Analyzing the Role of Part-of-Speech in Code-Switching: A Corpus-Based Study

Anonymous EACL submission

Abstract

Code-switching (CS) is a common linguistic phenomenon wherein speakers fluidly transition between languages in conversation. While the cognitive processes driving CS remain a complex domain, earlier investigations have shed light on its multifaceted triggers. This study delves into the influence of Part-of-Speech (POS) on the propensity of bilinguals to engage in CS, employing a comprehensive analysis of Spanish-English and Mandarin-English corpora. Compared with prior research, our findings not only affirm the existence of a statistically significant connection between POS and the likelihood of CS across language pairs, but notably find this relationship exhibits its maximum strength in proximity to CS instances, progressively diminishing as tokens distance themselves from these CS points.

1 Introduction

Code-switching (CS), the integration of two languages within a single utterance, is pervasive across diverse language pairs. This phenomenon presents the flexibility and adaptability of individuals in their language use and therefore serves as a testing ground for research into the cognitive mechanisms of bilingual language production. The studies emerging from this exploration have shown that CS involves multiple layers of linguistic processing and is influenced by the properties of the words, linguistic structures and socio-interactional considerations (Gardner-Chloros, 2009; Kootstra et al., 2020). In parallel, the practical implications of understanding CS extend to the development of Natural Language Processing (NLP) techniques tailored to meet the needs of multilingual communities. Recent research has seen attempts to integrate established linguistic theories of CS and harness machine-learning approaches for training Automatic Speech Recognition (ASR) and language identification models. However, these theories often originate from language pairs that exhibit syntactic similarities, and their practical application is often constrained by the efficacy of relevant dependency parsers (Berk-Seligson, 1986; Chi et al., 2023). While machine-learning approaches have demonstrated success in their targeted tasks, they have the potential in benefiting from the integration of linguistic features drawn from the corpus under examination (Adel et al., 2013; Attia et al., 2019). Thus, driven by the intrinsic role of word properties in bilingual language production and their potential utility in augmenting CS-related tasks, this paper explores the influence of part-of-speech (POS), a universal feature across all languages, on CS behaviors, aiming to provide valuable insights into their role in facilitating CS occurrences across language pairs, including those from the same (Spanish-English) and different (Mandarin-English) language family.

2 Related work

Numerous studies have been conducted to investigate the triggers for CS. Through the analysis of natural language corpora, it has been consistently observed that CS occurrences are more frequent when language-ambiguous words, primarily cognates, are in close proximity (Clyne, 1967; Broersma and De Bot, 2006; Kootstra et al., 2020). This observation aligns with the well-established notion that cognates lead to the simultaneous activation of both languages in speakers' minds, consequently influencing the use of both languages within a single utterance (Van Assche et al., 2012; Soares et al., 2019). However, it's essential to note that not all language pairs possess cognates, and even when they do, identifying these cognates requires linguistic expertise. Since the majority of CS triggers are nouns and proper nouns (Broersma and De Bot, 2006), the role of POS in identifying the constraints of CS has garnered attention from researchers (Soto et al., 2018). Similar to the ex-

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	SCONJ	VERB
BM	4.1	6.97	8.11	3.25	4.4	8.81	5.94	11.04	1.51	2.58	15.98	2.49	3.88	20.00
SEAME	3.11	5.24	16.94	1.59	1.47	3.97	1.71	15.42	2.95	4.87	14.05	5.73	1.26	21.70

Table 1: POS distribution (shown in percentage) in Bangor-Miami and SEAME corpus

periments on cognates, Soto et al. demonstrate the dependency of POS and CS, serving as an inspiration for our work. In this paper, we substantiate a more robust hypothesis that such dependency remains significant when considering the distribution of both POS and CS across word positions, and its strength diminishes as the POS moves further from the points of CS.

3 Methodology

3.1 Corpus

Two language pairs are investigated in this work. In the case of Spanish-English CS, we analyze the publicly available Bangor-Miami (BM) corpus, which features conversational speech recorded by bilingual speakers in the Miami, Florida region (Deuchar et al., 2014). The original Bangor-Miami data is automatically annotated using its native tagset, courtesy of the Bangor Autoglosser (Donnelly and Deuchar, 2011). For the sake of facilitating cross-linguistic comparisons, we opt for a version of the corpus that has been annotated with Universal POS tags (AlGhamdi et al., 2016). For Mandarin-English CS experiments, we explore the South East Asian Mandarin-English (SEAME) corpus. SEAME comprises conversations and interviews with bilingual speakers from Malaysia and Singapore (Lyu et al., 2010). We annotate SEAME utilizing the Spacy toolkit, following the methodology outlined in (Bhattacharya et al., 2023). The distribution of POS tags in both corpora is detailed in Table 1.

