Stranger than Paradigms
Word Embedding Benchmarks Don't Align With Morphology

Timothee Mickus
University of Helsinki
timothee.mickus@helsinki.fi

Maria Copot
Ohio State University
copot.1@osu.edu

Abstract

Word embeddings have proven a boon in NLP in general, and computational approaches to morphology in particular. However, methods to assess the quality of a word embedding model only tangentially target morphological knowledge, which may lead to suboptimal model selection and biased conclusions in research that employs word embeddings to investigate morphology. In this paper, we empirically test this hypothesis by exhaustively evaluating 1,200 French models with varying hyperparameters on 14 different tasks. Models that perform well on morphology tasks tend to differ from those which succeed on more traditional benchmarks. An especially critical hyperparameter appears to be the negative sampling distribution smoothing exponent: Our study suggest that the common practice of setting it to 0.75 is not appropriate: its optimal value depends on the type of linguistic knowledge being tested.

1 Introduction

Word embeddings have changed the NLP landscape by introducing a data-driven approach to meaning. They have found widespread application in NLP, computational semantics, and more recently, morphology (e.g. Zeller et al., 2014; Bonami and Guzmán Naranjo, 2023).

While architectures specifically intended to capture morphology exist (Cao and Rei, 2016; Cotterell et al., 2016; Cotterell and Schütze, 2015), embeddings with these properties are generally not employed because not available off the shelf pretrained on the languages of interest to the morphologist. A notable exception is fastText (Bojanowski et al., 2017), an architecture specifically tailored to factor in spelling information which has been tested on a diverse and wide collection of languages (Grave et al., 2018). Despite claims that this architecture is suited to model morphology due to its attention to subword information, this has rarely been properly tested on morphological benchmarking. Additionally, this type of embedding is explicitly avoided by researchers who do not wish to smuggle in the assumption that the units of morphology are primarily based on formal contrasts, rather than on more abstract contrasts of meaning (as argued by e.g. Štekauer, 2014).

The adoption of word embeddings in morphological research has therefore largely targeted general purpose embeddings, with architectures that are not optimised for capturing morphological structure. However, the evaluation of these models mostly relies on tasks that were not built with morphology in mind. Common NLP benchmarks used by models for morphological purposes generally target semantics: To take a concrete example, Lenci et al. (2022) provide an exhaustive evaluation of distributional semantics models on a wide array of tasks. They study a spate of benchmarks targeting target semantics, such as synonymy detection, analogy solving, sentiment analysis and natural language inference; but only two of their tasks involve morphology: the analogy task (whose methodological and ethical limitations are well documented, e.g., Linzen, 2016; Bolukbasi et al., 2016); and POS-tagging (where some morphological knowledge may be of use, although it is not explicitly required). This trend may be in part ascribed to the Anglo-centric approach of most NLP research: English is a language with relatively scarce inflectional morphology, which therefore has received comparatively little interest from morphologists interested in the subject.

The tension between the increasingly widespread use of general-purpose word embeddings in morphology and their evaluation on non-morphological benchmarks begs the engineering question of how to adapt the knowledge the community has developed for English to other languages, in a way that encompasses morphological applications in addition to semantic

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 173-189.
Irvine, California, June 27-29, 2024

ones. In the present paper, we investigate whether a discrepancy exists between NLP evaluation methodologies and morphology applications of word embeddings. We define eight tasks, probing for both inflection and derivation, evaluating both the geometry of the vector space and its usability in downstream scenarios, and exhaustively compare the behavior of 1200 continuous bag-of-words negative sample embedding models (Mikolov et al., 2013, "CBOW-NS") on traditional NLP semantic benchmarks as well as our proposed morphology tasks. We find that optimal hyperparameter settings are task-specific, and that there is a tradeoff between performance on tasks targeting different kinds of linguistic knowledge. We also stress the importance of the negative sampling distribution smoothing exponent hyperparameter, which we find to have a crucial role in our experiments despite its lack of notoriety.

2 Related works

Systematic studies of word embeddings. Works attempting to exhaustively evaluate word embeddings abound. These studies often delineate their area of focus to a specific architecture, language or hyperparametrization. For instance, Vulić et al. (2020) extensively study BERT models across six languages and five tasks. On the other hand, Lenci et al. (2022) provide an exhaustive overview of multiple English embeddings, across a diverse array of tasks and hyperparameters. Ulčar et al. (2020) and Grave et al. (2018) both limit their studies to fast-Text embeddings and the analogy task, but cover 9 and 10 languages respectively.Lastly, especially relevant to our present inquiry is the work of Köhn (2015), who focuses on the (morpho-)syntactic features captured in a diverse array of embedding architectures for Basque, English, French, German, Hungarian, Polish, and Swedish.

Architectures that capture morphology. A significant focus of interest concerns the development of embedding architectures designed to specifically capture some aspects of morphology. Chief of these is the fastText model of Bojanowski et al. (2017), which supplements the skip-gram model of Mikolov et al. (2013) with subword information. Cao and Rei (2016) propose an unsupervised character-level method that ranks each segment by its context-predictive power to capture information about morphological boundaries as well as morphological features. Cotterell et al. (2016) introduce a semisupervised architecture trained on a combination of raw and morphologically annotated text, which creates embedding spaces where morphologically similar words cluster together. Cotterell and Schütze (2015) present a latent-variable Gaussian graphical model trained on an embedding set and a lexical resource to smooth an existing set of word embeddings in a way that encourages the encoding of morphology. With the exception of Bojanowski et al.'s (2017) fasttext, these models have not yet reached widespread adoption among morphologists—in part due to their restricted typological coverage, as exemplified by the challenges non-concatenative morphology poses for subword-centric approaches (e.g., Amrhein and Sennrich, 2021).

Word embeddings and morphology. Word embeddings are a somewhat recent adoption in the study of morphology. A short survey of the literature outlines three main use-cases for embeddings.

The first case involves using the features of trained embeddings as input to prediction tools, with the aim to create resources or investigate the morphological system. One such instance is Zeller et al. (2014) employ embeddings to validate the construction of a derivational lexicon. Straka and Straková (2017) details the use of embeddings as input features for tasks where morphology is relevant, such as lemmatization or tokenization. Bafna and Žabokrtský (2022) study how subword embeddings can be used for cross-lingual transfer between morphologically similar, diachronically related languages. Another related approach is that of Marelli and Baroni (2015), who propose to learn linear maps to model affixation.

Related but distinct from this approach, a second set of works use embeddings as tools for gathering quantitative evidence about morphology. A variety of topics have been covered: Lapesa et al. (2018) quantitatively assess the difference between eventive and non-eventive *-ment* formations in French; Guzmán Naranjo and Bonami (2021) rely on embeddings to discuss overabundance; Varvara et al. (2021) addresses the question of semantic transparency; Bonami and Guzmán Naranjo (2023) derive quantitative evidence in favor of a paradigmatic conception of derivation from embeddings.

The third case is the use of embeddings for the purposes of defining a morphologically coherent group of items by the properties of the position they occupy in the geometrical space—the analysis of neighborhoods thus constructed may be qualitative (e.g. Wauquier, 2020) or quantitative (e.g. Huyghe and Wauquier, 2020). Varvara (2017) uses distributional representations to quantitatively evaluate neighborhood contents, and how they differ for event nominalizations and verbal nouns, A related trend of research involves performing operations on embeddings directly to derive quantifiable data—e.g., to study the difference between inflection and derivation (Bonami and Paperno, 2018; Rosa and Žabokrtský, 2019) or the status of specific morphological processes (Mickus et al., 2019).

