.012590471 | 0.20454262 | 7.870604992 | 60.64952087 | 24.25855064 | 6.543660641 | 3.200767756 | 1 | 1 | 1 | 1 | 1 |
| legible | illegible | ADJ | ADJ | 10 | 14 | 0.714285714 | -0.146128036 | 0.655937874 | 0.392646254 | 0.820389652 | 0.6094908 | 0.640687 | 6.696669102 | 36.3938179 | 20.07537842 | 4.225058079 | 3.157741308 | 1 | 1 | 1 | 1 | 0.993333333 |
| virile | virility | ADJ | NOUN | 31 | 41 | 0.756097561 | -0.121422163 | 0.32052624 | 0.156950483 | 0.36461354 | 0.27348202 | 0.5052413 | 8.628779411 | 45.71512604 | 29.43105316 | 7.679798126 | 3.592562914 | 1 | 1 | 1 | 1 | 1 |
| align | alignment | VERB | NOUN | 44 | 57 | 0.771929825 | -0.112422179 | 0.434639302 | 0.547905172 | 0.547905172 | 0.46956068 | 0.52094007 | 8.809389114 | 41.936409 | 41.936409 | 5.880772114 | 2.958320379 | 1 | 1 | 1 | 1 | 0.996666667 |
| diagonal | diagonally | ADJ | ADV | 29 | 36 | 0.805555556 | -0.093904503 | 0.250325703 | 0.669528981 | 0.805625797 | 0.40737808 | 0.72704506 | 10.79378033 | 31.72219467 | 20.6113281 | 6.368923187 | 2.551171064 | 1 | 1 | 1 | 1 | 0.996666667 |
| swift | swiftly | ADJ | ADV | 221 | 268 | 0.824626866 | -0.08374252 | 0.322392326 | 0.664117624 | 0.664117624 | 0.1430488 | 0.5054235 | 10.46981525 | 26.35445595 | 26.35445595 | 7.216631413 | 2.676677465 | 1 | 1 | 1 | 1 | 0.996666667 |
| equal | equally | ADJ | ADV | 1084 | 1303 | 0.831926324 | -0.079915134 | 0.167467131 | 0.55690192 | 0.55690192 | 0.6529767 | 0.31925362 | 11.59091949 | 45.42991638 | 45.42991638 | 4.576739311 | 2.962029219 | 1 | 1 | 1 | 1 | 1 |
| twine | entwine | NOUN | NOUN | 27 | 32 | 0.84375 | -0.073786214 | 0.525210949 | 0.415341336 | 0.802241728 | 0.2980164 | 0.17503107 | 7.740995407 | 47.2426796 | 21.15901947 | 5.587621212 | 4.549292088 | 1 | 1 | 1 | 1 | 1 |
| meek | meekly | ADJ | ADV | 41 | 47 | 0.872340426 | -0.059314001 | 0.495012559 | -0.077221752 | 0.326073036 | -0.06216051 | 0.5409096 | 7.53924942 | 52.07958984 | 28.15480804 | 7.977952003 | 2.907196999 | 1 | 1 | 1 | 1 | 0.996666667 |
| diligent | diligently | ADJ | ADV | 31 | 35 | 0.885714286 | -0.052706351 | 0.118492206 | 0.411962468 | 0.789815042 | 0.4792214 | 0.58334416 | 11.23854828 | 40.89390182 | 18.87889481 | 6.359957695 | 2.652206182 | 1 | 1 | 1 | 1 | 0.99 |
| recent | recently | ADJ | ADV | 1814 | 1676 | 1.082338902 | 0.034363268 | 0.316409741 | 0.694523808 | 0.694523808 | 0.66696435 | 0.52029103 | 11.1692028 | 37.93539047 | 37.93539047 | 4.477683067 | 2.055750847 | 1 | 1 | 1 | 1 | 0.993333333 |
| agile | agility | ADJ | NOUN | 38 | 34 | 1.117647059 | 0.04830468 | 0.277057729 | 0.417970716 | 0.549387767 | 0.513019 | 0.60552186 | 9.150637627 | 45.14403534 | 40.3048172 | 5.295179367 | 3.049808264 | 1 | 1 | 0.996666667 | 1 | 1 |
| direct | directly | VERB | ADV | 1472 | 1278 | 1.151799687 | 0.061376956 | 0.248962093 | 0.768704105 | 0.768704105 | 0.49159276 | 0.55068016 | 11.00771809 | 28.74817085 | 28.74817085 | 5.881885529 | 2.097105265 | 1 | 1 | 1 | 1 | 1 |