3.2 Triggering hypothesis

In their work, Soto et al. established a definition of CS words as the initial words following CS points. They convincingly demonstrated a robust statistical association between POS and the words preceding CS and the CS words themselves. However, this definition presents a problem that despite the χ^2 test affirming the dependence between POS and CS words, it remains plausible that this dependence may be influenced solely by word positions rather than the intrinsic nature of CS, because CS points are not uniformly distributed across all positions in a sentence and in particular, never occur at the start. This connection is shown in Figure 1. To illustrate, consider a scenario where a particular POS tag predominantly occurs at the start of a sentence, making it less likely to be CS words itself. This would indicate a significant distribution difference, even if the same POS tag is occasionally code-switched in other positions. In light of these considerations, we refine our hypothesis to assert that these POS tags maintain a statistically robust relationship with CS and the words surrounding it, even when accounting for specific word positions. Furthermore, we also posit that this relationship diminishes as it extends to more distant words.

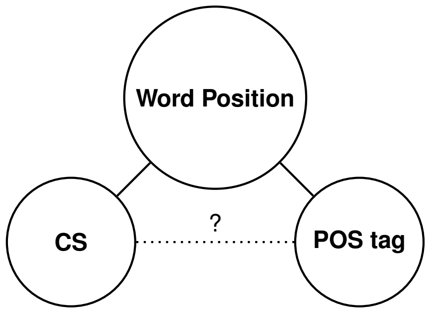


Figure 1: An undirected graph depicting the hypothetical connections between word position, CS, and POS.

4 Experiments

4.1 CS words

The relationship between the two variables, CS and POS, is examined using the χ^2 test for independence, with Yates' correction for continuity for small expected frequencies applied where necessary. To account for word positions, we classify words into three categories: Start, Mid, and End. In constructing contingency tables that tabulate the counts of all POS tags and their association with CS words, we compute the expected distribution based on Equation 1 under the null hypothesis that, given specific word positions, CS and POS are independent of each other. In this equation, $N(CS, ADJ)$ denotes the count of words being both CS and tagged as ADJ [1]. The variable i represents word positions, and P_i signifies the probability of a word being CS/ADJ at position i. It is important to note

[1]ADJ is used here for illustration, with all POS tags handled similarly

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	SCONJ	VERB
BM	-	-	-	√√ ↓	√√ ↑	√ ↓	√√ ↑	√√√ ↑	-	√√ ↓	√√ ↑	√ ↑	√ ↑	√√√ ↓
SEAME	√√√ ↑	√√√ ↓	√√ ↓	√√√ ↓	√√ ↑	√√√ ↓	√√ ↑	√√√ ↑	√√√ ↓	√√√ ↑	-	√√√ ↑	√ ↓	√√√ ↓

Table 2: The significance of running χ^2 statistical tests on each group of POS tags and CS words. One $\surd$ indicates $p < 0.01$, two indicate $p < 10^{-36}$ and three indicate $p < 10^{-100}$. $\uparrow$ and $\downarrow$ represent whether they more often or less often occur at the CS word.

that the earlier hypothesis proposed by (Soto et al., 2018), which does not account for word positions, can be regarded as a particular case where words are uniformly distributed across the Start, Mid, and End positions, affording them an equal likelihood of appearing at any point within a sentence.

$$\begin{aligned} N(CS, ADJ) &= \sum_{i \in s,m,e} P_i(CS, ADJ)Ni \\ &= \sum_{i \in s,m,e} P_i(CS, ADJ)Ni \\ &= \sum_{i \in s,m,e} P_i(CS)P_i(ADJ)Ni \end{aligned} \tag{1}$$

4.2 Neighbour words

The previous research primarily focused on investigating the presence of POS that directly precede and follow CS words, relying on distribution analysis and χ^2 tests to assess their associations. However, due to the inherent complexity of syntactic relationships within sentences, when examining CS holistically, the impact of various POS tags of CS words on neighboring words may result in intricate mutual offset or amplification effects. Since this analysis is grounded in count-based data, detecting significant changes can be challenging. To overcome this, we introduce a novel approach wherein we categorize CS based on the POS of CS words. For each CS category, we chart the distribution of POS in words immediately preceding and following the CS word, as well as those with a distance of two to four words away. These distributions are then compared to the overall POS distribution in the context of each POS category, enabling us to isolate the differences solely attributable to code-switching behaviors.

