Unsupervised Parsing by Searching for Frequent Word Sequences among Sentences with Equivalent Predicate-Argument Structures

Junjie Chen
The University of Tokyo
christopher@is.s.u-tokyo.ac.jp

Xiangheng He
Imperial College London
x.he20@imperial.ac.uk

Danushka Bollegala
The University of Liverpool
danushka@liverpool.ac.uk

Yusuke Miyao
The University of Tokyo
yusuke@is.s.u-tokyo.ac.jp

Abstract

Unsupervised constituency parsing focuses on identifying word sequences that form a syntactic unit (i.e., constituents) in target sentences. Linguists identify the constituent by evaluating a set of Predicate-Argument Structure (PAS) equivalent sentences where we find the constituent appears more frequently than non-constituents (i.e., the constituent corresponds to a frequent word sequence within the sentence set). However, such frequency information is unavailable in previous parsing methods that identify the constituent by observing sentences with diverse PAS. In this study, we empirically show that **constituents correspond to frequent word sequences in the PAS-equivalent sentence set**. We propose a frequency-based parser *span-overlap* that (1) computes the span-overlap score as the word sequence's frequency in the PAS-equivalent sentence set and (2) identifies the constituent structure by finding a constituent tree with the maximum span-overlap score. The parser achieves state-of-the-art level parsing accuracy, outperforming existing unsupervised parsers in eight out of ten languages. Additionally, we discover a multilingual phenomenon: participant-denoting constituents tend to have higher span-overlap scores than equal-length event-denoting constituents, meaning that the former tend to appear more frequently in the PAS-equivalent sentence set than the latter. The phenomenon indicates a statistical difference between the two constituent types, laying the foundation for future labeled unsupervised parsing research.

1 Introduction

Unsupervised constituency parsing focuses on identifying the constituent structure for target sentences. The constituent structure consists of a set of

We release the code and data at https://github.com/junjiechen-chris/ACL24-SpanOverlap
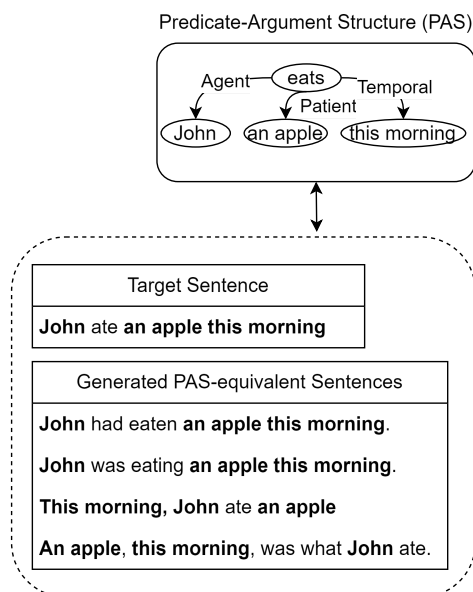

Figure 1: A target sentence and a set of PAS-equivalent sentences. We highlight in bold constituents that are frequent in the PAS-equivalent sentence set.

syntactically self-contained word sequences, commonly known as constituents. In addition to being syntactic units, those constituents encode semantic units in the sentence's semantic structure (Carnie, 2007; Fasold and Connor-Linton, 2014). Among all possible word sequences that can arise in natural language, the constituent is one of the few word sequences that can describe the semantic unit. As a result, we would have a higher chance of repeatedly observing the constituent than random word sequences given the semantic unit. In other words, we can observe the constituent as a frequent word sequence among semantically equivalent sentences. We refer to the frequent word sequence as a *word sequence pattern*.

In this paper, we empirically show a hypothesis: **constituents correspond to frequent word sequences in the set of sentences with equivalent Predicate-Argument Structures (PAS)**. PAS ab-

	Results were released after the market closed.
Penn Treebank	Markets stopped trading in many securities.
	Mr. Bandow is a Cato Institute fellow.
	It's an apple that John eats this morning.
Constituency Test	What John eats this morning is an apple.
	An apple, John eats this morning

Table 1: A comparison of data in Penn Treebank and data produced by constituency tests. The constituency test data are byproducts of testing the constituency of "an apple".

stracts the sentence semantics as an event with three elements: (1) predicates of the event; (2) participants of the event (i.e., arguments); (3) the semantic relationship between the predicates and arguments (i.e., thematic relations) (Palmer et al., 2005) (Figure 1). We refer to two sentences as *Predicate-Argument Structure (PAS) equivalent* when they have the same PAS elements. Controlling the PAS provides a good approximation to controlling the semantic structure. Figure 1 exemplifies our hypothesis with a target sentence, "John ate an apple this morning." Word sequences like "an apple" are constituents in the target sentence, whereas word sequences like "John ate" are non-constituents. Compared to the non-constituent, the constituent appears more frequently in the target sentence's PAS-equivalent sentence set, manifesting itself as a word sequence pattern.

However, such frequency information was not available to previous unsupervised parsing methods. Those previous methods focused on uncovering the constituent structure by either inducing a complex grammar (Kim et al., 2019; Yang et al., 2021) or evaluating whether a word sequence can withstand syntactic transformations (Cao et al., 2020; Wu et al., 2020; Li and Lu, 2023). Both methods operate upon sentences with diverse PAS, as shown in the upper half of Table 1. These sentences provide little information about the constituent structure, as evidenced by the low frequency of word sequences corresponding to constituents. Linguistic researchers perform constituency tests (e.g., movement) (Carnie, 2007) using handcrafted sentences with equivalent PAS, as shown in the lower half of Table 1. These PAS-equivalent sentences, as discussed above, provide significantly more information about the constituent structure, as evidenced by the high frequency of word sequences corresponding to constituents. The gap in data characteristics motivates our research on unsupervised parsing using PAS-equivalent sentences.

