

# Language Models and the Paradigmatic Axis

Timothee Mickus

University of Helsinki

timothee.mickus@helsinki.fi

## Abstract

The massive relevance of large language models, static, and contextualized word embeddings in today’s research in NLP implies a need for accounts of how they process data from the point of view of the linguist. The goal of the present article is to frame language modeling objectives in structuralist terms: Word embeddings are derived from models attempting to quantify the probability of lexical items in a given context, and thus can be understood as models of the paradigmatic axis. This reframing further allows us to demonstrate that, with some consideration given to how to formulate a word’s context, training a simple model with a masked language modeling objective can yield paradigms that are both accurate and coherent from a theoretical linguistic perspective.

## 1 Introduction

It is a truism to say that field of natural language processing (NLP) has seen profound changes over the past decade. The development of static neural word embeddings, the introduction of contextualized embeddings, and their re-branding as large language models are as many steps along this transition, and each have yielded many impressive technical advancements over the prior state of the art.

It is also a truism to say that this technical progress stems for the most part from an engineering culture, and that the concerns stressed as more prominent in NLP have primarily to do with the maturing technology of deep learning—much of the ongoing background discussion in NLP centers on questions such as scaling up (Sutton, 2019), or defining tasks to solve and metrics to optimize (Tedeschi et al., 2023; Ganesh et al., 2023). The current concerns of NLP pertain not to language, but to what can be achieved through language.

At the same time, there is a sizable body of work interested in discovering what aspects of language

are encoded in language models and word embeddings alike. Many adopt as their main angle of research treating language models as or comparing them to language speakers (e.g., Linzen et al., 2016)—to identify whether they encode some specific linguistic information (e.g., Hewitt and Manning, 2019; Chi et al., 2020); contrast these models with what actual speakers do (Bender and Koller, 2020); or characterize what they can and cannot capture (Merrill et al., 2022; Bouyamourn, 2023).

Underlying all this work on evaluating NLP models is the linguistic framework they are instances of—namely, distributional semantics. Tackling this subjects are technical accounts and surveys (a.o., Lenci, 2018), pieces discussing their usefulness to theoretical linguistics (e.g., Boleda, 2020), works underscoring the theoretical limitations of distributional models (e.g., Emerson, 2020), historical reviews of how this framework has evolved (Brunila and LaViolette, 2022). Yet, conceptual discussions of the distributional framework itself are surprisingly hard to find: Proposed extensions of distributional semantics more often than not focus on incorporating extraneous elements from more strictly formalized frameworks (e.g., Baroni et al., 2014; McNally, 2017; Herbelot and Copestake, 2021), rather than conceptualizing and formalizing distributional methods in and of themselves. This fact is all the more surprising once we factor in that the impressive successes of modern language models are achieved through purely distributional means.

In this paper, we build upon Sahlgren (2008), Gastaldi (2021) and Gastaldi and Pellissier (2021), who keenly analyzes the links between word embeddings, distributional semantics and structuralism. We argue here that systems trained on language modeling objectives can be understood in structuralist terms as *models of the paradigmatic axis*. Sahlgren, Gastaldi and Gastaldi and Pellissier also stress the link between structuralism and distributionalism. Unlike Sahlgren and Gastaldi, we do

not conflate distributional models with vector space semantics; and whereas [Gastaldi and Pellissier](#) connect paradigms and word embeddings through a reformalization of the concept of paradigm with the explicit goal of deriving structural representations, we argue that there is an obvious and immediate link between the language modeling objective and a paradigmatic axis and that this relationship can be attested empirically.

We first include a short historical account of distributionalism and a more substantiated description of our suggested framework for embeddings and language models in Section 2. We then provide empirical demonstrations of how basic linguistic considerations can shape the properties observed in language models in Section 3.

## 2 Language models and paradigms

We first start by gathering here some key elements of structuralist theory to provide the reader with all the relevant context; more thorough accounts can be found in [Brunila and LaViolette \(2022\)](#), [Sahlgren \(2008\)](#) and [Gastaldi \(2021\)](#).

**Structuralism and the paradigmatic dimension of language.** The birth of structuralism in linguistic is usually attributed to [Saussure \(1916\)](#). One chief concern underpinning it is the study of language for language’s sake ([Gastaldi, 2021](#)), which it achieves by making its central object of study the *structure* of the language. In short, the structuralist program, as framed by [Saussure \(1916\)](#), involves the following tenets: (i) that a language has a structure relating sound and meaning; (ii) that this structure can be established by isolating the signs of this language; and (iii) that to isolate signs, one needs to show that variation in sound (or meaning) entails variation in meaning (or sound).

Signs can be related to one another in a variety of ways; one we are especially vested in is that of a paradigmatic relation, as formalized by [Hjelmslev \(1971\)](#). Simply put, words that compete for the same position in a context are said to form a paradigm. Consider for instance ex. (1):

(1) I am teaching.

Notice how the word ‘teaching’ could have been replaced by some other word not attested in ex. (1), be it ‘writing’, ‘dancing’ or ‘fabulous.’ The relationship between ‘teaching’ and these other candidate words is one “*in absentia*,” that is, between

terms as members of the sign inventory of a language, rather than between terms co-occurring in a context. This contrasts with relationships that hold between terms in the same context, usually referred to as “syntagmatic”—consider for instance how in ex. (1) the word ‘I’ is necessary because of how it relates to the word ‘am,’ that is to say, this relationship holds “*in praesentia*.”