3 Methodology

We set out to answer the question of whether it is in fact problematic to evaluate models we use for morphology on tasks which chiefly target lexical semantics. We do so by evaluating the performance of the same model on a diverse range of tasks targeting different kinds of linguistic knowledge. Because of its rich morphology and availability of resources documenting morphological relations, we elect to work on French. We wish to make as few assumptions as possible about whether we expect any systematic differences in performance between tasks and about what they might be caused by should they manifest—we adopt a grid-search approach and evaluate models trained with an exhaustive combination of values for a wide range of hyperparameters.

Models We train Continuous Bag-Of-Word negative sample models (Mikolov et al., 2013, CBOW-NS). Models are implemented with gensim (Řehůřek and Sojka, 2010), trained on a 300M French sentences subset of Oscar (Ortiz Suárez et al., 2019) We include a presentation of the word2vec algorithm and a few remarks on the linguistic significance of its hyperparameters in Appendix A.

Models defined with varying hyperparameters:

(i) window size $w \in \{5, 10, 15, 20, 25\}$;

(ii) number of negative examples per positive example $N \in \{5, 10, 15, 20, 25\}$;

(iii) number of epochs $e \in \{1, 3, 5\}$;

(iv) negative sampling distribution exponent $\alpha \in \{-1.4, -1.0, -0.6, -0.2, 0.2, 0.6, 1.0, 1.4\}$;

(v) dynamic uniform sampling of window size $s \in \{\text{True}, \text{False}\}$.

All models have a dimension of $d = 50$, which we do not modify so as to avoid spurious concentration effects in higher-dimensional spaces.[1] All combination of hyperparameters are tested, for a total of 1200 different models. As hyperparameters (i), (ii) and (iii) are frequently encountered in the literature, we refer the reader to the original paper by Mikolov et al. (2013) as well as to Appendix A.2 for details.

The negative sampling smoothing hyperparameter α in (iv) is not frequently tuned, but Caselles-Dupré et al. (2018) suggest it might have application-specific relevance. It is used to define the probability distribution q under which negative examples are randomly sampled:

$$q(w) \propto p(w)^{\alpha}$$

where $p(w)$ is the relative frequency of each word in the training corpus. Mikolov et al. (2013) note that α allows one to mix unigram and uniform distributions over vocabulary items: Setting a value closer to 0 allows one to sample more from the tail of the vocabulary's frequency distribution. More precisely, remark that $\alpha = 0$ entails sampling negative examples uniformly over the entire vocabulary sorted by frequency; $\alpha = 1$ matches the unigram frequency distribution in corpus; $\alpha > 1$ overemphasizes frequent words, and $\alpha < 0$ overemphasises infrequent ones. The relative dearth of studies on the effects of α on CBOW-NS representations to this day motivates us to be particularly thorough when testing this hyperparameter.

The dynamic uniform sampling s in (v) is a gensim-specific re-implementation of the distance-based weighting of context words. It consists in randomly sampling, for each training example, an effective window size $\hat{w}$ uniformly between 1 and the maximum window size parameter allowed by the w hyperparameter, or more formally $\hat{w} \sim U(1, w)$. In practice, this entails that context words that are $k \leq w$ tokens apart from the target word are discarded in k/w of the training instances. Therefore, context words that are closer to the target word are more likely to be taken into account for prediction.

Common NLP benchmarks. All models are tested on the SimLex-999 French translation by Barzegar et al. (2018), the FEEL lexicon of Abdaoui et al. (2017), the automatic translation to French of the Google Analogy Test Set (GATS) provided by Grave et al. (2018), and a POS-tagging

[1] This low value of d mitigates the computational costs of running exhaustive experiments. For the same reason, models varying across epoch e only correspond to different checkpoints of the same training procedure.

task. For GATS results, we separately keep track of the accuracy on three groups of analogical relations: semantic, derivational and inflectional;[2] groups can be found in Appendix B.1. The POS-tagging data was selected from the French section of OMW (Bond and Paik, 2012), by selecting, for each lemma, all POS-tags it could correspond to.

Results for PoS-tagging and FEEL correspond to macro-F1 scores of multi-layer perceptrons[3] trained to predict the labels provided in the resource as binary vectors. Performance on SimLex-999 is evaluated as the Spearman correlation between human rating and cosine similarity. GATS performance corresponds to a 3CosAdd on a vocabulary restricted to the 300k most frequent words.

Inflectional tasks. To test a model's performance on inflectional morphology specifically, we set up four different tasks. Given that the community uses embeddings both as features (predictors to plug into another models, e.g., Straka and Straková, 2017) and as representations for manual exploration (e.g. Wauquier, 2020), we consider both classifier-based tasks and geometric evaluations. A second orthogonal distinction is whether these probing tasks involve one word form or multiple word forms at once: this, in essence, captures distinct approaches to morphology, depending on whether they focus on individual words or relationships between them. Data for all four of these tasks consisted of verbal paradigms taken from the GLàFF (Hathout et al., 2014), a large inflected lexicon of French. The data set used focused on words without homographs, and cells that are in current use in the French language. Only words that had more than 50 occurrences in our Oscar sample were included in the testing; cf. also Appendix B.2.

The first task involves a classifier over singular items: In our single cell prediction (SCP) task, we classify input verb forms depending on which paradigm cell they correspond to. The second task, a paired cells prediction (PCP) task, consists in predicting whether two input verb forms correspond to the same paradigm cell. We compare models on these two predictions tasks using macro-F1. In our third task, a single cell clustering (SCC) task, we assess with silhouette scores whether the embeddings of forms cluster according to their cell. Lastly, in our fourth task, a paired cell clustering (PCC) task, we report the silhouette score obtained by clustering pairs of forms depending on which relation they instantiate. For this last PCC task, we define pairs of verb forms as matrices of shape $[2 \times d]$, distance between two pairs P_A and P_B is then computed using the Froebenius norm $\|P_A - P_B\|_F$.

Derivational tasks. To evaluate how accurately models capture derivational morphology, we set up four tasks. Data for these tasks was taken from Namer et al. (2023), a database of French derivational relationships. They feature a variety of relationships between different parts of speech, reported in Appendix B.2. Semantics labels are attributed by grouping formal exponents in the resource following the clustering proposed by Guzman Naranjo and Bonami (2023). Only words that had more than 50 occurrences in our Oscar sample were kept; cf. Appendix B.2 for details.

Following the same logic as for our inflectional task, we consider two prediction tasks and two classification tasks. Independently from this, we also note that there is ongoing discussion in the theoretical morphology community about whether derivational relations should be defined by means of formal or semantic regularity (Štekauer, 2014). We therefore decide to consider as labels either the formal exponent of the derived form (e.g. *-ité*), or the semantics associated with it (e.g., adjective-to-property-noun conversion). The two predictions tasks are set up as simple logistic regression classifiers that predict the derivational cell (defined based on semantics vs formal exponents for DerPS and DerPF respectively); we report the corresponding macro-F1 scores. The two clustering tasks reemploy the same protocol as the PCC inflectional task: we construct $[2 \times d]$ matrices for each derivationally related pair in our dataset, and compute the silhouette score for clustering them along their exponents in the DerCF task or the semantics of the process in the DerCS task.

4 Results

Given the high number of models and tasks, we first study how and to what extent specific hyperparameters shape performance; we defer an overview of actual performances to Appendix C.2 to focus primarily on global trends. To attribute the observed variance across scores to specific factors, we apply gradient boosting trees (Friedman, 2001) to the set of all (task-standardized) scores, using as predic-

[2] The latter two are often grouped in a "syntactic" category; here we follow the taxonomy of Gladkova et al. (2016).

[3] One 25D layer with ReLU activation, optimized with Adam (lr. of 0.001, $\beta = (0.9, 0.999)$) for up to 10,000 iterations, implemented in `scikit-learn` (Pedregosa et al., 2011).

