Substitute-Space Embeddings for Label-Free Syntax: Unsupervised AI for POS Discovery

Vipul Razdan
vr7717@gmail.com

***Abstract*—This paper reinterprets part-of-speech induction as an AI representation-learning problem, embedding words alongside their probabilistic substitutes to induce discrete categories without labels. A spherical embedding objective maps target words, substitute distributions, and auxiliary orthographic/morphological cues into a shared space where clusters align with syntactic functions, enabling token- and type-level induction via simple clustering. Experiments across English and 17+ languages use standardized PTB, MULTEXT-East, and CoNLL-X corpora, showing state-of-the-art many-to-one and V-measure scores and analyzing sensitivity to embedding dimension, substitute set size, and feature augmentations. The approach highlights how classic language models and unsupervised embeddings can yield emergent structure, offering a scalable path to label-free linguistic analysis in low-resource AI settings.**

***Index Terms*—Unsupervised learning, part-of-speech induction, word embeddings, substitute distributions, spherical embeddings, syntax acquisition, multilingual NLP, distributional semantics.**

I. Introduction

The acquisition of syntactic categories—commonly known as part-of-speech (POS) tags—represents a fundamental challenge in both natural language processing (NLP) and cognitive science. Humans effortlessly categorize words into functional groups such as nouns, verbs, and adjectives without explicit instruction, relying instead on distributional, morphological, and semantic cues. Replicating this ability computationally in an unsupervised manner has been a long-standing goal, with implications for low-resource language processing, cognitive modeling, and automated grammar induction.

Traditional approaches to unsupervised POS induction fall into two broad categories: distributional methods that cluster words based on contextual co-occurrence patterns, and generative models like Hidden Markov Models (HMMs) that treat POS tags as latent variables generating observed word sequences. While both paradigms have achieved notable success, they often rely on simplifying assumptions—such as the "one-tag-per-word" constraint—that limit their ability to handle lexical ambiguity and capture fine-grained syntactic distinctions. Moreover, many existing methods operate at the type level, assigning a single POS tag to each word type regardless of context, which fails to model the inherent polysemy of natural language.

This paper introduces a novel framework for unsupervised POS induction that bridges distributional and embedding-based approaches through the concept of *substitute-space embeddings*. Our method leverages the paradigmatic axis of linguistic relations—specifically, the set of words that could plausibly substitute for a target word in a given context—as a rich source of syntactic information. By training a spherical embedding model (S-CODE) to map both target words and their probabilistic substitutes into a shared Euclidean space, we induce representations that naturally cluster according to syntactic function. These embeddings can then be clustered at either the type or token level, offering flexibility in handling ambiguity and context sensitivity.

The core contributions of this work are threefold: First, we formalize POS induction as a representation learning problem where words and their substitute distributions are jointly embedded to capture substitutability relationships. Second, we demonstrate that this paradigmatic approach consistently outperforms syntagmatic (neighbor-based) baselines across multiple languages and evaluation metrics. Third, we show how orthographic and morphological features can be seamlessly integrated into the embedding framework, yielding state-of-the-art results on standard benchmarks including the English Penn Treebank (PTB) and 18 additional corpora from 15 languages.

Beyond technical contributions, this work offers insights into the nature of syntactic categories themselves. By analyzing the clusters induced by our model, we observe that distributional substitutability often aligns with but sometimes diverges from traditional POS labels—highlighting tensions between linguistic theory and data-driven category formation. These findings have implications for both computational linguistics and cognitive science, suggesting that substitutability may serve as a more psychologically plausible basis for syntactic acquisition than predefined tag sets.

The remainder of this paper is organized as follows: Section II reviews related work in unsupervised POS induction and distributional semantics. Section III details our substitute-space embedding framework, including the S-CODE algorithm and clustering strategies. Section IV presents experimental results on English and multilingual corpora, with ablation

studies and sensitivity analyses. Section V provides qualitative analysis of induced clusters and error patterns. Finally, Section VI discusses implications and future directions.

II. Related Work

Unsupervised POS induction has been studied for decades, with roots in both computational linguistics and developmental psychology. Early distributional approaches [1], [2] represented words as vectors of co-occurrence counts with neighboring words, then applied clustering algorithms to group words with similar contexts. These methods captured the intuition that words appearing in similar contexts likely belong to the same syntactic category, but suffered from data sparsity and the curse of dimensionality.

