big data and cognitive computing

MDPI

Article

Part-of-Speech Context Vectors: Approximating Distributional Meaning of Syntactic Category Symbols

Xiaona Ma [1,*] and Carl Vogel [1,2,*]

[1] Trinity Centre for Computing and Language Studies, Computational Linguistics Group, School of Computer Science and Statistics, Trinity College Dublin, The University of Dublin, College Green, D02 PN40 Dublin, Ireland
[2] Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, 8 Fővám Square, 1093 Budapest, Hungary
* Correspondence: max4@tcd.ie (X.M.); vogel@cs.tcd.ie (C.V.)

Abstract

Words occurring in similar contexts have been observed to have similar meanings. A natural and established method within computational linguistics implements this observation by representing words as vectors with dimensions determined by words that are witnessed in fixed positions in relation to the target word. We generalize this context vector approach to part-of-speech (POS) sequences appropriate to word sequences. As with words, the context of a POS tag (considering the POS tags occurring before and after any target tag) reflects its syntactic constraints and may approximate the "meaning" of the target tag, from a distributional perspective. We use the 111-million-word British National Corpus (BNC) and the sequence of POS labels lifted from those texts to calculate POS context vectors. We observed significant agreement between the clusters of POS context vectors and the supercategories of corresponding POS tags, and examined potential categorization of the POS categories that emerged from the vector clusters. We also found that though vector measures partially align with the predictions of generativist linguistic theories, the approach suggests a more complex relation between syntactic categories. We conclude that a mutual-information-based approach better approximates the distributional "meaning" of syntactic categories than the conditional probability distribution of POS symbols.

Keywords: syntactic category; part of speech; context vector; generative grammar; notional theory of syntactic categories

Academic Editor: Ester Zumpano

Received: 14 April 2026
Revised: 6 June 2026
Accepted: 18 June 2026
Published: 23 June 2026

1. Introduction

A natural and established method within computational linguistics is the representation of texts as vectors of features (dimensions). The value for each dimension is typically based on observations, i.e., counts, of the feature—perhaps a zero, if it is absent, or a non-zero value representing a count or relative frequency. Such a vector may be used to implement the "bag of words" reduction of texts, representing texts by the observation of words that occur at ordered positions within the vector by some means, typically related to rank frequency. Where texts are represented using vectors of the same dimensionality, the same features in the same order, their vector representations support quantificational comparison using standard computations like cosine similarity or Euclidean distance. Such an approach is also used in estimating word meaning, noting for each target word the relative frequency of a fixed sequence of other words in some number of positions before and after the target word. These vectors are known as word context vectors [1], and given

suitable corpora (large, for example), techniques like clustering reveal patterns of word use that map naturally onto syntactic (e.g., determiners) and semantic (e.g., months) categories [2–4]. Here we introduce part-of-speech (POS) context vectors in the spirit of word context vectors. To do this, we apply the same approach that is used with word context vectors to texts that have been annotated with POS categories. We demonstrate the idea using the 111-million-word British National Corpus [5] (BNC) and the sequence of POS labels lifted from those texts rather than the words so labeled.

A POS context vector summarizes the distribution of POS tags in the context of a target POS tag. For example, in such a vector constructed for adjectives, it can be expected that the dimensions of nouns to the right of the target tag and determiners to the left of the target tag would obtain high values relative to other dimensions, while in a vector constructed for verbs, it would appear that determiners frequently occur to the right of the target, along with singular or plural nouns. Linguists think about syntactic categories through inspection of contexts in which exemplars may appear and reflection on constraints that prohibit appearance [6,7]. In distributional analyses of syntactic categories, the syntactic structures that are frequently shared by the words of a category, though probably not by every word in the category, are identified as the criteria and supply a distributional definition to the category. Representing words in a corpus with their POS tags, the syntactic categories of words and evident facts of their contextual distribution all amount to syntactic information. A POS context vector aggregates this information and approximates the distributional "meaning" of the target category. In what we report here, we analyze, for any target POS tag, the distributional facts of POS tags of neighboring words, not necessarily neighboring *constituents* that could appear as nodes in phrase parsing trees higher than POS tags—constituent-based contexts generalize what we explore here. Our purpose in introducing POS context vectors is to indicate two possible alternative constructions, which are frequently used in describing distributional patterns but may capture different aspects of syntactic structures, and to show how these constructions behave in relation to what may be expected from linguistic theories.

In referring to "linguistics" and "linguistic theory" here, we are thinking of the generativist paradigm and approaches to language that descend from or in some cases that react to that paradigm but remain committed to constituent structures and syntactic categories of both words and phrases. Although we have noted that it is a step beyond the current work to build context vectors that represent constituent positions, we are exploring the relations between the major syntactic categories that can be shared between words and phrases. The representation we explore inherently involves gradience of membership in major syntactic categories of granular part-of-speech categories as modeled in the proximity of individual POS vectors to the centers of their clusters. Linguistic theory also addresses gradience but typically not in an identical fashion. Work may be noted that conceives of attested relative frequency of association as informing components of grammar [8]. Some have constructed theoretical frameworks in which gradience is admitted across the components of grammar ontology, including syntactic categorization [9]. Frameworks for linguistic theorization may be noted in which syntactic categories associated with words are thought of as abstractions over constructions in which the items appear and are thus accompanied by gradience in assignment to lexical items [10]. Separately, it has been argued that part-of-speech syntactic categories associated with words are not required by linguistic ontology in a framework in which, instead, such word classes are at most shorthand for functions from structured semantic types to other structured semantic types [11,12], but without changing the population of the word classes.

In the linguistics literature, it has been observed that a certain level of arbitrariness is allowed in the selection of the distributional criteria for syntactic categories. For exam-

ple, ref. [13] gave the examples of *before* in (1) (Example (1) is example (85) (p. 19) in [13], but this, in turn, originates as part of example (66) (p. 181) of [14]) and suggested that the three instantiations of *before* are all prepositions, and the difference between the three cases is that *before* is a preposition taking a clausal complement in (1*a.*), a preposition taking a nominal complement in (1*b.*), and an intransitive preposition in (1*c.*).

Example 1.

a. John arrived before the last speech.
b. John arrived before the last speech ended.
c. John arrived before (hand).

Ref. [13] compared his analysis with the one by [15], whose criteria for prepositions and conjunctions clearly identify *before* in (1.b) as a conjunction. Ref. [16] pointed out that neither the criteria for prepositions given by [13] or [15] were chosen in a principled way, and clear evidence was not provided if one set of criteria functioned better than the other. In a more recent work, ref. [17] examined expressions of the form in (1) and suggested a definition for the relevant words in dictionaries similar to the argument by [13]. Selection of criteria is common in a distributional approach, as including every possible construction as a criterion may lead to a large number of categories—at one extreme, one word per category [18–20]. In practice, researchers may choose the criteria that they value the most, or, as [21] termed as "methodological opportunism", choose the criteria that would fit their expectations of the categories without a priori theoretical bases. It follows that the selection could overlook some similarities or dissimilarities between word groups, which in turn lead to inconvenience in the analyses of the construction of an expression longer than a single item. In this respect, given a corpus labeled with an established POS system, POS context vectors serve as a method to evaluate the distributional patterns of each category. For example, one may consider for any POS category whether the POS category that is the most similar to it is in the same superordinate category (referred to as "supercategory" hereafter) or in a distinct category; if the latter, one may reflect on whether the categorization scheme merits adjustment.

The term "lexical category" has been used in some linguistic work to describe categories like nouns, verbs, or adjectives, whose subordinate categories are usually open-class, containing many words, each used relatively infrequently, contrasting with functional classes like determiners or prepositions, which are closed-class and contain words used relatively frequently ("open" or "closed" indicating whether it is easy or difficult, respectively, to add new members to the class). Here we follow this tradition and use the term "word class" to cover both lexical and functional categories. "Syntactic category" and "grammatical category" can be used in a wider range than "word class" to encompass phrasal constituent categories like noun phrase or direct object. In this work, the three terms are used interchangeably when the discussion is limited to parts of speech. Parts of speech as labeled in a corpus are also taken as syntactic categories. In the corpus we are using, in most cases, part-of-speech tags label words in their inflected forms, e.g., *flights* or *sought*. In some cases, they also label markers, e.g., the genitive marker *'s*, symbols, e.g., an alphabetical symbol in a sentence, or multi-word expressions, e.g., *except that* or *of course*.

In this work, we seek to evaluate the efficacy of the POS context vectors in encoding syntactic information and observe the relationships among POS categories with their vector representations. We find that POS context vectors, as an approach exhausting distributional facts from a certain perspective, effectively encode the linguistic intuition underlying the syntactic categories. We observe significant agreement between the supercategories of POS tags and the clusters of context vectors of the same set of tags. Estimates from POS

context vectors partially align with the predictions of the relationships between syntactic categories predicted by landmark linguistic theories and at the same time point out aspects that cannot be explained by the theories when a large range of word use cases are included. In general, the POS context vectors based on the information-theoretic concept of mutual information agree more with the linguistic treatments than the vectors based on conditional probabilities, empirically demonstrating the strength of the mutual information measure in the construction of context vectors.

This work is motivated by the need for a method that is able to reliably and comprehensively calculate distributional patterns of syntactic categories and, as a consequence, enable quantitative evaluations of syntactic categories. The contribution of this work is twofold:

- We present POS context vectors and demonstrate their efficacy in encoding distributional "meaning" of syntactic categories, supporting their use in empirical observations;
- We apply the POS context vectors to evaluate the relationship between POS categories and identify where empirical consequences of the vectors can be compared with explicit or implicit aspects of those theoretical approaches.

The article is arranged as follows: Section 2 gives an overview of the background relevant to context vectors and linguistic theories with emphasis on syntactic categories and gradience in atomic categorization (not gradience of grammaticality); Section 3 describes the methods used in this work, including the dataset, construction of vectors, and the measurements used in the experiments; Section 4 present the results; we discuss in Section 5 the meaning of POS context vectors and the results, as well as the limitations of the work; and the main conclusions are summarized in Section 6. We note the tension between vector dimensionality and the interpretability of the dimensions, as well as the lack of an independent quantitative metric for the evaluation of the POS context vectors, which may be explored in future work.

2. Background

2.1. Context Vectors

Ref. [22] investigated the distributional structure in language, starting from the observation that each linguistic element occurs in certain positions relative to certain other elements. The work treats the distribution of a linguistic element, e.g., a word, as the sum of all its environments and proposes that the difference in the meaning of linguistic elements is a function of the difference in distributional structure. In practice, it is convenient to summarize the distributional facts with vectors. Substantial studies have made use of the vector representation and evaluated their effectiveness in approximating word meaning [23,24], shorter units such as affixes [25], or longer units such as phrases [26]. Comprehensive surveys of the applications of context vectors have been provided by [4,27,28].

Studies have shown that word context vectors are able to capture the syntactic properties shared by a group of words. Ref. [2] constructed context vectors with a 40-million-word corpus and examined whether syntactic categories would arise by clustering the vectors. The researchers recorded word co-occurrences, with the context set as two words to the left and two words to the right of a target. A hierarchical cluster analysis was then applied on the vectors, and the results showed that the nodes at certain heights overlap the orthodox syntactic categories. Labeling each node with the predominant syntactic category of the items governed by that node, it was found that only 5% of the 1000 most common words in the corpus were misclassified. Ref. [3] used a similar approach to construct context vectors and found that unambiguous past tense forms of verbs clustered together, separating from the cluster of unambiguous past participle forms of verbs. The relation between word context vectors and syntactic categories has also been observed in more recent studies [29,30].

Although the approach of context vectors is an established method to represent symbols in a multidimensional space and the distributional facts in word contexts have been proven to contain useful information about the syntactic category of the target words, we have not seen context vectors that are constructed directly for part-of-speech tags. Ref. [2] tagged a corpus with labels derived from clusters of word-level context vectors for the corpus (which approximates syntactic categories) and context vectors were constructed for bigrams and trigrams of the labels (with computational tractability gains from the smaller inventory of category labels than words as an evident contributing motivation), but the work aimed to examine the internal structure of phrasal categories, and it remains unclear how effectively a POS tag can be represented by such a vector or to what extent the vector representations can agree with the linguistic predictions of the tags. In any case, this idea is similar to what we propose: instead of using cluster identification labels for the clusters in which word context vectors lead to a word having its best fit as a category for a word, we use the category assigned to a word by a robust POS tagger [31], which was then processed and adopted by the linguists who prepared the BNC (the full annotation process is described in https://www.natcorp.ox.ac.uk/docs/URG/posguide.html (accessed on 13 April 2026)); both approaches then construct vectors over the labels that result. We think the vectors produced in this manner by [2] are best understood as word-cluster context vectors, and those produced in our analysis are POS context vectors.

Linguistic and POS tagging practices have suggested the potential of representing a POS tag with the POS tags in its contexts. In linguistic studies, the categorization of words frequently involves substituting items in the contexts with contextual words' category labels in order to allow certain semantic differences among the member words of the same category. For example, instead of comparing *hear* in *They will hear the bell* with *tear* in *They will tear the paper*, *bell* and *paper* can be replaced with a category label *N* and the two words *hear* and *tear* would appear in the same context ([7] p. 42). In distributional tests that are now widely used to assign words to syntactic categories, the test criteria typically consist of category labels and relations between categories instead of the original words in sentences [32,33]. In a POS tagging task, ref. [34] included the POS tags of the words preceding and following a word of interest as input features and saw a moderate increase in the performance of tagging unknown words (i.e., words that are not in the training dataset). In contrast, the words themselves before and after the word of interest contribute much less to the tagging accuracy. Given the close relation between a POS tag and its contextual tags, it follows that the vectors that have been constructed for words may be applied to POS sequences. Instead of approximating the features of a POS tag that have been explored in substantial theoretical and empirical studies, the vector representation of POS tags would focus on the distributional aspect based on co-occurrence pairs.

Among the various candidate methods that can be used to compute the values of the vectors, the information-theoretic measure *pointwise mutual information* (PMI) [35] is a particularly interesting one in finding the relation between word pairs. The measure evaluates how much more or less likely two items are to occur together than they would have been if the items were independent. A positive PMI value indicates that two items are likely to co-occur at an above-chance level, while a negative value indicates that the two items are rejecting each other in their contexts. Ref. [36] found that the metric was able to identify associated word pairs such as (*doctor*, *nurse*) or (*save*, *from*). The metric has been widely used in corpus linguistic studies to identify collocations [37,38]. Other count-based measurements, such as log-likelihood and odds ratio, have been used to identify keyword lists that can differentiate one corpus from another [39,40]. However, the measurements require two corpora to be compared, instead of profiling a single corpus. Ref. [41] evaluated several ways to compute the values of a word context vector, including

conditional probabilities, PMI, positive PMI (PPMI, reducing negative PMI values to 0 [42]), and the PMI values before taking the logarithm. Results showed that among three semantic tasks, PPMI vectors with a cosine similarity measure give the best performance. In this work, we also employ two computations for the value of POS context vectors, conditional probabilities and PPMI and evaluate to what extent the different relativization methods could encode linguistic intuitions.