The notion of paradigm found in [Hjelmslev \(1971\)](#) builds upon [Saussure’s \(1916\)](#) conception of associative series: [Saussure](#) highlights that we can associate series of words based on whether they share common formal elements (‘teaching’, ‘teaches’, ‘teacher’, ‘teach’, ...), have similar meanings (‘teaching’, ‘learning’, ‘education’, ...), or display formal similarities by happenstance (‘teach’, ‘peach’, ‘beach’, ...). As noted by [van Marle \(1984\)](#), this entails that [Saussure’s \(1916\)](#) view is “that the paradigmatic dimension of language is simply highly indefinite and undetermined” (p. 12). The position we defend here is that a fruitful application of the structuralist concept of paradigms or series to modern NLP only requires a Hjelmslevian take on paradigms. In practice, we will consider a paradigm to be a relationship in *absentia* between terms that are equally syntagmatically constrained.

**Distributionalism.** Distributionalism is a specific strand of American structuralism best exemplified by the figures of Bloomfield and Harris. Their main contribution to structuralism is a deeper focus on what the study of co-occurrences of items (be they signs, words, morphemes or phonemes) and their distributional regularities can highlight.

Harris, in particular, had a keen interest in formalizing linguistics as an empirical, objective science, for which he deemed imperative that observations be carried out as methodically as possible ([Léon, 2011](#)). A seminal example was provided in [Harris \(1954\)](#), where he argued that the analysis of co-occurrences of linguistic elements suffices to establish a structural description of a language.

One notion of interest in [Harris’s](#) work is that of *distributionally substitutable* elements: It consists in the iterative and methodological construction of sets of predictably interchangeable words. To take a concrete example, consider the context:

(2) On \_\_\_\_\_, the office is open from 9AM through 5PM.

Across a large corpus analysis, we expect that we

might attest several possible nouns referring to days of the week in the position left blank in ex. (2)—but nothing else. If, across all contexts we encounter them, these words are in fact substitutable, we can group them into a substitution set. This process can be iterated: For instance, if we have already established that days of the week form a substitution set, we can consider examples such as

- (3) The university is closed this Wednesday.  
 (4) The library is closed this Sunday.

Here, the contexts of the terms (underlined) can be equated as their differences only involve variation within a substitution set; which would therefore allow us to group the terms ‘university’ and ‘library’ in another substitution set. Remark that elements in a substitution set correspond to different paradigmatic choices (Sahlgren, 2008): In other words, distributional substitutability is an operationalization of the concept of paradigmatic relationships based on the distributions of words in context.

**Vector space semantics and distributional semantics models.** One early key success of the distributionalist approach was the discovery that distributional similarity correlates well with word similarity judgments (Rubenstein and Goodenough, 1965). This is often referred to as the *distributional hypothesis*: similar words will occur in similar contexts.<sup>1</sup> This novel perspective eventually gave rise to *distributional semantics*, the field studying how (word) distribution differences correlates with (word) meaning differences. However, to make good of this insight, one hurdle to overcome was the computational challenges entailed by a distributional analysis of an entire corpus. The advent of vector-based means of representing linguistic items (Salton et al., 1975; Landauer and Dumais, 1997)

<sup>1</sup>Harris himself was fundamentally invested in not relying on meaning and speaker cognition in linguistics (Brunila and LaViolette, 2022), and conceived distributional as strictly distinct from (though correlated with) meaning. This sheds an interesting light on literature surrounding the cognitive plausibility of distributional accounts of language (Miller and Charles, 1991; Landauer and Dumais, 1997; Mandera et al., 2017). Harris’s position is fundamentally at odds with many of the more successful and better studied linguistic frameworks: In particular, Chomsky (1965) frames linguistic as a branch of psychology, which has to be understood as a departure from distributionalism and structuralism. In that respect, approaches attempting to reconcile generativism and distributionalism (e.g., Baroni et al., 2014; Herbelot and Copestake, 2021), have to be put in the light of the distributional semantics enterprise, and have to be understood as departures from the purely distributional approach of Harris (1954).

provided the means necessary to carry out distributional analyses at this scale. As a result, modern expositions of distributional semantics often conflate vector space semantics and distributional models (e.g., Lenci, 2018; Boleda, 2020; though not always, e.g., Erk, 2012). The relation between vector representations and distributional analyses is, however, of a contingent nature—while the usefulness of high-dimensional space for semantic representations was established early on in computationally oriented research communities (Salton et al., 1975; Schütze, 1992), this need not be the sole means by which a distributional analysis can be carried out.

**The language modeling objective(s).** If vector space models and distributional models should not be conflated, why then should the current spate of embedding and language models be construed as distributional models? A number of the neural models that are discussed in NLP—and in particular most embedding and language models—are derived from word–context co-occurrences. In practice, they try to quantify the probability of a term given its context, or formally:

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

where  $t$  corresponds to a target *term*, and  $c$  stands for a *context*. What constitutes a term and a context can in principle vary quite a lot: Contexts have been defined by means of sentences, documents, paragraphs, or syntactic trees; whereas terms have been defined either as word, or increasingly commonly as word-pieces, and may or may not factor in spelling information.