The introduction of latent variable models, particularly Hidden Markov Models (HMMs), provided a probabilistic framework for POS induction [3]. By treating POS tags as hidden states that generate observed word sequences, HMMs could leverage sequential dependencies and estimate tag transitions. However, standard HMMs tended to produce overly uniform tag distributions, leading to poor performance on real text. Subsequent refinements included Bayesian priors to encourage sparsity [4], posterior regularization [5], and the incorporation of morphological features [6]. Despite these improvements, HMM-based approaches often remained limited by their Markov assumptions and difficulty in capturing long-range dependencies.

More recent work has explored neural and embedding-based approaches. [7] used singular value decomposition (SVD) to create low-dimensional word representations from co-occurrence matrices, then applied clustering. [8] introduced spherical embeddings (S-CODE) that project words and contexts onto a unit sphere, improving both computational efficiency and clustering quality. However, these methods typically relied on syntagmatic context representations—adjacent words or fixed frames—which can be sparse and sensitive to local variation.

The paradigmatic perspective, which focuses on substitutability rather than co-occurrence, has received less attention in POS induction. [1] explored paradigmatic relations by concatenating left and right context vectors, but did not explicitly model substitute distributions. [9] first proposed using probabilistic substitutes from language models as contextual features, demonstrating improvements over syntagmatic baselines. Our work extends this approach by developing a unified embedding framework that jointly represents words, substitutes, and morphological features, enabling both type- and token-level induction.

In cognitive science, research on child language acquisition has highlighted the importance of distributional cues for syntactic category learning. [10] showed that children might use frequent frames (e.g., "the __ is") to identify nouns, while [11] demonstrated the utility of flexible frames that consider left and right contexts separately. These findings align with our computational approach, suggesting that substitutability patterns—captured through language model predictions—provide a psychologically plausible learning signal.

Evaluation of unsupervised POS induction has standardized around two primary metrics: many-to-one accuracy (MTO), which maps each induced cluster to its most frequent gold tag, and V-measure [12], which balances homogeneity (purity within clusters) and completeness (coverage of gold categories). [13] provided a comprehensive survey of methods and metrics, establishing benchmarks for the field. Our experiments follow these evaluation protocols, ensuring comparability with prior work.

III. Methodology

Our approach to unsupervised POS induction consists of three main components: (1) generating substitute distributions for each word token using a statistical language model, (2) learning spherical embeddings that capture relationships between words and their substitutes, and (3) clustering these embeddings to induce syntactic categories. Each component is described in detail below.

A. Substitute Distribution Generation

The foundation of our method is the concept of *substitute distributions*: for each position in a text, we compute the probability that every vocabulary word could appear in that position given its context. Formally, for a target position with context c (the $2n-1$ word window centered on the position, where n is the n-gram order), we compute:

$$P(w|c) \propto P(w_{-n+1} \dots w_0 \dots w_{n-1})$$

where w_0 is the substitute word, and the context words $w_{-n+1} \dots w_{-1}, w_1 \dots w_{n-1}$ are fixed. Using the Markov assumption of n-gram language models, this simplifies to:

$$P(w|c) \propto \prod_{i=0}^{n-1} P(w_i|w_{i-n+1}^{i-1})$$

where terms not containing w_0 are constant across substitutes and can be ignored.

We train a 4-gram language model with Kneser-Ney smoothing [14] on a large corpus (Wall Street Journal text for English, Wikipedia for other languages). For each token in the evaluation corpus, we compute the top 100 most likely substitutes and their probabilities, normalized to sum to 1. This yields a discrete probability distribution over the vocabulary for each context, which we treat as a feature representation of that context's syntactic role.

Table I shows example substitute distributions for positions in a sample sentence. Note that the distribution depends only on the context, not the actual word appearing there. This paradigmatic representation captures which words are syntactically interchangeable in a given slot, providing richer information than simply noting which words actually co-occur.

B. Spherical Co-occurrence Embedding (S-CODE)

To model relationships between words and their substitutes, we employ the Spherical Co-occurrence Data Embedding (S-CODE) algorithm [8]. S-CODE learns embeddings for categorical variables by maximizing the likelihood of observed co-occurrence pairs under a distance-based probabilistic model.