In more recent studies on word vectors, dense vectors derived by neural network models are frequently used [43]. Intrinsic evaluations of the vectors, for example, the correlation between the cosine similarity computed with word vectors and human judgments of word pairs, show that dense vectors usually perform better than the sparse, count-based vectors [44–46]. However, the dimensions of dense vectors are difficult to explain. Ref. [45] shows that items with high values at the same dimension of a dense vector are not necessarily semantically related. Ref. [47] evaluated BERT embeddings and found that polarity and negation are associated with specific dimensions, but properties like synonymy exhibit complex patterns. In contrast, count-based vectors have dimensions that are inherently interpretable, which would be helpful to uncover the implicit systematic similarity between words or syntactic categories.

2.2. Binary Feature Distinctions in Generative Grammar

Syntactic categories have been regarded both as categories of the base in generative grammar (denoted with symbols such as N, V) and as sets of feature values common to the lexical items of a category (denoted with sets such as [+N, −V]) [48]. In the latter case, N and V are treated as abstract features. The "+" sign indicates that words in a category have the feature and the "−" sign indicates the absence of a feature. Ref. [49] pointed out that it was not necessary to retain the notion of categories in a grammar if the reliance on the analytic procedure of segmentation and classification is abandoned, and suggested treating categories as sets of feature values. In the binary feature system sketched by Chomsky, the category of nouns is described as [+N, −V], verbs are [−N, +V], and adjectives are [+N, +V] [32]. Category membership is unchanged, but the labels are made more granular.

Distributional evidence of the binary representations has been noted in the comparison of the verbs, derived nominals, and derived adjectives, as in examples (2*a.*) to (2*c.*) (example (2) is adapted from examples (3a) to (3c), p. 186 in [32]). It can be observed that the verb *indicate* in (2*a.*) takes a noun phrase as its complement but the derived nominal *indication* in (2*b.*) and the derived adjective *indicative* in (2*c.*) require a preposition *of* to introduce their complement; similarly, the adjective and the noun need a copula to form a finite clause, whereas the verb do not have such requirements. These observations are consistent with the idea that N is shared by both nouns and adjectives but is not available to verbs [32].

Example 2.

a. This indicates Mary's sincerity.
b. This is the indication of Mary's sincerity.
c. This is indicative of Mary's sincerity.

Ref. [50] extended this binary system to include adpositions. Adpositions may be assigned the feature set of [−N, −V], and the four categories can be organized as in Table 1. From the perspective of abstract binary features, nouns in the table are the least similar to verbs, and adjectives are the least similar to adpositions.

Table 1. Syntactic categories in a binary feature system, a natural extension of the binary sketch provided by Chomsky (N for nominal, and V for verbal).

	+N	**−N**
+V	adjectives	verbs
−V	nouns	adpositions

Ref. [50] also proposed an alternative feature system, as in Table 2, and argued that this was more appropriate for his study. In this system, nouns and adjectives switched their feature sets: nouns are [+N, +V] and adjectives are [+N, −V] now. Partial evidence supporting the alternative system was associated with two grammatical features: whether or not the category has a subject [±Subj] and whether or not the complement of the category contains a noun phrase as a direct object [±Obj]. The division was that nouns and verbs both have subjects (in a generalized sense), but adjectives and adpositions do not, and both verbs and adpositions take objects, but nouns and adjectives do not, making nouns the least similar to adpositions and verbs the least similar to adjectives.

Table 2. Syntactic categories in the alternative binary feature system provided by [50].

	+N	**−N**
+V	nouns	verbs
−V	adjectives	adpositions

Although the syntactic commonalities of and differences between the categories mentioned above are consistent with their binary feature representations, the features themselves may not be understood as a set of syntactic behaviors. Ref. [32] noted that a feature is part of the essence shared by the categories that have the feature, which inspires more than syntactic behaviors and is an abstract formal property. In this sense, the relation between the feature representations and the distributional patterns of items in a category is indirect. We have seen inferences of similarity (or dissimilarity) relations arising from the binary representations. Whether or not these inferences are applicable to the distributional patterns of the categories when expanding the range of syntactic contexts beyond the evidence above, for example, to the contexts that are available in language use, may be tested with POS context vectors.

2.3. Gradience of Notional Features

In addition to abstract and distributional properties, researchers also characterize word classes with semantic features, i.e., notional features, especially in languages that are not heavily inflected [7]. For example, nouns typically denote entities or abstract concepts; adjectives typically associate a property with the referent of a noun they modify. The membership of a category defined in terms of notional features is often understood from a prototypical view: there are prototypes of a category that exhibit the features of the category most explicitly, and items that are less focal to the category are less semantically appropriate to the category.

The notional description of syntactic categories often allows a certain ranking of the categories. Refs. [51,52] found that on the time-stability scale, nouns frequently denote stable phenomena, verbs typically denote rapid changes, and adjectives stretch over a wide range on the scale, from stable physical qualities abstracted from typical nouns (especially the items that are frequently included in the category of adjectives) to temporary states. Stassen [53] (p. 168-9) categorized words by property concepts (e.g., material, color, human propensity) and ordered the categories in a hierarchy with one end being the most likely to

be expressed as nouns and the other end being the most likely to be expressed as verbs. Ref. [54] proposed a notional theory of syntactic categories, covering the basic categories and arranging them on a clear gradient. Two notional features were used in the theory: predicativity (denoted as **P**) and referentiality (denoted as **N**). The feature **P** is inherent to verbs: it is "event-specific", inherently relational, and situation-defining. Categories labeled with **P** can be used as predicators. The feature **N** is inherent to nouns and names: it is entity-specific and typically labels elements participating in situations. The items of the categories that have **N** in their representations can be used as arguments.

In the notional theory, a category can be represented with a combination of the two features. The six combinations for the six syntactic categories basic to a language (assume this language has adjectives) are listed in Table 3. (Although the representation {P;N} is used to label verbs by [54], it is also discussed in the work how the finite form of verbs is categorically complex and should be represented as a combination of {P} and {P;N}. Thus, we only include the infinitive form of verbs in the {P;N} category.) The relative preponderance of **P** and **N** of a category is indicated by the way the two features are combined: {P} means the category presents only the **P** feature; {P;N} means the category presents both the **P** and the **N** feature, but the level of **P** is greater than that of **N**; {P:N} means the magnitude of **P** is the same as the magnitude of **N** in the category. (Here ":" indicates equal magnitude of the two symbols and ";" indicates that the magnitude of the feature to the left of the symbol is greater than the magnitude of the feature to the right of the symbol. Adjectives are typically predicative (and thus "verb-like") but typically not associated with change (and thus "noun-like"). Hence the category is described as equal with both features). {N;P} indicates that **N** is more preponderant than **P**; {N} means the category only exhibits the **N** feature; {} is a category that presents neither of the two features. The author also suggested an alternative representation of the notional categories, using the internal weightings of the features, as listed in the second column in Table 3. As shown in the table, if excluding functors, a decrease of one unit in **P** is accompanied by a one-unit increase in **N**. With respect to the five non-empty categories, **P** or **N** alone would be able to categorize the difference among them (for example, 3P entails 1N and vice versa). In this sense, the category of functor is notionally complex, as both **P** and **N** are reduced to 0 and both features would be required in its representation.

Table 3. The notional representation, internal weightings of predicativity (P) and referentiality (N) of each category, and the corresponding syntactic categories.

Notional Representation	Weights of P and N	Syntactic Category
{P}	4P, 0N	auxiliary
{P;N}	3P, 1N	infinitives
{P:N}	2P, 2N	adjective
{N;P}	1P, 3N	noun
{N}	0P, 4N	name
{}	0P, 0N	functor

The judgment of the presence of **P** and **N** for a word, that is, whether or not the word can function as a predicator or an argument, usually refers to a constituent, head-modifier analysis of a clause. For example, in the sentence *I saw her eat the pizza*, the infinitive *eat* heads the phrase *eat the pizza* and the phrase as a whole functions as an argument to the predicator *saw*. Given its head role in the argument, the feature **N** is part of the infinitive's notional representation, although the word alone does not function as an argument [55]. Ref. [56] also mentioned the interaction between notional features and syntactic structures

in a discussion of nouns. Bare nouns are typically not predicators or arguments. When a noun, for example, *doctor*, is used as an argument, it usually requires a determiner, e.g., *the doctor*, or when used as a predicator, it requires a copula, e.g., *being a doctor*. Ref. [56] noted that the requirement of other items when functioning as a predicator or argument is evidence that nouns are not strong in **P** or **N**. It is also clear from the example that the features are associated with phrasal constituents, which are then attached to the head components (though controversy may exist regarding which part functions as the head of a constituent in different formalisms).

The notional theory by Anderson also allows secondary combinations of the features. For example, a suggested representation of gerunds is {P;(N;P)}, as gerunds present properties of both verbs and nouns. Gerund is also one of the categories that are not among the six basic categories listed in Table 3, but nevertheless, it is suggested that a gradience analysis may be generalized to more syntactic categories, given the internal weights of the two features of the categories [54]. Others have developed this framework further, but for our purposes (e.g., [57]), the initial idea and cross-category similarity relations it implies suffice. We have presented this theory because it explicitly addresses gradience in syntactic categories, although it does not directly indicate pairwise similarity relations among categories. We have here drawn out implications of the theory in this direction, and we use these later in evaluating the POS supercategory similarity relations that follow from our approach to POS context vectors.

2.4. Summary

The theory of language has made reference to the syntactic categories of words for millennia, and theories of the internal structure of parts of speech continue to be formulated. Within these, gradients of membership and typicality are noticed. Prominent approaches to these ideas within linguistic theory have been indicated. We see a lack of a computable, exhaustive method to empirically evaluate the implications about syntactic categories from the approaches, and the POS context vectors we present in this work would help to address this problem. The basic assumptions of POS context vectors are described, and count-based and dense vectors are compared, with the count-based vectorization providing an interpretable means of exploring category membership and typicality through clustering.

3. Methods

3.1. Dataset

To demonstrate the idea of POS context vectors, we use the full British National Corpus (BNC), the xml edition released in 2007. The corpus consists of 111,044,604 lexical items and multi-word sequences. Multi-word sequences, referred to as "multiword" in the BNC xml files, are the multiple-word combinations that function as one word class; for example, *of course* is labeled as an adverb and *except that* is labeled as a subordinating conjunction. In the case of sequences labeled "multiword", we treat the sequences as a single unit and discard POS annotations on the constituent items.

The BNC labeled the syntactic category of a lexical item or a multiword with both a set of granular categories (the "c5" tagset (for a guide to the tagset, please see https://ucrel.lancs.ac.uk/bnc2/bnc2guide.htm (accessed 13 April 2026))) and a set of supercategories (denoted as SPOS in this work). The c5 tagset labels categories such as singular nouns and plural nouns, and a supercategory covers a set of c5 labels (e.g., singular noun and plural noun are both labeled as the supercategory of noun). Table 4 gives the counts and percentages of the 4 types of c5 tags in the corpus. Fifty-five of the total 91 tags are word class tags, 4 are punctuation tags, 2 are unclassified items, and 30 are ambiguity tags. (The ambiguity tags are applied wherever the probabilities assigned by the part-of-speech

tagger used in the BNC (CLAWS) to its first and second choice tags were considered too low for a reliable disambiguation.) Word class tags are the most commonly used tags, labeling 83.4% of the instances in the corpus. There are 11 supercategories included in the BNC: adjective, article, adverb, conjunction, pronoun, interjection, noun, preposition, punctuation, verb, and unclassified items. In the corpus, the same c5 tag may be paired with different SPOS tags for different lexical items. Here in the analysis, we group the c5 tags into SPOS categories with the most common pairing. The mapping between c5 tags and SPOS tags is included in Table A1 in the appendix. While supercategories might have been defined differently, splitting prepositions into subcategories or merging articles and determiners, for this work, we utilize the distinctions recorded in the BNC. (In the BNC, the category of prepositions contains three granular categories: (1) PRP, most of the words (as well as multi-word expressions), which may be frequently used as prepositions are labeled as PRP, in which case they typically take a complement; (2) PRF, a category for the word *of* alone, for its frequency and post-nominal function; and (3) TO0, the infinitive marker *to*. Words that may be frequently used as prepositions are tagged as adverb particles (AVP in the tagset) when they do not have complements. For example, the word *down* in *run her friend down* is labeled as an AVP while *down* in *run down the hill* is labeled as a preposition (examples taken from the BNC).)

Table 4. Counts of c5 tag types in each category. The percentages in parentheses are the percentages of instances of the tags in the British National Corpus (BNC).

Word Classes	Punctuation Symbols	Unclassified Items	Ambiguity Tags	Total
55 (83.4%)	4 (12.3%)	2 (0.08%)	30 (3.5%)	91

The naming convention of c5 tags provides another system of word classes. The first two letters of a c5 tag indicate an intermediate category (between SPOS and c5), while the third letter indicates a subcategory (for details, please see http://www.natcorp.ox.ac.uk/docs/c5spec.html (accessed 13 April 2026)). We re-label the intermediate categories in relation to their proportion in the corresponding SPOS class. If the instances of an intermediate class take more than 50% of the total instances of the corresponding SPOS, the intermediate class is re-labeled as the SPOS tag affixed with "_M" (indicating its position of being the "main" to the SPOS, for example "ADJ_M"), while the rest of the intermediate categories in the same SPOS category are re-labeled as the SPOS tag affixed with "_O" ("O" for "other", e.g., "ADJ_O"). There are 16 intermediate classes in total. The mapping for the adjective and noun categories is presented in Table 5. As shown in the table, the main part of the supercategory ADJ contains items labeled with "AJ0", "AJC", and "AJS", and the main part of the supercategory NOUN contains only common nouns. The full mapping between c5, intermediate class, and SPOS is included in Table A1 in the appendices.

Table 5. Examples of mapping between part-of-speech (POS) categories.