We apply the hypothesis to unsupervised parsing and propose a frequency-based method, *span-overlap*, to exploit the frequency information. The span-overlap method (1) generates a PAS-equivalent sentence set using a Large Language Model (LLM), (2) computes the span-overlap score for word sequences as the sequence's frequency in the sentence set, and (3) identifies the constituent structure by finding the constituent tree with maximum span-overlap scores. The span-overlap parser obtains higher parsing accuracy than state-of-the-art unsupervised parsers in eight out of ten languages evaluated and by a large margin. The span-overlap score can effectively separate constituents from non-constituents, especially for long constituents. These results highlight the utility and applicability of the frequency information to unsupervised parsing. In addition to the hypothesis, we discover a multilingual phenomenon: participant-denoting constituents tend to have higher span-overlap scores than equal-length event-denoting constituents, meaning that the former tend to appear more frequently in the PAS-equivalent sentence set than the latter. The phenomenon indicates a statistical difference between the two constituent types, laying the foundation for labeled unsupervised parsing research.

2 Background

Unsupervised constituency parsers seek to identify unlabeled constituents for any given sentence. A constituent in a sentence $w := (w[1], ..., w[n])$ is a consecutive word sequence $w^{i,j} := (w[i], ..., w[j])$ that functions as an independent unit in the structure. The word sequence can be represented as a two-tuple (i, j), referred to as a span. We measure the parsing accuracy by the F1 score between the set of predicted spans and the set of gold spans (Klein and Manning, 2002).

Unsupervised parsers make predictions by (1) evaluating a span score for each word sequence and (2) finding a tree optimizing the span score. Grammar-based unsupervised parsers (Yang et al., 2021; Zhao and Titov, 2020) compute the span score as the posterior span probability $p_{\mathcal{G}}((i, j)|w)$ with respect to a probabilistic grammar $\mathcal{G}$ (i.e., how likely the grammar selects (i, j) as a constituent). The grammar-based parsers predict constituents by maximizing the span score, seeking to maximize the probability of the prediction containing constituents. Language model-based parsers (Cao

et al., 2020; Wu et al., 2020; Li and Lu, 2023) compute the span score by whether the language model reacts to syntactic transformations involving the span. The language model-based parsers predict constituents by minimizing the span score, searching for constituents most independent of the remainder of the sentence.

The language model-based unsupervised parsing methods have attracted much attention lately due to the language model's applicability in various syntactic (Hu et al., 2020; Futrell et al., 2019) and semantic tasks (Huang and Chang, 2023). Given a word sequence $w^{i,j}$, the method evaluates the span score as the distance d between w (the original sentence) and $f(w, (i, j))$ (the transformed sentence designed to test the constituency of $w^{i,j}$), as shown in Equation 1. Cao et al. (2020) measures the distance by the change in the acceptability between w and $f(w, (i, j))$ using an external acceptability model trained with synthetic data. Wu et al. (2020) and Li and Lu (2023) measure the difference by the distance in language model embeddings between w and $f(w, (i, j))$.

$$\text{score}(w, (i, j)) := d(w, f(w, (i, j))) \quad (1)$$

Recent advances in Large Language Models (LLM) have led to a surge in the application of LLM-enhanced data to Natural Language Processing systems (Yang et al., 2023; Perez et al., 2023; Liu et al., 2022). Given an instruction (typically in the form of prompts), the LLM iteratively generates its answer based on the instruction and partially generated answers. Recent studies (Zhou et al., 2023; Chia et al., 2023) point out that state-of-the-art LLMs can follow difficult instructions and generate the correct answer with over 90% accuracy. This excellent instruction-following capability enables us to sample PAS-equivalent sentences from the LLM.

3 Related Works

Grammar-based unsupervised parsers (Klein and Manning, 2002; Kim et al., 2019; Zhao and Titov, 2020) learn probabilistic grammar by maximizing the likelihood of sentences observed in a corpus. However, these sentences have diverse PAS to the degree that constituent word sequences can only be observed a limited number of times. The limited observation provides insufficient positive evidence on whether a word sequence is a constituent through the sequence's frequency. Also, the grammar has to reject non-constituent but frequent word

sequences tied to certain expressions. For example, the word sequence "it is" is a non-constituent but frequent word sequence in sentences like "(it (is (a book)))". The lack of positive evidence and the abundance of false-positive evidence make up the main challenges in unsupervised constituency parsing. The two issues likely contribute to the low performance of the grammar-based method in non-English languages (see Table 3 and Table 4). Our span-overlap method is immune to the two issues. Our method detects the constituent-related word sequence pattern, utilizing the positive evidence revealed in the PAS-equivalent sentences. The PAS-equivalent sentences express the same PAS in diverse ways, effectively suppressing the frequency of the expression-tied non-constituent word sequences (i.e., false-positive evidence). Our method achieves significantly higher parsing performance than the grammar-based methods in non-English languages.

Previous language model-based methods conduct constituency tests by applying handcrafted templates to word sequences. However, different languages implement the syntactic transformation differently, requiring templates tailored to each language. To circumvent the problem, Cao et al. (2020) limit the scope of languages to English, which results in an English-only parser. Wu et al. (2020) and Li and Lu (2023) limit the transformation to movement and standalone tests only, which limits their parser's performance. By leveraging LLMs for performing the transformation, our span-overlap method can cover a diverse set of transformations while being applicable to many languages. Furthermore, the previous method has to perform the constituency test for every word sequence, incurring $O(n^2)$ language model evaluations for sentences with n words. Our method only generates a few dozen acceptable PAS-equivalent sentences for each target sentence, incurring only $O(1)$ evaluations of the LLM. The low complexity allows for more efficient language model-based parsing. For example, our method takes 14 minutes to parse the development set of Penn Treebank (Marcus, Mitchell P. et al., 1999) whereas Li and Lu (2023)'s method takes about an hour.