Models that do not directly capture the above often instead compute a related quantity, or an information-theoretic variant thereof. For instance, while the CBOW objective of Mikolov et al. (2013) is explicitly eq. (1), the counterpart skip-gram architecture instead models  $p(c|t)$ ; moreover, in practice, the exact objectives used to trained word2vec, the negative sampling and hierarchical softmax objectives, differ from eq. (1). Note however that the former is simply a reformulation of the probability definition, whereas the latter has already been the subject of much analysis, starting with Levy and Goldberg (2014) who related it to PMI-based models. Looking at more recent works, it is also straightforward to identify the masked language modeling introduced by Devlin et al. (2019) as an instance of eq. (1); it also corresponds to the sentinel-based objective of T5 architectures (Raffel

et al., 2020); whereas the ELECTRA architecture of (Clark et al., 2020) is explicitly linked to the negative sampling objective. As for causal language models, it can be identified as a formulation of the usual autoregressive objective  $p(w_i|w_{<i})$ .

In short, many neural and non-neural NLP systems, as they can be construed as word generators conditioned on other text, fall within the scope of eq. (1). That similar objectives have been used to develop the most prominent tools across the last decade, from static word embeddings to language models,<sup>2</sup> appears an obvious consequence of the very limited amount of annotations necessary to set up this objective: The sole requirement is that terms be identified within their context—i.e., that the corpus be presegmented in linguistic units.

**A definition of distributional models.** In what follows, we consider a distributional model to be any system that satisfies the following criteria:

- (i) given a context, it produces a distribution of terms, following eq. (1);
- (ii) this distribution is derived from corpus data;
- (iii) this distribution is applicable beyond the corpus data it was derived from.<sup>3</sup>

One could consider, as a fourth criterion, requiring that the context does not contain the term—out of concern that the probability  $p(t|c)$  would degenerate to assigning 1 to the attested term  $t$  and 0 to all other terms. Such a case can only occur if the context is itself segmented (or segmentable) in linguistic units. Document models (e.g., Salton et al., 1975; Landauer and Dumais, 1997) would be ruled out by this fourth criterion.

**Distributional models are models of the paradigmatic axis.** This can be established by considering the following three facts.

First, that the language modeling objective is fundamentally ambiguous: While it is reasonable to expect that a well-formed model of eq. (1) tends

<sup>2</sup>One family of models conspicuously absent are those trained with human feedback, such as ChatGPT.

<sup>3</sup>This third criterion might seem somewhat trivial, but it both reflects the actual practices of the community that builds said models (assessing generalization capabilities on held-out data is a central tenet of the NLP methodology), and constitutes a departure from strict corpus-based accounts of distributional semantics, including Harris (1954) as well as more recent developments. For instance, Baroni et al. (2014) state (p. 247) that “the meaning of content words lies in their distributions over large spans of texts.”

to assign greater probabilities to the terms that are indeed attested in their respective contexts, this expectation is however defeasible, since speakers may elect to use terms that are less common or surprising. Consequently, a model will assign non-zero probability scores to words other than the actual attested term: If we were to provide ex. (2) to a language model, we would not expect it to assign all its mass to a single term (say “Tuesday”) as some other terms could also fit this context (unless we are faced with an acute case of overfitting).

Second, that the model’s learned distribution should be syntagmatically (and semantically) constrained. If we assume our distributional model assigns probabilities in a manner that reflects what humans are likely to produce, then, while we might expect some fundamental ambiguity between possible terms, this ambiguity is not absolute. Going back to what a model would do of ex. (2), we can strongly conjecture that its probability mass would indeed be accumulated on a narrow class of terms, including mostly days of the weeks. Words belonging in this class will necessarily share a number of semantic traits—since by construction all of them are equally adequate in this context, they also have to be semantically compatible with it: In short, the relationship between terms described by the contextual distribution in eq. (1) should in principle capture some aspect of their semantics, as per the distributional hypothesis. We can also point out that the distribution for this context ought to characterize determiners as much more unlikely than nouns, i.e., this contextual constraint is not just semantic in nature, but rather syntagmatic.

Third, that the learned distribution is a relationship in absentia. Which actual term  $t$  is attested in a given context  $c$  is in fact somewhat irrelevant, as we are dealing a distribution over ambiguous terms. The relation between the output probability distribution and the attested word is thus only a loose indicator of our model’s validity. What we really expect of a language model is that it properly encodes the underlying ambiguity of possible terms in a manner that is coherent with the syntagmatic constraints of the context. As a consequence, the probability distribution therefore encodes a relationship between abstract terms that compete for a given position, and not the relation between the one attested term and its context.

In short, the objective of eq. (1) entails (i) associating a series of ambiguous terms (ii) with similar semantics constrained by the syntagmatic relation-



ships encoded in the context (iii) as a relationship in absentia. Thus, the output probability distribution of a language model describes a relationship between words that is conceptually similar to Saussure’s (1916) associative series, Hjelmslev’s (1971) paradigms and Harris’s (1954) distributionally substitutable elements—or more simply put, distributional models are models of the paradigmatic axis.<sup>4</sup>

**Connections with prior works.** That word embedding models are related to the structuralist concept of a paradigmatic axis is not an entirely novel idea: Sahlgren (2008) already identified that some (non-neural) word embedding models, especially those which define contexts as windows of words around the target term, instantiate paradigmatic relations. A very similar connection between distributional models and paradigms was also established by Gastaldi and Pellissier (2021), but they do not equate the model’s objective with the structuralist concept. Instead, Gastaldi and Pellissier identify paradigms as a supplementary construct to explain why specific terms co-occur across varied contexts. Their notion of paradigms departs from the usual structuralist concept in two ways: (i) they propose to formalize paradigms by means of syntactic, informational and characteristic content; and (ii) they explicitly formulate paradigms as sets (rather than terms that may be more or less directly associated) that can exhibit some form of hierarchical subclass structure. These theoretical additions are more than justified when considering what they yield: some means of deriving a linguistic structure from pure distributional analysis. However, they also obfuscate the relationship between language modeling objectives and paradigms, which limits the applicability of their conception of paradigmatic relation as an analytical tool for modern NLP systems.