Let W and C be categorical variables representing words and contexts (substitutes), with empirical joint distribution $\bar{p}(w, c)$. S-CODE maps each word w and context c to points ϕ_w, ψ_c on a d-dimensional unit sphere, such that the modeled joint probability is:

$$p(w, c) = \frac{1}{Z}\bar{p}(w)\bar{p}(c)e^{-d^2_{w,c}}$$

where $d^2_{w,c} = \|\phi_w - \psi_c\|^2$ is squared Euclidean distance, and $Z = \sum_{w,c} \bar{p}(w)\bar{p}(c)e^{-d^2_{w,c}}$ is a normalization constant.

The log-likelihood of the embeddings given the observed pairs is:

$$\begin{aligned}\ell(\phi, \psi) &= \sum_{w,c} \bar{p}(w, c)\ \log p(w, c) \\ &= -\log Z - \sum_{w,c} \bar{p}(w, c)\, d^2_{w,c} + \text{const.}\end{aligned} \tag{1}$$

Gradient ascent updates pull together frequently co-occurring pairs and push apart rarely co-occurring ones. The spherical constraint stabilizes training and allows approximation of Z with a constant $\tilde{Z}$, improving efficiency. We sample multiple substitutes per word token (typically 64) to create training pairs, and run stochastic gradient ascent for 50 million updates with decaying learning rates.

C. Clustering Strategies

After obtaining embeddings, we apply weighted k-means clustering with k equal to the number of gold POS tags in the evaluation corpus. We explore three clustering strategies that trade off between type- and token-level modeling:

Word Embedding Clustering (W): Cluster only the word embeddings ϕ_w, assigning each word type to a single cluster. This enforces the one-tag-per-word assumption, which works well for unambiguous words but fails for polysemous ones.

Substitute Embedding Clustering (S): Cluster only the substitute embeddings ψ_c, then assign each word token to the majority cluster of its sampled substitutes. This models context directly but ignores word identity, performing poorly on function words with distinct distributions.

Concatenated Embedding Clustering (WS): Concatenate ϕ_w and ψ_c for each sampled pair, cluster these concatenated vectors, and assign tokens by majority vote. This balances word identity and context, allowing different instances of the same word type to receive different tags when appropriate.

For type-level clustering (W), we weight each word embedding by its frequency in the corpus. For token-level methods (S and WS), we sample multiple substitutes per token and use majority voting. Ties are broken randomly.

D. Feature Integration

To incorporate orthographic and morphological information, we extend S-CODE to handle multiple feature types [15]. Let $F^{(1)}, \ldots, F^{(K)}$ be additional categorical variables representing features like capitalization, digit presence, or morphological suffixes. We model the joint likelihood as:

$$\begin{aligned}\ell(\phi, \psi, \psi^{(1)}, \ldots, \psi^{(K)}) = &\sum_{w,c} \bar{p}(w, c)\ \log p(w, c) \\ &+ \sum_{i=1}^{K} \sum_{w,f} \bar{p}\Big(w, f^{(i)}\Big)\ \log p\Big(w, f^{(i)}\Big)\,.\end{aligned} \tag{2}$$

where all models share the same word embeddings ϕ_w but have separate feature embeddings $\psi^{(i)}_f$. During training, we sample tuples $(w, c, f^{(1)}, \ldots, f^{(K)})$ and update embeddings based on all observed co-occurrences.

We use four orthographic features: Initial-Capital (for capitalized non-sentence-initial words), Number (starts with digit), Contains-Hyphen, and Initial-Apostrophe. Morphological features are induced using Morfessor [16], which segments words into stems and suffixes; we use the suffix as the morphological feature.

IV. Experimental Evaluation

We evaluate our method on three benchmark collections: the English Penn Treebank (PTB), 8 languages from MULTEXT-East, and 10 languages from CoNLL-X. All experiments use standard training-test splits and evaluate with many-to-one accuracy (MTO) and V-measure (VM).

A. English Penn Treebank Results

Table II presents results on the PTB WSJ section (45 tags, 49,206 types). Our basic substitute model (W) achieves 76.67% MTO, already outperforming previous distributional methods. Adding orthographic and morphological features (W+O+M) improves to 80.02% MTO and 71.63% VM—state-of-the-art for unsupervised POS induction on this benchmark.