SPOS	Intermediate Class	c5	Description	Examples
ADJ	ADJ_M	AJ0	Adjective (general or positive)	*old*, *beautiful*
ADJ	ADJ_M	AJC	Comparative adjective	*older*, *better*
ADJ	ADJ_M	AJS	Superlative adjective	*oldest*, *best*
ADJ	ADJ_O	CRD	Cardinal number	*3*, *fifty-five*
ADJ	ADJ_O	ORD	Ordinal number	*third*, *fifth*
ADJ	ADJ_O	DT0	general determiner	*some*, *these*
NOUN	NOUN_M	NN0	common noun, neutral for number	*committee*, *data*
NOUN	NOUN_M	NN1	singular common noun	*pencil*, *goose*
NOUN	NOUN_M	NN2	plural common noun	*pencils*, *geese*
NOUN	NOUN_O	NP0	proper noun	*London*, *Mars*
NOUN	NOUN_O	ZZ0	alphabetical symbols	*A*, *a*

3.2. Construction of POS Context Vectors

We construct context vectors on the c5 tags, the granular POS tags in the BNC. The entire corpus is treated as one long discourse, with the texts in the corpus concatenated in sequence. Punctuation tags are preserved in the construction of context vectors. For each occurrence of a c5 tag, its context in this work consists of two c5 tags to its left and two to its right. We refer to the c5 type that a context vector is representing as the *target tag* and the c5 types available in the corpus as the set of *context tags*, as each c5 type has appeared in the context of some targets in the corpus. The possible dimensions of the vectors are the combinations of the context tags with the 4 contextual positions. For example, the c5 tag "**AJ0**" occupies 4 dimensions: **AJ0_L1** (AJ0 occurs as the first item to the left of a target), **AJ0_L2** (AJ0 is the second item to the left of a target), **AJ0_R1** (AJ0 is next to a target, to the right of the target), and **AJ0_R2** (AJ0 is the second item to the right of a target). The 91 context tags yield 364 possible dimensions (91 tags × 4 contextual positions). We use the dimensions of the 20 most frequent c5 types in the corpus and include a complementary POS type to cover the remaining 71 c5 types by taking the average of the relevant values. (The choice of 20 is somewhat arbitrary, and it is meant to illustrate the outcome of using an interesting number of actual dimensions plus one dimension accounting for all the remaining dimensions. See Table A2 in Appendix B for indicative evaluations from fewer actual dimensions to the entire set of dimensions individuated by POS tags. The choice of average was meant to keep the values of the complementary dimensions at the same scale as the actual dimensions. By averaging dimensions, we risk potential loss of information; however, applying dimensionality reduction algorithms may yield dimensions that are difficult to interpret. We include a discussion about the tension between dimensionality and interpretability in Section 5.3.1 and an indicative result of applying principle component analysis (PCA) [58] for dimension reduction in Table A3 in Appendix B). Just as for the frequent POS types, there are also 4 contextual positions for this complementary type. In total, there are 84 dimensions in a context vector. The structure of the context vectors is depicted in (1):

$$CV(t) = \langle g(t, c^1_{-j}), ..., g(t, c^n_{-j}), g(t, \overline{C}_{-j}), ..., g(t, c^1_{+j}), ..., g(t, c^n_{+j}), g(t, \overline{C}_{+j}) \rangle \quad (1)$$

where $CV(t)$ is the context vector of a target POS tag t; $c^1 ... c^n$ are the n most frequent context tags in a corpus ($n = 20$ in the main results in this work—in Appendix B, we include results of the vectors with a higher or lower n); $\overline{C}$ is a POS type included to aggregate the values of the 71 context types that are not among the 20 most frequent context tags (BNC has 91 c5 types); $_{-j}$ and $_{+j}$ indicate contextual positions, j items to the left or the right of a target, and j ranging from 1 to m ($m = 2$ in this work); and $g(t, c^i_{-j})$ is the value of t having c^i at $_{-j}$ position, similar for $g(t, c^i_{+j})$.

We compute the values of context vectors in two ways. Both involve counting the number of co-occurrences of a target t and a context tag c^i at a position x relative to t (the range of x is $(-1, -2, +1, +2)$, with negative values indicating positions to the left of t and positive values indicating positions to the right of t) but differ in the relativization of the raw counts. The first one relativizes the counts by the frequency of the target tag, as in (2), where $f(t)$ is the count of t and $f(t, c^i_x)$ is the count of c^i at the x position relative to t:

$$g(t, c^i_x) = \frac{f(t, c^i_x)}{f(t)} \quad (2)$$

Given a target t, the relativization estimates $p(c^i_x|t)$, the probability of having a context tag c^i at the position x relative to t. Compared to raw counts, the relativization avoids detecting large differences between two c5 tags caused by a difference in the frequency of the two

targets. It also allows cross-corpus comparison when needed. This method is similar to the method used by [2]. We refer to the relativization as **Prob** in this work.

The second relativization of raw counts is based on pointwise mutual information (PMI). PMI for a target t and a context c^i at a contextual position x is defined as in (3), where $p(c^i_x)$ is the probability of c^i occurring at position x in a corpus, with any c5 tag as its target. (In most cases, $p(c^i_x)$ is the same as $p(c^i)$, the probability of observing c^i in a corpus. Exceptions would be the c5 types occurring at the start or end of the corpus, which could lose a few occurrences in estimating $p(c^i_x)$ because the target does not exist.) The term $p(t)$ is the probability of observing t in a corpus; $p(t|c^i_x)$ is the probability of observing t as the target given c^i at position x, and $p(t, c^i_x)$ is the probability of the co-occurrence of t and c^i, with c^i at position x in the context of t.

$$PMI(t, c^i_x) = \log_2 \frac{p(c^i_x|t)}{p(c^i_x)} = \log_2 \frac{p(t|c^i_x)}{p(t)} = \log_2 \frac{p(t, c^i_x)}{p(t) * p(c^i_x)} \tag{3}$$

PMI measures, given a target t, how much more (or less) likely for c^i to occur at the position x in the context of t than would have been expected assuming the occurrence of c^i at position x is independent of t. As shown in (3), PMI can also be understood as, given c^i at a position x, how much more (or less) likely it is for t to occur as its target than would have been expected if the occurrence of t is independent from c^i_x. Thus, like the definition by [35], the association between t and c^i is mutual but different from the original measure, in which the order of two events is irrelevant; we include the additional factor of fixed position of c^i relative to t. We can also interpret PMI as in the last part of (3), that is, given the contextual position x, how much more (or less) likely for t and c^i to co-occur than would have been expected assuming the occurrences of the two tags are independent.

The PMI of t and c^i_x can be estimated as in (4), where $f(c^i_x)$ is the count of c^i at the contextual position x in a corpus and N_x is the total number of co-occurrence pairs involving the position x:

$$PMI(t, c^i_x) = \log_2 \frac{N_x * f(t, c^i_x)}{f(t) * f(c^i_x)} \tag{4}$$

The value of PMI ranges from negative infinity, when $f(t, c^i_x)$ is 0, to $log_2 \frac{N_x}{f(t)}$, when $f(t, c^i_x)$ equals $f(t)$ and $f(c^i_x)$. The analytical upper limit is positive infinity when $\frac{N_x}{f(t)}$ is infinitely large. In the computation we report here, the effect of the change in N_x on the value is minimal (N_x is only different by 1 across 4 positions, caused by the border cases in the corpus) and the upper limit monotonically decreases with $f(t)$, faster when $f(t)$ is small. The range of the analytical upper limit of PMI in this work is (7.80, 23062.22).

We compute positive PMI (PPMI) in the context vectors, which retain the positive PMI values and reduce the negative values to 0, as in (5):

$$g(t, c^i_x) = PPMI(t, c^i_x) = \max(PMI(t, c^i_x), 0) \tag{5}$$

The main reason for removing the negative values is that the logarithm would exaggerate the difference between vectors when the actual probabilities of co-occurrence are low. For example, if the ratio in (4) for one target tag at a dimension is 0.0001 and for another target tag at the same dimension is 0.001, the two target tags may be better considered similar with respect to this dimension, as the corresponding co-occurrences are not likely to be observed. However, after taking the logarithm, the first value is -13.29 and the second is -9.97. The increased difference would bring unexpected dissimilarity between the two vectors. In addition, negative infinities exist when the co-occurrence of two items is not observed in the corpus. Thus, we adopted the PPMI instead of the PMI measures.

The distributional criteria that assign words to syntactic categories frequently involve the phrasal structures in which the word of interest may be observed. For example, nouns are words that can occur as the head of a noun phrase, and pronouns may be considered a subset of nouns because the distribution of pronoun phrases is similar to noun phrases [33]. One may think of context vectors that attend to m context positions before and after a target POS category as providing summary statistics of all the $(2m + 1)$-grams involving the target POS category at the center. A POS n-gram ($n > 1$) encodes the pre-terminal sequence of nodes that contributes to some cut of a syntax tree, but without an encoding of the non-terminal higher nodes. Thus, the same POS sequence may be the pre-terminal fringe of two distinct syntax trees, but aspects of the syntactic structure are nonetheless encoded. The inflections of words have also been used as criteria for certain syntactic categories. Some inflectional information would be left out in the vectorization process—for example, English pronouns inflect for nominative and accusative case while nouns do not—but some inflections are reflected in the POS sequences; e.g., infinitives may be preceded by an auxiliary while finite verbs are not and are captured by the POS context vectors.

The value in a Prob vector is higher if the relevant tag occurs more frequently at the relevant position than other tags in the context of t. It records the distribution of c5 tags in the context of a target. In contrast, a PPMI vector emphasizes the mutual association between a target and a context tag; that is, the target c5 frequently has the context c5 at a certain position, and when the context c5 appears at that position, it frequently accompanies the target c5, among all the items that could have appeared as a target. Translating this constructional requirement into syntax, one may speculate that Prob vectors highlight any preponderant linear ordering of a target and a context, while PPMI vectors highlight the within-constituent sequential order. We are interested in comparing the aspects of syntactic categories that the relativization methods can inform; thus, both methods are included in this work and used in the following tests.

3.3. Clustering and Agreement

Assume that the POS context vectors capture adequate syntactic information to characterize the c5 tags and that c5 tags in the same supercategories share more syntactic similarities with each other than with tags from another category; it is expected that a clustering result of the POS context vectors should overlap with the categorization of the c5 tags to an observable extent. Here we cluster the context vectors into K groups, where K equals the number of supercategories of the target c5 tags involved in the clustering. Then we compute to what extent the clusters agree with the supercategories. The clusters are computed with the k-means function in `R` [59], and the agreement is measured with two metrics: Cohen's Kappa and Adjusted Rand Index (ARI). (The function and its parameters are `kmeans(df, K, iter.max = 25, nstart = 10000`. Here, `nstart` is the parameter for random initialization. For each kmeans clustering, 10,000 random centroids are chosen to increase the stability of the results).

We exclude the ambiguity tags, punctuation symbols, and unclassified items from the clustering, leaving 55 c5 tags and 9 supercategories ($K = 9$). The ambiguity tags fall between 2 granular syntactic categories, and thus their supercategories remain undetermined. An alternative treatment for these tags could be to first cluster and then assign the tags to the clusters to which they are closest. Different from word classes, punctuation symbols and unclassified items can be identified by their form and meaning alone. (Unclassified items are items that cannot be appropriately classified as items of the English lexicon, such as non-English words, typographical symbols, and hesitation fillers.) It was shown in a preliminary test that instead of being clustered into two distinct clusters, punctuation symbols and unclassified items usually merged into clusters of other tags, meaning there

are practically 9 instead of 11 supercategories. Thus, these two supercategories are also removed from the clustering.

Cohen's Kappa measures inter-rater reliability. The supercategories of the c5 tags may be taken as judgments made by one rater, and features like word inflection, syntactic structure, or semantic properties may factor into the complex decision process, while the clusters of context vectors could be taken as ratings from another rater, which relies solely on the POS contexts of a tag. Cohen's Kappa requires that the two raters assign items to the same set of categories and thus is sensitive to the mapping between cluster IDs and supercategories. In this work, we use the mapping that can maximize the Kappa value. In the cases where there are several mappings available, the mapping used in this work will be randomly chosen. It has also been noted that given an $m \times m$ cross-classification table, where m is the number of categories and $m > 2$, the Cohen's Kappa value of the 2×2 table obtained by merging the $m - 1$ categories into one category can be used to assess the inter-rater agreement on the category that is not merged [60,61]. We compute this value for each supercategory as an indicator of the extent to which the traditional categorization relies on the items' syntactic contexts. We adopt conventional interpretations of ordered categorical levels of κ values but also present the κ values and significance levels, agreeing that interpretation as "fair" or "moderate", etc., is influenced by contextual norms and expectations [62], which are not available to a novel task such as ours.

Another set of metrics that are frequently used to evaluate agreement between two partitions of subjects is based on object pairs [63–66]. This set of metrics is typically based on 4 counts of pairs of objects:

1. The number of object pairs (in this work, pairs of c5 tags) that are placed in the same group by both partitions (in the case of this work, both c5 tags are in the same syntactic category and in the same cluster—it is not required for the cluster to match the syntactic category), denoted as a.
2. The number of object pairs that are placed in the same group by the first partition but in different groups by the second partition, denoted as b;
3. The number of object pairs that are placed in different groups by the first partition but in the same group by the second partition, denoted as c;
4. The number of object pairs that are placed in different groups by both partitions, denoted as d.

Here we adopt the metric of Adjusted Rand Index (ARI). The Rand Index [67,68] is computed as in (6), that is, the proportion of agreed pairs in all object pairs:

$$RI = \frac{a + d}{a + b + c + d} \tag{6}$$

ARI is the Rand Index corrected for chance, following the general form of correction, as in (7). $E(RI)$ is the expectation of Rand Index if two partitions are independent of each other:

$$ARI = \frac{RI - E(RI)}{1 - E(RI)} \tag{7}$$

The correction measures, of the maximum level of agreement that cannot be realized by chance, how much is achieved by the two partitions. Cohen's Kappa is computed in the same form, where the metric being corrected is the proportion of items along the diagonal of a cross-classification table. ARI is computed as in (8) [66]:

$$ARI = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)} \tag{8}$$

It is also possible to compute the pair-based agreement analogously to Kendall's rank correlation, as in (9). This has been proposed by [63] and measures the difference between the probability of two partitions agreeing on an object pair and the probability that the two partitions disagree. The metric is an increasing function of Rand Index and [65] shows that after correction for chance with (7), the metric is the same as ARI.

$$\frac{(a+d)-(b+c)}{a+b+c+d} \tag{9}$$

We run a significance test for ARI. The variance of the metric is derived by [69]. The distribution of ARI under the null hypothesis (i.e., the two partitions are independent from each other) is approximated by a normal distribution. Although [70] showed that the distribution is positively skewed with a small sample size (e.g., 16 subjects), ref. [69] presented a series of simulations in which the type I error rate with a normal distribution is close to 0.05 (0.53 on average for sample sizes from 5 to 15 and 0.048 on average for sample sizes from 20 to 10,000), making the normal distribution a valid assumption for ARI under null hypothesis.