It is worth stressing that the objective eq. (1) also entails some differences with respect to the traditional notion of a paradigm. In particular, the inclusion of a term in an associative series is quantified by the probability assigned to it through eq. (1). While this is in line with the “highly indefinite and underdetermined” view of Saussure (1916), this also starkly contrasts with later developments of this concept—chief of which Harris’s (1954)—

<sup>4</sup>It is tempting to include syntagmatic relations in what distributional models describe (e.g., Sahlgren, 2008). Yet syntagmatic relations are expected to hold between words in the context, given as input. A more appropriate characterization would be that they constrain paradigmatic series: Syntagmatic relations are implicitly captured to explicitly model paradigms.

where for any term we may say whether or not it is part of a paradigm. Distributional models, in contrast, construe the relevance of a term to a specific paradigm as a matter of fuzzy set membership: Some terms are more likely members than others.

### 3 Empirical confirmation

While the notion that systems designed to satisfy the language modeling objective are models of the paradigmatic axis is an appealing one, we still require some empirical confirmation of its validity.

Our approach will be as follow: train neural networks with a language modeling objective; and then verify whether their output distributions over terms describe reasonable paradigms. To showcase whether this re-framing of language models as models of the paradigmatic axis can be helpful to the linguist, we can also discuss whether manipulating what linguistic information is provided as context modifies performances in a theoretically coherent way. In practice, our focus will be on *positional* information: This has been one of the features separating static embedding models such as word2vec from contextual embedding models such as BERT, and we can strongly expect that models where context is captured as a bag-of-words yield much less accurate representations of the paradigmatic axis than models that properly factor word order. Very relevant prior work by Sinha et al. (2021) already found this positional information to be necessary for high downstream performances.

A direct comparison of off-the-shelf static and contextual embedding models is somewhat meaningless to our particular endeavor, since they vary on many aspects—including but not limited to the data they have been trained on, the number of parameters they contain and the complexity of the computations they perform. As such, we will start by describing in Section 3.1 two closely related architectures for position-aware and position-agnostic language models which we will then train on the same data, so as to provide a meaningful comparison of their outputs in Sections 3.2 and 3.3.

#### 3.1 Architectures

To facilitate our empirical investigation of whether language modeling objectives lead to models of the paradigmatic axis, let us lay out a few design requirements as to how our language models should be conceived. First, to simplify any judgments on the resulting distributions over terms, it is prefer-

able to study models trained on data pre-segmented in words, rather than word-pieces or other types of linguistic units. Second, it is preferable to keep the model conceptually simple so that its computations remain interpretable, although it is also necessary to ensure that the model is expressive enough to produce non-trivial representations of the paradigmatic axis. Third, the model needs to be lightweight enough to guarantee the replicability of our experiments. Fourth and last, as we focus on positional information, we should make sure that ablating all position information does not require a massive overhaul of the network.

Factoring in all these design requirements, we propose two architectures loosely inspired from the Transformer architecture (Vaswani et al., 2017), one *position-agnostic* and the other *position-aware*. In both cases we consider words as terms, contexts are defined as all other words in a sentence (i.e., we consider some form of masked language modeling). Formally, our position-agnostic network can be described as:

$$p(t_i|c, \theta) = \text{softmax} \left( \mathbf{W}^{(\text{proj})} \mathbf{o} \right) \quad (2)$$

$$\mathbf{o} = \phi \left( \mathbf{W}^{(\text{out})} \phi(\mathbf{h}) \right) \quad (3)$$

$$\mathbf{h} = \text{softmax} \left( \frac{\mathbf{q} \cdot \mathbf{K}^T}{\sqrt{d}} \right) \mathbf{V} \quad (4)$$

$$\mathbf{q} = \text{LayerNorm} \left( \mathbf{W}^{(\text{query})} \phi(\mathbf{t}) \right) \quad (5)$$

$$\mathbf{K} = \text{LayerNorm} \left( \mathbf{W}^{(\text{key})} \phi(\mathbf{X}) \right) \quad (6)$$

$$\mathbf{V} = \text{LayerNorm} \left( \mathbf{W}^{(\text{value})} \phi(\mathbf{X}) \right) \quad (7)$$

where  $\mathbf{W}^{(\text{out})}$  is of shape  $[d \times 2d]$ ,  $\mathbf{W}^{(\text{proj})}$  is of shape  $[d \times V]$  (with  $V$  the number of word types in our vocabulary), and all other matrices of shape  $[d \times d]$ ;  $\phi$  is a nonlinear activation function. The input  $\mathbf{X}$  corresponds to layer-normalized input embeddings for the words in the context of the attested word  $t$ , i.e., all tokens  $t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_n$  in the sentence except for  $t_i$ :

$$\mathbf{X} = \text{LayerNorm} \left( \begin{bmatrix} \mathbf{x}_{t_1} \\ \vdots \\ \mathbf{x}_{t_{i-1}} \\ \mathbf{x}_{t_{i+1}} \\ \vdots \\ \mathbf{x}_{t_n} \end{bmatrix} \right) \quad (8)$$