Researchers have long pointed out that language models capture shallow syntax, such as subject-verb agreements (Hu et al., 2020) and phrase segmentations (Luo et al., 2019; Downey et al., 2022). However, the shallow syntax covers only a small subset of information that hierarchical constituent
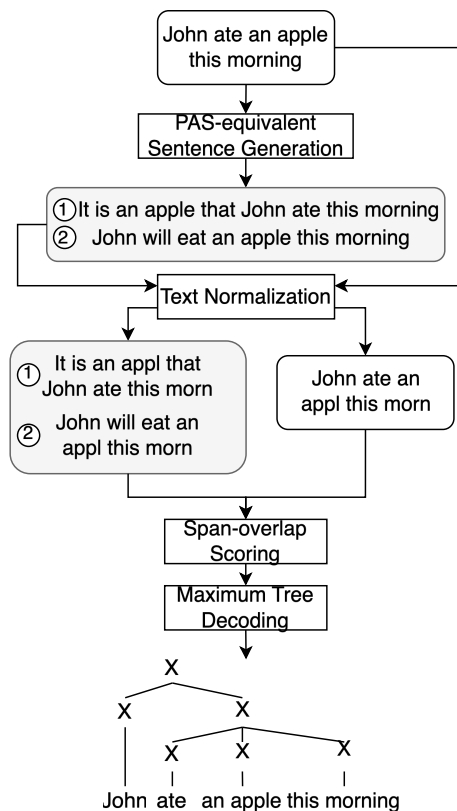
Figure 2: Overview of the span-overlap method. The grey box indicates the set of PAS-equivalent sentences.

Original Sentence	
John ate an apple this evening	
Instructions	PAS-equivalent Sentences
Movement	This evening, John ate an apple.
Question Conversion	Did John eat an apple this evening
Tense Conversion	John will eat an apple this evening
Clefting	It is John who ate an apple this evening
Passivization	An apple has been eaten by John this evening
Heavy NP Shift	John ate this evening an apple.

Table 2: Example of PAS-equivalent sentences produced by the six instruction types.

structures provide. For example, knowing "a beautiful New York City" as an NP does not provide any information about the NP's internal structure (i.e., "(a (beautiful ((New York) City)))"). Language model-based methods utilize the language model's ability to identify shallow syntax for reconstructing the hierarchical constituent structure. Our span-overlap method represents an improvement over previous language model-based methods, as discussed in the above paragraph. As we will see in Section 5, both our span-overlap parser and the baseline language model-based parser outperform the GPT-3.5 parser in all languages. This result indicates that LLM alone cannot recover the full constituent structure, prompting further research on language model-based unsupervised parsing.

4 Method

This section introduces the proposed span-overlap method (Figure 2). Given a target sentence w, the span-overlap method performs parsing in four steps: (1) Generating a set of sentences $s = \{s_1, s_2, ...\}$ where all s_k are PAS-equivalent to w; (2) Normalizing words in w and $s_k \in s$; (3) Mea-

suring the frequency of all word sequences of w within the sentence set s. We denote each word sequence that extends from the i-th word to the j-th word as $w^{i,j}$. We refer to the frequency of $w^{i,j}$ within s as the *span-overlap score* for the span (i, j); (4) Performing maximum tree decoding over the span-overlap score.

Firstly, we generate a PAS-equivalent sentence set s for w using an instruction-following LLM (e.g., GPT-3.5). We instruct the LLM with six types of PAS-preserving instructions listed in Table 2. The movement instruction contains four sub-instructions common in most languages (topicalization, preposition, postposition, and extraposition). The instruction models that constituents can be moved around as an independent unit. The question conversion instruction asks the LLM to convert w to a yes-no question or a wh-question. The tense conversion instruction asks the LLM to convert w to another tense. The instruction models the closer relationship between the verb and object via adding or removing auxiliaries between the verb and subject. The clefting instruction asks the LLM to construct a cleft sentence based on w, modeling subject and object extractions in embedded clauses. The passivization asks the LLM to convert w from an active voice to a passive voice. The language-specific instruction contains three sub-instructions: scrambling, ba-construction, and heavy NP shift. We present the instruction configuration in Appendix A.2.

Secondly, we normalize words in w and $s_k \in s$ to reduce incorrect word sequence mismatches due to unmatching inflectional suffixes or misaligned tokenization settings. We perform stemming and contraction normalization in this step. The stemming removes the inflectional suffixes in a word (e.g., working → work). The contraction normalization expands one contracted word to two separate words (e.g., didn't → did n't), aligning with the treebank's tokenization.

Thirdly, we compute the span-overlap score as

the frequency of the normalized $w^{i,j}$ within the normalized sentence set s (Equation 2). The span-overlap score is an exact matching score, requiring $w^{i,j}$ to be a substring of sentences $s_k \in s$. We also present results in appendix A.1 for a span-overlap variant using fuzzy matching scores.

$$\text{score}(w, (i, j)) := \frac{\sum_{s_k \in s} \mathbb{1}(w^{i,j} \text{ is a substring in } s_k)}{|s|} \tag{2}$$

Lastly, we decode the constituent structure by finding a tree with maximum span-overlap scores. We apply the Viterbi algorithm (Cohen et al., 1971) for this purpose.

5 Experiment

5.1 Experimental Settings

We conduct experiments in 10 languages, namely English (Penn Treebank (PTB)) (Marcus, Mitchell P. et al., 1999), Chinese (Penn Chinese Treebank 5.1 (CTB)) (Palmer, Martha et al., 2005)[1], German, French, Hungarian, Swedish, Polish, Basque, Hebrew, and Korean (SPMRL) (Seddah et al., 2013). We apply the same training/development/testing split as Li and Lu (2023). We use the development section for analysis and the testing section for comparison with existing unsupervised parsers. As per previous research (Shen et al., 2018; Kim et al., 2019), we evaluate the parsing accuracy using the unlabeled sentence-F1 (SF1) score on sentences of length ≥ 2. We drop punctuation (e.g., comma and period) and trivial spans (i.e., spans containing a single word and the sentence level span) when computing the SF1 score.

We generate PAS-equivalent sentences using an instruction-following LLM GPT-3.5-turbo-0125. We apply English instructions for all languages but add a specific directive asking the LLM to generate native samples for non-English languages. We manually examined 100 samples in English and Chinese and found that most samples had PAS identical to or similar to the target sentence.[2] We apply the snowball-stemmer[3] for English, French, German, Basque, Hungarian, and Swedish, and the stempel-stemmer[4] for Polish. We skip the stemming process for Chinese, Korean, and Hebrew. Chinese words do not take inflections. Hebrew

and Korean do not have a readily available stemmer. We normalize contractions only in English and French because of the common use of contractions in these languages. While not implementing the full normalization may harm the parsing accuracy in many languages, the span-overlap parser achieves state-of-the-art level performance without the full normalization in those languages.