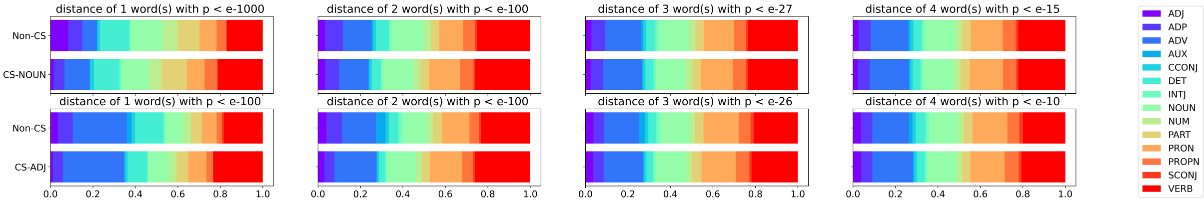
5 Results

5.1 CS words

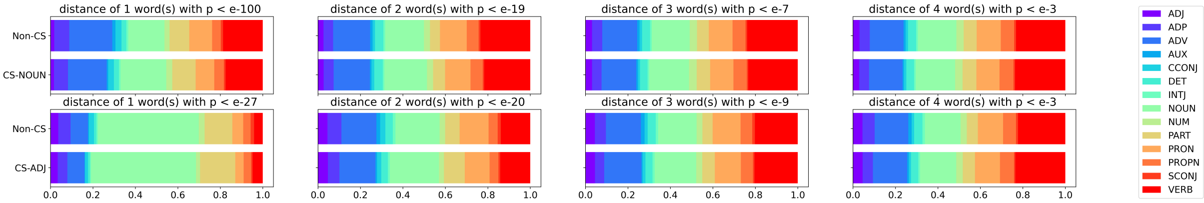
Table 2 presents the results of χ^2 statistical tests on each group of POS tags and CS words where a single $\surd$ indicates a significance level of $p < 0.01$, two indicate $p < 10^{-36}$ and three indicate $p < 10^{-100}$. $\uparrow$ and $\downarrow$ represent whether these tags occur more or less frequently at CS words based on our observations. The analysis reveals a strong statistical relationship for most of the POS tags. Notably, in contrast to (Soto et al., 2018), where CONJ and SCONJ, PRON, and NOUN exhibit distinct effects on CS words in the BM corpus, we find that they exhibit similar behaviors. One potential explanation can be our different assumptions about word positions. PROP and CONJ tags are more likely to appear at the beginning of sentences, significantly influencing our calculations. It is also worth noting that SEAME generally exhibits a stronger statistical relationship when compared to BM. This suggests that Mandarin and English have a more diverse syntactic structure compared to Spanish and English, leading to less flexibility in CS. Additionally, an interesting finding is the infrequency of switches on VERB or AUX in both language pairs. This can be attributed to the fact that these verbs are typically preceded by pronouns and require agreement in terms of person and number, which imposes constraints on the act of CS.

5.2 Neighbour words

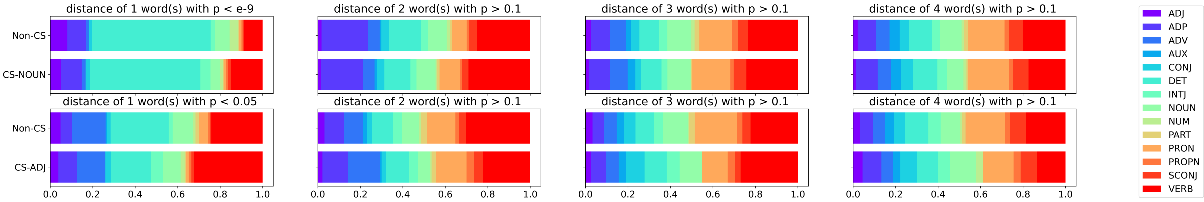
In the interest of space, Figure 2 exclusively depicts the distribution of POS for words positioned at 1-4 words away from CS points which are categorized as NOUN and ADJ, while the complete set of results can be found in the Appendix. It can be observed that as words distance themselves from CS points, the difference in the distribution of POS between words near CS and non-CS words diminishes, especially in SEAME. This trend is supported by decreasing p-values from χ^2 tests. The difference is still significant for the closest words in BM, while further words show no significance at all. Additionally, it can be found that the preceding words generally have more influence compared to the following words, which is consistent with (Soto et al., 2018). Notably, in SEAME even the largest p-value among these tests is smaller than e^{-3}. This result can be attributed to the linguistic principle



