Glottometrics

52/2022

International Quantitative Linguistics Association

Glottometrics

Vol. 52, 2022

Glottometrics is an open access scientific journal for the quantitative research of language and text published twice a year by International Quantitative Linguistics Association (IQLA). The journal was established in 2001 by a pioneer of Quantitative Linguistics Gabriel Altmann.

Manuscripts can be submitted by email to glottometrics@gmail.com. Submission guideline is available at https://glottometrics.iqla.org/.

International Quantitative Linguistics Association (IQLA)

Friedmanngasse 50
1160 Vienna
Austria

eISSN 2625-8226

Contents

Dynamics of language in social emergency: investigating COVID-19 hot words on Weibo

Yikai Zhou[1], Rui Li[2], Guangfeng Chen[2], Haitao Liu[1*]

[1] Department of Linguistics, Zhejiang University
[2] College of Foreign Languages, Hunan University
[*] Corresponding author's email: lhtzju@yeah.net

DOI: https://doi.org/10.53482/2022_52_395

ABSTRACT

Drawing on word embeddings techniques and tracking the frequency and semantic change of hot words on Sina Weibo during the COVID-19 pandemic, this study investigates how language and discourse change during crisis. More specifically, correlation tests were conducted between word frequency ranks, pandemic data, and word meaning change ratio. Results indicated that the frequency of some hot words changed with both pandemic data and the frequency of other hot words, which were significantly correlated with the American pandemic data rather than that of China. Moreover, February of 2020 saw the most distinctive semantic changes marked by a large part of the nearest neighbors for WAR metaphors. The correlations between changes in the frequency and nearest neighbors of COVID-19 related hot words exhibited some acceptable peculiarities. This study proves the availability of studying discourse through language change by observing minor semantic change on connotation level from social media, which adds a new perspective to the impact of the COVID-19 pandemic.

Keywords: semantic change, social media, word embeddings, COVID-19 pandemic, discourse.

1 Introduction

Characterized as a pandemic by WHO (World Health Organization) on March 11, 2020, COVID-19 (Coronavirus Disease-19) has left impacts on various aspects of human life and society. Previous studies (e.g., Martikainen and Sakki 2021; Wicke and Bolognesi 2021) have investigated the social impacts of COVID-19 with discourse analysis. And metaphorical framings have been identified by many scholars in this case. For example, Wicke and Bolognesi (2021) discovered a decrease in war framing; Hafner and Sun (2021) investigated the role of metaphorical framings of the pandemic as a fight in the success of leadership in New Zealand. While these studies may shed new light on the pandemic's impacts, they did not particularly investigate COVID-19 discourse through the very fundamental semantic change. Since the semantic change of language use could inform social issues in turn, hot words, widely applied

to reflect social opinion and social trends (Liu et al. 2016), should imply the interplay between discourse and society during the COVID-19 pandemic.

With the advancement of information technology and artificial intelligence, linguistic and discourse studies are now faced with new challenges and opportunities. First, online social media has emerged as a new platform for communication, providing a large size of real language and discourse materials. As Sindoni and Moschini (2021) put, "materiality of media and semiosis of communication have both changed to the point where labels, such as new media, new literacies, new technologies, new genres, prioritize the newness of interactional phenomena and communicative events that take place in, and are shaped by, digital environments." Second, linguistic studies applying data-driven methods are now flourishing and detecting more detailed changes in the language system, resulting in deeper understandings of its major features such as being complex, human-driven, and self-adaptive (Liu 2018). Moreover, for discourse is "language as social practice determined by social structures" (Fairclough 1989, p. 17), the study of discourse therefore should not be restricted to language itself, but language in context.

Since word is a fundamental form of language and therefore discourse, the evolution of word meaning, or semantic change, warrants exploration for a better understanding of the dynamics of language in context and discourse. For word meaning, Firth (1957, p. 11) put it that "we shall know a word by the accompany it keeps". In other words, it is constructed within a context (Cruse 1986). In addition, word meaning is slippery (Saeed 2016). Especially, semantic changes on connotative levels are more difficult to capture since connotation is associated emotions, values, and differences according to the speaker's social standing and the term's social use (Bloomfiled 1933), the change of which cannot be immediately detected and concluded into dictionaries.