| adequate | inadequate | ADJ | ADJ | 540 | 399 | 1.353383459 | 0.131420864 | 0.572139483 | 0.85035818 | 0.85035818 | 0.6046255 | 0.6669251 | 8.780157089 | 20.17670441 | 20.17670441 | 4.899864197 | 2.315942287 | 1 | 1 | 1 | 1 | 1 |
| moral | immoral | ADJ | ADJ | 143 | 94 | 1.521276596 | 0.182208184 | 0.850880805 | 0.55696191 | 0.685027639 | 0.265902 | 0.41752362 | 4.467960358 | 45.74933243 | 40.31386566 | 6.890714645 | 3.357193708 | 1 | 1 | 1 | 1 | 0.996666667 |
| adorn | adornment | VERB | ADJ | 75 | 41 | 1.829268293 | 0.262277407 | 0.532993355 | 0.699422789 | 0.336968815 | 0.114089586 | 0.39548928 | 8.482800484 | 32.86373901 | 39.88184357 | 8.035006523 | 3.638370037 | 1 | 1 | 1 | 1 | 0.996666667 |
| general | generally | ADJ | ADV | 4624 | 1663 | 2.780517138 | 0.444125576 | 0.151666225 | 0.681065703 | 0.681065703 | 0.30626163 | 0.20057958 | 12.42634296 | 34.19142914 | 34.19142914 | 6.367324352 | 2.771512032 | 1 | 1 | 1 | 1 | 1 |
| affected | unaffected | ADJ | ADJ | 169 | 54 | 3.12962963 | 0.495492945 | 0.657261148 | 0.999999995 | 0.896112944 | 0.32393596 | 0.61385715 | 6.386886597 | 0 | 23.44345474 | 6.440442562 | 2.588334799 | 1 | 0 | 1 | 1 | 0.996666667 |
| puff | puffy | NOUN | ADJ | 159 | 48 | 3.3125 | 0.520155887 | 0.422849662 | 0.99999998 | 0.744467736 | 0.4369579 | 0.31916857 | 8.721008301 | 0 | 26.13147354 | 5.586453915 | 3.543226719 | 1 | 0 | 1 | 1 | 1 |
| soft | softly | ADJ | ADV | 1464 | 440 | 3.327272727 | 0.5220884 | 0.376974362 | 0.585929109 | 0.585929109 | 0.42656302 | 0.37807295 | 9.839198112 | 46.36457062 | 46.36457062 | 6.516010284 | 3.222293377 | 1 | 1 | 1 | 1 | 1 |
| kindle | rekindle | VERB | VERB | 41 | 11 | 3.727272727 | 0.571391172 | 0.714437097 | 0.129296598 | 0.551361037 | 0.07387713 | 0.64079165 | 6.166191578 | 50.82941437 | 27.09479141 | 8.253250122 | 2.776978254 | 1 | 1 | 1 | 1 | 1 |
| screw | unscrew | NOUN | VERB | 187 | 44 | 4.25 | 0.62838893 | 0.275425373 | 0.999999983 | 0.768262643 | 0.32713297 | 0.51421106 | 10.22406292 | 0 | 29.39017105 | 6.423710346 | 3.339606581 | 1 | 0 | 1 | 1 | 0.993333333 |
| woe | woeful | NOUN | ADJ | 68 | 14 | 4.857142857 | 0.686380877 | 0.275524197 | 0.040549769 | 0.18123938 | 0.15879224 | 0.39544463 | 9.256235123 | 101.6122437 | 37.58177567 | 5.518935204 | 3.409828424 | 1 | 1 | 1 | 1 | 0.996666667 |
| liberal | illiberal | ADJ | ADJ | 55 | 11 | 5 | 0.698970004 | 0.794424948 | 1.00000005 | 0.871113866 | 0.26176828 | 0.52847385 | 5.195296288 | 0 | 24.19586754 | 5.903927326 | 3.129012108 | 1 | 0 | 1 | 1 | 1 |
| kind | unkind | ADJ | ADJ | 390 | 72 | 5.416666667 | 0.733732111 | 0.616784411 | 0.348000011 | 0.547168573 | 0.37142575 | 0.1738415 | 7.587397099 | 52.24323654 | 43.49697495 | 6.885253906 | 3.284067631 | 1 | 1 | 1 | 1 | 1 |
| fragile | fragility | ADJ | NOUN | 207 | 36 | 5.75 | 0.759667845 | 0.519347999 | 0.381490678 | 0.610923117 | 0.23166524 | 0.6663977 | 7.518525124 | 37.80799103 | 27.36410713 | 6.670958996 | 2.558635235 | 1 | 1 | 1 | 1 | 0.996666667 |
| arrogant | arrogantly | ADJ | ADV | 116 | 17 | 6.823529412 | 0.834009068 | 0.313581507 | 0.999999962 | 0.782935935 | 0.35991865 | 0.56145155 | 10.47408867 | 0 | 17.29563332 | 6.123917103 | 2.68116045 | 1 | 0 | 1 | 1 | 0.996666667 |