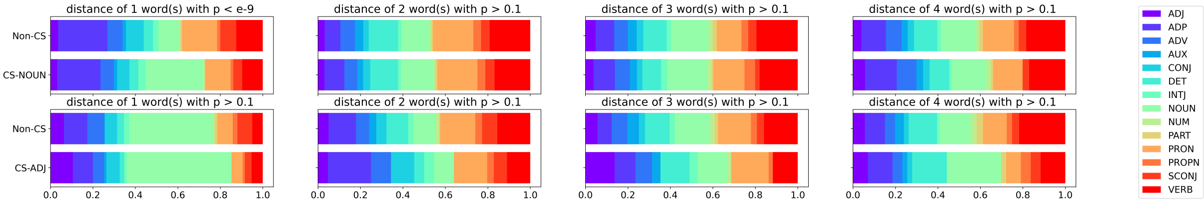
(a) POS of words positioned at 1-4 words ***before*** CS words tagged as ***NOUN*** (top) and ***ADJ*** (bottom) in ***SEAME***



(b) POS of words positioned at 1-4 words ***after*** CS words tagged as ***NOUN*** (top) and ***ADJ*** (bottom) in ***SEAME***



(c) POS of words positioned at 1-4 words ***before*** CS words tagged as ***NOUN*** (top) and ***ADJ*** (bottom) in ***BM***



(d) POS of words positioned at 1-4 words ***after*** CS words tagged as ***NOUN*** (top) and ***ADJ*** (bottom) in ***BM***

Figure 2: The visualization of the distribution of POS for words positioned at 1-4 words away from CS points, specifically those categorized as NOUN and ADJ in both corpora.

that every word's usage is influenced by its context. The displayed results for SEAME also reveal that ADJ occurs less frequently preceding switched NOUNs, which aligns with the tendency for noun phrases to be switched together. A similar rationale can be applied to the observation that VERB and AUX are more common before switched NOUNs.

6 Conclusion

With a thorough analysis of two language pairs, we extend prior work by incorporating the impact of word positions and robustly confirm the statistically significant connection between POS and CS. The significance level is higher for Mandarin-English, suggesting a more diverse syntactical structure leads to less flexibility in CS. By categorizing CS words and investigating neighboring POS, we observe that this relationship is strongest in close proximity to CS instances, gradually diminishing as words move farther from CS points. In order to validate the practical utility of our findings, we intend to integrate these observed features into the design of CS generation models, enabling us to compare the model outcomes with established theories in future research.

7 Limitations

The calculation in our study relies on external NLP tools for POS tagging, while it is a challenging task for CS. Additionally, the scarcity of available CS data necessitates our selection of only two language pairs, despite our efforts to choose pairs with varying syntactic characteristics. It is also worth noting that the syntactic intricacies within a sentence may be far more complex than what has been addressed in this paper. Although we extend prior work by incorporating word positions into our analysis, it's possible that other factors not covered in this study, such as topic relevance and prosodic elements, also influence CS behaviors to some extent.

References

Heike Adel, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–211, Sofia, Bulgaria. Association for Computational Linguistics.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107, Austin, Texas. Association for Computational Linguistics.

Mohammed Attia, Younes Samih, Ali Elkahky, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2019. POS tagging for improving code-switching identification in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 18–29, Florence, Italy. Association for Computational Linguistics.

Susan Berk-Seligson. 1986. Linguistic constraints on intrasentential code-switching: A study of spanish/hebrew bilingualism. *Language in Society*, 15(3):313–348.

Debasmita Bhattacharya, Jie Chi, Julia Hirschberg, and Peter Bell. 2023. Capturing Formality in Speech Across Domains and Languages. In *Proc. INTERSPEECH 2023*, pages 1030–1034.

Mirjam Broersma and Kees De Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9(1):1–13.

Jie Chi, Brian Lu, Jason Eisner, Peter Bell, Preethi Jyothi, and Ahmed M. Ali. 2023. Unsupervised Code-switched Text Generation from Parallel Text. In *Proc. INTERSPEECH 2023*, pages 1419–1423.

Michael G. Clyne. 1967. Transference and triggering; observations on the language assimilation of postwar german-speaking migrants in australia. 2010.

Margaret Deuchar, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto, and Diana Carter. 2014. *5. Building Bilingual Corpora*, pages 93–110. Multilingual Matters, Bristol, Blue Ridge Summit.