Cohen's Kappa and ARI are also computed and tested between the clusters obtained by the two vector relativization methods.

3.4. Computing the Dissimilarities Between Groups of c5 Tags

Euclidean distance and cosine between two context vectors are frequently used to measure how distant or similar two vectors with equal dimensions are—in our case, the same judgments of the two corresponding c5 tags. When it comes to the measure between two groups of c5 tags, which is required in the evaluation of the predictions made by linguistic theories, an extra consideration is about how to aggregate the information provided by each member tag. Here we aggregate the information in two ways: (a) representing a group with a central vector, in which, at each dimension, the value of the central vector is the arithmetic mean of the values of the member vectors at the same dimension, and computing the Euclidean distance or the reciprocal of cosine between the central vectors of two groups ($dist_c$ and $rcos_c$); (b) computing the Euclidean distance or cosine similarity between pairs of c5 tags and computing the arithmetic mean of the pairwise values ($dist_p$ and $rcos_p$), as in (10), where C_a and C_b are two groups, m is the number of c5 tags in C_a, n is the number of c5 tags in C_b, V_i is the context vector for i^{th} c5 in C_a, V_j is the context vector for j^{th} c5 in C_b, and k is the number of pairwise similarities computed for the 2 groups.

$$\begin{aligned} dist_p(C_a, C_b) &= \frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{n} dist(V_i, V_j) \qquad (i \neq j \quad if\ a = b) \\ rcos_p(C_a, C_b) &= \frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{cosine(V_i, V_j)} \qquad (i \neq j \quad if\ a = b) \\ k &= \begin{cases} m * n & if\ a \neq b \\ m * (n-1) & if\ a = b \end{cases} \end{aligned} \tag{10}$$

The four measurements above are then normalized, and the dissimilarity between two groups of the c5 tags is the arithmetic mean of the normalized measurements, computed as in (11):

$$\begin{aligned} dissim(C_a, C_b) &= (Z(dist_c) + Z(rcos_c) + Z(dist_p) + Z(rcos_p))/4 \\ Z(x) &= (x - \bar{x})/sd(x) \end{aligned} \tag{11}$$

3.5. Summary

We use the British National Corpus (BNC) to construct context vectors for part-of-speech (POS) tags for English evidence. Two methods are used to compute the vector values: one estimates the probability distribution of c5 tags in the context of a target c5 (**Prob**), and the other is based on the mutual information between a context tag and a target tag (**PPMI**). We cluster the POS context vectors and measure their agreement with natural supercategories with two metrics, Cohen's Kappa and ARI. The computation of dissimilarity between groups of c5 tags is also introduced and will be used to evaluate the predictions of landmark linguistic theories. We have approached linguistic theories in the generativist paradigm that explore the concept of gradience in syntactic categories, and the nature of their approaches is that they have not always directly theorized about the questions that we study. Our interest here is in making contact with those frameworks, where possible, to identify where empirical consequences of our approach can be compared with explicit or implicit aspects of those theoretical approaches. Along the way, we isolate example categorizations of note arising from our method and relate them through examples to prior theoretical approaches. This treatment is not exhaustive. Indeed, it leaves open a great deal of future evaluations (not to mention the array of specific parameters in the method we have reported that could instructively be tested with other settings), including, for example, attaching our model of POS vectors with past proposals to build vector-based grammars (e.g., [71]).

4. Results

This section presents the results of our evaluation of POS context vectors using the two methods of computation, PPMI and Prob. We examine the internal clusters and the similarity relations implied between POS tags. These correspond to aspects of POS "meaning". As an initial anecdote, we recall an approach taken to demonstrate that word-level context vectors encode word meanings. A typical example is the demonstration of a systematic relationship between *king* and *queen*, mediated by the meanings of *male* and *female*:

$$cosine([V(\text{king}) - V(\text{male}) + V(\text{female})], [V(\text{queen})]) \approx 1. \quad (12)$$

We may see similar relationships among POS context vectors. For example, we also observe that the finiteness obtained by subtracting the vector of infinitive lexical verbs from the vector of finite lexical verbs is similar to the vector of the finite forms of *do* minus the infinitive *do*. Given a proper threshold, for example, the similarity level within a supercategory, groups of analogous pairs may arise with the POS context vector. We see this as a potential application of the POS context vectors and leave it for separate future exploration. Our main purpose here is to explore the qualities of the clusters as detailed above and in relation to what is anticipated by what landmark theories of linguistics suggest about syntactic categories.

4.1. Summary Statistics of the POS Context Vectors

As defined, the PPMI context vectors may contain a large number of zeros as a result of reducing negative mutual associations to 0, and the Prob context vectors may contain zeros if the occurrence of two c5 tags is not observed. Table 6 lists the number of zeros (and non-zero values) at each position and in the entire vector for each sort of construction, averaged over the vectors for 91 c5 tags. We do this to give an indication of the distribution of zero and non-zero values across the vector space.

As shown in Table 6, on average, 65% of the values in a PPMI context vector are 0 (54.82/84), of which 18% appear at the L1 position, 15% at the L2 position, 18% at the R1 position, and 15% at the R2 position. This aligns with the expectation that the restriction posed by a target tag is stronger in contexts closer to it. The paired Wilcoxon rank sum test confirmed the differences: the number of zeros at the L1 position is significantly higher than at the L2 position, and at the R1 position is significantly higher than at the R2 position ($p < 0.0001$ for both tests). Tests also show that the difference between the numbers at L1 and R1, or between the numbers at L2 and R2, are not significant ($p > 0.05$ for both tests). In contrast, on average, there is less than 1 zero in Prob context vectors, indicating a rich range of possible co-occurrences that can be observed in actual linguistic data.

Table 6. Summary statistics of zeros values in context vectors (Ln—n positions to the left of the target; Rn—n positions to the right of the target).

Vector Type	Count (Mean)	4 Positions	L1	L2	R1	R2
PPMI	zeros	54.82	14.89	12.69	14.87	12.37
	non-zeros	29.18	6.11	8.31	6.13	8.63
Prob	zeros	0.53	0.19	0	0.32	0.02
	non-zeros	83.47	20.81	21.00	20.68	20.98

Table 7 presents the means and standard deviations of the values in each sort of context vector. The computation for PPMI vectors is limited to the non-zero values. We first compute the mean and standard deviation for each context vector and report in the table the averages of the means and the averages of the standard deviations.

Table 7. Summary statistics of the vector values (Ln—n positions to the left of the target; Rn—n positions to the right of the target).

Vector Type	Statistics	4 Positions	L1	L2	R1	R2
PPMI	mean (non-zeros)	0.83	1.15	0.65	1.13	0.68
	sd (non-zeros)	0.76	0.94	0.53	0.95	0.59
Prob	mean	0.038	0.037	0.038	0.038	0.038
	sd	0.061	0.072	0.045	0.072	0.045

We observe that the PPMI non-zero values at L1 tend to be higher and more variable than the values at L2, and the same for the values at R1 and R2 ($p < 0.0001$ in paired Wilcoxon rank sum tests), while the differences between L1 and R1 and between L2 and R2 are not significant ($p > 0.05$ in paired Wilcoxon rank sum tests). In the case of Prob context vectors, the mean value at one position is not significantly different from the mean at any other position, as the low probabilities at each position are also included in the computation. However, standard deviations give a result similar to the PPMI values: the standard deviation at the L1 positions is significantly higher than at the L2 positions ($p < 0.0001$) and the same for R1 and R2; the differences between L1 and R1 and between L2 and R2 are insignificant ($p > 0.05$). Like the observations with zeros in Table 6, the tests are consistent with the expectation that given a target tag, there is less uncertainty with the positions closer to the target.

4.2. Cross Classification with Supercategories

Fifty-five context vectors of word classes (recall from above: punctuation symbols, unclassified items, and ambiguous tags are excluded in clustering) are clustered into nine

groups, the same as the number of supercategories of the 55 c5 tags for which the vectors are constructed.

Cohen's kappa and ARI between any two of the groupings (PPMI clusters, Prob clusters, and supercategories) are listed in Table 8. We include a weighted version of the two measurements: instead of counting the c5 types that were clustered, we count the number of instances labeled by those c5 types. This is equivalent to assigning each c5 type a weight of its proportion in the corpus in the computations. We use this to account for the fact that sizes of the c5 types in the corpus are not balanced.

Table 8. Cohen's Kappa and Adjusted Rand Index (ARI) between a set of clusters and the supercategories or between the two sets of clusters. The asterisks indicate significance after Bonferroni correction ($\alpha_1 = 0.05/12 = 0.004$).

	Cohen's Kappa		**ARI**	
	Types	**Tokens**	**Types**	**Tokens**
supercategories & PPMI clusters	0.474 *	0.545 *	0.295 *	0.667 *
supercategories & Prob clusters	0.401 *	0.577 *	0.189 *	0.563 *
Prob clusters & PPMI clusters	0.526 *	0.570 *	0.542 *	0.742 *

Cohen's Kappa and ARI between each set of clusters and the supercategories are significant after Bonferroni correction (the original α is 0.05; after Bonferroni correction, $\alpha_1 = 0.05/12 = 0.004$ for each test), demonstrating that the identification of the c5 tags and their categorizations partially incorporated the syntactic information encapsulated in the POS context vectors. The PPMI clusters agree with the supercategories to a higher level than the Prob clusters, except for Cohen's kappa computation over tokens rather than types. Though some distributional information is shared between the two relativization methods, the difference indicates that the two methods are not mutually redundant as representations of syntactic categories, empirically demonstrating the claim mentioned above (see Section 3.2). Furthermore, the difference in agreement levels suggests that the natural syntactic categories attend more to the mutual association between a target and its context than to the simple probability distribution of the context tags.

The agreement between the two sets of vector clusters is also significant and higher than the agreement with supercategories, except for Cohen's Kappa between Prob clusters and supercategories computed over tokens. This is consistent with the type of the information based on which the categorizations are obtained: context vectors only rely on the distributions of the tags in the pre-terminal fringes of phrase parsing trees, while the natural supercategories could utilize information beyond the flat distributional patterns, which may include, among others, semantic properties, nested syntactic structures, or word inflections.

We have noticed a particularly high value in Table 8 of Cohen's Kappa between Prob clusters and supercategories computed over tokens. This is a result of the imbalanced sizes of the POS tags in the corpus. In particular, Prob clusters and the supercategories agree on two of three preposition tags, while PPMI clusters agree with supercategories on one of the three, and the two sets of clusters agree on none. The three tags account for 5.5% of 55 tags, but the instances of prepositions account for 13.3% of all the linguistic items in the BNC. Thus, when computed over tokens, the one more preposition tag shared between Prob clusters and supercategories increased the overall agreement level. This imbalance in tag sizes also contributes to the generally higher values of the weighted agreement measures. For example, there are only five noun tags (less than 10% of all POS tags), but the supercategory of nouns is the largest category in terms of tokens (25.9% of total tokens are

nouns) and the corresponding cluster is the largest cluster with both PPMI and Prob context vectors (27.8% of the total tokens are in the PPMI cluster and 37.5% of the total tokens are in the Prob cluster). The number of types and tokens in the clusters and supercategories are included in Appendix C.

The agreement on each supercategory (under the assumption that each cluster corresponds to one supercategory) is listed in Table 9, computed with Cohen's Kappa. We see that the agreement varies among categories. Between PPMI clusters and the supercategories, the agreements on adjectives, conjunctions, pronoun, and nouns are above 0.6, meaning all all are significant and are usually deemed as having good agreement [72]; the agreements on adverbs and prepositions are insignificant (between 0.2 and 0.3), indicating heterogeneity in the distribution patterns of the items in these two categories; articles and interjections obtain negative values and also insignificant agreement. Prob clusters agree with supercategories significantly on five of the nine categories, but only the category of adjectives obtains a value greater than 0.6. The third row again demonstrates that the directional co-occurrence provides different information from the mutual association: the two methods show perfect agreement on adjectives and interjections, good agreement on articles and pronouns, but moderate agreement on nouns and verbs and disagreement on the rest. The category of verbs obtains moderate to fair agreement in all three comparisons. This is consistent with the observations that the category consists of several radically different sub-groups, for example, finite verbs, participles, and auxiliaries.

Table 9. Agreement on each syntactic category, computed with Cohen's Kappa. * indicates that the agreement is significant after Bonferroni correction ($\alpha_2 = 0.05/27 = 0.019$).

	ADJ	ADV	ART	CONJ	INTERJ	PREP	PRON	NOUN	VERB
PPMI & SynCat	0.739 *	0.262	−0.032	0.732 *	−0.030	0.295	0.673 *	0.899 *	0.380 *
Prob & SynCat	0.739 *	0.299	−0.028	−0.058	−0.030	0.543 *	0.460 *	0.453 *	0.421 *
PPMI & Prob	1.000 *	−0.062	0.641 *	−0.073	1.000 *	−0.066	0.769 *	0.528 *	0.555 *

Table 10 lists the items cross-classified by the two sets of clusters and supercategories. The blue and black tags are the cross-classification between PPMI clusters and supercategories. The gray and black tags are between Prob clusters and supercategories. That is, items in black are c5 tags that are shared by a cluster according to both the PPMI and Prob methods, so blue highlights clustering unique to the PPMI method and gray highlights clustering unique to the Prob method. The clusters identified the distributionally different items in a supercategory and cluster them with the items with which they share more syntactic contexts. Some of the shift in the grouping identities is supported by the practice of part-of-speech tagging in the BNC. For example, the tag CRD is used to tag cardinal numbers, which can function as nouns (e.g., *annual report of* ***1868***) or as determiners (e.g., *about* ***a*** *dozen*). Though combining two different uses, CRD is in the supercategory of adjectives. The resemblance between CRD and nouns identified by both PPMI and Prob clusters indicates that the nominal use of numbers is more frequent than its prenominal use in the corpus. Another example is the clustering of AVQ and the pronominal c5 tags. AVQ labels adverbs such as *when* and *where*. Both interrogative and relative uses of the adverbs are labeled as AVQ, as in (3). (The examples included in this paragraph are extracted from the BNC and its tagging guidelines (https://ucrel.lancs.ac.uk/bnc2/bnc2guide.htm (accessed on 13 April 2026)), distributed by the University of Oxford on behalf of the BNC Consortium. All rights in the texts cited are reserved. In the square brackets in Examples (3) to (5), we included the text identifier and sentence number for each example.) We use a set of regular expressions to examine the words that are labeled as AVQ in the corpus and found that 61.48% of all the AVQs (149357 of 242926) were identified as initiating an

interrogative or relative clause. (The code used to identify the relevant cases is included in the online data repository.) DTQ labels the wh- (interrogative) determiner pronoun and relative pronoun (e.g., *what* and *which*, as in (4)) and PNQ labels interrogative and relative pronouns (e.g., *who* and *whoever*, as in (5)). Our regular expressions identified that 80.54% of all the instances of DTQ (515,323 of 639,864) and 89.57% of all the instances of PNQ (191,394 of 242,926) were used in the form of Examples (4) and (5). From a semantic point of view, AVQ covers words like *when* and *where*, which denote time and place and may be taken as adverbs, while *what* and *who* are pronouns as they can substitute a noun phrase, but the preponderant uses of the instances suggest a syntactic similarity between AVQ and the two other tags in the pronominal category. This similarity is captured by PPMI and Prob method, both placing the three tags in the same cluster rather than clustering AVQ together with AV0 (general adverbs such as *often* and *well*) and AVP (adverb particles such as *off* and *up*). In addition, XX0 (the negation *not*) and TO0 (the infinitive marker *to*) are also separated from the other members in their respective supercategories and placed in the same category by both methods for their distinctive behavior of preceding an infinitive verb. We discuss the categorization of the items involving "to" or negation below.