As a promising diachronic tool in semantic change analysis, word embeddings can capture minor semantic change on connotation levels (e.g., Čech et al. 2019; Hamilton et al. 2016). It is instructed by the distributional hypothesis, positing that word meanings are embedded in co-occurrence relationships (Firth 1957; Harris 1954). This technology has considerably enhanced the accuracy of investigations on semantic change, and therefore allows linguists to conclude some regularities. For instance, the rate of semantic change negatively correlates with word-usage frequency (Englhardt et al. 2019; Hamilton et al. 2016). Research applying word embeddings has obtained bountiful findings, however, further innovations are needed in the choice of materials (e.g., Grag et al. 2018; Mou et al. 2015), which were mostly formal texts from historical corpora (Englhardt et al. 2020) and less dynamic compared with instant data from online social media. By using word embeddings and retrieving real linguistic data relevant to certain influential social events on social media, more changes in connotative meaning, as well as traces of associated discourse may be found.

Given the aforementioned, despite the numerous studies published to date, it remains unclear whether language data from online social media are suitable for investigating language change identified in

changing context and discourse during social emergencies. With more than 500 million active users in 2020[1], Sina Weibo, a Chinese social media site from which big data can be collected, is an ideal real-world source of language data. Applying word embeddings, this study aims to investigate the COVID-19-related hot words on Weibo during the first half of the year 2020 and describe both the frequency and semantic change of these words, which can inform us on how to conceptualize the dynamism of pandemic and how to react to its development (Wicke and Bolognesi 2021). On achieving this purpose, this study is supposed to understand the subtle changes of language system as well as related discourse phenomena to help sketch the picture of our conception of the COVID-19 pandemic. Three research questions are to be addressed:

1) How does the use of relevant hot words on social media change in China during the COVID-19 pandemic?

2) Can the change of COVID-19 hot words reveal any discourse? If yes, what is it?

3) What are the possible causes of the above changes to online hot words and discourse during the COVID-19 pandemic?

2 Material

Two data sources, microblog texts on COVID-19 crawled from Sina Weibo (https://weibo.cn) using a self-edited python project and COVID-19 data from WHO's website (https://COVID19.who.int) were involved. For language materials, 64,453 pieces of valid texts were posted from December 31, 2019, to June 30, 2020, using "新冠" (*xin'guan*, COVID-19) or "肺炎" (*feiyan*, pneumonia) as keywords on Weibo to ensure to the relevance to COVID-19 pandemic. Details including the number and the average length of posts are presented in Table 1. Furthermore, since China and the US (United States) are the largest economies in the world with a high level of interdependence in trade and economy as well as a sound Sino-US relationship based on cooperation rather than conflict (Supadhiloke 2012) - changes in one's society would be quickly sensed by the other, we chose the pandemic data of these two countries as social situations that might affect people's mental representation of the pandemic. COVID-19 data obtained from WHO's website cover the new daily confirmed cases and deaths in China and the US during January 3, 2020, and June 30, 2020.

[1] Data accessed to *Weibo 2020 User Development Report* https://data.weibo.com/report/reportDetail?id=456.

Table 1: Details of language materials.

Month	Number of posts	Average post length (words per piece)
January	10,653	60.74
February	8,341	64.96
March	10,536	64.32
April	10,999	63.97
May	11,062	67.68
June	12,862	56.23

3 Methodology

3.1 Instruments

The data collection tool used in this study was a self-edited Python project, which could retrieve Weibo posts containing the defined keywords during the selected period. Information including the text, user id, post time could be collected at the same time. The database linked to this python project was MongoDB, in which all the information was stored. Jieba, a Python module for Chinese word segmentation was used for word segmentation and word counting.