Kevin Donnelly and Margaret Deuchar. 2011. The bangor autoglosser: A multilingual tagger for conversational text.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge University Press.

Gerrit Jan Kootstra, Ton Dijkstra, and Janet G. van Hell. 2020. Interactive alignment and lexical triggering of code-switching in bilingual dialogue. *Frontiers in Psychology*, 11.

Dau-Cheng Lyu, Tien Ping Tan, Chng Eng Siong, and Haizhou Li. 2010. Seame: A Mandarin-English code-switching speech corpus in South-East Asia. In *INTERSPEECH*.

Ana Paula Soares, Helena Oliveira, Marisa Ferreira, Montserrat Comesaña, António Filipe Macedo, Pilar Ferré, Carlos Acuña-Fariña, Juan Hernández-Cabrera, and Isabel Fraga. 2019. Lexico-syntactic interactions during the processing of temporally ambiguous l2 relative clauses: An eye-tracking study with intermediate and advanced portuguese-english bilinguals. *PLOS ONE*, 14(5):1–27.

Víctor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, pos tags and entrainment in code-switching. In *Interspeech*.

Eva Van Assche, Wouter Duyck, and Robert Hartsuiker. 2012. Bilingual word recognition in a sentence context. *Frontiers in Psychology*, 3.

A Appendix

Figures 3 and 4 depict the POS distribution for words positioned 1-4 words before and after all CS points in SEAME, while Figures 5 and 6 present the corresponding results for BM. As discussed in the paper, we observe that the disparity in POS distribution between words near CS and non-CS words diminishes as words move away from CS points, particularly in SEAME. It's worth mentioning that, for BM, certain CS categories like PART suffer from small sample sizes, some even reaching zero counts. Due to this limitation, we do not provide the results of the χ^2 test for them, as it is not applicable in these cases.