α	task	e	N	w	s
0.69	0.23	0.18	0.08	0.08	0

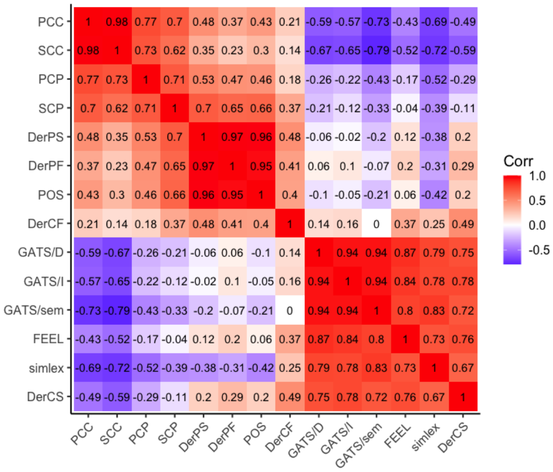
Table 1: mean absolute SHAP value from boosting trees predicting performance.

Task	**top 1**	**top 10**		**top 100**	
		mean	**median**	**mean**	**median**
simlex	0.20	0.20	0.20	0.20	0.20
FEEL	0.20	0.24	0.20	0.36	0.20
GATS/sem	0.20	0.20	0.20	0.09	0.20
GATS/D	0.20	0.20	0.20	0.24	0.20
DerCF	0.60	0.76	0.60	0.70	0.60
DerCS	0.20	0.24	0.20	0.35	0.20
DerPF	0.60	0.64	0.60	0.73	0.60
DerPS	0.60	0.80	0.80	0.80	1.00
POS	0.60	0.88	1.00	0.84	1.00
GATS/I	0.20	0.28	0.20	0.36	0.20
SCP	1.40	1.40	1.40	1.20	1.40
PCP	1.40	1.40	1.40	1.32	1.40
SCC	1.00	1.24	1.40	1.06	1.00
PCC	1.00	0.88	1.00	0.86	1.00

Table 2: The mean and median α of the top 1, top 10 and top 100 performing models for each task.

tors the hyperparameters as well as the task, before computing SHAP values (Lundberg and Lee, 2017). Corresponding results can be seen in Table 1: The model had a residual mean standard error (RMSE) of 0.17 on the test set (one third of the data). Remarkably, the most important predictor was found to be the negative sampling distribution smoothing exponent α, with a mean absolute SHAP value of 0.69. Across most tasks, we find that values of α tend to produce natural clusters of model scores (see Figure 2 in Appendix C).

A summary of the distribution of performance for α values by task is reported in Table 2. We observe that common NLP benchmarks (FEEL, simlex, as well as all categories of analogies in GATS), appear to benefit from an α value of 0.2, while the tasks we devised to target inflectional morphology fare best with $\alpha \geq 1$. Derivational tasks lie somewhere in between: in DerCS, where processes are grouped by their semantics pattern close to semantic tasks, the optimal α is slightly higher than 0.2; in the three other tasks, optimal α values range from 0.6 to 0.8. POS appears to perform best with values in between those of inflection and derivation, with α slightly lower than 1. Data from GATS does not pattern as expected given the analogical relation type.



	PCC	SCC	PCP	SCP	DerPS	DerPF	POS	DerCF	GATS/D	GATS/I	GATS/sem	FEEL	simlex	DerCS
PCC	1	0.98	0.77	0.7	0.48	0.37	0.43	0.21	-0.59	-0.57	-0.73	-0.43	-0.69	-0.49
SCC	0.98	1	0.73	0.62	0.35	0.23	0.3	0.14	-0.67	-0.65	-0.79	-0.52	-0.72	-0.59
PCP	0.77	0.73	1	0.71	0.53	0.47	0.46	0.18	-0.26	-0.22	-0.43	-0.17	-0.52	-0.29
SCP	0.7	0.62	0.71	1	0.7	0.65	0.66	0.37	-0.21	-0.12	-0.33	-0.04	-0.39	-0.11
DerPS	0.48	0.35	0.53	0.7	1	0.97	0.96	0.48	-0.06	-0.02	-0.2	0.12	-0.38	0.2
DerPF	0.37	0.23	0.47	0.65	0.97	1	0.95	0.41	0.06	0.1	-0.07	0.2	-0.31	0.29
POS	0.43	0.3	0.46	0.66	0.96	0.95	1	0.4	-0.1	-0.05	-0.21	0.06	-0.42	0.2
DerCF	0.21	0.14	0.18	0.37	0.48	0.41	0.4	1	0.14	0.16	0	0.37	0.25	0.49
GATS/D	-0.59	-0.67	-0.26	-0.21	-0.06	0.06	-0.1	0.14	1	0.94	0.94	0.87	0.79	0.75
GATS/I	-0.57	-0.65	-0.22	-0.12	-0.02	0.1	-0.05	0.16	0.94	1	0.94	0.84	0.78	0.78
GATS/sem	-0.73	-0.79	-0.43	-0.33	-0.2	-0.07	-0.21	0	0.94	0.94	1	0.8	0.83	0.72
FEEL	-0.43	-0.52	-0.17	-0.04	0.12	0.2	0.06	0.37	0.87	0.84	0.8	1	0.73	0.76
simlex	-0.69	-0.72	-0.52	-0.39	-0.38	-0.31	-0.42	0.25	0.79	0.78	0.83	0.73	1	0.67
DerCS	-0.49	-0.59	-0.29	-0.11	0.2	0.29	0.2	0.49	0.75	0.78	0.72	0.76	0.67	1

Figure 1: Spearman correlation for performance of models with $\alpha > 0$ on the different tasks.

We can further observe that values of $\alpha < 0$ tend to yield lower scores: Mann-Whitney U-tests indicate that with the sole exception of GATS/sem,[4] scores for positive values of α are significantly greater ($p < 10^{-30}$, common language effect size: $0.6902 \leq f \leq 0.9993$). Training boosting trees only on models with $\alpha > 0$ degrades the fit (RMSE $= 0.26$ on the test set) but also redistributes the importance of the predictors, with tasks playing a dominant role (0.44 mean absolute SHAP) and α remaining a close second (0.37 mean absolute SHAP).

These different optimal settings strongly suggest that models that fare well on traditional NLP tasks likely do not dominate on morphology tasks. To establish whether this expectation is borne out, we compute the correlation of scores for each pair of tasks. Given the low scores for negative values of α we have established, we restrict our observations to values of $\alpha > 0$; we refer the reader to Appendix C.4 for related results across all models. Results are shown in Figure 1. We observe that NLP benchmarks (all subsets of GATS, SimLex-999, FEEL) correlate well with each other, but less well with the morphological tasks we devised (SCC, PCC, SCP, PCP for inflection; DerCF, DerCS, DerPF and DerPS for derivation), where correla-

[4] In fact, for GATS/sem, we find the opposite trend: Higher values of α lead to lower scores ($p < 10^{-5}$, $f = 0.4266$). This is due to the fact that the highest values of α lead to even greater decreases in performance than the lowest values of α. Values of $\alpha \in \{-0.2, 0.2\}$ yield the highest scores.

tion is lower and occasionally even negative. With the exception of DerCS, derivational tasks pattern in the middle, being highly correlated with each other, and having middling levels of correlation with both inflectional and semantic tasks.

5 Discussion

Two types of distributional information. The results we observe in Section 4 suggest that models exhibit a range of behaviours between two poles, defined by whether the task is testing knowledge of semantics or of morphosyntactic properties. Performances on inflectional morphology on the tasks we devised were uncorrelated or even anticorrelated to SimLex-999 and GATS/sem results, the tasks that targets lexical semantics in the most narrow sense in our set. Why is that we observe such extreme trade-offs—where better performances on semantic similarity entail lower performances on inflection, with derivation and POS-tagging patterning in the middle? One possible answer lies in the theoretical framework underpinning static word embeddings such as word2vec, i.e., distributional semantics.