To get the precise change of meanings of the target words in contexts, we applied the Word2vec model proposed by Mikolov et al. (2013a), which is now widely used in the field of natural language processing. With a given corpus, the Word2ec model represents each word in it with a list of (sometimes hundreds of) numbers called a vector. Originally, word vectors are encoded in a way called one-hot vector such that if we have a sentence (vocabulary) "Thank you very much", the vectors of each word should be: vector ("*thank*") = [1, 0, 0, 0], vector ("*you*") = [0, 1, 0, 0], vector ("*very*") = [0, 0, 1, 0] and vector ("much") = [0, 0, 0, 1]. But such vectors cannot represent the different distances (interpreted as similarities) between the semantics of each pair of words when they are projected to a two-dimensional space. To solve this problem, Word2vec model was designed to represent words in distributed vectors based on word dependence and minimize computational complexity (Mikolov et al. 2013b). It can do two things originally with two architectures (Mikolov et al. 2013a): the continuous bag-of-words model where all words get projected into the same position to predict the word based on the context, and the continuous skip-gram model where a word is put into a log-linear classifier with continuous projection layer to predict words with a certain range before and after the current one. A typical way to test and compare different word vectors is finding semantically similar words. For example, to find a word similar to *long*, we can compute vector X = *vector* ("*longest*")-*vector*("*long*")+*vector*("*short*"). Then a word in the vector space closet to X measured in terms of cosine is the result (Turney and Pantel 2010). This operation can be further done in high dimensional word vectors on a large amount of data, and the results can be used to answer very subtle semantic relationships between words (Mikolov et al. 2013a).

For example, the diachronic change of nearest neighbors that have the largest cosine values of the target words can be used to tack the semantic change of them (Shi and Lei 2019). Here we have the same application of calculating similar words (or nearest neighbors) with the Gensim library of Python, which provides all the features of a Word2vec model (Jatnika et al. 2019). Its implementation is a case where an open-source implementation is more efficient than the original Word2vec coding (Srinivasa-Desikan 2018). Once a cleaned text was input into it, Gensim worked out a vector map as well as nearest neighbors of the defined word in minutes.

In the current study, correlation tests were conducted on Statistical Product and Service Solutions (SPSS 26.0).

3.2 Research Procedures

Research procedures were observed as follows. First, after data collecting, the repetitive reposted Weibo texts were filtered out, and emojis, as well as URL strings were deleted before word counting. The cleaned Weibo texts were then processed by Python to get word frequencies and nearest neighbors.

Second, Weibo texts were compiled with their post date for Python to calculate word frequencies. In this regard, daily word frequencies and ranks were recorded. For identifying hot words and their nearest neighbors, texts were compiled into 6 monthly recorded files[2]. According to Gao and Wang (2017), hot words are widely used in actual network communication and are extensively spread in social life. Therefore, in our study, words that kept monthly top rankings were candidates for hot words. Then, Gensim ran on these 6 prepared files and printed the nearest neighbors of selected words.

Then, we conducted two sets of Pearson correlation tests. This was inspired by van Dijk's discourse-cognitive-social triangle (van Dijk 2009). In this model, the mental representations of language users as a major cognitive factor and social factors including social interaction, social situations, and social structures should be considered when explaining the relations between discourse and society mediated by cognition. In our case, pandemic data present social situation during the pandemic, and Weibo posts contain discourse as well as social cognition. First, informed by the finding of Li et al. (2020) that the declaration of COVID-19 in China changed frequency of words of emotions on Weibo, we tried to explore the correlation between pandemic data as social situation after the declaration and hot word frequency as public response. Secondly, inspired by co-occurrence rates measuring the correlation strength between any two highlighted words (Zhu et al. 2020), we also conducted correlations between frequency ranks of hot words in pairs. And thirdly, informed by Englhardt et al. (2020), we inspected correlations between word meaning change ratio (1) and monthly mean word frequency ranks. As defined in (1), the word meaning change ratio is calculated based on the number of new nearest neighbors

[2] Weibo corpus is available one line https://github.com/eddiezhou99/COVID-19-Weibo-Corpus.

of a certain word. It is noteworthy that generally, genism would output at most ten nearest neighbors of a word, but when there are not enough nearest neighbors, there would be fewer outputs.

$$\text{Word meaning change ratio (month)} = \frac{\text{number of unprecedented nearest neighbors}}{\text{number of nearest neighbors of that month}} \quad (1)$$