We analyze the span-overlap score for two types of constituents: participant-denoting and event-denoting constituents. The span-overlap score measures the word sequence frequency within the PAS-equivalent sentence set. The objective is to identify how the language model behaves when generating constituents of different characteristics. We refer to a constituent as participant-denoting when it denotes a participant in the PAS. Likewise, we refer to a constituent as event-denoting when it denotes an event in the PAS. For example, Noun Phrases (NP) and Prepositional Phrases (PP) are common participant-denoting constituents, whereas Verb Phrases (VP) and Inflectional Phrases (IP) are common event-denoting constituents. We compare the span-overlap score for phrase types with more than twenty occurrences. We normalize the phrase type annotation for three languages (French, Polish, and Basque) whose annotation does not align with the PTB's annotation. In French, we merge the Srel, Sint, and Ssub phrases into S, and the VPpart and VPinf into VP. In Polish, we treat the zdanie phrase as S, the fwe as VP, and the fno as NP (Woliński et al., 2018). In Basque, we treat the SN phrase as NP and the SP phrase as PP (de Ilarraza Sánchez et al., 2008).

5.2 Comparison with the State-of-the-Art

Table 3 and Table 4 showcase the effectiveness and applicability of the frequency information revealed in the PAS-equivalent sentence set. The two tables compare the span-overlap parser with three PCFG-based parsers (N-PCFG, C-PCFG (Kim et al., 2019), and TN-PCFG (Yang et al., 2021)) and the latest language model-based parser (MLM-IP (Li and Lu, 2023)). We also include a GPT-3.5 parser that predicts PTB style parse tree in its serialized form (e.g., "(S (NP (DT The) (NN value)) (VP (VBD was) (VP (VBN disclosed)))")"). We drop trivial and duplicate spans when computing the unlabelled SF1 score for the GPT-3.5 parser.

The span-overlap parser represents a significant improvement over previous unsupervised parsers, outperforming the best baseline parser in eight out

---

[1]The PTB and CTB datasets are available under the LDC User Agreement for Non-Members license.

[2]More details in Appendix A.3

[3]Available as a part of the NLTK package.

[4]Available as a part of the egothor project.

SPMRL	German	French	Hungarian	Swedish	Polish	Basque	Hebrew	Korean	Mean
GPT-3.5	29.4	25.0	28.7	35.2	40.5	21.1	2.1	4.0	23.2
N-PCFG†	37.8	42.2	37.9	14.5	31.7	30.2	41.0	25.7	32.6
C-PCFG†	37.3	40.5	38.3	23.7	32.4	27.9	39.2	27.7	33.3
TN-PCFG†	47.1	39.1	43.1	40.0	48.6	36.0	39.2	35.4	40.8
MLM-IP‡	40.8	**48.7**	39.1	46.3	53.3	44.0	**50.4**	43.7	45.7
Span Overlap	**49.5**	48.5	**48.1**	**51.2**	**61.6**	**49.7**	43.8	**48.1**	**50.1**

Table 3: SF1 score on SPMRL's test set. The best results are highlighted in bold. †: Results from (Yang et al., 2021). ‡: Results from (Li and Lu, 2023).

	English	Chinese
GPT 3.5	36.2	5.1
N-PCFG†	50.8 (50.8)	29.5 (34.7)
C-PCFG†	55.2 (52.6)	39.8 (37.3)
TN-PCFG†	**57.7** (51.2)	39.2 (39.3)
MLM-IP‡	49.0 (46.4)	- (37.8)
Span Overlap	52.9	**48.7**

Table 4: SF1 score on the English (PTB) and the Chinese (CTB) test set. The best results are highlighted in bold. †: results from (Yang et al., 2021); ‡: results from (Li and Lu, 2023); -: result not reported. Numbers in parentheses () are our reproduction with the released code and data.

of ten languages. The span-overlap parser outperforms the MLM-IP parser by 3.99 SF1, 10.9 SF1, and 4.32 mean SF1 scores in the English PTB, the Chinese CTB, and the multilingual SPMRL dataset, respectively. It respectively outperforms the best grammar-based parser (TN-PCFG) by 9.5 SF1 and 9.31 mean SF1 scores in the Chinese CTB and the multilingual SPMRL dataset but underperforms the TN-PCFG parser by 4.7 SF1 score in the English PTB dataset. The high parsing accuracy establishes the span-overlap parser as a simple but competitive multilingual unsupervised parser.

The strong parsing performance is not solely due to the application of the GPT-3.5 model. As shown in the table, the GPT-3.5 parser is the worst-performing baseline parser in the PTB, CTB, and SPMRL dataset, achieving 36.2 SF1 in the PTB, 5.1 SF1 in the CTB, and 23.2 mean SF1 score in the SPMRL dataset. The low parsing accuracy indicates that the GPT-3.5 model can only recover a small portion of the constituent structure. In contrast, the proposed span-overlap parser uncovers significantly more constituent structures, outperforming the GPT-3.5 parser by 16.7 SF1 in the PTB, by 43.6 SF1 in the CTB, and by 26.86 mean SF1 score in the SPMRL dataset.

5.3 Participant-denoting Constituents Tend to Have Higher Span-overlap Scores than Equal-length Event-denoting Constituents

In this section, we demonstrate a statistical difference between equal-length participant-denoting

and event-denoting constituents through the span-overlap score. We focus on four participants-denoting constituents (i.e., Noun Phrases (NP), Prepositional Phrases (PP), Adpositional Phrases (AP), and Quantifier Phrases (QP)) and four event-denoting constituents (i.e., Declarative Clauses (S), Verb Phrases (VP), Inflectional Phrases (IP), and Complementizer Phrases (CP)). We also include the span-overlap score for random word sequences with the same sequence length statistics as gold constituents (i.e., For each random sequence, there is a gold constituent with the same sequence length).