As Sahlgren (2008) and Gastaldi (2021) outline, the distributional semantics framework of Harris (1954) has historical ties with linguistic structuralism, through the works of Bloomfield (Bloomfield, 1933) and indirectly those of de Saussure (de Saussure, 1916). If we consider the objective function of neural embeddings such as word2vec, we see that these models broadly attempt to predict a target word given its context: Embeddings attempt to capture a conditional probability $p(t|c)$ of targets t given their context of occurrence c.[5] This is the hallmark of a "paradigmatic model", as Sahlgren (2008) puts it: In short, these models are trained to guess which word might appear in a given context. To hearken back to linguistic structuralism, we can say these models attempt to fill in a given paradigmatic slot in a syntagmatic context, or that they try to establish associative series—which can involve either formal relations or semantic relations.

From a distributionalist point of view, contexts of occurrence constrain words in two different manners: through morphosyntactic dependencies and through lexical semantic requirements. In fact, these two different types of constraints are obvious if we compare the following examples:

(i) You know, this is the way we eat in _____.

(ii) I think this game is really _____.

One can easily surmise that the blanked word in Example (i) has to refer to a place: In other words, the distributional cues around this gap constrain the lexical semantics of words that can fit this specific context. On the other hand, Example (ii) leaves the semantics very much unconstrained, but requires specific morphosyntactic features—valid inserts range from "easy" to "stupid" to "dark" but their validity hinges on their adjectival nature.

If we now return to our embeddings evaluation, we can observe that the two different types of distributional constraints entail that it is logically possible that some tasks may show uncorrelated behavior, as they measure a model's ability to capture one or the other of these types of constraints. Success on our inflectional tasks requires a proper modeling of morphosyntactic cues, whereas success on SimLex-999 requires a proper modeling of lexical semantics. These two sets of tasks are extreme positions in a trade-off situation: for SimLex-999, morphosyntactic cues are irrelevant; likewise for our inflectional tasks, capturing lexical semantics is much less important—and may actually be detrimental to performance. That these two sets of tasks correspond to extreme positions does suggest that most distributional representation evaluations tasks can be classified along two continuums, depending on the extent to which they probe lexical semantics and morphosyntactic modeling. These two aspects are not orthogonal, but it is nevertheless useful to consider them as distinct—especially given the intermediate position of derivational tasks and POS tagging, as shown by their optimal α values (Appendix C.1) and correlation patterns (Figure 1).

Derivation in the middle. Derivational tasks inherently rely on a combination of morphosyntactic and lexical semantics knowledge: French deverbal nouns in *-eur* can denote human professions (*recruteur* 'recruiter'), properties of human agents (*fumeur*, 'smoker') or inanimate instruments (*compteur* 'counter'), among others. Properly handling *-eur* forms requires that models capture on the one hand the morphosyntactic regularities surrounding agent or instrument nouns (e.g. often preceded by an article, often within short distance of a transitive verb), and on the other hand the different possible relationships of lexical semantics between a verbal base and its noun in *-eur* (agent, instrument etc).

[5] In practice, word2vec models can involve the related probabilities $p(c|t)$ (for skip-gram models) or $p(t \in c)$ (for negative sampling models). Both of these can be related to the probability of interest through renormalization or Bayes' rule.

Further strengthening our analysis of distributional constraints as morphosyntactic or semantic, we find that POS tagging, a task that is inherently about capturing morphosyntactic relationships, but which abstracts over individual relationships in order to uncover regularities of a different nature, patterns in between inflection and derivation.

Morphological clustering tasks. An important point to stress is that the clustering tasks always return negative average silhouette scores. In other words, on average, any datapoint in the SCC, PCC, DerCF and DerCS tasks could be better assigned to some other cluster. This would suggest that morphosyntactic contrasts do not shape the vector space landscape in an intuitive, meaningful way, even when the model has optimal hyperparameters for the task. This is perhaps because paradigm cells, despite their foundational role in morphological theory, need not describe a coherent group of usages: despite both being plural nouns, *chairs* refers to more than one instance of CHAIR while *scissors* refers to a singular object. This is an extreme example of a state of affairs that plagues the concept of paradigm cell. Our tasks are also defined on imbalanced classes, which intuitively makes the tasks at hand more challenging. Furthermore, the performances we observe on prediction tasks (SCP, PCP, DerPF, DerPS) are clearly above random chance or majority label heuristics:[6] This again confirms that morphosyntactic cues are properly encoded in suitably hyperparametrized models, suggesting that poor performance of clustering tasks is a consequence of the geometry of vector spaces being defined by permeable boundaries between paradigm cells rather than a result of models failing to capture existing patterns.

This fact also explains the behavior of the DerCS task: while we can expect the information necessary for solving the task to be present in models that capture morphosyntactic features (as evidenced by DerPS), the layout of the space makes this information hard to retrieve by clustering means. In addition, the use of semantic labels also entails that the derivational relation we selected have lexical semantic correlates, which can be exploited to perform well on the task. DerCS performance would then be reliant on the same cues as semantics tasks, explaining why it unexpectedly patterns with them.

The deal with GATS. Analogy solving is another case where prior assumptions are not borne out by our experiments. GATS/I and GATS/D should in principle pattern with inflectional and derivational tasks respectively—however, all GATS tasks behave more in line with semantic tasks.

One possible source of this unexpected result is the frequency of the words employed in GATS. GATS contains only fairly frequent lexemes, which are more likely to have more senses, and irregular semantic and morphological relationships to their base (Patterson et al., 2001; Baayen and del Prado Martín, 2005; Wu et al., 2019)—all of which place GATS/I further along on the lexical semantics gradient than our tasks, which contain words from all parts of the frequency gradient. GATS morphological analogies do however occupy a median position: results on I and D analogies are not as unrelated to morphological tasks as results for the semantic-type analogies.

It is also worth noting that one can trivially obtain high results on morphological analogies through linear offset methods without having to encode morphosyntactic features. If vectors only track lexical semantic distributional constraints, then we can expect two inflected forms of a given lemma to have roughly equivalent embeddings. In such a scenario, morphology-based analogies like *danse*:*dansait*::*mange*:*mangeait* would entail that $\vec{danse} - \vec{dansait} \approx \vec{mange} - \vec{mangeait} \approx \vec{0}$, and therefore solving these analogies through linear offsets would devolve into a trivial solution, e.g.:

$$\begin{aligned} \vec{x} &= \vec{danse} - \vec{dansait} + \vec{mangeait} \\ &\approx \vec{0} \qquad\qquad\qquad\quad + \vec{mangeait} \quad \approx \vec{mange} \end{aligned}$$

In other words, it is in principle possible for models that do not encode morphological traits in any relevant way (i.e., that only consider lexical semantic distributional constraints) to succeed on this supposedly morphological benchmark. Linzen (2016) raises similar concerns and stresses that cues often lie close to one another in word2vec space, which is only one of the major points for which the analogy task has been criticized (e.g. Rogers et al., 2017; Schluter, 2018; Garg et al., 2018).

Why α? This gradient take on distributional benchmarking tasks also explains why shaping the negative sampling distribution is found to be so impactful. If what is needed to succeed on inflectional tasks is a good representation of the morphological contrasts instantiated by the language of interest,

[6] Macro-F1 for majority baselines: SCP: 0.019; PCP: 0.370 DerPF 0.006; DerPS: 0.051.

negative evidence for learning these contrasts can be easily found at the very top of the vocabulary's frequency list: Contrasting the word of interest with the full paradigm of a handful of frequent lexemes in the language would get one most of the way to a working representation of morphological contrasts. Such extreme selection based on frequency is not suited for semantic tasks, which benefit from having a wider variety of negative examples and thus prefer lower values of the exponent compared to more purely morphological tasks.