The position-aware model is highly similar to the position-agnostic model, except that we replace

eq. (5) with

$$\mathbf{q} = \text{LayerNorm} \left( \mathbf{W}^{(\text{query})} \phi(\mathbf{p}_i) \right) \quad (9)$$

and the input  $\mathbf{X}$  in eq. (8) is now defined as

$$\mathbf{X} = \text{LayerNorm} \left( \begin{bmatrix} \mathbf{x}_{t_1} + \mathbf{p}_1 \\ \vdots \\ \mathbf{x}_{t_{i-1}} + \mathbf{p}_{i-1} \\ \mathbf{x}_{t_{i+1}} + \mathbf{p}_{i+1} \\ \vdots \\ \mathbf{x}_{t_n} + \mathbf{p}_n \end{bmatrix} \right) \quad (10)$$

In detail, these models are centered on the use of a scaled-dot attention mechanism (Bahdanau et al., 2016; Vaswani et al., 2017) as shown in eq. (4): the hidden representation  $\mathbf{h}$  in eq. (4) is an average of the value representations in eq. (7), weighted by how similar key and query representations are (eqs. (5), (6) and (9)). Keys and values are computed from the context (eqs. (6) to (8) and (10)), whereas the query is derived from minimal input information about the term: In our position-aware architecture, this input is simply the index of the term (eq. (9)); in the position agnostic model, we use a default input vector  $\mathbf{t}$  for all terms, learned along with the other model parameters (eq. (5)).<sup>5</sup> To further bolster the expressiveness of these language models, we include specific subnetworks linked to the computations of keys, values and queries, as well as a final computation block after the attention head (eq. (3)) and before projection onto the vocabulary space (eq. (2)).

As a useful reference point, we also include a word2vec CBOW model (Mikolov et al., 2013)—which, while not directly comparable, has been extensively studied in prior literature. For each model (including word2vec), we replicate training with three different seeds. Models are trained on a corpus of 20M sentences, half of which are sampled from Wikipedia, whereas the other half comes from BookCorpus (Zhu et al., 2015). Further details are available in Appendix A.

### 3.2 Accuracy

The first item we focus on is whether our models are accurate: How often is the most likely term according to  $p(t|c)$  in fact the one we attest in our held out evaluation set?

<sup>5</sup>Using an attention mechanism allows us to dynamically weight the different value vectors based on the query and keys' vectors. This is therefore more expressive than the basic CBOW scheme of Mikolov et al. (2013), where all context items are always averaged with equal weights.

arch.	dataset	acc.	$\mathbb{E}[p(t c)]$
pos	bookcorpus	$0.450 \pm 0.001$	$0.346 \pm 0.003$
	wikipedia	$0.397 \pm 0.001$	$0.290 \pm 0.003$
nopus	bookcorpus	$0.289 \pm 0.001$	$0.200 \pm 0.002$
	wikipedia	$0.193 \pm 0.000$	$0.103 \pm 0.002$
w2v	bookcorpus	$0.033 \pm 0.000$	$0.003 \pm 0.000$
	wikipedia	$0.033 \pm 0.001$	$0.005 \pm 0.000$

Table 1: Model accuracy and mass assigned to the attested term (average of 3 runs).

Corresponding results are displayed in Table 1, which lists performances both in terms of accuracy (the proportion of terms ranked as first by the language model) and average probability assigned to the attested term  $t$ , noted  $\mathbb{E}[p(t|c)]$ . First, metrics on BookCorpus are always higher than their counterpart on Wikipedia—this likely stems from the higher average sentence length in the latter, along with the more diverse vocabulary it uses. None of the model pass the threshold of 50% accuracy, suggesting that most of the time, the most probable term (as ranked by our models) is not in fact the one we attest in the corpus. Second, we find a clear distinction between the three models considered: Word2vec fares significantly worse than the other two more complex models, but the addition of position also clearly improves both accuracy probability mass metrics as compared to the position-agnostic model. Third, we can see a fairly low standard deviation across all three runs—i.e., results are generally stable.

Overall, these results suggest a nuanced take: We do not find these models to be highly accurate, but we do see some confirmation of our hypothesis that linguistically informed context (in our case, positionally informed contexts) fare better.

### 3.3 Syntagmatic compatibility

It is however worth remembering that model accuracy is a flawed metric, and should not serve as a means of evaluating language models as models of the paradigmatic axis—since speakers and writers can and do elect to use unlikely terms. Instead, we ought to look at whether the words highlighted as relevant for a paradigm are compatible with the syntagmatic constraints of its context. As a simplified first step towards answering this, we consider looking at part of speech information: If the term we attest in our context is a noun, we should expect that the most likely terms according to  $p(t|c)$

should all be nouns.<sup>6</sup>

A first technical question to solve, then, concerns how to establish *which set of likely terms* one should focus on: Given that paradigms retrieved from language models are probabilistic in nature, we need some means of deciding which words to rule in or out of a paradigmatic set. In practice, we need some manner of restricting the output vocabulary to the most likely terms. In the present work, we consider two simple approaches. The first consists in simply taking the top  $k = 10$  most likely terms according to the model. The second, consists in using conformal prediction sets (CPS; Vladimir Vovk, 2005), a principled way of selecting a subset of the possible output terms so as to guarantee a coverage of  $N = 80\%$ . Simply put, a coverage of 80% entails that that selected subsets each have 80% chances of containing the attested term. In practice, we use a least-ambiguous set-valued classifier method (Sadinle et al., 2019): We (i) measure the probability mass assigned to each attested term on a held out calibration set; (ii) compute the  $1 - N^{\text{th}}$  quantile  $q$  of these probability scores; and (iii) build sets from term distributions  $p(t|c)$  by considering all values above that threshold quantile  $q$ , or  $\mathcal{T} = \{t' : p(t'|c) \geq q\}$ . Assuming symmetry and iid. between test and calibration data, the probability of the attested term  $t$  should be greater than  $q$  for  $N\%$  of the test examples, and thus included in  $\mathcal{T}$  with a likelihood of  $N\%$ .