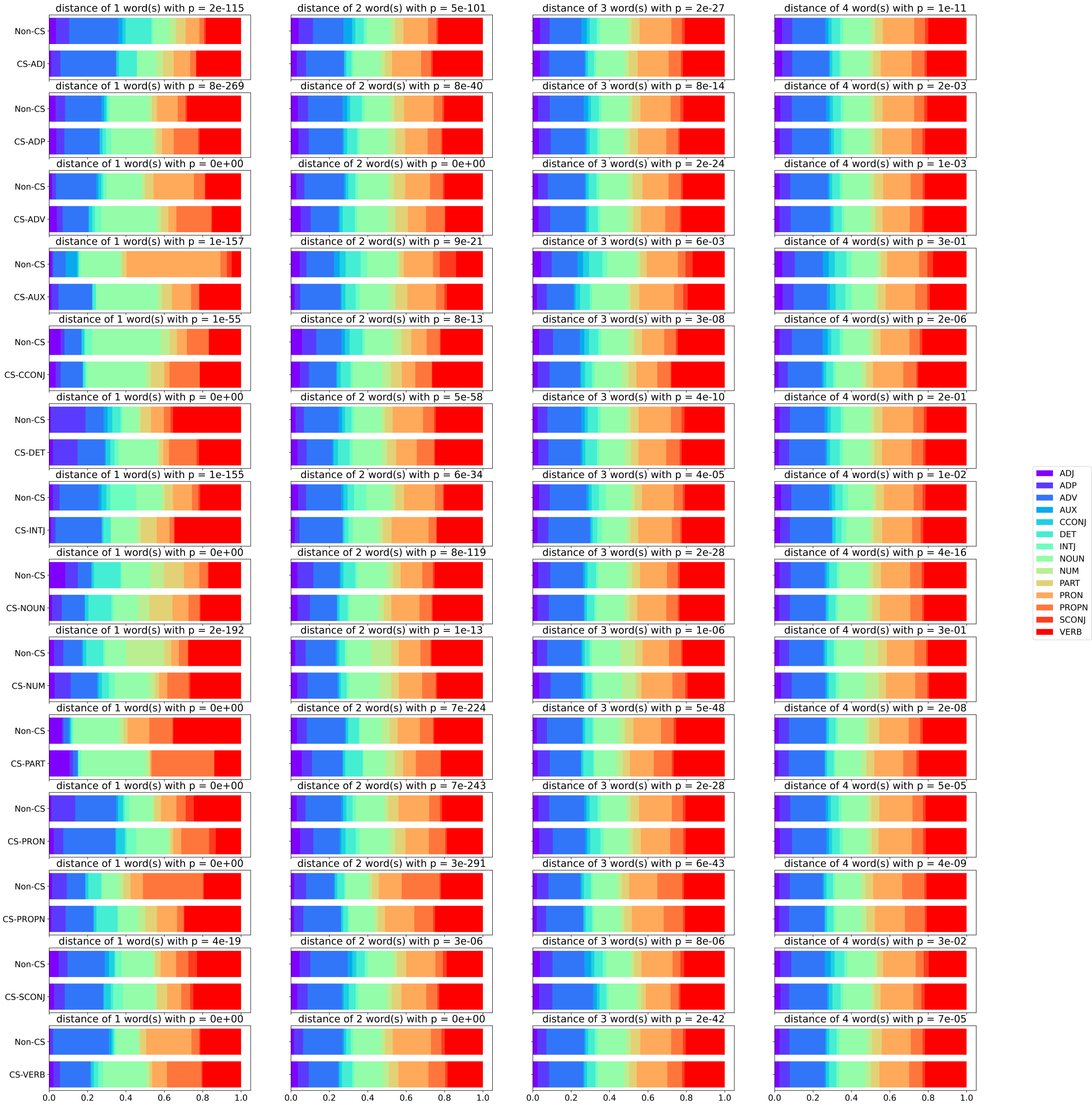


Figure 3: The visualization of the distribution of POS for words positioned at 1-4 words before CS points in SEAME.