Table 10. Cross classification: cluster index vs. super categories. PPMI clusters: blue & black; Prob clusters: gray & black. Items in black are c5 tags that are shared by a cluster according to both the PPMI and Prob methods, so blue highlights clustering unique to the PPMI method and gray highlights clustering unique to the Prob method. The total number of c5 tags included here is 55. Ambiguous tags (30 in total), punctuation symbols (4 in total), and unclassified items (2 in total) are excluded.

Super Category	Cluster Index								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
ADJ	AJ0 AJC AJS DT0 ORD							CRD	
ADV		AV0 AVP XX0				XX0	AVQ	AV0 AVP	
ART	AT0								
CONJ				CJT CJC CJS		CJT	CJC CJS		
INTERJ							ITJ		
PREP		TO0		PRF PRP		TO0 PRF PRP			
PRON	DPS	PNX					DTQ EX0 PNP PNQ PNI	PNI PNX	
NOUN								NN0 NN1 NN2 NP0 ZZ0	

Table 10. *Cont.*

Super Category	Cluster Index								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
VERB		VDG VDN VVG VVN VHN	VDB VDD VDZ VVB VVD VVZ	VBG VBN VHG	VBI VDI VHI VVI	VM0 VHN		VDG VDN VVG VVN	VBB VBD VBZ VBG VBN VHB VHD VHZ VHG VM0 VVB VVD VVZ

Example 3.

a. When (AVQ) do your courses start? [A0F.3117]
b. . . . if you let me know when(AVQ) the police are called in. [BMU.2291]

Example 4.

a. Which (DTQ) country do you live in? [A7N.979]
b. Give 4 details which(DTQ) should appear on an order form [HBP.417]

Example 5.

a. Who (PNQ) is this speaking? [CN3.258]
b. My social worker who(PNQ) gave me advice asked... [FR5.1420]

Within the clusters arising from PPMI vectors, negation (XX0) clusters with infinitive markers (TO0) and modals (VM0). Within Prob clusters, negation patterns with infinitive markers. We note that linguistic theory has analyzed the infinitive "to" as a kind of auxiliary [73–75]. With respect to negation, first notice a property of classical negation demonstrated in Table 11. Within this, the symbol $\perp$ is taken to be a proposition that is always false (i.e., a contradiction); the proposition p may be true or false. The usual semantics of the material conditional is such that $p \rightarrow \perp$ is truth-functionally equivalent to $\neg p$ (both statements are true or false in exactly the same states of affairs).

Table 11. Truth table demonstration of the classical bivalent logical relationship between negation and implication.

$\perp$	p	$\neg p$	$p \rightarrow \perp$
0	0	1	1
0	1	0	0

Thus, negation may be thought of as an implicit conditional (one could devise a language that lacked an explicit word for negation if it has a way to form material conditionals and express contradictions). Following reasoning such as that described by [76], negation may be characterized as modal, a covert epistemic operator.

Example 6.

a. If p is true then it is known that a contradiction exists
b. If p is true then it means there is a contradiction
c. If p is true then there is a contradiction

The first conditional, Example (6*a.*), unpacks the negation as conveying epistemic force; (6*b.*) conveys indication. Even without such explicit modality in the consequent of the conditional, the bare conditional truth, functionally equivalent to $\neg p$ (6*c.*), removes clear assertion of falsity or truth for the content of the antecedent—"if p then q" asserts neither p nor the negation of p—in the context of utterance. Certainly, explicit negation is a non-veridical operator: $\neg p$ does not entail p. Such semantic reasoning places "not" in the same conceptual space as temporal "will", deontic "ought", and so on. Further, negation has been provided a semantics as a modality [77]. Reflecting on the usual Kripke semantics for modal logic as well as the intuitions due to Kratzer about the multiplicity of model orderings of possible worlds, one may see that negation as a modality requires an anti-reflexive accessibility relation [78,79]. We concede that modal analysis of negation is disputed [80]. Ref. [76], in an aside (p. 449), supposes that negation is not modal; however, as described here (and there), explicit negation satisfies diagnostics of modality. We do not deny that the syntactic patterns differ: (7*a.*) is grammatical, but (7*b.*) is not.

Example 7.

a. Lee may arrive
*b. *Lee not arrive*

Rather, here we point out that "not" sometimes does align with clear modals (see (8*a.*) and (8*b.*)).

Example 8.

a. Val may not participate. Sal may.
b. Val may advance. Sal not.

In any case, the relevant cluster reveals that some syntactic contexts of modals are shared with explicit negation, and sufficiently so to be cluster-mates, given their own similarity and the intra-cluster similarities of items in other clusters. Thus, because of both distributional facts and semantic considerations, we see a kind of elegance to the clusters that link explicit negation to infinitival "to" and modals.

4.3. Compare POS Context Vector Similarity Relations to Predictions of Binary Classification Systems

The results of the last section may be understood as a comparison between POS vector representations and linguistic theory regarding *membership* of particular categories (the c5 inventory) and supercategories. Here we consider *gradients of similarity* that may be seen as predicted consequences of linguistic theories (with some of the conflicting alternatives having been noted earlier; see Section 2.2). We compute the dissimilarity among the four categories described with the binary features in generative grammar (that is, adjective, noun, verb, and preposition) and evaluate to what extent the least similar category pairs estimated by context vectors agree with the ones predicted by the binary systems.

The granular categories in the same supercategory may be vastly different from each other with respect to the context they can appear in; for example, content verbs such as *eat* and *forget* can immediately precede a noun phrase or a relative clause, but auxiliaries require a content verb in its bare form to follow them in most cases. To highlight the

dissimilarities between categories, we use the c5 tags that are core to the supercategories (i.e., the intermediate categories suffixed with "_M" identified in Section 3.1, rather than the full supercategories or the clusters of c5 types obtained by the POS context vectors), as listed in Table 12.

Table 12. c5 tags of each category in the binary systems.

Main-POS	c5
NOUN	NN1, NN2, NN0
VERB	VVD, VVG, VVN, VVI, VVZ, VVB
ADJ	AJ0, AJC, AJS
PREP	PRP, PRF

We present the results in Table 13. The four categories in the binary feature systems are in the first column. The second column lists at each row the category that are the least similar to the category in the corresponding row in the first column, predicted by the natural extension of the sketch from Chomsky (see Table 1 in Section 2.2), and the third column lists the predictions derived from the system proposed by [50] (see Table 2 in Section 2.2). The least similar categories computed with context vectors are in the last two columns. Given a category, the category that is the least similar to it is selected from the remaining three categories. For example, in the first row of Table 13, among the three categories of verbs, prepositions, and adjectives, the verb may be seen as the least similar to the noun, according to the natural extension, but the preposition is estimated to the be least similar among the three by both types of POS context vectors, which is the same as the prediction by the alternative system by [50]. Categories on which the vector estimates agree with the natural extension are marked in blue, and those on which the vector estimates agree with the alternative system are in bold.

Table 13. Predictions and estimates of each syntactic category's least similar (most distant) category, derived from binary feature systems and computed with POS context vectors. Blue marks the agreements between the two vector approaches and the naturally extended binary system (see Table 1), and bold marks the agreement between the two vector approaches and the alternative binary system (see Table 2).

Category	Prediction from Table 1	Prediction from Table 2	Estimate by PPMI Vectors	Estimate by Prob Vectors
nouns	verbs	**prepositions**	**prepositions**	**prepositions**
verbs	nouns	**adjectives**	nouns	**adjectives**
adjectives	prepositions	verbs	prepositions	preppositions
prepositions	adjectives	nouns	adjectives	adjectives

We see in Table 13 that the vector estimates show a preponderance of agreement with the predictions of the binary systems: the least similar category computed with the vectors agrees either with the natural extension or with the alternative system. PPMI vectors agree more with the natural extension (same results on three of the four categories) and the Prob estimates split evenly between the two systems. The binary feature systems present a symmetric least similar relation. When a rich range of contexts is considered, as with the vector approach, this symmetric relation is preserved only in the adjective–preposition pair: both relativization methods show that the preposition is the least similar to the adjective, and the adjective is the least similar to a preposition. In comparisons involving the category of nouns and verbs, we see that sorts of vectors estimate that the preposition among three categories is the least similar to the noun, though the adjective instead of the noun is the

least similar to the preposition; the behavior of verbs is more variable: either nouns or adjectives may share the least syntactic constraints with verbs.

4.4. Refer to the Gradience of Notional Features

The representation of syntactic categories with **P** and **N** in the notional theory by [54] provides another set of predictions of similarities between categories. Table 14 reproduces Table 3 and lists for each category the member c5 tags. (We restrict the range of auxiliaries to modal verbs, as the auxiliary use of *have*, *be*, and *do* is not separated from their usage as common verbs. It is also possible that the words in one c5 category belong to more than one notional category. For example, the word *proposal* is a nominalization and is in the second-order notional category {(N;P);(P;N)}, as the word presents properties of verbs and nouns at the same time, and the nominal properties preponderate [7]. However, the word is labeled with NN1 in the BNC, the same as other singular common nouns, which usually belong to the category of {N;P}. Thus, the c5 tags listed in the table function as an approximation of the membership of the notional categories.) According to the table, it may be predicted that modals are more similar to infinitives than to nouns, as the difference in the former pair is one unit of **P**, while in the latter it is three units of **P**. Similar predictions can be made for other pairs of categories. In this section, we test the Spearman's rank correlation between the notional difference of a c5 pair (difference measured with notional features) and the syntactic similarity of the pair estimated with POS context vectors. As noted in Section 2.3, the functor is different from the other five categories in the table in that it breaks the inter-definition of the two features, so we compute the notional difference in different methods for the pairs that contain a functor and the pairs that do not, and test the correlation separately. Pairs of c5 tags from the same notional category are excluded from the correlation tests.

Table 14. c5 tags of each notional category. **VM0**: modal auxiliary verb; **VVI**: the infinitive form of lexical verbs; **AJ0**: general adjective; **AJS**: superlative adjective; **NN0**: common noun, neutral for number; **NN1**: singular common noun; **NN2**: plural common noun; **NP0**: proper noun; **PNP**: personal pronoun; **PRP**: preposition, except *of*; **PRF**: preposition *of*; **CJS**: subordinating conjunction.

Notional Category	Weights of P and N	Word Class	c5 Tags
{P}	4P, 0N	modal	VM0
{P;N}	3P, 1N	infinitives	VVI
{P:N}	2P, 2N	adjectives	AJ0, AJS
{N;P}	1P, 3N	nouns	NN0, NN1, NN2
{N}	0P, 4N	names	NP0, PNP
{}	0P, 0N	functor	PRP, PRF, CJS

Table 15 lists the correlations between the two vector measurements and the absolute difference in **P** (Δ_P) for the c5 pairs that do not contain a functor. **N** is not used here because the absolute difference in **N** equals Δ_P for these pairs. The tests with both measurements of PPMI vectors and with the cosine similarity of Prob vectors are significant. The direction of the significant correlations is as expected: as the difference in **P** increases, the syntactic contexts of two categories become less similar. We see that the identification of **P** for a class of words made use of the constituent, head-modifier structure of clauses (see Section 2.3). One may think of the connection between this structural information and the POS sequences underlying the context vectors in terms of the connection between phrase parsing trees and the pre-terminal fringe of POS tags. More than one phrase-parsing

tree may be possible for a particular sentence, given only the pre-terminal POS sequence, but the sequence nevertheless limits the range of potentially suitable trees. The significant correlations obtained with the PPMI context vectors empirically demonstrate that the vectors are capable of capturing the structural information considered by the notional theory. The magnitudes of the correlations obtained with the PPMI vectors are stronger than those obtained with Prob vectors, indicating that the mutual association between POSs tags models the structures better than a directional association, which is summarized with Prob context vectors.

Table 15. The Spearman rank correlation ($n = 31$) between the absolute difference in **P** (Δ_P) and POS context vector measures ($dist_v$ denotes Euclidean distance between two vectors and cos_v denotes cosine similarity between two vectors) for pairs of c5 tags, excluding the functor category. The asterisks indicate significant correlations after Bonferroni correction ($\alpha_3 = 0.05/4 = 0.0125$).