We begin by observing the difference between the two constituent types regardless of the constituent's sequence length. Figure 3 shows the mean span-overlap score along with the score distribution for NP, PP, QP, S, and VP in the English PTB dataset. The AP, IP, and CP are not annotated in the dataset and, therefore, cannot be shown in the figure. The three participant-denoting constituents (i.e., NP, PP, and QP) have higher mean scores than the two event-denoting constituents (i.e., S and VP). The distribution figure shows that most participant-denoting constituents have high scores, whereas most event-denoting constituents do not. We can characterize this feature using the skewness value of the score distribution (Russo, 2020). A lower skewness value indicates more constituents with high scores than those with low scores. As shown in the figure, the participant-denoting constituents have a negative skewness value, whereas the event-denoting constituents have a positive skewness value. This skewness value indicates that participant-denoting constituents tend to appear more frequently in the PAS-equivalent sentence set than event-denoting constituents. The comparison establishes the use of the mean score and skewness value as a metric to reflect the statistical difference between the two constituent types.

Figure 4 demonstrates the statistical difference between equal-length participant-denoting and event-denoting constituents in English and Chinese. Comparing constituents of different lengths, we see
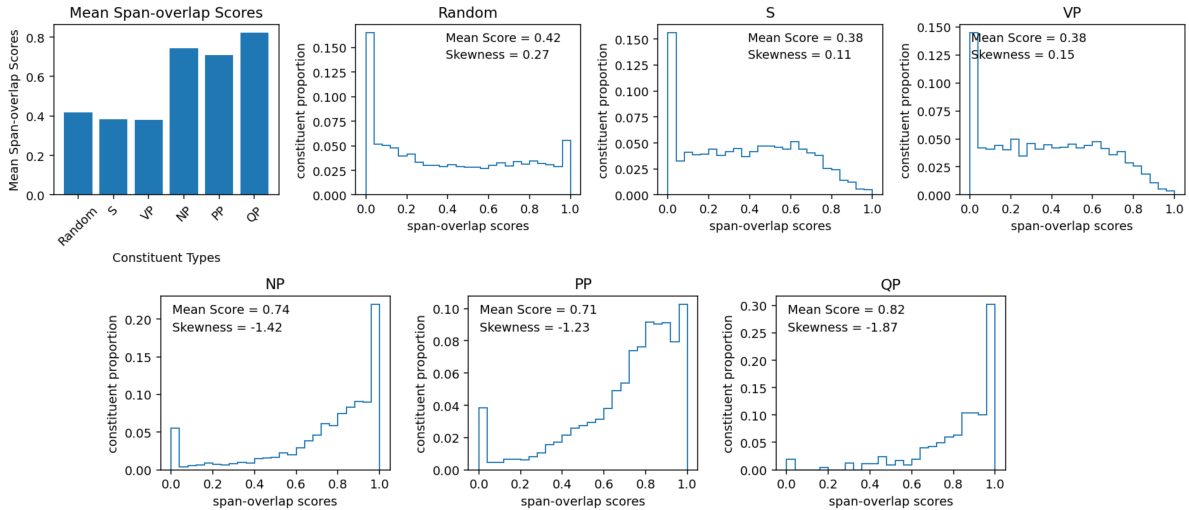
Figure 3: The mean span-overlap score and their distributions for S, VP, NP, PP, QP, and Random in the PTB development set. The distribution is approximated with histograms.
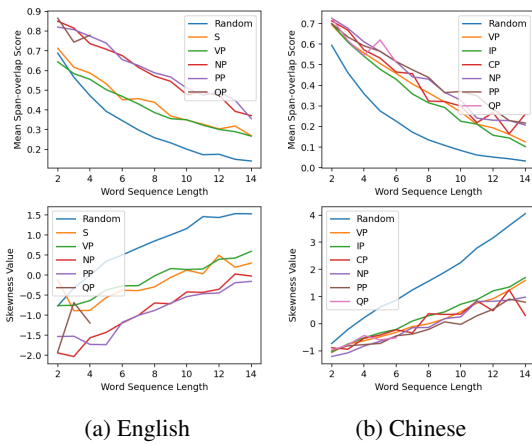

(a) English   (b) Chinese

Figure 4: Mean and skewness values of span-overlap scores.

that short constituents tend to have higher scores than long constituents. The correlation between span-overlap scores and sequence lengths prompts whether the statistical difference is due to the length difference. This is unlikely the case. As shown in the figure, participant-denoting constituents have higher mean span-overlap scores and lower skewness values than equal-length event-denoting constituents *over all sequence lengths*. In other words, participant-denoting constituents tend to appear more frequently in the PAS-equivalent sentences than equal-length event-denoting constituents. We include the mean score and skewness value for the other eight languages in Figure 6, which shows the same trend as in Chinese and English. These results highlight a multilingual phenomenon: participant-denoting constituents tend to have higher span-

overlap scores than equal-length event-denoting constituents. The phenomenon indicates a statistical difference between the two constituent types, shedding light on future unsupervised parsing research with phrase labels.

5.4 Span-overlap Score Distinguishes between Constituents and Non-constituents

This experiment provides a glimpse of the discriminating power of the span-overlap score. We compare the span-overlap score with random scores in terms of whether the score can identify the constituent in a set of mutually incompatible word sequences. The random score is uniformly sampled from [0, 1], the same range as the span-overlap score. We measure the proportion of gold constituents to which the score assigns a strictly higher value than all incompatible non-constituents. We refer to a non-constituent as incompatible with a gold constituent when the two sequences partially overlap (i.e., they cannot co-exist in the same tree).