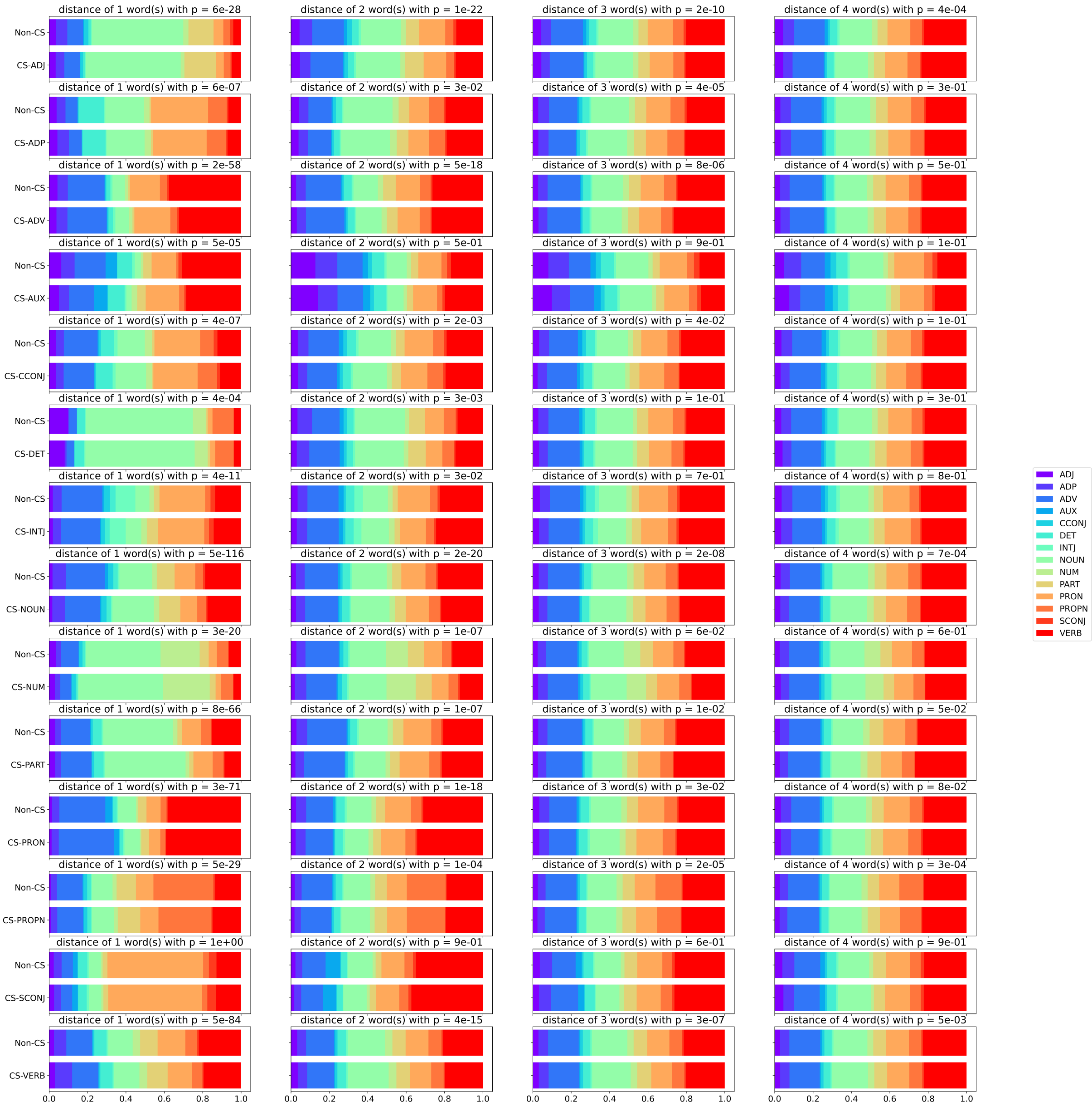


Figure 4: The visualization of the distribution of POS for words positioned at 1-4 words after CS points in SEAME.

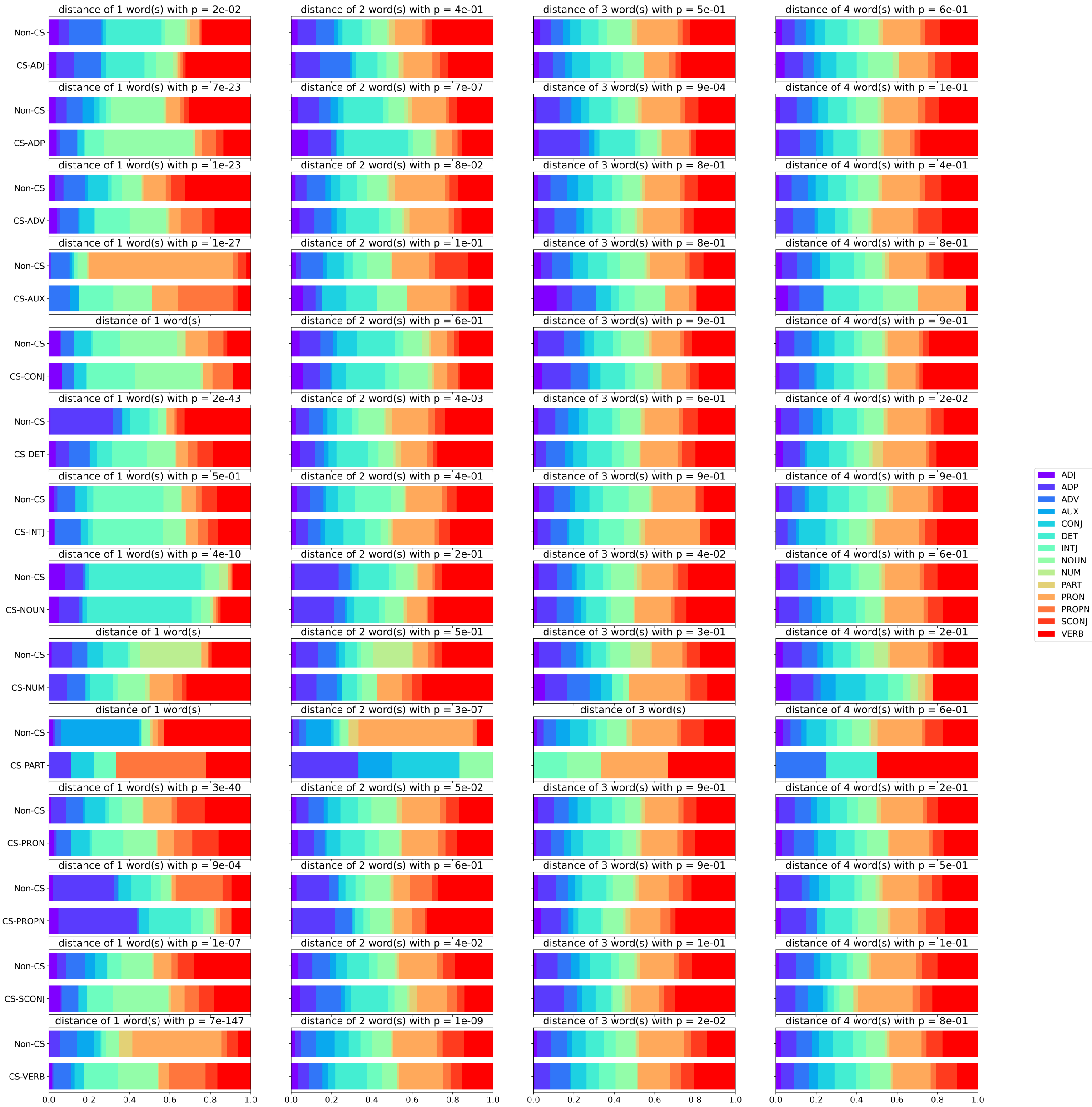


Figure 5: The visualization of the distribution of POS for words positioned at 1-4 words before CS points in BM.