	PPMI Vectors		**Prob Vectors**	
	$\boldsymbol{dist_v}$	$\boldsymbol{cos_v}$	$\boldsymbol{dist_v}$	$\boldsymbol{cos_v}$
cor with Δ_P	0.492 *	−0.494 *	0.342	−0.488 *

In Table 16, we correlate the vector measures with the notional distance ($dist_{N,P}$) for pairs of c5 tags that contain a functor. We form a two-dimensional vector for each of the c5 tags, one dimension for the **P** feature and the other for the **N** feature (thus, we have, for example, $\langle 0, 0 \rangle$ for PRP and $\langle 4, 0 \rangle$ for VM0), and $dist_{N,P}$ is computed as the Euclidean distance between two two-notional vectors. None of the correlations are significant, indicating that along the **P** and **N** gradience, treating functors as not presenting any of the two features does not align with the syntactic behavior of the functional words. Ref. [7] observed that some prepositions share properties of items from lexical categories. For example, the preposition phrase *on the shelf* headed by *on* in *Put the book on the shelf near John's photograph* (taken from [7] (p. 302)) can be replaced by *here* in a way similar to a noun phrase replaced by a pronoun, and it licenses modifiers (*near John's photograph* modifies *on the shelf*). According to Rauh, prepositions can be subcategorized into three classes: a class that mainly exhibits properties of functional items, a class that exhibits properties of other lexical categories, and an intermediate category. Spatial, temporal, and scalar prepositions form the second category. Rauh argues that Anderson ignores the lexical properties of some of the prepositions by assigning prepositions into the functor category without restriction. In the analyses we present here, following the POS labeling in the BNC, the prepositions we used do not contain the adverbial use of preposition-like words, for example, *up* in *pick up* (which is tagged as an adverb particle) or *before* in *I have visited there before* (which is tagged as a general adverb). However, the category does merge the heads of preposition phrases (e.g., *in* in *in the room*) and some case markers (e.g., *to* in *give it to him*), and merges the lexical functors and functional functors as Rauh categorized. From this perspective, our results supply empirical evidence for the argument from [7] that treating functors as devoid of any notional features could undervalue the properties of this group of linguistic items.

Table 16. The Spearman rank correlation ($n = 27$) between the notional distance ($dist_{n,p}$) and POS context measures ($dist_v$ denotes Euclidean distance between vectors and cos_v denotes cosine similarity between vectors) for pairs of c5 tags which contain one functor tag and a tag from other categories. None of the correlations are significant after Bonferroni correction ($\alpha_4 = 0.05/4 = 0.0125$).

	PPMI Vectors		**Prob Vectors**	
	$\boldsymbol{dist_v}$	$\boldsymbol{cos_v}$	$\boldsymbol{dist_v}$	$\boldsymbol{cos_v}$
cor with $dist_{N,P}$	0.299	0.168	−0.264	0.197

4.5. Summary

Relations between POS categories are computed with POS context vectors and compared with the ones that are evident in linguistic treatment of syntactic categories. We evaluate the agreement between the clusters of POS context vectors and the natural supercategories of the POS tags. Significant agreements are observed, and implications of the differences are analyzed. We also compared the vector estimates with predictions that are derived from two approaches to syntactic categories in linguistic theories: the binary feature system in generative grammar and the notional representation by [54]. We see both agreement between the vector estimates and the predictions and aspects that the categories' history of use may suggest differently.

5. Discussion

Context vectors are constructed on part-of-speech (POS) sequences in this work. Given a corpus annotated with POS tags and a target POS tag, we note the contexts of the target, that is, n items to the left and n items to the right of each occurrence of the target POS, and record for the target the value of a POS type occurring at each of the contextual positions relative to the target. The values are count-based, and we used two methods to relativize the raw counts: one is to divide the raw counts by the frequency of the target, estimating the conditional probabilities of observing a context tag given a particular target, and the other is to compute PPMI, based on the information-theoretic concept of mutual information. In this work, we include in a vector the dimensions of the 20 most frequent context tags in the BNC, and a complementary POS type that aggregates the co-occurrence values of the less frequent context items.

A POS context vector may be taken as an approximation of the distributional meaning of a POS tag, and we use this computable representation to categorize the tags and estimate the similarity between syntactic categories. In general, we find that the results obtained with context vectors based on mutual information agree more with the practices within linguistic theories. The theory of dependency locality argues that the linear order of word pairs in dependency relations should be minimized [81] to facilitate language processing (notwithstanding evidence that people frequently use language with qualities that manifest inefficiency [82]). Ref. [83] demonstrated that word pairs and pairs of POS tags in direct dependency relations are higher in mutual information than pairs in grandparent-dependent relations or in sibling relations. The relation between mutual information, word order, and the dependency structure suggests that the context vector based on mutual information may encode the dependency structure better than the method based on conditional probability, which contributes to the higher agreement between the PPMI context vectors and linguistic investigations.

5.1. POS Context Vectors as a Representation of POS Tags

Context vectors applied to word sequences are frequently used to represent words in multi-dimensional spaces. The relation between the distributional pattern of a word and its meaning was noted by [22]: certain linguistic elements co-occur with certain other linguistic elements, and the difference in distributional structure correlates with the difference in meaning. It is convenient to generalize this vector representation to other sequences that have elements that are not randomly organized. By collecting the environments of a particular item, i.e., the distribution of the elements that co-occur with the target, we obtain a representation of the target based on its history of use. In this work, we extend this representation to sequences of part-of-speech tags.

It is useful to reflect on what a POS context vector stands for. In the case of word context vectors, the vectors are frequently taken as an approximation of the meaning of

the words, and the reasoning about the similarity (or differences) of vector representations for items strongly aligns with synonymy judgments of word meaning. Synonymy is an empirical consequence of the relative proximity of the contexts of use [84], while context vectors provide a proper summarization of this contextual information [85]. The method of context vectors also parallels lexicographical work. Lexicographers identify entries for a word, each corresponding to a set of language samples among which the word is used in a particular way. Context vectors also make use of contexts to distinguish words, though they merge the contexts that may be differentiated by a lexicographer and provide an overall representation. In the case of part-of-speech tags, we argue that it is still possible to treat a POS context vector as an approximation of certain aspects of the meaning of a POS, which are also empirically determined by contexts of use. The structuralist practice assigned words to syntactic categories with certain criteria; for example, adjectives and nouns can both precede verbs, but a noun can immediately precede a verb while a component is usually required between an adjective and a verb. From this distributional perspective, a POS tag is partially defined in terms of the tags with which it could co-occur, and the definition is approximated by a POS context vector. A POS context vector may also be understood in relation to the meaning of the POS sequences lifted from word sequences. A POS sequence of a sentence is the pre-terminal fringe of the tree structure of the sentence, and it poses certain constraints on what trees could be possible. These constraints may be treated as constituting the meaning of the sequence. If representing this meaning as phrase structure nodes with their probabilities, one may expect that a certain correlation can be identified between the node probabilities and the values in the context vectors of the component POS tags. Thus, the meaning of the POS sequence is broken down to the meaning of its component tags (and potentially the inter-tag relations), which is represented as a context vector. As mentioned in Section 2, a syntactic category can be defined with notional features or abstract semantic properties, and the meaning of a POS sequence may be understood from other perspectives. It remains unclear what is captured by a POS context vector with respect to these aspects of meaning, but the meaning of a tag from a distributional perspective may be seen as approximated by a context vector based on its history of use.

Vector representation of POS tags has certain advantages: it can exhaustively account for the salient linguistic phenomena that are available in a corpus and is consistent in the treatment of POS tags. The exhaustive design serves as one solution to the issue of methodological opportunism raised by [21]; i.e., when there is more than one criterion that may be deemed possible for assigning a word to a category, researchers may choose a subset of the criteria that fits their purpose the most, which may bias the categorization by a particular hypothesis and ignore certain patterns presented in linguistic data. Contrary to the selective approach, POS context vectors utilize the observed POS sequences and weigh the usage of a POS tag by frequency, thus forming a relatively comprehensive representation with distributional facts. Another issue that has been pointed out by linguists is that a practice or the definition of a feature may shift through the course of linguistic analyses. Ref. [7] noted that in the notional theory by [54], the feature of predicativity is described as event-specific and typical to verbs, and the infinitive form of lexical verbs is described as high in the feature, but the auxiliaries are treated as the highest in the features because of their inflection for finiteness, instead of event-relevant features. The construction of the POS context vector is not affected by such issues of inconsistencies by design.

5.2. Vector Representation and Practice Within Linguistic Theories

As noted in the above section, POS context vector approximates the meaning of syntactic categories from the perspective of their frequency of use in relation to other

categories, while the natural and theoretic linguistic arguments may utilize information of various sources. Thus, we evaluate to what extent two lines of investigation agree with each other and what insights the structural representation provides.

We cluster the POS context vectors and evaluate the agreement between vector clusters and a natural categorization of the granular POS tags. We observe significant agreement between the clusters and the natural categories: Cohen's Kappa of 0.474 between the PPMI clusters and natural categories and 0.401 between the Prob clusters and natural categories. The agreements measured with ARI are also significant. Conventional interpretations of Cohen's Kappa would judge *moderate* agreement between the supercategories and both PPMI clusters and Prob clusters, for both absolute types and token-count-weighted types. We take this to indicate that POS context vectors of either sort succeed in capturing empirical consequences of linguistic intuitions inherent in the initial category structure (supercategories, c5 subcategories, and c5 categories applied to individual lexical tokens). We see a large element of the disagreement following from the force of evidence supplied by the substantial amount of data embodied by the BNC, through which support for equivalences and distinctions are readily available but not foreseen in the flow of linguistic analysis. The relation between data-driven and intuition-driven analysis is validated in the significant values that obtain for ARI: the analysis of Cohen's Kappa holds the linguistic theory to the claim that some selections of c5 categories all belong together, a different set for each supercategory; ARI devolves this to claims related to pairs of c5 categories (in the same supercategories). Additional disagreements emerge with the ARI metric, bringing in the off-diagonal pairs, but the agreement remains at a significant level. We do not claim that all of the disagreement amounts to genuine linguistic knowledge made plain by the data; rather, some follows from measurement details (as is evident in PPMI and Prob disagreement), such as the vector dimensionality and data treatments that have been described.

Where the clusters of POS tags do not agree with the syntactic categories, we observe results consistent with the decisions made in the POS tagging process in the construction of the BNC. We also observe that the disagreements appear to be similar to records of frequent POS tagger errors and changes in word categorizations. Ref. [86] recorded that when a tagger incorrectly tagged a preposition, the most frequent incorrect tag is subordinate conjunctions (and vice versa). In PPMI clusters, subordinate conjunctions and prepositions are placed in the same cluster, and in Prob clusters, the subordinate *that* and prepositions are in the same cluster. Ref. [87] noted that in traditional grammar, articles (*a*, *the*), demonstratives (*this*, *that*, *these*, *those*), possessives (*my*, *your*, *his*), quantifiers (*every*, *some*, *many*), and numbers (*one*, *two*) are conflated into the broader class of adjectives. With our vector representations, the first four of the five classes are placed together in the same cluster as general adjectives, though cardinals are in the cluster with nouns.

The clustering also suggests an adjustment to the traditional categorizations of granular POS tags. Model auxiliaries are usually taken as a subset of verbs, the infinitive marker *to* is a preposition in the BNC tagset, and the negation *not* belong to the family of adverbs. POS context vectors based on mutual information identified the three POS tags as a distinct category that is not available to other tags. This classification agrees with the treatment by [73,75] of *to* as an uninflected verb that takes a verb in its bare form as a complement. We have also provided evidence in Section 4.2 that, following the reasoning by [76], negation may be characterized as a modal. The result of grouping the three tags together by context vectors is largely informed by the fact that infinitives instead of other forms of verbs could follow the instances of the tags. Though modal auxiliaries and *to* or *not* may be used differently, for example, modals can initiate interrogatives while the other two cannot, the clustering result shows that when the various contexts are integrated

and weighted by frequency, the three items strongly resemble each other, and analyzing them in terms of taking infinitives as complements may benefit further syntactic analysis. As an approach anchored in substantial linguistic data and arguably a solution to the issue of methodological opportunism raised by [21], we see here that the POS context vector provides a distinct perspective of viewing the categorization of items and could be of help in discovering distributional patterns of syntactic categories.

Some linguistic theories represent categories of linguistic items with a set of features, which allows certain ordering of the categories. In this work, we view the ordering as predictions of the level of similarity between the categories and compare the predictions with the estimates obtained with POS context vectors. The binary feature systems in generative grammar characterize a syntactic category with the values of two abstract features N and V, and we evaluated whether the two categories with opposite values of the two features (e.g., [+N, +V] vs. [−N, −V], or [+N, −V] vs. [−N, +V]) would be the least similar to each other, compared to pairs of two other categories. Two binary representations are included here: a natural extension of the original representation proposed by [49] and an alternative characterization proposed by [50]. We found that neither of the context vector constructions agrees fully with either of the binary systems. In particular, the context vector estimates break the symmetric relation between the categories—one category among several categories being the least similar to another category does not indicate that the latter among other categories is the least similar to the former.

One syntactic category of particular interest in this work is the category of prepositions. Prepositions have been documented for their flexible behavior and resemblance to words of other categories, and subcategorization or adjustment to what use cases should be labeled as prepositions has been suggested [17,88,89]. It follows that employing a subcategorization at a particular granularity will omit detailed distinction to a certain extent, and the context vectors constructed based on the POS tagging may serve as an averaged representation of the underlying variable syntactic behaviors. Here we follow the distinction made by the BNC and compare the implication we draw out from the notional representation proposed by [54] with the distributional patterns that are captured by the POS context vectors. We see that disharmony between the two representations arises mainly from functors, of which prepositions form a main subset. As described in Section 2.3, we were able to show that a category that at least merges two types of prepositions, words that function as the head of a preposition phrase and words that assign cases, as well as subordinate conjunctions (which is also treated as functor by [54]), do not line in a gradient with other categories as what was implied in the notional representation. This result is subject to the categories we used here. With more granular or different categorizations of prepositions, we might be able to see significant correlations between the vector approach and the notional representation and thereafter be able to infer what prepositions in particular syntactically behave as their position in the notional gradience. It is also possible to generalize the co-occurrence of POS tags in the context vectors to constituents (e.g., at the phrase level) or head-dependent relations. Here we see that the POS context vector functions as a tool that is able to draw representations for syntactic categories in actual, large corpus and allows a numerical evaluation of the similarity and dissimilarity between syntactic categories.

5.3. Limitations

5.3.1. Interpretability and Dimension Reduction

We use the count-based context vectors in this work to illustrate that the distributional approach to categorizing words may be implemented with vectors, and the co-occurrence patterns between a POS tag and its neighboring tags provide meaningful information on the identity of the tag, at least to the extent that the tags clustered together reflect

the linguistic intuition in categorizing them. The count-based vectors benefit from their explicit dimensions: when two vectors that are not usually treated as belonging to the same supercategory are clustered together, one can examine the systematic similarities between the two target tags by directly investigating the dimensions at which both tags obtain high values, for example. Studies with word vectors have shown that dense vectors or dimensionality-reduction algorithms may output word vectors that could perform better on intrinsic evaluations (for example, identifying similar word pairs or synonyms) [46,90]. Analogously, dense vectors for POS tags or vectors obtained after dimensionality reduction may also obtain higher agreement scores with respect to the supercategories (as we examined in this work) or suit certain tasks better. However, such algorithms present certain difficulties in explaining the dimensions of output vectors. Typically, the explanation of dense word vectors relies on finding the words that obtained high values in one dimension [45,91], or isolating the dimensions responsible for a particular semantic property, for example, negation or polarity [47]. This type of explanation relies on the meaning of the word groups or the tasks that are known to language users, while in terms of syntactic categories, it is of interest to explore what brings two categories together and whether such distributional evidence merits certain adjustments in the categorization of the granular categories. In this case, inherently interpretable dimensions, such as the dimensions that are explicitly coded with contextual tags and their positions in this work, suit these needs.