4 Results

4.1 Descriptive Statistics

In general, hot words are those with high frequencies within a period of time (Gao and Wang 2017; Liu et al. 2016), and semantics should also be considered (Wang et al. 2017). We therefore gave three criteria to COVID-19 hot words: (a) the words should be of high frequencies among all content words from the weibo posts during the observed time (January 2020 to June 2020), (b) they should be semantically relevant to the COVID-19 pandemic, and (c) for reducing redundancy of discussion, they shall be semantically distinctive or representative of a group of semantically similar words. With these three criteria in mind, 5 frequent words, viz. 新冠 *COVID-19*, 疫情 *pandemic situation*, 防控 *prevention and control*, 中国 *China* and 美国 *US* were identified as the COVID-19 hot words for inspection among the top 30 frequent words (see Table 2) in Weibo posts. Note that although there are some other possible candidates, we dropped them according to criterion (c) because they share more or fewer similarities to the chosen words that are more highly ranked (for example, 病例 *case* and 确诊 *confirmed* were often used to describe 疫情 *pandemic situation*).

Table 2: Top 30 frequent words in COVID-19-related weibo texts.

Rank	Word	Token	Frequency (%)	Rank	Word	Token	Frequency (%)
1	的 of	214173	4.3003	16	人 person	20527	0.4122
2	新冠 COVID-19	72687	1.4595	17	中国 China	20514	0.4119
3	肺炎 pneumonia	60747	1.2197	18	例 case, quantifier	19593	0.3934
4	了 end mark	59053	1.1857	19	美国 US	19447	0.3905
5	在 at	58752	1.1797	20	也 also	18830	0.3781
6	疫情 pandemic situation	54635	1.0970	21	为 for/be	17660	0.3546
7	是 be	39075	0.7846	22	不 negation	17488	0.3511
8	和 and	35214	0.7070	23	都 all	17451	0.3504
9	日 day	33341	0.6694	24	防控 prevention & control	16862	0.3386
10	月 month	32946	0.6615	25	患者 patient	15297	0.3071
11	病例 case	28721	0.5767	26	对 to	15290	0.3070
12	病毒 virus	27172	0.5456	27	感染 infection	15039	0.3020
13	确诊 confirmed	26146	0.5250	28	将 will	13610	0.2733
14	我 I	23808	0.4780	29	就 at once	13487	0.2708
15	有 have	20559	0.4128	30	工作 work	13399	0.2690

4.2 Frequency and Semantic Change of Hot Words

Word frequency helps analyze the public trend in certain social events (Rustam et al. 2021). By ranking word frequencies among our data, we may capture some social and psychological trends during the pandemic. As can be seen from Figure 1, although the frequency ranks of the 5 hot words in 6 months were above 30, their ranks changed drastically over time. At beginning of January, only 疫情 *pandemic situation* and 防控 *prevention and control* reached the top 50. And later on, the frequency of other words increased enormously. This is a representation of the growing trend of COVID-19 to enter the center of public attention.