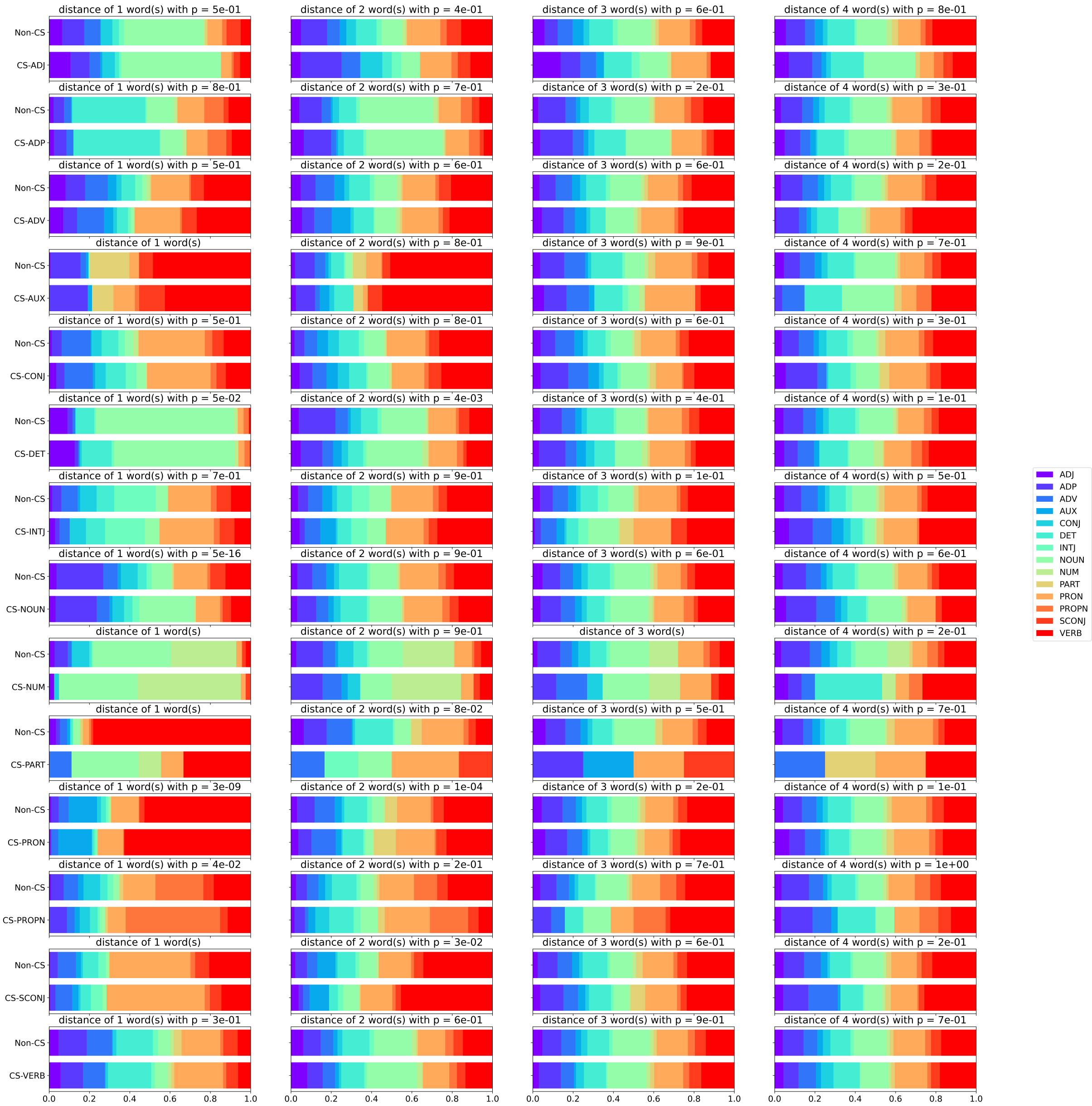


Figure 6: The visualization of the distribution of POS for words positioned at 1-4 words after CS points in BM.