To take a concrete example, consider the word *is*. This word is highly frequent, and an exceptionally poor disambiguator of aptitude to continue a particular sentence: *is* can be used to express any property intrinsic to the subject or circumstantial (*she is good* vs *he is here*), to imply existence (*she thinks therefore she is*), as an auxiliary to convey the tense, aspect and mood of another verb (*he is going out*, *she is to go there tomorrow*). Because of its wide variety of uses, *is* may take any noun as its subject or object, it may be modified by several adjectives and adverbs, and may be found in a wide variety of grammatical constructions. The sheer frequency of the verb exacerbates this feature of its usage. The distributional representation of *is* will therefore collapse all of these uses into the same representation, leading to a word embedding which is itself not necessarily helpful in pinning down the meaning and usage of *is*, but which is a good representation of which cues are not particularly informative about a word's meaning, since they may co-occur with many outcomes.

Hence we expect word frequency to be an accurate correlate of words that are poor disambiguators: Not only do frequent words by definition occur in a large amount of linguistic contexts, they also tend to have more senses (Zipf, 1942) and to occur in more varied contexts (Dennis and Humphreys, 2001). It is therefore unsurprising that disproportionately taking frequent words as negative examples is helpful for morphological tasks: because of the variety of contexts they occur in, they are going to be particularly useful in warding off unwarranted associations that are not important for creating a representation of the target word.

Furthermore, frequent words are more likely to have irregular morphology, while infrequent words are much more likely to behave regularly (Wu et al., 2019). While both regular and irregular words may be frequent, it is very rare to find infrequent irregular words: If a word does not follow regular patterns, this information must be explicitly encoded in the mental lexicon, which is only possible if the word is frequent enough to have a sufficiently strong mental representation.[7] Calling morphological behavior "regular" amounts to saying that the morphological pattern applies to many words, most of which will be infrequent. Conversely, one expects irregular patterns to apply to a few frequent words (Beniamine, 2018)—i.e., frequent words have more varied behavior than infrequent words. Hence, in order for a model to learn morphology, it must focus on frequent words, which are the locus of the greatest variety of patterns in the system.

This hypothesis correctly predicts that tasks in which knowledge of morphosyntactic information about specific words is being targeted will benefit from the highest values of α: in our case, inflectional tasks, closely followed by POS tagging (which targets more abstract morphosyntactic properties that aggregate over larger groups of words), followed by derivational tasks (which target morphosyntactic and lexical semantics information simultaneously) and lastly by those targeting lexical semantics alone (SimLex-999, FEEL, GATS/Sem). It also predicts that while tasks targeting lexical semantics might benefit from lower values of α, no linguistic task will benefit from oversampling from the tail of the vocabulary with $\alpha \leq 0$.

6 Conclusions

In this paper, we showed that the performance of static embeddings on morphological tasks need not correlate with their performance on lexical semantic tasks, which constitute most major NLP benchmarks. Morphological tasks can be shown to benefit from different hyperparameters than semantic tasks; optimal settings for derivational and inflectional processes also differ.

This is all the more crucial in theoretical morphology approaches aiming to use distributional representations as meaning proxies: our findings highlight that the exact hyperparametrization can affect the outcome we observe. Choosing hyperparameters is not theoretically neutral, and different conclusions may emerge from different settings. In particular, works in theoretical morphology that rely on embeddings to compare derivation and inflection (e.g. Bonami and Paperno, 2018; Rosa and

[7] Work on language change supports this statement: Words taking irregular patterns disappear from the language, or regularize, unless they are frequent enough to have their irregularity memorized (e.g. Lieberman et al., 2007).

Žabokrtský, 2019) are at risk of reporting conclusions biased in favor of inflection or derivation, depending on the exact hyperparametrization of their embeddings.

This methodological point ties in to another contribution of this work, namely that we experimentally underscore that distributional representations are not purely lexical semantic representations, but also incorporate morphosyntactic features. This contrasts with the often held position that distributional models are to be construed as meaning representations (e.g. Schütze, 1992; Lenci, 2018; Boleda, 2020; Apidianaki, 2023). The historical structuralist roots of distributionalism highlighted by Sahlgren (2008) and Gastaldi (2021) are especially useful to understand the limits inherent to this position.

Beyond theoretical remarks, this work also offers perspectives for other applications of distributional models: Applications of (contextualized) embedding architectures to morphology may have interest in manipulating the frequency of the examples shown to the model. In particular, modeling inflection benefits from paying close attention to the head of the unigram distribution of words in a corpus: We plan to explore whether sampling from different smoothed vocabulary distributions also helps models such as BERT (Devlin et al., 2019) to capture inflectional patterns more accurately.

In all, the growing number of applications of NLP to morphology makes it imperative that we think more carefully about the data and tasks we use for evaluation. Research attempting to construct tools for morphology and morphologically rich languages might be hindered by the Anglo-centric approach prevalent in NLP. Here, we have demonstrated for French CBOWs that the common practice of setting the α hyperparameter to 0.75 following Mikolov et al. (2013) is in fact inappropriate—not only for morphology modeling but also for classical NLP benchmarks. This is all the more concerning given that French is a well-documented, resource-rich language with a vibrant NLP research community, and begs the question of how inappropriate are Anglo-centric choices for typologically more distinct languages.

Acknowledgements

We thank Olivier Bonami, Denis Paperno, and the multiple anonymous reviewers for comments on this work that substantially bettered it, as well as James Robert Jarmusch for his creative input.

This work is part of the FoTran project, funded by the European Research Council (ERC) under the EU's Horizon 2020 research and innovation program (agreement № 771113). We also thank the CSC-IT Center for Science Ltd., for computational resources and NVIDIA AI Technology Center (NVAITC) for the expertise in distributed training.

References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.

Chantal Amrhein and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marianna Apidianaki. 2023. From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation. *Computational Linguistics*, pages 1–59.

R Harald Baayen and Fermín Moscoso del Prado Martín. 2005. Semantic density and past-tense formation in three germanic languages. *Language*, pages 666–698.

Niyati Bafna and Zdeněk Žabokrtský. 2022. Subword-based cross-lingual transfer of embeddings from Hindi to Marathi and Nepali. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 61–71, Seattle, Washington. Association for Computational Linguistics.

Siamak Barzegar, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and Andre Freitas. 2018. SemR-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sacha Beniamine. 2018. *Typologie quantitative des systèmes de classes flexionnelles*. Ph.D. thesis, Université Paris Diderot.

Leonard Bloomfield. 1933. *Language*. Henry Holt.

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

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Olivier Bonami and Matías Guzmán Naranjo. 2023. *Distributional evidence for derivational paradigms*, pages 219–258. De Gruyter, Berlin, Boston.

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio, Rivista semestrale*, (2/2018):173–196.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. pages 64–71.

Kris Cao and Marek Rei. 2016. A joint model for word embedding and word morphology. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 18–26, Berlin, Germany. Association for Computational Linguistics.

Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 352–356, New York, NY, USA. Association for Computing Machinery.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660, Berlin, Germany. Association for Computational Linguistics.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.

Simon Dennis and Glyn W Humphreys. 2001. A context noise model of episodic word recognition. *Psychological Review*, 108(2):452–478.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Juan Luis Gastaldi. 2021. Why can computers understand natural language? *Philosophy & Technology*, 34(1):149–214.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.

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).

Matías Guzmán Naranjo and Olivier Bonami. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa*, 6.

Matías Guzman Naranjo and Olivier Bonami. 2023. Distributional assessment of derivational semantics. Presented at the 53rd Annual Meeting of the Societas Linguistica Europaea. Bucharest, Romania.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Nabil Hathout, Franck Sajous, and Basilio Calderone. 2014. GLÀFF, a large versatile French lexicon. In *Proceedings of LREC 2014*.