Figure 5 plots the proportion by word sequence lengths. The span-overlap score distinguishes between constituents and non-constituents, identifying more constituents than the random score for all sequence lengths and in all languages. In the European languages, the span-overlap score reliably identifies around 50% of constituents, regardless of the constituents' length. In contrast, the random score identifies around 40% of short constituents but only around 10% of long constituents. The comparison highlights the effectiveness of the span-overlap score in identifying constituents, especially
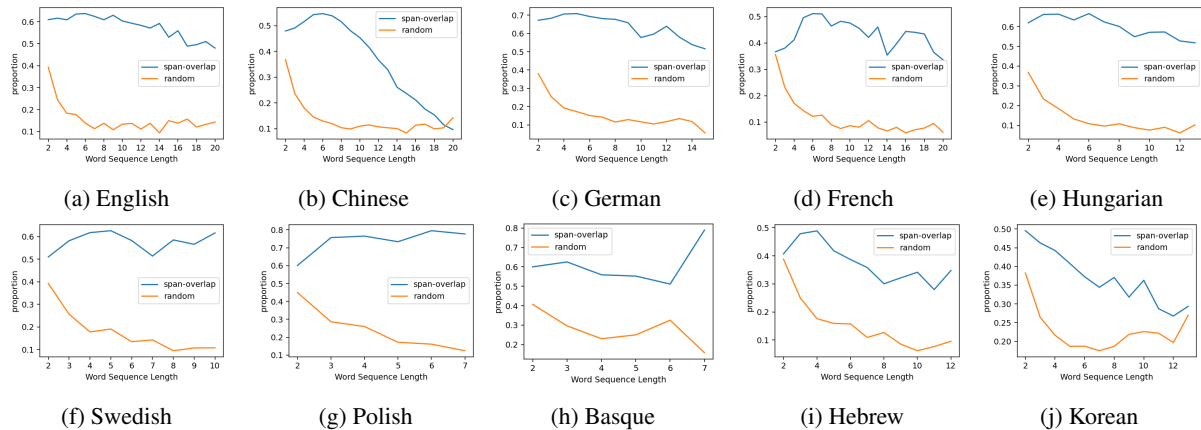
Figure 5: Proportion of gold constituents to which the scores assign strictly higher values than non-constituents

Ablation	English	Chinese	German	French	Hungarian	Swedish	Polish	Basque	Hebrew	Korean
Random	13.33	21.13	24.50	12.45	9.91	14.87	26.83	16.90	18.48	22.59
Right-branching	37.61	29.79	28.14	27.62	11.24	35.55	37.33	13.63	31.37	16.59
GPT-3.5	36.20	5.50	29.43	25.04	19.07	39.88	39.21	17.52	2.20	4.90
Movement	53.08	44.39	48.55	48.55	34.68	48.63	59.77	42.50	42.59	40.50
Question Conversion	50.95	40.90	42.31	42.78	34.01	45.25	53.29	34.96	38.57	35.09
Tense Conversion	42.93	36.81	33.05	30.53	19.65	37.20	44.40	18.77	35.83	24.22
Clefting	50.21	0	0	0	0	0	0	0	0	0
Passivization	52.15	37.09	39.11	42.65	26.57	41.20	51.47	25.11	39.92	29.06
Language-Specific	0	35.95	43.22	0	29.73	0	50.59	34.29	37.13	36.23
Best Instruction Type	53.08	44.39	48.55	48.55	34.68	48.63	59.77	42.50	42.59	40.50
Full Instructions	53.98	48.53	50.47	49.26	43.48	49.44	60.15	45.43	43.28	50.15

Table 5: SF1 scores for span-overlap parser with only one instruction type and with all 6 instruction types. Results are obtained using the development set.

for long constituents. However, Chinese and Koreans are exceptions. While the span-overlap score still identifies more constituents than the random score, the proportion drops for long constituents. This is because no long word sequences appear in both the target sentence and the generated PAS-equivalent sentences. The problem might be related to the sample quality problem discussed in Section 7.

5.5 Ablation of PAS-Preserving Instructions

Table 5 showcases the benefit of aggregating multiple PAS-preserving instructions. The table compares the span-overlap parser with only one instruction type and with all six instruction types. Even with a single instruction type, the span-overlap parser outperforms the random, right-branching, and GPT-3.5 parsers. This result suggests that the span-overlap method uncovers non-trivial constituent structures. The best parser with a single instruction type is always the one with the movement instruction. The phenomenon suggests a fundamental role of movable constituents in identifying constituent structures. Combining all instruc-

tion types sees significant improvements in parsing accuracy. The average improvement is 3.09 SF1 score between the parser with all instruction types and the best parser with a single instruction type. This result indicates the benefit and necessity of aggregating multiple instructions in generating the PAS-equivalent sentence set.

6 Conclusion

In this paper, we approached the unsupervised constituency parsing problem by exploiting word sequence frequency information revealed in a set of PAS-equivalent sentences. Firstly, we empirically verified a hypothesis: **constituents correspond to frequent word sequences in the PAS-equivalent sentence set**. We proposed a frequency-based method *span-overlap* that computes the span-overlap score as the word sequence's frequency in the PAS-equivalent sentence set. The span-overlap parser outperformed state-of-the-art parsers in eight out of ten languages evaluated and significantly outperformed the GPT-3.5 parser in all languages. The span-overlap score can effectively separate constituents from non-constituents, especially for long

constituents. These results highlight the effectiveness and applicability of the frequency information to unsupervised parsing. Additionally, our study uncovers a *multilingual* phenomenon: participant-denoting constituents tend to have higher span-overlap scores than equal-length event-denoting constituents, meaning that the former tend to appear more frequently in the PAS-equivalent sentence set than the latter. The phenomenon indicates a statistical difference between the two constituent types, laying the foundation for future labeled unsupervised parsing research. Finally, our ablation analysis shows a significant improvement by aggregating multiple PAS-preserving transformations, indicating the necessity of the aggregation in comprehensively exposing the constituent structure.

7 Limitations

The span-overlap method employs a diverse but non-exhaustive set of instructions to generate PAS-equivalent sentences. We did not cover PAS-preserving transformations that are unique to some languages. For example, Korean and Chinese have a null subject transformation that can remove the subject from sentences. The span-overlap method may miss out on word sequence patterns that those transformations cover. However, this paper aims to demonstrate the utility of word sequence frequency information and verify that constituents correspond to frequent word sequences. We will leave the incorporation of more complex language-specific transformations to future work.

Additionally, the span-overlap method faces challenges with sample quality, particularly in non-English languages. Previous research (Rosól et al., 2023; Seghier, 2023) has pointed out that the GPT-3.5 model performs significantly better in English than in other languages. Similarly, we observed that English samples are of better quality than Chinese samples. We refer the readers to Appendix A.3 for more details. We also observed that using GPT-4 instead of GPT-3.5 can significantly improve the parsing accuracy in a preliminary experiment. These results suggest that the sample quality currently limits the span-overlap parser. That being said, the span-overlap parser's performance indicates the utility of the word sequence frequency estimated with even low-quality samples. We expect the sample quality issue to be mitigated in the future with the rapid development of LLM technologies.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Unsupervised parsing via constituency tests. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4798–4808, Online. Association for Computational Linguistics.