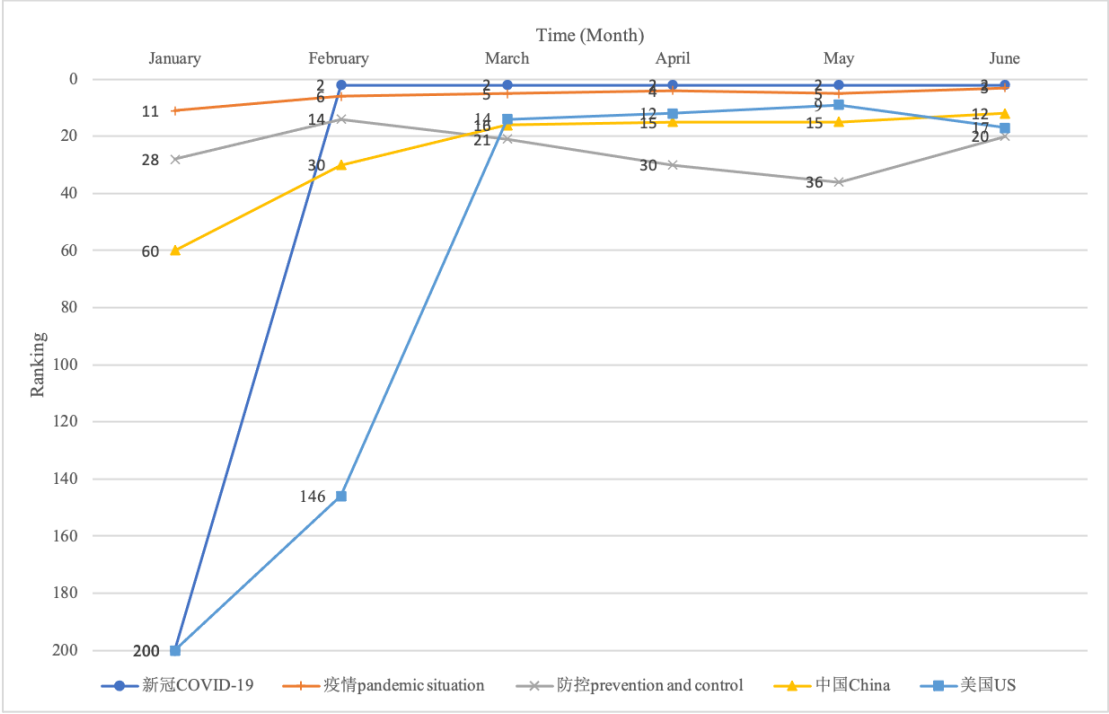

Figure 1: Diachronic change of the frequency rankings of hot words. Rankings exceeding 200 taken as 200 for visual clarity.

Figure 1 shows that in late February all the five words entered the range of the top 100. This is in line with the result by Rajput, Grover and Rathi (2020), who analyzed the word frequency of COVID-19-related tweets from January to April of 2020 and found peaks in February and March, with Coronavirus, Covid19, and Wuhan being the most frequent words. And these two months, in fact, also saw a peak in China and a rise in the US in the pandemic (see Figure 2). Therefore, it seems necessary to investigate the correlations between word frequency and pandemic data (specifically case number), which is shown in subsection 4.3.

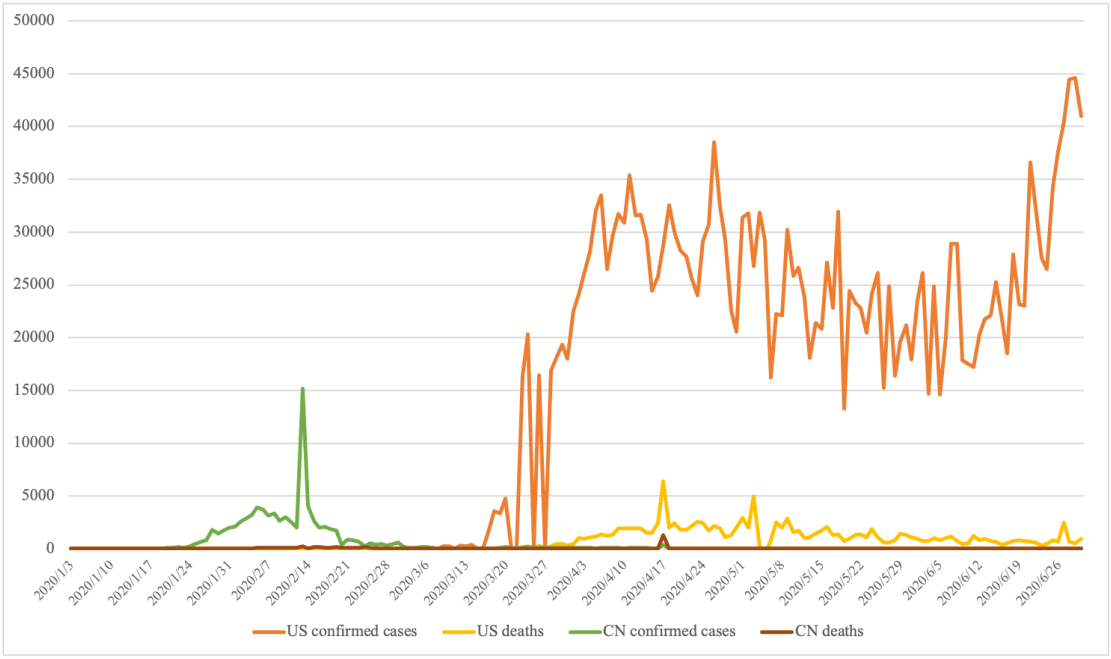

Figure 2: COVID-19 data in China and the US during Jan 3, 2020, and June 30, 2020.