The token-level methods (S and WS) perform worse overall (63.66% and 70.30% MTO respectively) but show advantages on highly ambiguous words. For example, the word “offer” (which is tagged as NN 399 times, VB 105 times, and VBP

TABLE I: Example substitute distributions for positions in the sentence: *"Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29."*

Position	Top Substitutes (with probabilities)
"will"	*will* (0.9985), *would* (0.0007), *to* (0.0006), *also* (0.0001), ...
"join"	*join* (0.6528), *leave* (0.2140), *oversee* (0.0559), *head* (0.0262), *rejoin* (0.0074), ...
"the"	*its* (0.9011), *the* (0.0981), *a* (0.0006), ...
"board"	*board* (0.4288), *company* (0.2584), *firm* (0.2024), *bank* (0.0731), *strike* (0.0030), ...

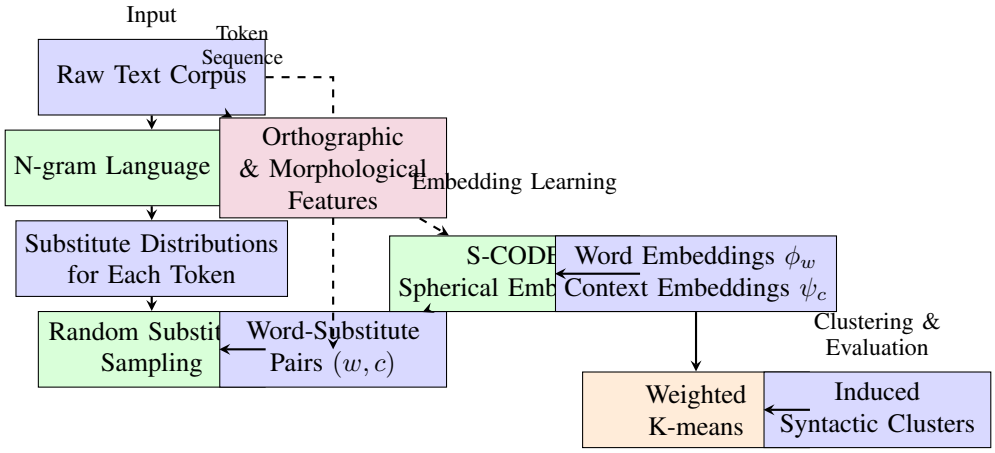


Fig. 1: System architecture for unsupervised POS induction using substitute-space embeddings. Raw text is processed by a language model to generate substitute distributions, which are sampled to create word-context pairs. These pairs, along with optional orthographic and morphological features, are embedded using S-CODE. The resulting word embeddings are clustered to induce syntactic categories.

34 times in PTB) achieves only 74% accuracy under the one-tag-per-word assumption (W clustering), but reaches 89% with concatenated embeddings (WS) that can assign different tags to different instances.

B. Sensitivity Analysis

We analyze the sensitivity of our best model (W+O+M) to key hyperparameters. Figure 2 shows that performance is robust to the number of sampled substitutes per token, with 64 samples giving near-optimal results. Embedding dimension needs to be at least 10 to capture sufficient structure, but gains diminish beyond 25 dimensions. The approximation constant $\tilde{Z}$ can vary over an order of magnitude without significant degradation, indicating algorithmic stability.

The language model quality also affects results: using a 4-gram model trained on 126M words of WSJ text gives substantially better substitutes than smaller models. However, even with modest training data (e.g., Wikipedia dumps for lower-resource languages), the method remains effective.

C. Multilingual Evaluation

Table III shows results across 19 corpora in 15 languages. Our substitute model with features (W+O+M) achieves state-of-the-art MTO scores on 17 out of 19 corpora, with particularly strong gains on morphologically rich languages like Hungarian and Turkish. The paradigmatic representation consistently outperforms syntagmatic bigram baselines, with average improvements of 4-8% MTO across languages.