Andrew Carnie. 2007. *Syntax: A Generative Introduction*. Blackwell, Malden, MA.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: towards holistic evaluation of instruction-tuned large language models. *CoRR*, abs/2306.04757.

A. Cohen, J. Heller, and A. Viterbi. 1971. A new coding technique for asynchronous multiple access communication. *IEEE Transactions on Communication Technology*, 19(5):849–855.

Arantza Díaz de Ilarraza Sánchez, Enrique Fernández-Terrones, Izaskun Aldezabal, and María Jesús Aranzabe. 2008. From dependencies to constituents in the reference corpus for the processing of basque (EPEC). *Proces. del Leng. Natural*, 41.

C.m. Downey, Fei Xia, Gina-Anne Levow, and Shane Steinert-Threlkeld. 2022. A masked segmental language model for unsupervised natural language segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–50, Seattle, Washington. Association for Computational Linguistics.

Ralph W. Fasold and Jeff Connor-Linton. 2014. *An Introduction to Language and Linguistics*, 2 edition. Cambridge University Press.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Florence, Italy. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jiaxi Li and Wei Lu. 2023. Contextual distortion reveals constituency: Masked language models are implicit parsers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5208–5222, Toronto, Canada. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongyin Luo, Lan Jiang, Yonatan Belinkov, and James Glass. 2019. Improving neural language models by segmenting, attending, and predicting the future. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1483–1493, Florence, Italy. Association for Computational Linguistics.

Marcus, Mitchell P., Santorini, Beatrice, Mary Ann Marcinkiewicz, and Taylor, Ann. 1999. Treebank-3. Linguistic Data Consortium.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, Martha, Chiou, Fu-Dong, Xue, Nianwen, and Lee, Tsan-Kuang. 2005. Chinese treebank 5.0. Linguistic Data Consortium.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit,

Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Maciej Rosól, Jakub S. Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. 2023. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Scientific Reports*, 13.

Riccardo Russo. 2020. *Statistics for the Behavioural Sciences: An Introduction to Frequentist and Bayesian Approaches*. Routledge.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA. Association for Computational Linguistics.

Mohamed L. Seghier. 2023. Chatgpt: not all languages are equal. *Nature*, 615:216.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron C. Courville. 2018. Neural language modeling by jointly learning syntax and lexicon. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Marcin Woliński, Elżbieta Hajnicz, and Tomasz Bartosiak. 2018. A new version of the składnica treebank of Polish harmonised with the walenty valency dictionary. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*

(LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. Association for Computational Linguistics.

Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021. PCFGs can do better: Inducing probabilistic context-free grammars with many symbols. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1487–1498, Online. Association for Computational Linguistics.

Yanpeng Zhao and Ivan Titov. 2020. Visually grounded compound PCFGs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4369–4379, Online. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

A Appendix

A.1 Span-overlap Scoring with Fuzzy Matching

The application of exact matching (i.e., the span only matches identical substrings in PAS-equivalent sentences) rather than fuzzy matching (i.e., the span matches similar substrings) requires further discussions. In a preliminary experiment, we found that the fuzzy matching variants of the span-overlap parser underperform the span-overlap parser utilizing exact matching. We experimented with two fuzzy matching variants, one based on the BLEU score (Papineni et al., 2002) and the other based on the METEOR score (Banerjee and Lavie, 2005). The two fuzzy matching scores allow the matching between two textually different word sequences that share the same meaning. For example, "give John a book" and "give a book to John".

$$\text{score}(w, (i, j)) := \frac{\sum_{s_k \in s} \max_p f(w^{i,j}, s_k^{p, p+j-i})}{|s|} \quad (3)$$

To introduce the fuzzy matching variants, we generalize Equation 2 as Equation 3. When the scoring function f is an indicator function, Equation 3 is the span-overlap score we introduced in the main text. In the BLEU and METEOR score-based variants, we employ the BLEU and METEOR scores for f, respectively. Table 7 compares the parsing accuracy of the two fuzzy matching variants against the exact matching variant on a 10-word subset of the development set of the PTB and the SPMRL German dataset. The 10-word subset contains sentences of length ≤ 10, as per previous research (Klein and Manning, 2002). As shown in the table, the fuzzy matching variants fall behind the exact matching variant in parsing accuracy by a large margin. Upon further investigation, we found that fuzzy matching scores not only match "give John a book" with grammatical strings (e.g., "give a book to John"), but also match with ungrammatical strings (e.g., "give John a book in"). The matching with ungrammatical strings might introduce unnecessary noise to the span-overlap model, which is detrimental to the parsing performance.

A.2 PAS-preserving Instructions

Table 6 shows the specific instruction we used in our experiment. To generate samples for non-English languages, we append an additional directive, "The generated sentences must be in <language>.", to the above instructions. For example, we add "The generated sentences must be in German." to generate German sentences.

A.3 Examining Sample Qualities

We manually examined 100 random pairs of (target sentence, generated PAS-equivalent sentence) in English and Chinese. We rate the PAS-equivalence in the range of 0-10 (i.e., whether two sentences express the same PAS information). We rate 10 if the two sentences express the same PAS information, 5 if half of the information from the target sentence is missing in the generated sentence, and 0 if the two sentences are completely irrelevant. We present some rating examples in Table 8.

We can make two observations from Figure 7: (1) Most generated sentences have a PAS that is reasonably similar to the PAS of the original sentence; (2) English samples tend to have higher PAS-equivalence scores than Chinese samples. These observations, on the one hand, confirm the plausibility of generating PAS-equivalent sentences with the GPT-3.5 model. On the other hand, the observa-

	Prompts
Movement	Generate diverse sentences by performing the topicalization transformation to the below sentence.
	Generate diverse sentences by performing the preposing transformation to the below sentence.
	Generate diverse sentences by performing the postposing transformation to the below sentence.
	Generate diverse sentences by performing the extraposition transformation to the below sentence.
Question Conversion	Generate diverse sentences by converting the below sentence to a yes-no question.
	Generate diverse sentences by converting the below sentence to a wh-question.
Tense Changing	Generate diverse sentences by converting the below sentence to other tenses.
Passivization	Generate diverse sentences by passivizing the below sentence.
Clefting	Generate diverse sentences by converting the below sentence to a cleft sentence.
Question Conversion	Generate diverse sentences by performing the scrambling transformation to the below sentence.
	Generate diverse sentences by performing heavy NP shift.
	Generate diverse sentences by converting the below sentence into ones with ba-construction.