The complementary type we included in the vector reduces the co-occurrence with some specific tags to a value of "to what extent does the target tag co-occur with the collection of tags that are not frequent in a corpus". Though certain information is still available from the complementary dimensions (in our case, four dimensions from the complementary type), this gives rise to the question of whether meaningful, discriminative information with the less frequent tags is evened out in the aggregation. We include in Table A3 in Appendix B a preliminary test of reducing the 364 actual dimensions to 44 dimensions with principal component analysis (PCA) [92] while preserving 90% of the variation in the data, which obtained agreement results similar to the one we obtained with the count-based vectors, indicating that it is not likely that meaningful information was ignored by averaging the less-frequent tags. However, these are only indicative results with one algorithm. Other available dimensionality-reduction algorithms and the question of how to interpret the output dimension may be explored in future work.

5.3.2. Quantitative Syntactic Behavior Support of the Vector Results

We argue that a POS context vector functions as a set of summary statistics of the distribution pattern of the target tag, and thus the comparison of POS context vectors provides quantitative measurements of syntactic similarity between two tags. However, we also noticed that, though we have evaluated the performance of the context vectors with respect to the groups of tags arising from the theory of linguistic syntax, considerations from semantic theory are also noted in the c5 tagging process and the grouping of the c5 tags to more coarse-grained categories. Thus, the agreement between the context vector clusters and the supercategories may not serve as a purely syntactic metric, and in this work, we do not strive for maximal agreement scores. Instead, we aim to show that the significant agreement between the clusters and the supercategories suggests an overlap in the information underlying the two categorization approaches, i.e., the syntactic aspect of the word classes, and this syntactic aspect is captured by count-based context vectors.

The limitation here is the lack of an independent quantitative syntactic metric against which the performance of the POS context vectors can be evaluated. As a first step towards such evaluation, we used a set of regular expressions to extract the relative and interrogative use of three tags and illustrated the agreement between the syntactic structures and the

performance of the POS context vectors, but the approach remains limited—for example, the use of cardinal numbers similar to nouns may not be easily separated from their adjective use with regular expressions. It is possible that such a metric may be developed from constituency parsing, for example, whether two POS tags are similar in terms of their appearance in similar constituents or from dependency parsing and resulting dependency relations. It may also be possible to rank syntactic similarity for pairs of POS tags with language users' judgment, the same as the semantic similarity test sets with word vectors. The development of the metrics in future studies will be helpful to examine what syntactic information is encoded in the context vectors and what is not encoded in the vectors.

5.4. Summary

In summary, the approach of POS context vectors may be understood as approximating the distributional meaning of syntactic categories, which benefits from its design of being exhaustive and consistent. In this work, we make use of the linear order of POS tags, leaving the constituency structure or other nested structures for further research. Experiments comparing the relations between syntactic categories estimated with POS context vectors and informed by natural supercategories or landmark linguistic theories yield significant agreements as well as discrepancies. The differences between the two lines of investigation remain interesting: the POS context vectors could suggest certain reorganizations of the categories of POS tags, and though linguistic theories have organized syntactic categories with two-feature systems, the relation between the categories may be more complex if one tries to accommodate the various syntactic behaviors of the categories.

6. Conclusions

We present a vector representation of POS tags in this work. Given a corpus of POS tags lifted from the word sequences, context vectors can be constructed as a representation based on the category-tag distributions of category "meanings". We have demonstrated two methods of computing the syntactic context vectors, one based on probabilities anchored in relative frequencies of occurrence and one based on mutual information, using the 111 million words of the British National Corpus as the data source. We have demonstrated relationships between the context vector representations and operationalizations of approaches to syntactic category structure as evident in linguistic theory. We have noted instances of harmony and disharmony between predictions from the linguistic theory of syntactic categories and what is visible for either approach to context vectors. We see that the mutual information method of context vector computation is more often in harmony with the expectations that arise from linguistic theory. However, we also note that the context vector and linguistic theory approaches, in the formulation explored here, are anchored in a fully shared ontology: both linguistic theory and the context vector approaches commit to the classification of words with syntactic categories, but linguistic theory also assigns categories to word sequences and phrasal units, and that is missing from the version of context vectors we report here. The label for a phrasal unit may agree in some respect with the label on some words within it (for example, a noun phrase may be thought of as a nominal and contain within it a word that is well labeled as nominal) but disagree on others (for example, a noun phrase may contain a preposition within it, but a preposition is not generally well characterized as a nominal). Thus, some disagreements between predicted and computed similarity relations are to be expected. Some disagreements may also follow from differing sources of estimates: categories from linguistic theory are informed by expert impressions of likely frequencies in actual data and are able to anticipate uses that are logically possible but perhaps never tested, as well as constrained by impossibilities; context vectors are anchored in the counts of items that actually appear, and impossible items

are indistinguishable from unattested items. Given that both approaches are informed by textual distributions (even with differing methods of estimating counts) and the fact that the context vectors avail of expertise from linguistic theory in the initial classification of words with part-of-speech categories, some level of agreement is also to be anticipated. We have shown both similarities and differences, as discussed above. More profound tasks remain in quantifying the extent to which observed differences are tantamount to noise, which are predictable differences, and which arise from syntactic context vectors encoding information not hitherto incorporated into linguistic theory.

While others explore the ramifications of the frequency of association of part-of-speech categories (e.g., [93]) or vectors derived from clusters of word context vectors [2], or even propose vector-based grammars [71], we do not know of other research that proposes or uses part-of-speech context vectors as we have described here. We have demonstrated that the concept is useful in illuminating the theory of syntactic categories for natural language and anticipate that others will join us in extending this concept, not least by applying this approach to the part-of-speech tag sequences appropriate to very large corpora for other natural languages.

Author Contributions: Conceptualization, C.V.; methodology, X.M. and C.V.; software, X.M. and C.V.; validation, X.M. and C.V.; analysis, X.M. and C.V.; data collection, X.M. and C.V.; data analysis, X.M. and C.V.; literature review, X.M. and C.V.; initial drafting, X.M.; editing, X.M. and C.V. All authors have read and agreed to the published version of the manuscript.

Funding: We are grateful to the China Scholarship Council and Trinity College Dublin which have funded the PhD studies of X.M., supervised by C.V.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used here is that of the British National Corpus (BNC). The PPMI and Prop context vectors and other data that arise from our processing of the BNC are available via https://clg.scss.tcd.ie/SyntacticCategoriesAndPOSContextVectors/.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript (see Appendix A for the full list of POS categories):

POS	Part of Speech
BERT	Bidirectional Encoder Representations from Transformers
BNC	British National Corpus
PCA	Principle Component Analysis
PPMI	Positive Pointwise Mutual Information
Prob	Probability estimated by relative frequency
ARI	Adjusted Rand Index
P	Predicativity
N	Referentiality

Appendix A. List of Part-of-Speech Categories

Table A1. Mapping between SPOS, intermediate class, and c5 tags.

SPOS	Intermediate Class	c5
ADJ	ADJ_M	AJS
		AJ0
		AJC
	ADJ_O	CRD
		DT0
		ORD
ADV	ADV_M	AV0
		AVQ
		AVP
	ADV_O	XX0
CONJ	CONJ_M	CJT
		CJC
		CJS
PRON	PRON_O	DPS
		DTQ
		EX0
	PRON_M	PNP
		PNX
		PNI
		PNQ
PREP	PREP_M	PRP
		PRF
	PREP_O	TO0
SUBST	SUBST_M	NN1
		NN2
		NN0
	SUBST_O	NP0
		ZZ0
ART	ART_M	AT0
VERB	VERB_M	VVD
		VVG
		VVN
		VVI
		VVZ
		VVB

Table A1. *Cont.*

SPOS	Intermediate Class	c5
VERB	VERB_O	VBB
		VBG
		VBI
		VBN
		VBD
		VBZ
		VDB
		VDD
		VDN
		VDI
		VDG
		VDZ
		VHN
		VHB
		VHG
		VHZ
		VHI
		VHD
		VM0
PUNCT	PUNCT_M	PUL
		PUN
		PUQ
		PUR
INTERJ	INTERJ_M	ITJ

Appendix B. Results of POS Context Vectors with Different Numbers of Dimensions

Table A2. The agreement between context vector clusters and supercategories of c5 tags (computed on c5 types), varying the n most frequent c5 tags (n recorded in the column of num_c5) used in the dimensions. The asterisks indicate significance after Bonferroni correction ($\alpha_5 = 0.05/128 = 0.00039$).

num_c5	PPMI ConVec		Prob ConVec	
	Kappa	ARI	Kappa	ARI
5	0.346 *	0.203 *	0.466 *	0.333 *
6	0.348 *	0.217 *	0.466 *	0.333 *
7	0.475 *	0.335 *	0.475 *	0.356 *
8	0.343 *	0.222 *	0.475 *	0.357 *
9	0.331 *	0.242 *	0.475 *	0.357 *
10	0.341 *	0.220 *	0.475 *	0.356 *
11	0.311 *	0.208 *	0.475 *	0.356 *
12	0.422 *	0.385 *	0.375 *	0.301 *

Table A2. *Cont.*

num_c5	PPMI ConVec		Prob ConVec	
	Kappa	ARI	Kappa	ARI
13	0.407 *	0.338 *	0.375 *	0.301 *
14	0.335 *	0.227 *	0.357 *	0.201 *
15	0.335 *	0.217 *	0.357 *	0.201 *
16	0.301 *	0.214 *	0.357 *	0.201 *
17	0.377 *	0.255 *	0.357 *	0.201 *
18	0.320 *	0.156 *	0.400 *	0.183 *
19	0.453 *	0.278 *	0.401 *	0.189 *
20	0.474 *	0.295 *	0.401 *	0.189 *
21	0.456 *	0.301 *	0.401 *	0.189 *
22	0.443 *	0.251 *	0.401 *	0.189 *
23	0.410 *	0.306 *	0.401 *	0.189 *
24	0.346 *	0.254 *	0.401 *	0.189 *
25	0.401 *	0.263 *	0.401 *	0.189 *
26	0.358 *	0.177 *	0.401 *	0.189 *
27	0.384 *	0.202 *	0.401 *	0.189 *
28	0.297 *	0.189 *	0.401 *	0.189 *
29	0.299 *	0.140 *	0.401 *	0.189 *
30	0.257 *	0.105	0.401 *	0.189 *
40	0.327 *	0.191 *	0.401 *	0.189 *
50	0.298 *	0.172 *	0.401 *	0.202 *
60	0.305 *	0.179 *	0.401 *	0.202 *
70	0.336 *	0.155 *	0.401 *	0.202 *
80	0.387 *	0.195 *	0.401 *	0.202 *
90	0.275 *	0.062	0.401 *	0.202 *

Table A3. The agreement between context vector clusters and supercategories of c5 tags (computed on types). The vectors used here are the PPMI vectors and Prob vectors of 44 dimensions, reduced from 364 dimensions (91 contextual tags × 4 contextual positions) by PCA while preserving 90% variation in the data. The asterisks indicate significance after Bonferroni correction ($\alpha_6 = 0.05/6 = 0.0083$).

	Cohen's Kappa	ARI
supercategories & PPMI clusters	0.491 *	0.288 *
supercategories & Prob clusters	0.362 *	0.178 *
Prob clusters & PPMI clusters	0.666 *	0.555 *

Appendix C. Counts of Types and Tokens in Clusters and Supercategories

Table A4. Counts of c5 types for the cross classification between PPMI clusters and supercategories. The row and column margins give the counts of instances (in millions) labeled with the corresponding cluster IDs or supercategories.

Supercategory	Cluster ID									Instance Count (m)
	1	2	3	4	5	6	7	8	9	
ADJ	5							1		1.1
ADV		2				1	1			6.3
ART	1									8.6
CONJ				3						5.3

Table A4. *Cont.*

Supercategory	Cluster ID									Instance Count (m)
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
INTERJ							1			0.4
PREP				2		1				12.3
PRON	1	1					5			7.9
NOUN								5		24.0
VERB		5	6		4	1			9	16.8
instance count (m)	19.3	8.7	4.2	16.0	3.3	3.8	7.0	25.8	4.6	92.7

Table A5. Counts of c5 types for the cross-classification between Prob clusters and supercategories. The row and column margins give the counts of instances (in millions) labeled with the corresponding cluster IDs or supercategories.

Supercategory	Cluster ID									Instance Count (m)
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
ADJ	5							1		1.1
ADV		1					1	2		6.3
ART	1									8.6
CONJ						1	2			5.3
INTERJ							1			0.4
PREP		1				2				12.3
PRON	1						4	2		7.9
NOUN								5		24.0
VERB			3	3	4	1		4	10	16.8
instance count (m)	19.3	2.4	0.4	0.4	3.3	11.4	11.4	34.8	9.4	92.7

References

1. Billhardt, H.; Borrajo, D.; Maojo, V. A context vector model for information retrieval. *J. Am. Soc. Inf. Sci. Technol.* **2002**, *53*, 236–249. [CrossRef]
2. Finch, S.; Chater, N. Bootstrapping Syntactic Categories. In Proceedings of the 14th Annual Conference of the Cognitive Science Society, Bloomington, IN, USA, 29 July–1 August 1992; pp. 820–825.
3. Burgess, C.; Lund, K. Modelling Parsing Constraints with High-dimensional Context Space. *Lang. Cogn. Process.* **1997**, *12*, 177–210. [CrossRef]
4. Erk, K. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Lang. Linguist. Compass* **2012**, *6*, 635–653. [CrossRef]
5. BNC Consortium. The British National Corpus (XML Edition) [Data Set], Oxford Text Archive. 2007. Available online: http://hdl.handle.net/20.500.14106/2554 (accessed on 13 April 2026).
6. Radford, A. *Transformational Grammar: A First Course*; Cambridge Textbooks in Linguistics; Cambridge University Press: Cambridge, UK, 1988.
7. Rauh, G. *Syntactic Categories: Their Identification and Description in Linguistic Theories*; Oxford University Press: Oxford, UK, 2010.
8. Müller, G.; Englisch, J.; Opitz, A. Extraction from NP, frequency, and minimalist gradient harmonic grammar. *Linguistics* **2022**, *60*, 1619–1662. [CrossRef]
9. Torrens-Urrutia, A.; Novák, V.; Jiménez-López, M.D. Fuzzy Property Grammars for Gradience in Natural Language. *Mathematics* **2023**, *11*, 735. [CrossRef]
10. Espree-Conaway, D.A. Prepositions in Mixtec? The case of gradient word classes in Tu'un Javi. *LACUS Forum* **2025**, *5*, 35–50.