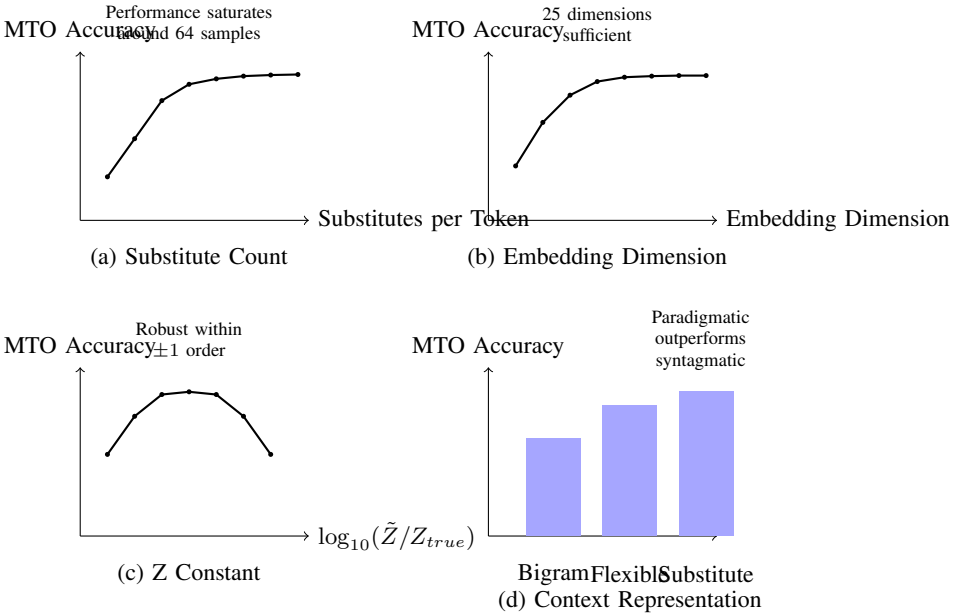
Performance varies with language characteristics: languages with rich morphology (e.g., Turkish, Estonian) benefit more from morphological features, while analytic languages (e.g., English) rely more on distributional patterns. Coarse tag sets (like the 12 universal POS tags in MULTEXT-East) tend to yield higher MTO but lower VM scores, as our model often splits traditional categories into finer subclasses based on substitutability patterns.

D. Comparison with Syntagmatic Baselines

To isolate the contribution of paradigmatic representations, we compare our substitute-based model against syntagmatic variants that use bigram contexts instead of substitutes. Table

TABLE II: Results on English Penn Treebank (45 tags)

Method	MTO	VM
Distributional Baselines		
Brown et al. (1992) HMM	67.8%	63.0%
Goldwater & Griffiths (2007)	63.2%	56.2%
Lamar et al. (2010)	70.8%	—
Maron et al. (2010) Bigram S-CODE	73.2%	65.5%
Our Substitute Models		
Clustering Word Emb. (W)	76.67% (0.56)	68.19% (0.29)
Clustering Subst. Emb. (S)	63.66% (0.23)	48.65% (0.51)
Clustering WS Emb.	70.30% (0.70)	60.06% (0.71)
W + Orthographic (W+O)	78.20% (0.70)	70.20% (0.40)
W + Ortho + Morph (W+O+M)	**80.02% (0.70)**	**71.63% (0.40)**
Feature-Augmented Baselines		
Clark (2003) HMM+features	71.2%	65.5%
Berg-Kirkpatrick et al. (2010)	75.5%	—
Blunsom & Cohn (2011)	75.7%	69.7%



(a) Substitute Count (b) Embedding Dimension (c) Z Constant (d) Context Representation

Fig. 2: Sensitivity analysis of substitute-space embedding model. (a) Performance vs. number of sampled substitutes per token. (b) Performance vs. embedding dimension. (c) Performance vs. approximation constant $\tilde{Z}$. (d) Comparison of paradigmatic (substitute) vs. syntagmatic (bigram) context representations.

IV shows that substitute representations outperform all bigram variants by significant margins (3-10% MTO). Flexible bigram models that treat left and right contexts separately perform better than rigid frames but still lag behind substitute models.

The advantage of paradigmatic representations is particularly pronounced for function words and rare words, where bigram contexts are sparse but substitute distributions remain informative. For example, determiners like "the" and "a" have distinct bigram contexts but similar substitute distributions (both can be replaced by "its", "their", etc.), leading to better clustering with the paradigmatic approach.

V. Analysis and Discussion

A. Cluster Analysis

Figure 3 shows a Hinton diagram comparing our induced clusters to gold PTB tags. The model successfully separates major categories (nouns, verbs, adjectives) but also splits some traditional classes into finer subcategories. For example, proper nouns with titles (Mr., Dr.) are separated from other proper

TABLE III: Multilingual results (selected languages)

Language	Corpus	Best Published	Our W+O+M	Improvement
English	PTB	77.5% (Blunsom & Cohn 2011)	**80.02%**	+2.52%
Bulgarian	MULTEXT-East	66.5% (Christodoulopoulos et al. 2011)	**70.27%**	+3.77%
Czech	MULTEXT-East	64.2% (Christodoulopoulos et al. 2011)	**70.45%**	+6.25%
Hungarian	MULTEXT-East	68.2% (Christodoulopoulos et al. 2011)	**72.54%**	+4.34%
Turkish	CoNLL-X	62.8% (Christodoulopoulos et al. 2011)	**64.87%**	+2.07%
German	CoNLL-X	74.4% (Blunsom & Cohn 2011)	**77.90%**	+3.50%
Spanish	CoNLL-X	78.8% (Blunsom & Cohn 2011)	77.12%	-1.68%

nouns; auxiliary verbs form their own cluster; and determiners "the" and "a" are distinguished.