Richard Huyghe and Marine Wauquier. 2020. What's in an agent? *Morphology*, 30:185–218.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Gabriella Lapesa, Lea Kawaletz, Ingo Plag, Marios Andreou, Max Kisselew, and Sebastian Padó. 2018. Disambiguation of newly derived nominalizations in context: A distributional semantics approach. *Word Structure*, 11(3):277–312.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Lang. Resour. Eval.*, 56(4):1269–1313.

Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychol Rev*. PMID: 26120909.

Timothee Mickus, Olivier Bonami, and Denis Paperno. 2019. Distributional effects of gender contrasts across categories. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 174–184.

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

Fiammetta Namer, Nabil Hathout, Olivier Bonami, Georgette Dal, Dany Amiot, Lucie Barque, Gilles Boyé, Stéphanie Caët, Basilio Calderone, Christine Da Silva Genest, Alexander Delaporte, Guillaume Duboisdindien, Achille Falaise, Natalia Grabar, Pauline Haas, Frédérique Henry, Mathilde Huguin, Nyoman Juniarta, Loïc Liégeois, Stéphanie Lignon, Lucie Macchi, Grigoriy Manucharian, Caroline Masson, Fabio Montermini, Nadejda Okinina, Alexndre Roulois, Franck Sajous, Daniele Sanacore, Mai Thi Tran, Juliette Thuilier, Yannick Toussaint, Delphine Tribout, and Marine Wauquier. 2023. Demonette-v2. April 2, 2023.

Pedro Javier Ortiz Suárez, Benoît Sagot, and 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) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Karalyn Patterson, Matthew A Lambon Ralph, John R Hodges, and James L McClelland. 2001. Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. *Neuropsychologia*, 39(7):709–724.

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

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. `http://is.muni.cz/publication/884893/en`.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

Rudolf Rosa and Zdeněk Žabokrtský. 2019. Attempting to separate inflection and derivation using vector space representations. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 61–70, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Magnus Sahlgren. 2008. The distributional hypothesis. *The Italian Journal of Linguistics*, 20:33–54.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, New Orleans, Louisiana. Association for Computational Linguistics.

Hinrich Schütze. 1992. Word space. In *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann.

Pavol Štekauer. 2014. *Štekauer, P. 2014. 'Derivational paradigms.' In: Lieber, R. – Štekauer, P. (eds.) The Oxford Handbook of Derivational Morphology. Oxford: Oxford University Press, 354-369*, pages 354–369.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Rossella Varvara. 2017. *Verbs as nouns: empirical investigations on event-denoting nominalizations*. Ph.D. thesis.

Rossella Varvara, Gabriella Lapesa, and Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology*, 31(4):409–446.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Marine Wauquier. 2020. *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. Ph.D. thesis. Thèse de doctorat dirigée par Hathout, Nabil Sciences du langage Toulouse 2 2020.

Shijie Wu, Ryan Cotterell, and Timothy J. O'Donnell. 2019. Morphological irregularity correlates with frequency. *CoRR*, abs/1906.11483.

Britta Zeller, Sebastian Padó, and Jan Šnajder. 2014. Towards semantic validation of a derivational lexicon. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1728–1739, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

George Kingsley Zipf. 1942. The unity of nature, least-action, and natural social science. *Sociometry*, 5(1):48–62.

A Word2vec and what it means to the linguist

The first difficulty that comes to the linguist with the adoption of NLP tooling is that of understanding and interpreting the mechanics of the software at hand. In this section, we start by providing a brief technical overview of how the CBOW model of Mikolov et al. (2013) functions in Appendix A.1, and move on to a linguistics-oriented characterization of its hyperparameters in Appendix A.2.

A.1 Algorithmic overview

At their core, distributional semantics models attempt to characterize the distribution of words. For neural-based models, this almost always entails estimating the probability of a token t in its context c:

$$\Pr(t|c) \tag{1}$$

where c corresponds to some notion of context: For CBOW, the context is construed as a sliding window of words co-occurring in a sentence; For BERT, contexts are equated to sentences; for causal language models such as GPT, the context is understood as all preceding words.

The CBOW architecture models probabilities such as Equation (1) by means of learned vector representations for words and contexts:

$$\Pr\left(t_i|c = (t_1 \dots t_m)\right) \propto \vec{t_i} \cdot \vec{c_i} \tag{2}$$

Context representations correspond to sums of word-level features:

$$\vec{c_i} = \sum_{j=\min(1,\ i-w)}^{i-1} \vec{e_j} \quad + \sum_{j=i+1}^{\max(m,\ i+w)} \vec{e_j} \tag{3}$$

As such, the CBOW model consists in two sets of vector representations: target vectors $\vec{t_i}$, which are solely used for estimating the probability of a word in context, and input embeddings $\vec{e_j}$ which serve both as a means to model the context and as input features for downstream applications. Most studies, this one included, concerns themselves with the latter embeddings.

In the specific implementation we rely on (viz. the `gensim` implementation, Řehůřek and Sojka, 2010), the window size w can be either fixed or stochastically determined for every training example. In details, this sampling corresponds to replacing the window size w in Equation (3) with an effective window size $\hat{w}$ uniformly sampled between 1 and w:

$$\hat{w} \sim U(1, w) \tag{4}$$

In practice, this entails that context words that are $k \leq w$ tokens apart from the target word are discarded in k/w of the training instances. Therefore, context words that are closer to the target word are more likely to be taken into account for prediction.

To estimate what probability to assign for a given token in a given context, a practical approach consists in training the model using both positive and negative evidence, through a procedure known as “negative sampling." This is equivalent to maximizing the objective $\mathcal{O}$ listed in Equation (5):

$$\mathcal{O} = \Pr\left(t_i|c\right) - \sum_{t_n \in N} \Pr(t_n|c) \tag{5}$$

Simply put, a negative-sampling CBOW model is trained to maximize the likelihood of an attested word t_i in its context c, and minimize the likelihood of all words in a set N of negative examples (not attested in this context c). The negative examples $t_n \in N$ are randomly sampled for each positive examples, using a distribution derived from the raw frequency distribution $\Pr(t)$ of word types t in the training corpus:

$$p_n(t) \propto \Pr(t)^{\alpha} \tag{6}$$

As is usual with neural networks, the parameters $\vec{t_1} \dots \vec{t_V}$ and $\vec{e_1} \dots \vec{e_V}$ are estimated through stochastic gradient descent, with the goal of maximizing the objective $\mathcal{O}$ in Equation (5). Rather than testing for convergence of this objective on a held-out validation set, it is more usual to expose all of the available training data to the model for a pre-determined number of times, or ‘epochs.’

A.2 Interpretation of hyperparameters

A keen reader, having suffered through Appendix A.1, might notice that the algorithm of a CBOW negative-sampling model is—at least in part—linguistically interpretable. In the present paper, we specifically discuss five hyperparameters.

The *window size* w controls how contexts of occurrences are modeled. A large window entails that more word tokens intervene in the definition of a context representation $\vec{c_i}$, whereas a smaller window narrows the relevant context to the more immediate surrounding of the target word. Likewise, whether or not to employ a *dynamic window size sampling* algorithm, as detailed in Equation (4) also interest the linguist, as this window size sampling is equivalent to assigning a greater weight to context words closer to the target words. In other words, to re-purpose Firth’s (1957) famous quip, the window w controls what company a word keeps.

The *number of negative examples*, $\#N$, determines how to weigh positive and negative evidence. As a consequence, a larger sample set N of negative examples entails that the model will be more penalized for assigned non-negligible probability mass to negative evidence. Too large a N can however lead to a detrimental effect, as the model could be incentivized to focus solely on minimizing the negative evidence, thereby leading to an incoherent modeling of the positive evidence. In short, the size of the negative sample establish a position in a trade-off between ensuring that spurious associations between negative examples and attested contexts do not arise (when $\#N$ is large), and emphasizing the importance of fitting to the attested data (when $\#N$ is small).