Having decided on how to select paradigm subsets, we can now turn to a second technical question: how to measure whether terms in a paradigm have the correct part-of-speech. POS-tagging systems that rely on full sentences to label words are not suitable to our purposes, since they could bias the labeling of terms in a paradigm towards the part-of-speech of the attested term by sheer virtue of the syntagmatic constraints of the context. Instead, our inquiry requires a context-independent means of establishing possible parts-of-speech for selected terms. We therefore fall back to a lexical resource—namely Wiktionary, owing to its large coverage;

<sup>6</sup>It is perhaps more common to evaluate distributional models on semantic tasks, given the distributional hypothesis expects contextual similarity to be linked to semantic similarity (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Hill et al., 2015). We depart from this tradition as this aspect of distributional representations seems to be somewhat consensual. While assessing the POS-tagging capabilities of language and embedding models alike has been studied extensively prior to this work (e.g., Elman, 1990; Lenci et al., 2022), little has been done to study whether the full output distribution of a language model is syntagmatically coherent.

method	arch.	dataset	% valid POS
baseline		bookcorpus	47.135
		wikipedia	47.385
CPS	pos	bookcorpus	$87.250 \pm 0.207$
		wikipedia	$84.087 \pm 0.239$
	nopos	bookcorpus	$76.895 \pm 0.198$
		wikipedia	$71.981 \pm 0.093$
	w2v	bookcorpus	$60.337 \pm 0.097$
		wikipedia	$60.775 \pm 0.079$
Top 10	pos	bookcorpus	$81.929 \pm 0.254$
		wikipedia	$80.490 \pm 0.367$
	nopos	bookcorpus	$72.074 \pm 0.180$
		wikipedia	$68.820 \pm 0.233$
	w2v	bookcorpus	$71.961 \pm 0.111$
		wikipedia	$71.551 \pm 0.058$

Table 2: Proportion of syntagmatically compatible likely paradigm terms, according to the POS tag of the attested term (average of 3 runs).

we rely on the English RDF parse by [Sérasset and Tchechmedjiev \(2014\)](#). This wide coverage, however, comes at the expense of leniency and accuracy. We therefore consider as a baseline using the full vocabulary as a paradigm subset: This gives us a strict lower bound for model performances. For simplicity, we ignore terms (both attested and in the paradigms) for which we find no Wiktionary entry; any term in a given paradigm is counted as syntagmatically compatible as long as one of its reported parts of speech could match one of the reported parts of speech of the attested term. We then report the average proportion of paradigm members that are syntagmatically compatible.

An overview of the corresponding results is displayed in Table 2. A few key observations need to be made. First, we can take notice of the very high lower bound suggested by our baseline—this can be explained in part by the leniency of our procedure as well as the noisiness of the POS-tag inventory derived from Wiktionary, although the categorical flexibility exhibited by the English lexicon may also play a role. We also highlight that all our experiments are clearly on average more compatible than this baseline—suggesting that, although our methodology suffers from its limitations, we can observe some evidence that the language modeling objective corresponds to establishing linguistically meaningful paradigms.

Furthermore, we see that terms in paradigms are generally more syntagmatically compatible

for BookCorpus paradigms rather than Wikipedia paradigms. This nuances our earlier discussions with respect to accuracy: Our language models appear indeed fundamentally less adequate when it comes to modeling paradigms in Wikipedia. A wider lexicon might entail a lesser ability to construct lexically meaningful representations of paradigmatic distributions: Exposing a language model to more numerous but rarer words might lower its average performance.

Lastly, we see that positional information significantly improves the syntagmatic compatibility of terms in paradigms. In a few cases, the word2vec baseline models are comparable to the position-agnostic language models. This hinges on the criterion used to establish paradigms: Selecting the top-10 highest probability scores yields less compatible sets than the quantile-based conformal set approach, except for word2vec. This should come as no surprise, given that the conformal sets are constructed based on the likelihood of an attested term. Word2vec models, as shown in Table 1, are generally not accurate in this regard; in particular, the probability mass they assign to the attested term tends to be low. Less accurate models therefore yield larger conformal sets, which we expect to be less syntagmatically compatible. This can be verified by looking at the average size of the conformal prediction sets: While the position-aware models yield conformal sets containing  $\approx 42$  terms in average, and the position agnostic  $\approx 285$ , this number rises to  $\approx 26\,441$  for wordvec—i.e., more than a quarter of the vocabulary is included in the conformal set.