Table 6: Instructions used in generating PAS-equivalent sentences.

(a) German   (b) French   (c) Hungarian   (d) Swedish
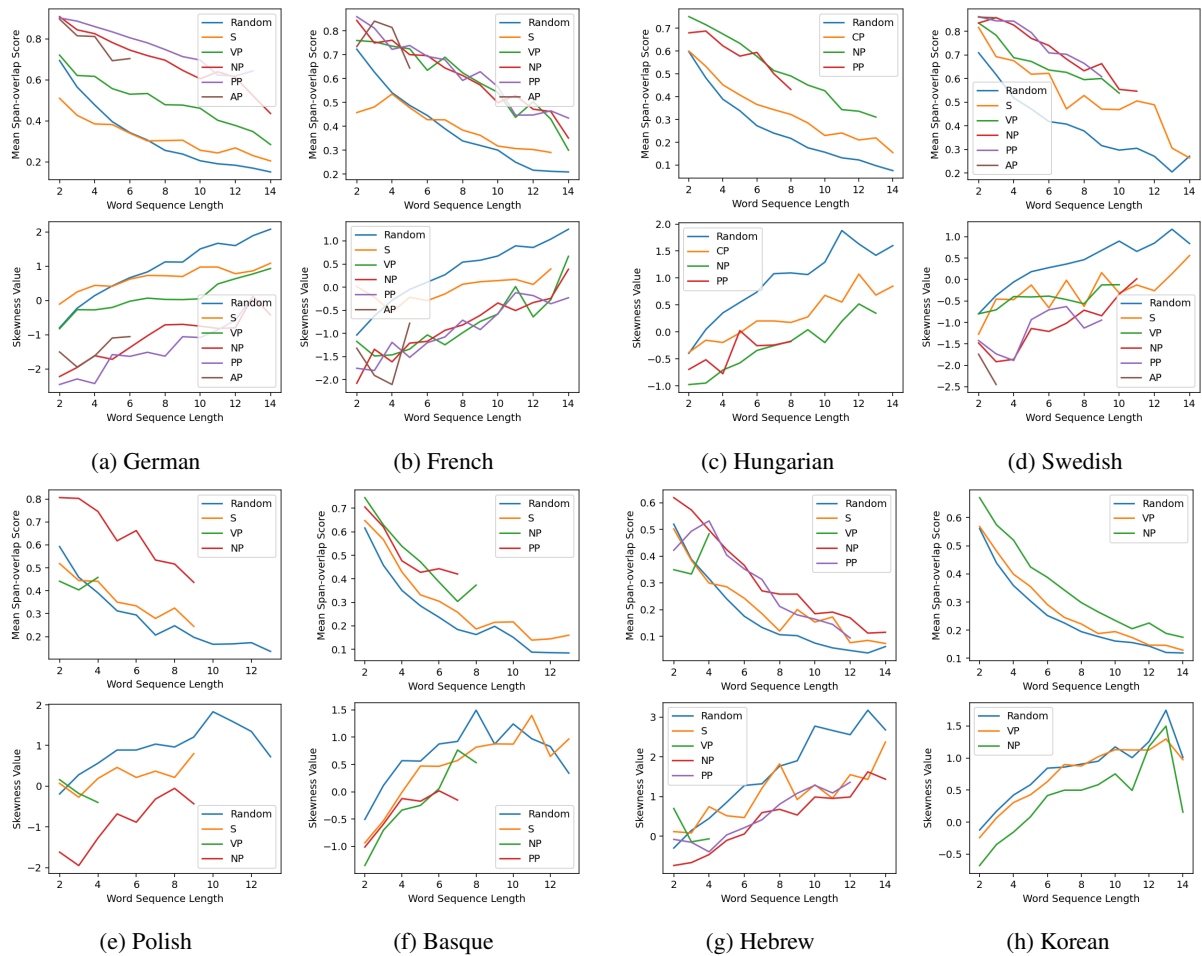
(e) Polish   (f) Basque   (g) Hebrew   (h) Korean

Figure 6: Mean and skewness values of span-overlap scores. The values are measured over development sets and are grouped by their sequence lengths.

	METEOR	BLEU	EM
PTB-DEV-10	59.7	62.9	66.6
SPMRL-DE-DEV-10	63.7	64.4	67.9

Table 7: Sentence-F1 score on the 10-word subset of the PTB and SPMRL German development set.

tions indicate a limiting factor for the span-overlap method: the sample quality.

| Original Sentence	Generated Sentence	Rating
Idle Thought	The child, lost in the park, found solace in a world filled with color and wonder.	0
BANKERS ACCEPTANCES : 8.40 % 30 days ; 8.35 % 60 days ; 8.27 % 90 days ; 8.20 % 120 days ; 8.15 % 150 days ; 8.02 % 180 days .	Among the options are bankers acceptances at 8.02% for 180 days, catering to different investment durations.	6
They did n't panic during the first round of selling this morning .	They had not panicked during the first round of selling this morning.	10

Table 8: Examples of PAS-equivalence rating
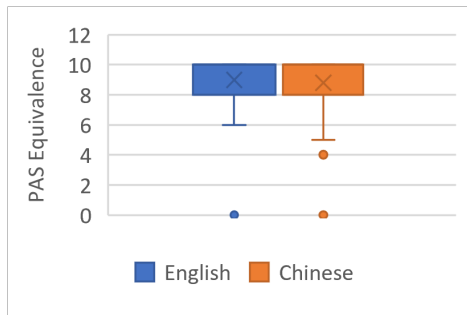
Figure 7: PAS-equivalence scores in English and Chinese