A related point that will interest the linguist concerns how to sample negative evidence; as we detailed in Equation (6), the CBOW architectures provide a *negative sampling smoothing hyperparameter* α to control this sampling process. Setting a value closer to 0 allows one to sample more from the tail of the vocabulary’s frequency distribution. More precisely, remark that $\alpha = 0$ entails sampling negative examples uniformly over the entire vocabulary sorted by frequency; $\alpha = 1$ matches the unigram frequency distribution in corpus; $\alpha > 1$ over-emphasizes frequent words, and $\alpha < 0$ over-emphasises infrequent ones.

Lastly, an import hyperparameter to consider is the *number of epochs*: Given that this controls how often the same positive evidence is used to adjust the model’s parameters, it has natural implications for the reach of any claim derived from the use of a CBOW model. From a practical point of view, we also remark that a lower number of epochs might result in a model that does not properly capture all the intricacies of the positive evidence used for its training—whereas a higher number of epochs can lead to a model that “over-fits" its training data, i.e., does not generalize properly to novel data.

Remark that we have ignored some key hyperparameters that are often discussed in the NLP literature. In particular, we do not discuss the dimension of the trained embedding as it has no obvious simple linguistic interpretation.

B Data used in experiments

B.1 Analogical relations in GATS

Subset	Section
Inflection	gram3-present-participle
Inflection	gram4-past-participle
Inflection	gram5-plural
Inflection	gram6-nationality-adjective
Inflection	gram7-past-tense
Inflection	gram8-plural-verbs
Derivation	gram1-adjective-to-adverb
Derivation	gram2-opposite
Semantic	antonyms-adjectives

(*Continued on next column*)

(*Continued from previous column*)

Subset	Section
Semantic	capital-common-countries
Semantic	capital-world
Semantic	city-in-state
Semantic	currency
Semantic	family

Table 3: Analogical relations in GATS, grouped as inflection, derivation or semantics.

B.2 Morphological processes in Demonette

Type	Process	N. pairs
Sem.	1A>N	1324
Sem.	1A>V	423
Sem.	1N>A	3870
Sem.	1N>V	2631
Sem.	1V>A	2960
Sem.	action	7520
Sem.	agent	2302
Sem.	el:N>A	279
Sem.	eur:V>A	356
Sem.	ième:NUM>A	57
Form.	CONVERSION:N>A	182
Form.	CONVERSION:N>V	2353
Form.	CONVERSION:V>N	2345
Form.	PST.PART:V>A	317
Form.	Vble:V>A	324
Form.	age:V>N	1625
Form.	aire:N>A	424
Form.	al:N>A	449
Form.	ance:V>N	95
Form.	ant:V>A	915
Form.	el:N>A	279
Form.	erie:A>N	99
Form.	erie:V>N	85
Form.	eur:V>A	356
Form.	eur:V>N	1580
Form.	euse:V>N	526
Form.	eux:N>A	402
Form.	ien:N>A	98
Form.	ier:N>A	201
Form.	if:N>A	372
Form.	if:V>A	132
Form.	ifier:A>V	50
Form.	ion:V>N	1946
Form.	ique:N>A	1742
Form.	iser:A>V	373
Form.	iser:N>V	278
Form.	itude:A>N	62
Form.	ité:A>N	1082
Form.	ième:N>A	57
Form.	ment:V>N	1285
Form.	rice:V>N	196
Form.	té:A>N	81
Form.	ure:V>N	73
Form.	é:V>A	1272
Form.	ée:V>N	66

(*Continued on next column*)

(*Continued from previous column*)

Table 4: Processes from Démonette

task	α	e	N	w	s
SimLex-999	0.2	1	15	15	False
FEEL	0.2	5	25	5	False
GATS/sem	0.2	5	25	10	False
GATS/D	0.2	3	25	15	False
DerCF	0.6	3	10	5	False
DerCS	0.2	3	20	5	False
DerPF	0.6	5	5	5	False
DerPS	0.6	5	5	5	False
POS	1.0	5	10	5	False
GATS/I	0.2	5	25	10	False
PCC	1.4	3	25	5	True
SCC	1.4	3	10	5	False
PCP	1.0	5	5	5	False
SCP	1.0	5	20	10	False

Table 5: Hyperparameters of best performing model by task.

C Supplementary results

C.1 Optimal hyperparameters for each task

We provide the optimal hyperparameters for each task in Table 5 for replication purposes. As noted in the main text, the most obvious trend we can identify is the α hyperparameter. We can also remark that most task benefit from training across multiple epochs (with the exception of SimLex-999), and most do not benefit from the shrinking s (with the exception of SCC). Also worth highlighting is that we do not observe that large windows favor semantic tasks.

task	highest	deciles								
	score	9th	8th	7th	6th	5th	4th	3rd	2nd	1st
SimLex-999	0.310	0.292	0.282	0.272	0.261	0.250	0.242	0.232	0.223	0.213
FEEL	0.399	0.378	0.372	0.366	0.359	0.348	0.333	0.302	0.266	0.225
GATS/S	0.330	0.283	0.247	0.229	0.200	0.173	0.153	0.130	0.101	0.071
GATS/D	0.176	0.132	0.114	0.098	0.086	0.073	0.060	0.039	0.019	0.008
DerCS	−0.005	−0.012	−0.015	−0.019	−0.026	−0.034	−0.050	−0.080	−0.099	−0.112
DerCF	−0.026	−0.052	−0.062	−0.071	−0.083	−0.098	−0.135	−0.170	−0.193	−0.210
DerPF	0.549	0.502	0.486	0.471	0.456	0.425	0.355	0.272	0.206	0.157
DerPS	0.746	0.700	0.686	0.674	0.658	0.621	0.523	0.425	0.340	0.275
POS	0.744	0.716	0.704	0.694	0.680	0.651	0.587	0.524	0.462	0.392
GATS/I	0.379	0.332	0.305	0.284	0.259	0.237	0.204	0.183	0.158	0.119
SCC	−0.101	−0.168	−0.203	−0.251	−0.303	−0.331	−0.346	−0.354	−0.360	−0.373
PCC	−0.099	−0.157	−0.188	−0.221	−0.261	−0.292	−0.304	−0.313	−0.320	−0.329
SCP	0.817	0.778	0.752	0.719	0.600	0.434	0.374	0.325	0.273	0.228
PCP	0.526	0.486	0.475	0.464	0.449	0.403	0.394	0.391	0.389	0.387

Table 6: Maximum and deciles of scores per task

C.2 Highest performances per task

In Table 6, we summarize our models' scores on each of the task, by looking at both the maximum score achieved and deciles. We can make two key observations: First, as stressed in the main text scores for morphological clustering tasks are systematically negative, meaning that embeddings do not form homogeneous, well-delineated clusters according to morphological features. Second, the spread between the first and ninth deciles tends to be be much more extreme with morphological tasks (both inflectional and derivational) than with semantic task. Whether these results suggest that morphological distinctions are not adequately captured by distributional models in general, or whether the blame is to be pinned on word2vec more specifically is an intriguing question we intend to pursue in future work.