Sizes of the conformal prediction sets can interest us for another reason. We can expect that conformal prediction sets should be larger when paradigms can contain more words. In terms of parts-of-speech, we therefore expect that open grammatical categories like noun, verbs and adjectives should yield larger sets than closed categories, such as articles, conjunctions and prepositions.<sup>7</sup> An overview of the CPS sizes, broken down per part-of-speech, is provided in Table 3, along with the number of relevant conformal sets. Open categories (verbs, nouns, proper nouns, adjectives) tend to yield the largest sets, whereas closed categories

<sup>7</sup>[Angelopoulos and Bates \(2022\)](#) suggest that conformal prediction set sizes can be used as proxies for model uncertainty: A larger conformal set is more ambiguous as to what the target should be. In short, we expect CPSs to capture the uncertainty inherent to the ambiguity of different parts of speech.



	number of CPSs	avg. CPS size		
		w2v	nopos	pos
<b>adjective</b>	106 908	26 386.5	306.1	46.1
<b>adverb</b>	91 442	26 897.5	297.0	34.7
<b>article</b>	27 229	26 693.2	294.5	21.3
<b>conjunction</b>	28 683	26 840.7	286.8	34.7
<b>determiner</b>	31 388	27 024.4	284.5	30.9
<b>infix</b>	7	26 917.3	314.4	71.0
<b>interjection</b>	30 691	26 765.7	287.9	35.6
<b>noun</b>	241 535	26 443.9	308.6	46.6
<b>numeral</b>	19 047	26 226.1	237.1	22.3
<b>particle</b>	27 708	26 855.1	208.4	19.0
<b>phr. unit</b>	6563	26 832.1	294.3	28.7
<b>postposition</b>	868	26 596.1	316.8	47.9
<b>prefix</b>	1400	26 721.6	82.1	8.2
<b>preposition</b>	80 831	26 824.7	291.5	27.3
<b>pronoun</b>	45 210	26 922.3	277.7	28.7
<b>proper noun</b>	13 776	26 489.7	303.4	42.6
<b>suffix</b>	19 321	26 583.3	297.6	26.7
<b>symbol</b>	19 906	26 414.7	282.0	24.4
<b>verb</b>	159 314	26 489.1	319.0	51.0
<b>all</b>	354 388	26 440.6	285.0	42.0

Table 3: Conformal prediction sets size per part of speech (averages of 3 runs).

(aside from the two least represented, infixes and postpositions) yield smaller conformal prediction sets. In fact, the difference in CPSs sizes between nouns, verbs, adjectives, adverbs and proper nouns vs. those for all other parts of speech is statistically significant.<sup>8</sup>

## 4 Conclusion

In the present article, we have argued that language models and word embeddings can be understood through a structuralist lens as models of the paradigmatic axis, as long as we factor in the inherent ambiguous nature of language modeling objectives. We have highlighted how this conception builds upon prior work (Sahlgren, 2008; Gastaldi, 2021; Gastaldi and Pellissier, 2021), and where it distinguishes itself from these prior approaches—in terms of the range of models it considers, as well as by explicitly embracing the departures from the earlier formulations of this structuralist concept. The position we endorse here is to minimize the assumptions necessary to frame language models in structuralist terms: With fewer assumptions comes broader application. In contrast, Gastaldi

<sup>8</sup>Mann-Whitney  $U$  tests:  $p < 10^{-32}$ , common-language effect size  $f > 0.66$  in position-aware models and  $f > 0.57$  in position-agnostic models

and Pellissier’s (2021) position can be understood as a narrower form of the present argument designed to allow the emergence of structural representations of the context—but it is worth asking whether one should really expect of distributional models that they yield explicit structural representations (Rumelhart and McClelland, 1986; Buder-Gröndahl, 2023).



One crucial point we have left out of our discussion concerns whether purely linguistic paradigms actually exist. The data we use to train distributional models are not in fact linguistic in nature, but sociolinguistic; they encode social variation and biases, and consequently distributional models do as well (Bolukbasi et al., 2016; Garg et al., 2018). We should expect the paradigms that the language modeling objective obtains to not purely encode linguistic relationships. As such, it is crucial to evaluate the extent to which we can abstract away from the sociolinguistic aspect of the training data.

Hence, one contribution of the present work is to propose a preliminary empirical verification of whether this conception of language models (and therefore word embeddings) as models of the paradigmatic axis is coherent. To that end, we have demonstrated how manipulating the linguistic information in the input contexts of conceptually simple architectures yields predictable effects, and how conformal prediction sets can be leveraged to select paradigm terms in a linguistically meaningful way—in that selected terms are syntagmatically compatible with the context from which we derive them.

In the present work, we have striven to provide a basis that is easy to comprehend and straightforward to build upon—which comes at the cost of our experiments and models being simplistic in many regards. This work also leaves a number of research questions open for future inquiries: Do larger models yield more accurate representations of the paradigmatic axis? What other linguistic information should we include or remove from our contexts? How do these models behave with respect to other pre-segmentations of the training corpora—and especially the ubiquitous word-piece segmentations? How can a model of the paradigmatic axis be leveraged to study other linguistic phenomena, and what methodological steps should we take to mitigate its potential lack of accuracy?

## Acknowledgements

We thank Timothée Bernard, Tommi Buder-Gröndahl, Mathilde Huguin, Jussi Karlgren and Denis Paperno, as well as the three anonymous reviewers for discussions and comments on this work that substantially bettered it.

 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 N° 771113).  We also thank the CSC-IT Center for Science Ltd., for computational resources. This work is also supported by the ICT 2023 project “Uncertainty-aware neural language models” funded by the Academy of Finland (grant agreement N° 345999).