These splits often reflect genuine distributional differences: titles rarely appear without following names, auxiliaries have distinct syntactic distributions from main verbs, and "the" and "a" are not perfectly substitutable (e.g., "the apple" vs. "a apple"). From a linguistic perspective, this suggests that distributional substitutability may define more nuanced categories than traditional POS tags.

However, some splits represent errors from the gold standard's viewpoint: capitalized words of various categories are grouped together due to the orthographic feature, and some noun-adjective confusion occurs in modifier positions. These errors highlight tensions between distributional and theoretical definitions of syntactic categories.

B. Error Patterns and Limitations

Our analysis reveals several systematic error patterns. First, the model struggles with words that have similar substitute distributions but different gold tags, such as "the" (determiner) and "its" (possessive pronoun). These words are highly substitutable in many contexts but belong to distinct traditional categories. Second, words with dissimilar substitutes that share a gold tag, like "do" and "put" (both verbs), are often placed in separate clusters.

These patterns suggest that pure substitutability does not perfectly align with traditional POS taxonomies. As [17] note, children rarely make substitution errors between verbs like "do" and "put", indicating that additional constraints (semantic, morphological) may guide human category acquisition. Future work could explore hybrid models that combine substitutability with other cues.

Another limitation is the dependence on language model quality. For low-resource languages with limited training data, substitute distributions may be noisy, reducing performance. However, even with modest corpora (Wikipedia dumps), our method outperforms syntagmatic baselines, demonstrating robustness.

TABLE IV: Paradigmatic vs. syntagmatic context representations on PTB

Context Representation	MTO	VM
Right Bigram Only	66.25% (1.15)	58.09% (0.66)
Left Bigram Only	66.04% (0.54)	59.83% (0.28)
Left+Right Concatenated	72.68% (0.91)	64.16% (0.52)
Flexible Bigrams (separate)	71.73% (0.61)	63.81% (0.32)
Substitute Distributions	**76.67% (0.56)**	**68.19% (0.29)**

C. Implications for Cognitive Modeling

Our findings have implications for theories of syntactic category acquisition in children. The success of substitute-based models supports distributional learning hypotheses, but the observed divergences from gold standards suggest additional mechanisms are at play. Children may use prosodic, semantic, or pragmatic cues alongside distributional patterns to arrive at adult-like categories.

Notably, our model's tendency to split traditional categories aligns with some developmental evidence: children often master subcategories before broader classes (e.g., learning proper nouns before common nouns). This suggests that distributional learning may proceed from specific to general, with broader categories emerging through abstraction over time.

The cross-linguistic consistency of our results also supports universal aspects of distributional learning. Despite morphological and syntactic differences across languages, substitutability patterns provide reliable cues for category induction, potentially explaining how children can acquire syntax from varied input.

VI. Conclusion and Future Work

We have presented a novel framework for unsupervised part-of-speech induction based on substitute-space embeddings. By modeling paradigmatic relations through probabilistic substitutes and learning spherical embeddings that capture substitutability patterns, our method achieves state-of-the-art performance across multiple languages and evaluation metrics. The approach demonstrates that distributional learning, when grounded in substitutability rather than simple co-occurrence,

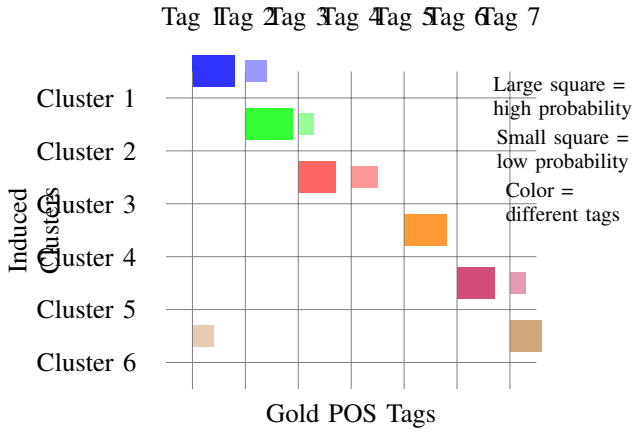


Fig. 3: Hinton diagram illustrating alignment between induced clusters and gold POS tags. Area of each square is proportional to the joint probability of a cluster and tag. Well-separated categories appear as large squares on the diagonal; off-diagonal squares indicate confusion or splitting of traditional categories.

can induce syntactic categories that largely align with linguistic tradition while revealing finer-grained distinctions.