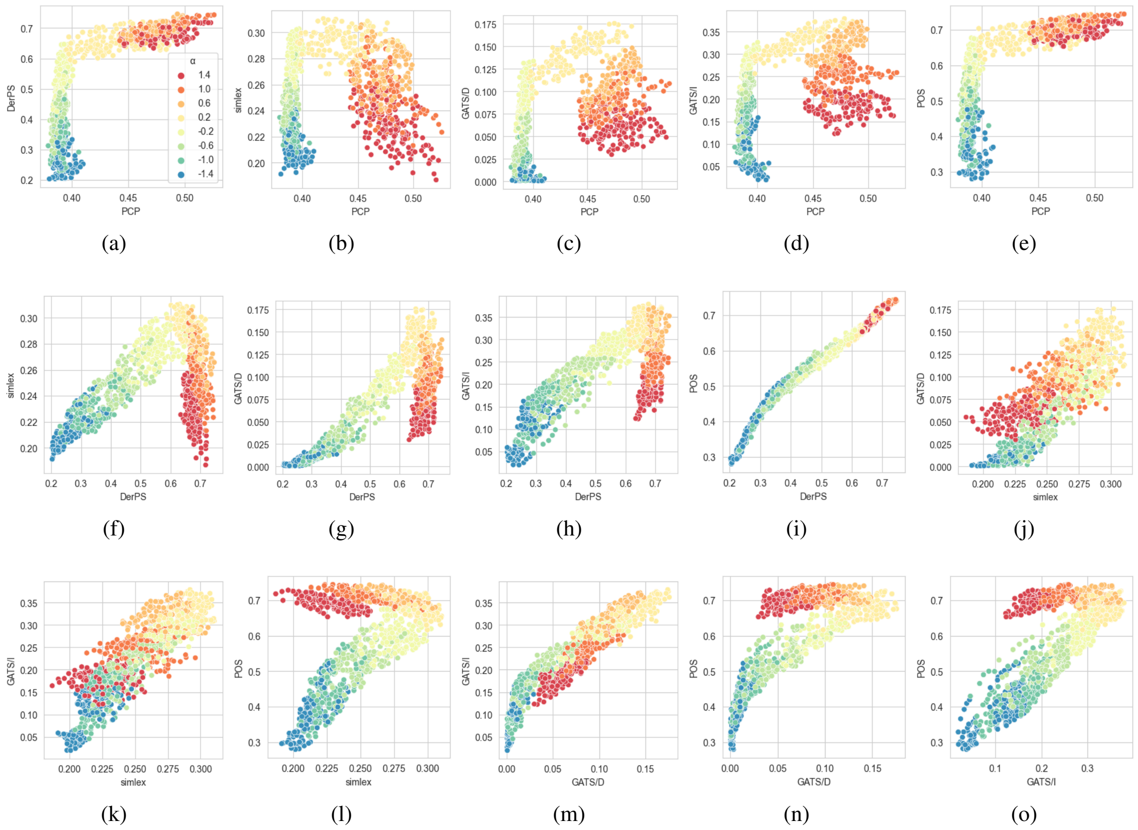
C.3 Correlation matrices by values of α

We can visualize the difference of quality induced by the α hyperparameter. can be visualized by plotting, for each pair of task, how individual model scores relate to one another and what value of α they use, as shown in Figure 2 for five of the tasks (simlex, DerPS, PCP, POS, GATS.D and GATS.I). Correlation in performance across pairs of tasks tends to be monotonic between our morphological tasks as well as between traditional NLP benchmarks, however our morphological tasks do not appear to align well with traditional benchmarks. The sole exception to that is the POS-tagging task, which is found to correlate very strongly with our morphological derivation prediction tasks (shown in Figure 2i) and entertains a complex, non-linear relationship with all other NLP benchmarks. The α hyperparameter also accounts for much of the variation we observe: different values of α tend to produce easily delineated clusters of models, except when comparing GATS and SimLex-999 (see Figures 2j and 2k). In this latter case, note that values of α produce poorer results on both benchmarks the further away they stray from the optimal value of $\alpha = 0.2$, suggesting that here as well α determines much of the attested behavior.

C.4 Trends when including $\alpha < 0$

There are some interesting trends that emerge from looking at models with $\alpha < 0$ which we have not discussed in the main text so as to focus our argument on more successful models.

One interesting empirical approach that we can take to highlight the effect of these negative α hyperparametrizations consists in performing clustering analyses as shown in Figures 3a and 3b: inflectional, derivational and semantic tasks reliably clustered closely with tasks of the same linguistic type but, depending on the specific clustering algorithm, derivational tasks formed superclusters with inflection (e.g. full linkage clustering) or with semantics (e.g. UPGMA), confirming the intermediate status of derivational tasks. This matches with the argument we lay out in Section 5. However, if we instead only focus on $\alpha > 0$ as in Figures 3c and 3d, this effect is no longer observed, and deriva-



(a) (b) (c) (d) (e)

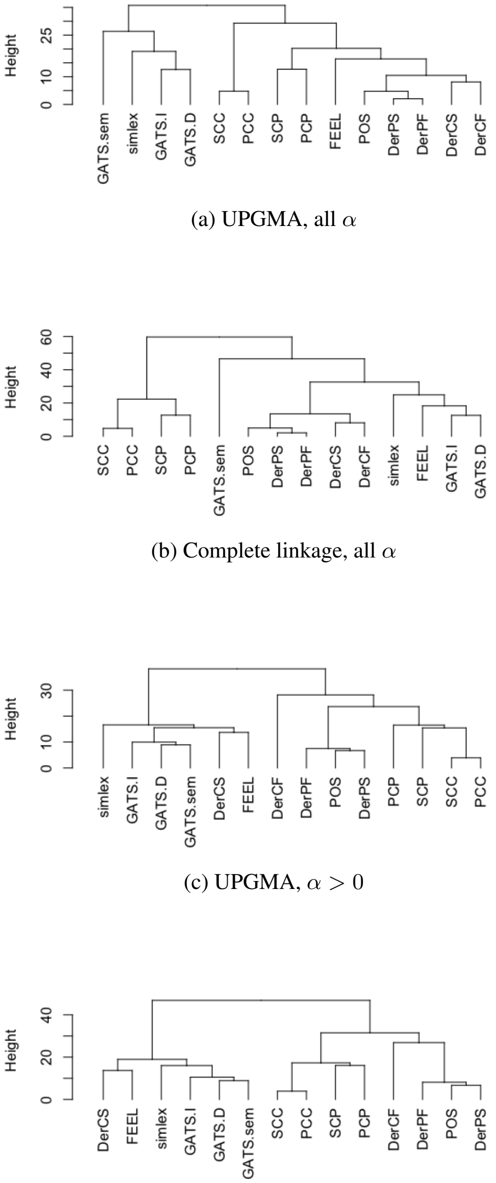
(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

Figure 2: Selected examples of the correlation patterns found in our task set. α can be seen to account for most variation in performance.

tion tasks (with the exception of DerCS) always cluster with inflection tasks.

Another factor to point to is that the core observations from Section 4 also hold when looking at all models. For instance, that tasks cluster depending on the type of linguistic knowledge they target is reflected in Figure 4, although the general picture is overall less clear than when restricting the analyses to $\alpha > 0$.



(a) UPGMA, all α

(b) Complete linkage, all α

(c) UPGMA, $\alpha > 0$

(d) Complete linkage, $\alpha > 0$

Figure 3: Task hierarchical clustering based on observed scores.

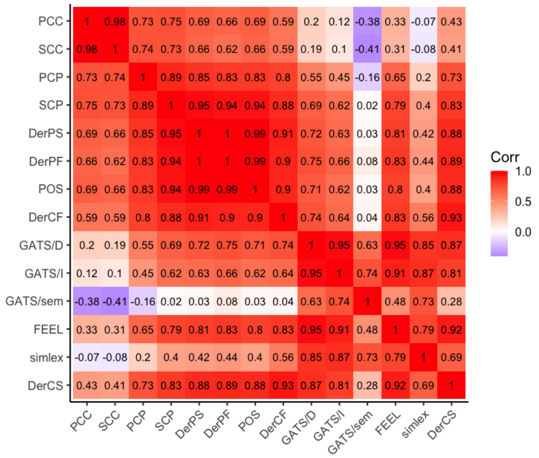


Figure 4: Spearman correlation for performance of models with all values of α on the different tasks.