## References

- Anastasios N. Angelopoulos and Stephen Bates. 2022. A gentle introduction to conformal prediction and distribution-free uncertainty quantification.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.
- Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for composition distributional semantics. *Linguistic Issues in Language Technology*, 9.
- Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.
- 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.
- Adam Bouyamourn. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Singapore. Association for Computational Linguistics.
- Mikael Brunila and Jack LaViolette. 2022. What company do words keep? Revisiting the distributional semantics of J.R. Firth & Zellig Harris. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4403–4417, Seattle, United States. Association for Computational Linguistics.
- Tommi Buder-Gröndahl. 2023. The ambiguity of BERTology: what do large language models represent? *Synthese*, 203(1):15.
- Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.
- 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.
- Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.
- Guy Emerson. 2020. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453, Online. Association for Computational Linguistics.
- Katrin Erk. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.
- Ananya Ganesh, Jie Cao, E. Margaret Perloff, Rosy Southwell, Martha Palmer, and Katharina Kann. 2023. Mind the gap between the application track and the real world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1833–1842, Toronto, Canada. Association for Computational Linguistics.

- 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.
- Juan Luis Gastaldi and Luc Pellissier. 2021. The calculus of language: explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews*, 46(4):569–590.
- Zellig Harris. 1954. [Distributional structure](#). *Word*, 10(2-3):146–162.
- Dan Hendrycks and Kevin Gimpel. 2016. [Gaussian error linear units \(GELUs\)](#).
- Aurélie Herbelot and Ann Copestake. 2021. [Ideal words](#). *KI - Künstliche Intelligenz*, 35(3):271–290.
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. [SimLex-999: Evaluating semantic models with \(genuine\) similarity estimation](#). *Computational Linguistics*, 41(4):665–695.
- Louis Hjelmslev. 1971. *Prolégomènes à une théorie du langage. suivi de "La structure fondamentale du langage"*. Éditions de Minuit.
- Thomas K. Landauer and Susan T. Dumais. 1997. [A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge](#). *Psychological Review*, 104(2):211–240.
- 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.
- Omer Levy and Yoav Goldberg. 2014. [Neural word embedding as implicit matrix factorization](#). In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jacqueline Léon. 2011. [Z. S. Harris and the semantic turn of mathematical information theory](#). In *History of Linguistics 2008*, pages 449–458. John Benjamins.
- Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. [Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation](#). *Journal of Memory and Language*, 92:57–78.
- Louise McNally. 2017. *Kinds, descriptions of kinds, concepts, and distributions*, pages 39–62. Düsseldorf university press, Berlin, Boston.
- William Merrill, Alex Warstadt, and Tal Linzen. 2022. [Entailment semantics can be extracted from an ideal language model](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 176–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *arXiv preprint arXiv:1301.3781*.
- George A. Miller and Walter G. Charles. 1991. [Contextual correlates of semantic similarity](#). *Language and Cognitive Processes*, 6(1):1–28.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21(1).
- Herbert Rubenstein and John B. Goodenough. 1965. [Contextual correlates of synonymy](#). *Commun. ACM*, 8(10):627–633.
- David E Rumelhart and James L McClelland. 1986. [On learning the past tenses of english verbs](#).
- Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. [Least ambiguous set-valued classifiers with bounded error levels](#). *Journal of the American Statistical Association*, 114(525):223–234.
- Magnus Sahlgren. 2008. [The distributional hypothesis](#). *The Italian Journal of Linguistics*, 20:33–54.
- Gerard Salton, Anita Wong, and Chun-Shu Yang. 1975. [A vector space model for automatic indexing](#). *Commun. ACM*, 18(11):613–620.
- Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris.
- Hinrich Schütze. 1992. [Word space](#). In *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann.

Gilles Sérasset and Andon Tehechmedjiev. 2014. [Dbnary : Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations](#). In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, pages 67–71, Reykjavik, Iceland.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. [Masked language modeling and the distributional hypothesis: Order word matters pre-training for little](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rich Sutton. 2019. [The bitter lesson](#).

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. [What’s the meaning of superhuman performance in today’s NLU?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Jaap van Marle. 1984. *On the Paradigmatic Dimension of Morphological Creativity*. Foris Publications, Dordrecht, The Netherlands.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Glenn Shafer Vladimir Vovk, Alexander Gammerman. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *The IEEE International Conference on Computer Vision (ICCV)*.

## A Implementation details

For the position-aware and position-agnostic models, we use a latent dimension of  $d = 256$  and a GELU activation function (Hendrycks and Gimpel, 2016). We optimize cross-entropy between the model output and the attested term at each position, using the Adam optimization algorithm with decoupled weight decay (Loshchilov and Hutter, 2019), using a learning rate of 0.001,  $\beta = (0.9, 0.999)$ , and a weight decay of 0.01.

Models are trained on a corpus of 20M sentences, half of which are sampled from Wikipedia, whereas

the other half comes from BookCorpus (Zhu et al., 2015): These corpora corresponds to the sources used for training BERT (Devlin et al., 2019), but the amount of data we consider here is orders of magnitude lower. We also select 20k sentences for testing, and 2k for further calibration in Section 3.3; likewise, half of the sentences in both sets are sampled from Wikipedia and half from BookCorpus. We pre-segment the corpus in words using nltk (Bird and Loper, 2004), using a vocabulary comprising the 100k most frequent words; we pre-process all sentences by lowercasing, stripping accents, and normalizing to the NFKD unicode norm. Models are trained for one epoch over these data, by minibatches of 50 sentences truncated to a maximum length of 128 tokens.

The word2vec baselines are trained on the same data using a vector size of 100, window of 5, and 5 negative examples per target. For our language models, training requires 12 to 16h hours on a RTX 3080 GPU, and about half an hour on CPUs for the word2vec baseline.