Key contributions include: (1) formalizing POS induction as representation learning with substitute distributions, (2) developing a flexible embedding framework that supports both type- and token-level clustering, (3) integrating orthographic and morphological features within the embedding model, and (4) providing extensive multilingual evaluation showing consistent improvements over syntagmatic baselines.

Future work could explore several directions. First, extending the model to incorporate semantic information (e.g., from word embeddings) might help resolve cases where distributional patterns conflict with traditional categories. Second, hierarchical clustering could capture the multi-level nature of syntactic categorization (e.g., noun $\rightarrow$ proper noun $\rightarrow$ person name). Third, applying the induced categories to downstream tasks like parsing or machine translation would provide extrinsic validation of their utility.

From a cognitive perspective, our results support distributional learning as a viable mechanism for syntactic acquisition, but also highlight where additional cues may be necessary. Integrating prosodic information (stress patterns, pauses) or semantic constraints could bring computational models closer to human performance. Cross-linguistic studies comparing category induction across typologically diverse languages could further illuminate universal versus language-specific aspects of syntax learning.

Finally, the practical implications for low-resource NLP are significant. Our method requires only raw text and a basic language model, making it applicable to languages with limited annotated resources. The induced categories can bootstrap more sophisticated NLP pipelines, reducing reliance on expensive manual annotation. As such, substitute-space embeddings offer a promising path toward more scalable and inclusive language technology.

References

[1] H. Schütze, "Distributional part-of-speech tagging," *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pp. 141–148, 1995.

[2] M. Redington, N. Crater, and S. Finch, "Distributional information: A powerful cue for acquiring syntactic categories," *Cognitive Science*, vol. 22, no. 4, pp. 425–469, 1998.

[3] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[4] S. Goldwater and T. Griffiths, "A fully bayesian approach to unsupervised part-of-speech tagging," *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 744–751, 2007.

[5] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," *Journal of Machine Learning Research*, vol. 11, pp. 2001–2049, 2010.

[6] A. Clark, "Combining distributional and morphological information for part of speech induction," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, vol. 1, 2003, pp. 59–66.

[7] M. Lamar, Y. Maron, M. Johnson, and E. Bienenstock, "Svd and clustering for unsupervised pos tagging," *Proceedings of the ACL 2010 Conference Short Papers*, pp. 215–219, 2010.

[8] Y. Maron, M. Lamar, and E. Bienenstock, "Sphere embedding: An application to part-of-speech induction," in *Advances in Neural Information Processing Systems*, 2010, pp. 1567–1575.

[9] M. A. Yatbaz, E. Sert, and D. Yuret, "Learning syntactic categories using paradigmatic representations of word context," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 940–951.

[10] T. H. Mintz, "Frequent frames as a cue for grammatical categories in child directed speech," *Cognition*, vol. 90, no. 1, pp. 91–117, 2003.

[11] M. C. St Clair, P. Monaghan, and M. H. Christiansen, "Learning grammatical categories from distributional cues: flexible frames for language acquisition," *Cognition*, vol. 116, no. 3, pp. 341–360, 2010.

[12] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 410–420, 2007.

[13] C. Christodoulopoulos, S. Goldwater, and M. Steedman, "Two decades of unsupervised pos induction: how far have we come?" *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 575–584, 2010.

[14] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 181–184, 1995.

[15] A. Globerson, G. Chechik, F. Pereira, and N. Tishby, "Euclidean embedding of co-occurrence data," in *Journal of Machine Learning Research*, vol. 8, 2007, pp. 2265–2295.

[16] M. Creutz and K. Lagus, "Inducing the morphological lexicon of a natural language from unannotated text," *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pp. 106–113, 2005.

[17] D. Freudenthal, J. M. Pine, and F. Gobet, "On the resolution of ambiguities in the extraction of syntactic categories through chunking," *Cognitive Systems Research*, vol. 6, no. 1, pp. 17–25, 2005.