| accurate | inaccurate | ADJ | ADJ | 377 | 53 | 7.113207547 | 0.852065481 | 0.750249836 | 0.81572196 | 0.815721966 | 0.65017635 | 0.56794995 | 5.527614594 | 25.32097626 | 25.32097626 | 4.84975481 | 2.956447601 | 1 | 1 | 1 | 1 | 0.996666667 |
| cream | creamy | NOUN | NOUN | 540 | 74 | 7.297297297 | 0.863162094 | 0.599306801 | 0.758661717 | 0.756862117 | 0.60005559 | 0.5562466 | 6.486506939 | 34.20965958 | 34.20965958 | 6.603127602 | 3.110988825 | 1 | 1 | 1 | 1 | 1 |
| eternal | eternally | ADJ | ADV | 355 | 38 | 9.342105263 | 0.970444756 | 0.276856295 | 0.304612496 | 0.715238433 | 0.48295608 | 0.529075 | 10.89361286 | 47.55542755 | 26.23868179 | 5.507172108 | 2.764129877 | 1 | 1 | 1 | 1 | 1 |
| taste | tasteless | VERB | ADJ | 402 | 30 | 13.4 | 1.127104798 | 0.402387941 | 1.00000004 | 0.733808198 | 0.3955234 | 0.30990288 | 8.631839752 | 0 | 35.64149094 | 6.116044044 | 3.649265528 | 1 | 0 | 1 | 1 | 1 |
| vulnerable | invulnerable | ADJ | ADJ | 400 | 23 | 17.39130435 | 1.240332155 | 0.811461572 | 1.000000009 | 0.888015207 | 0.25314298 | 0.46685767 | 5.276731968 | 0 | 19.37677956 | 6.656951427 | 3.44847703 | 1 | 0 | 1 | 1 | 0.996666667 |
| organize | reorganize | VERB | VERB | 1118 | 61 | 18.32786885 | 1.263111969 | 0.869652789 | 1.000000005 | 0.771968408 | 0.48857716 | 0.39803097 | 4.147880554 | 0 | 25.22073936 | 6.013104916 | 3.218162537 | 1 | 0 | 1 | 1 | 1 |
| entice | enticement | VERB | NOUN | 64 | 3 | 21.33333333 | 1.329058719 | 0.681762127 | 0.26792013 | 0.652533106 | 0.31264114 | 0.5677519 | 5.926108837 | 51.74614334 | 25.91586685 | 5.639873981 | 2.763174772 | 1 | 1 | 1 | 1 | 1 |
| perfect | imperfect | ADJ | ADJ | 1131 | 50 | 22.62 | 1.354492601 | 0.634852777 | 1.000000024 | 0.72893664 | 0.48157975 | 0.3998459 | 6.753213406 | 0 | 29.32304955 | 6.424180031 | 3.022644043 | 1 | 0 | 1 | 1 | 0.993333333 |
| practical | impractical | ADJ | ADJ | 1228 | 47 | 26.12765957 | 1.417100509 | 0.761258465 | 1.000000029 | 0.879780925 | 0.28355852 | 0.2977402 | 5.865210056 | 0 | 21.50602722 | 7.540131092 | 3.60287714 | 1 | 0 | 1 | 1 | 1 |
| common | uncommon | ADJ | ADJ | 3376 | 114 | 29.61403509 | 1.471497587 | 0.809960027 | 0.798199116 | 0.798199116 | 0.5582302 | 0.42494074 | 5.339118481 | 32.89084244 | 32.89084244 | 5.115360737 | 2.935920238 | 1 | 1 | 1 | 1 | 0.996666667 |
| modest | immodest | ADJ | ADJ | 521 | 13 | 40.07692308 | 1.602894371 | 0.59055727 | 0.999999927 | 0.584791733 | 0.1986642 | 0.2945947 | 7.766177654 | 0 | 27.20295715 | 5.993225098 | 3.681729794 | 1 | 0 | 1 | 1 | 1 |
| tool | retool | NOUN | VERB | 800 | 10 | 80 | 1.903089987 | 0.603875274 | 1.000000018 | 0.595746703 | 0.15780097 | 0.06716147 | 6.324433327 | 0 | 58.21119308 | 8.055106163 | 4.27487278 | 1 | 0 | 1 | 1 | 1 |
| top | topless | NOUN | ADJ | 3089 | 27 | 114.4074074 | 2.058454144 | 0.653141689 | 0.005902749 | 0.066670457 | 0.37071112 | 0.09676645 | 6.165837765 | 103.8796616 | 92.97896576 | 6.109016895 | 4.034013748 | 1 | 1 | 1 | 1 | 1 |