11. Elbourne, P. A program for eliminating syntactic categories. In *Linguistic Inquiry*; The MIT Press: Cambridge, MA, USA, 2025; pp. 1–26. [CrossRef]
12. Elbourne, P. Adjectives without syntactic categories. *Nat. Lang. Linguist. Theory* **2026**, *44*, 17. [CrossRef]
13. Aarts, B. Modelling linguistic gradience. *Stud. Lang. Int. J. Spons. Found. Found. Lang.* **2004**, *28*, 1–49. [CrossRef]
14. Emonds, J.E. Root and Structure-Preserving Transformations. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1970.
15. Jacobsson, B. Adverbs, Prepositions, and Conjunctions in Englisth: A Study in Gradience. *Stud. Linguist.* **1977**, *31*, 38–64. [CrossRef]
16. Croft, W.A. Beyond Aristotle and gradience: A reply to Aarts. *Stud. Lang. Int. J. Spons. Found. Found. Lang.* **2007**, *31*, 409–430. [CrossRef]
17. Reynolds, B. Prepositions in (English) Dictionaries. *Dict. J. Dict. Soc. North Am.* **2025**, *46*, 31–60. [CrossRef]
18. Bloomfield, L. *Language*; George Allen & Unwin Ltd.: London, UK, 1935.
19. Harris, Z.S. From Morpheme to Utterance. *Language* **1946**, *22*, 161–183. [CrossRef]
20. Croft, W. Word classes, parts of speech, and syntactic argumentation. *Linguist. Typology* **2005**, *9*, 431–441.
21. Croft, W. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*; Oxford University Press: Oxford, UK, 2001. [CrossRef]
22. Harris, Z.S. Distributional structure. *Word* **1954**, *10*, 146–162. [CrossRef]
23. Landauer, T.K.; Dumais, S.T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **1997**, *104*, 211–240. [CrossRef]
24. Sagi, E.; Kaufmann, S.; Clark, B. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*; Basili, R., Pennacchiotti, M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 104–111.
25. Marelli, M.; Baroni, M. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychol. Rev.* **2015**, *122*, 485–515. [CrossRef] [PubMed]
26. Boleda, G.; Baroni, M.; Pham, T.N.; McNally, L. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)—Long Papers*; Koller, A., Erk, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 35–46.
27. Turney, P.D.; Pantel, P. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.* **2010**, *37*, 141–188. [CrossRef]
28. Boleda, G. Distributional Semantics and Linguistic Theory. *Annu. Rev. Linguist.* **2020**, *6*, 213–234. [CrossRef]
29. Sahlgren, M. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces. Ph.D. Thesis, Stockholm University, Department of Linguistics, Stockholm, Sweden, 2006.
30. Westbury, C.; Hollis, G. Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behav. Res. Methods* **2019**, *51*, 1371–1398. [CrossRef] [PubMed]
31. Leech, G.; Garside, R.; Bryant, M. CLAWS4: The Tagging of the British National Corpus. In *Proceedings of the COLING 1994*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1994; Volume 1.
32. Baker, M.; Croft, W. Lexical Categories: Legacy, Lacuna, and Opportunity for Functionalists and Formalists. *Annu. Rev. Linguist.* **2017**, *3*, 179–197. [CrossRef]
33. Aarts, B. Syntactic argumentation. In *The Oxford Handbook of English Grammar*; Aarts, B., Bowie, J.B., Popova, G., Eds.; Oxford University Press: Oxford, UK, 2019; pp. 21–39.
34. Nakagawa, T.; Kudoh, T.; Matsumoto, Y. Unknown word guessing and part-of-speech tagging using support vector machines. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*; Hitotsubashi Memorial Hall, National Center of Sciences: Tokyo, Japan, 2001; pp. 325–331.
35. Fano, R.M. *Transmission of Information: A Statistical Theory of Communications*; The MIT Press: Cambridge, MA, USA, 1961.
36. Church, K.W.; Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
37. Ballance, O.J. Methodological considerations for the use of mutual information: Examining the role of context in collocation research. *Res. Methods Appl. Linguist.* **2022**, *1*, 100024. [CrossRef]
38. Evans, C.; Lorenzo-Dus, N. A corpus-assisted discourse analysis of children's and groomers' talk in online grooming interactions. *Appl. Corpus Linguist.* **2025**, *5*, 100147. [CrossRef]
39. Rayson, P.; Garside, R. Comparing corpora using frequency profiling. In *WCC '00: Proceedings of the Workshop on Comparing Corpora–Volume 9*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2000; pp. 1–6. [CrossRef]
40. Pojanapunya, P.; Todd, R.W. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguist. Linguist. Theory* **2018**, *14*, 133–167. [CrossRef]

41. Bullinaria, J.A.; Levy, J.P. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav. Res. Methods* **2007**, *39*, 510–526. [CrossRef] [PubMed]
42. Niwa, Y.; Nitta, Y. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *COLING '94: Proceedings of the 15th Conference on Computational Linguistics–Volume 1*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1994; pp. 304–309. [CrossRef]
43. Johnson, S.J.; Murty, M.R.; Navakanth, I. A detailed review on word embedding techniques with emphasis on word2vec. *Multimed. Tools Appl.* **2024**, *83*, 37979–38007. [CrossRef]
44. Baroni, M.; Dinu, G.; Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*; Toutanova, K., Wu, H., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; Volume 1, pp. 238–247. [CrossRef]
45. Guillot, S.; Prouteau, T.; Dugue, N. Sparser is better: One step closer to word embedding interpretability. In *Proceedings of the 15th International Conference on Computational Semantics*; Amblard, M., Breitholtz, E., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 106–115.
46. Ascari, R.; Giabelli, A.; Malandri, L.; Mercorio, F.; Mezzanzanica, M. A Fistful of Vectors: A Tool for Intrinsic Evaluation of Word Embeddings. *Cogn. Comput.* **2024**, *16*, 949–963. [CrossRef]
47. Karwa, S.; Singh, N. Disentangling Linguistic Features with Dimension-Wise Analysis of Vector Embeddings. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*; Cao, T., Das, A., Kumarage, T., Wan, Y., Krishna, S., Mehrabi, N., Dhamala, J., Ramakrishna, A., Galystan, A., Kumar, A., et al., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 461–488. [CrossRef]
48. Chomsky, N. *Aspects of the Theory of Syntax*, 50th ed.; The MIT Press: Cambridge, MA, USA, 1965.
49. Chomsky, N. Remarks on Nominalization. In *Readings in English Transformational Grammar*; Jacobs, R.A., Rosenbaum, P.S., Eds.; Ginn: Waltham, MA, USA, 1970; pp. 184–221.
50. Jackendoff, R. *X-Bar Syntax: A Study of Phrase Structure*; The MIT Press: Cambridge, MA, USA, 1977.
51. Givón, T. *On Understanding Grammar*; Academic Press: New York, NY, USA, 1979.
52. Givón, T. *On Understanding Grammar*; John Benjamins: Amsterdam, The Netherlands, 2018.
53. Stassen, L. *Intransitive Predication*; Oxford University Press: Oxford, UK, 1997.
54. Anderson, J.M. *A Notional Theory of Syntactic Categories*; Cambridge Studies in Linguistics; Cambridge University Press: Cambridge, UK, 1997.
55. Anderson, J. On the status of auxiliaries in notional grammar. *J. Linguist.* **1990**, *26*, 341–362. [CrossRef]
56. Anderson, J.M. Syntactic Categories and Notional Features. In *Fuzzy Grammar: A Reader*; Aarts, B., Denison, D., Keizer, E., Popova, G., Aarts, B., Denison, D., Keizer, E., Popova, G., Eds.; Oxford University Press: Oxford, UK, 2004; pp. 225–237. Available online: https://academic.oup.com/book/0/chapter/421302262/chapter-pdf/52203685/isbn-9780199262564-book-part-15.pdf (accessed on 13 April 2026).
57. Wu, Z. A quantitative method for syntactic gradience: Words, phrases, and the constructions in between. *Linguist. Vanguard* **2025**, *11*, 131–138. [CrossRef]
58. Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
59. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2025.
60. Chmura Kraemer, H.; Periyakoil, V.S.; Noda, A. Kappa coefficients in medical research. *Stat. Med.* **2002**, *21*, 2109–2129. [CrossRef] [PubMed]
61. Warrens, M.J. New Interpretations of Cohen's Kappa. *J. Math.* **2014**, *2014*, 203907. [CrossRef]
62. Hanegraaf, P.; Wondimu, A.; Mosselman, J.J.; de Jong, R.; Abogunrin, S.; Queiros, L.; Lane, M.; Postma, M.J.; Boersma, C.; van der Schans, J. Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: A mixed-methods review. *Br. Med. J. Open* **2024**, *14*, e076912. [CrossRef] [PubMed]
63. Hubert, L. Nominal scale response agreement as a generalized correlation. *Br. J. Math. Stat. Psychol.* **1977**, *30*, 98–103. [CrossRef]
64. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
65. Albatineh, A.N.; Mihalko, M.N.B. On Similarity Indices and Correction for Chance Agreement. *J. Classif.* **2006**, *23*, 301–313. [CrossRef]
66. Warrens, M.J.; van der Hoef, H. Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs. *J. Classif.* **2022**, *39*, 487–509. [CrossRef]
67. Rand, W.M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846–850. [CrossRef]
68. Warrens, M.J.; van der Hoef, H. Understanding the Rand index. In *Advanced Studies in Classification and Data Science*; Springer: Singapore, 2020; pp. 301–313. [CrossRef]
69. Douglas Steinley, M.J.B.; Hubert, L. The variance of the adjusted Rand index. *Psychol. Methods* **2016**, *21*, 261–272. [CrossRef] [PubMed]

70. Qannari, E.M.; Courcoux, P.; Faye, P. Significance test of the adjusted Rand index. Application to the free sorting task. *Food Qual. Prefer.* **2014**, *32*, 93–97. [CrossRef]
71. Stolcke, A. Syntactic category formation with vector space grammars. In *Proceedings of the Annual Meeting of the Cognitive Science Society*; University of California: Oakland, CA, USA, 1991; Volume 13, pp. 908–912.
72. Altman, D.G. *Practical Statistics for Medical Research*; Chapman & Hall/CRC: Boca Raton, FL, USA, 1991.
73. Pullum, G.K. Syncategorematicity and English infinitival to. *Glossa* **1982**, *16*, 181–215.
74. Gazdar, G.; Pullum, G.K.; Sag, I.A. Auxiliaries and Related Phenomena in a Restrictive Theory of Grammar. *Language* **1982**, *58*, 591–638. [CrossRef]
75. Levine, R.D. Auxiliaries: To's company. *J. Linguist.* **2012**, *48*, 187–203. [CrossRef]
76. Klecha, P. Diagnosing Modality in Predictive Expressions. *J. Semant.* **2013**, *31*, 443–455. [CrossRef]
77. Berto, F. A Modality Called 'Negation'. *Mind* **2015**, *124*, 761–793. [CrossRef]
78. Kripke, S.A. Semantical Considerations on Modal Logic. *Acta Philos. Fenn.* **1963**, *16*, 83–94.
79. Kratzer, A. The Notional Category of Modality. In *Formal Semantics: The Essential Readings*; Kratzer, A., Eikmeyer, H.J., Rieser, H., Eds.; Blackwell: New York, NY, USA, 1981; pp. 289–323.
80. De, M.; Omori, H. There is More to Negation than Modality. *J. Philos. Log.* **2018**, *47*, 281–299.
81. Grodner, D.; Gibson, E. Consequences of the Serial Nature of Linguistic Input for Sentenial Complexity. *Cogn. Sci.* **2005**, *29*, 261–290. [CrossRef] [PubMed]
82. Martínez, E.; Mollica, F.; Gibson, E. Even laypeople use legalese. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2405564121. [CrossRef] [PubMed]
83. Futrell, R.; Levy, R. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*; Lapata, M., Blunsom, P., Koller, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; Volume 1, pp. 688–698.
84. Quine, W.V.O. Two Dogmas of Empiricism. In *From a Logical Point of View*; Harper & Row: New York, NY, USA, 1963; pp. 20–46.
85. Baroni, M.; Bisi, S. Using cooccurence statistics and the web to discover synonyms in a technical language. In *Proceedings of LREC2004*; Centro Cultural de Belem: Lisbon, Portugal, 2004; pp. 1725–1728.
86. Macklovitch, E. Where the tagger falters. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1992; pp. 25–27.
87. LaTerza, I. Adjectives and determiners. *Lingua* **2015**, *168*, 85–103. [CrossRef]
88. Rauh, G. On the grammar of lexical and non-lexical prepositions in English. In *The Semantics of Prepositions*; Zelinsky-Wibbelt, C., Ed.; De Gruyter Mouton: Berlin, Germany; Boston, MA, USA, 1993; pp. 99–150. [CrossRef]
89. Baldwin, T.; Kordoni, V.; Villavicencio, A. Prepositions in Applications: A Survey and Introduction to the Special Issue. *Comput. Linguist.* **2009**, *35*, 119–149. [CrossRef]
90. Raunak, V.; Gupta, V.; Metze, F. Effective Dimensionality Reduction for Word Embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*; Augenstein, I., Gella, S., Ruder, S., Kann, K., Can, B., Welbl, J., Conneau, A., Ren, X., Rei, M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 235–243. [CrossRef]
91. Béranger, A.; Dugué, N.; Guillot, S.; Prouteau, T. Filtering Communities in Word Co-Occurrence Networks to Foster the Emergence of Meaning. In *Proceedings of the Complex Networks & Their Applications XII*; Cherifi, H., Rocha, L.M., Cherifi, C., Donduran, M., Eds.; Springer: Cham, Switzerland, 2024; pp. 377–388.
92. Lê, S.; Josse, J.; Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **2008**, *25*, 1–18. [CrossRef]
93. Mansfield, J.; Krapp, L.S. A simple explanation for harmonic word order. *Cogn. Sci.* **2025**, *49*, e70056. [CrossRef] [PubMed]