As for semantic change, shown in Figure 3, all five hot words have experienced semantic change to different degrees, which witnessed the largest meaning change ratio in February. However, on comparing the nearest neighbors (see Table 3), it is clear that most changes to 中国 *China* and 美国 *US* were country or city names indicating the pandemic situations in different places. Therefore, we decided to exclude them in our further discussions. Then the three active hot words are 新冠 *COVID-19*, the newly coined term referring to the pandemic, 疫情 *pandemic situation*, the one that always came with xin'guan when people talked about the situation and news about the pandemic, and 防控 *prevention and control*, which have been constantly called by the government. Generally, the feature of the identified semantic change is that the changes in nearest neighbors of three hot words were more diversified in February and March (see Figure 3), and remarkable traces of the war-related terms (or war metaphors) were found in the nearest neighbors in February (see Table 3).

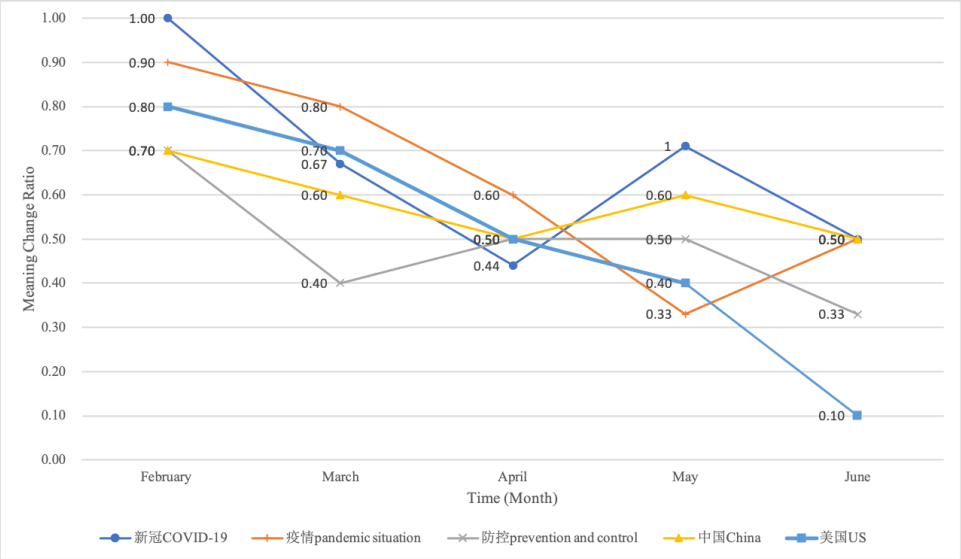

Figure 3: Word meaning change ratio of hot words.

In what follows, we will discuss the semantic change of 新冠 *COVID-19*, 疫情 *pandemic situation* and 防控 *prevention and control*, respectively.

Semantic change of 新冠 *COVID-19*

The word 新冠 *COVID-19* experienced a semantic change in connotation in two phases: from neutral to negative (January to May) and negative to neutral (May to June). At the beginning of the pandemic in January, as a newly coined word, it was associated with uncertainty to the new virus with nearest neighbors such as 无关 *irrelative* and 不明 *unknown*. This was probably due to the lack of knowledge about the virus at that time (Alkandari et al. 2021). In February and March, the word 爆发 *breakout* emerged and became its top nearest neighbor, indicating that the society was well aware of the rapid development of the pandemic. In April and May, approximately 30% (5 out of 15) of *xin'guan*'s nearest neighbors were negative words: 斗争 *struggle*, 遏制 *contain*, 肆虐 *rage* and 扩散 *spread*. It seems that in this period, people regarded COVID-19 as a monster or enemy, which was also detected by Wicke and Bolognesi (2020) on Twitter. However, in June, the word lost its negative nearest neighbors. In this period, nearest neighbors including 疫情 *pandemic situation*, 早期 *early phase*, and 或致 *might cause* presented a neutral emotion of COVID-19 by Chinese Weibo users, which conformed to the slowdown of domestic pandemic. This finding shows that the social perception of COVID-19 has experienced a slightly positive shift since the spreading of it was well under control after a tough time.

Semantic change of 疫情 *pandemic situation*

The word 疫情 *pandemic situation* generally experienced a negative trend in its connotation, and it had more war-related nearest neighbors than the other two active hot words. To be more specific, January and February witnessed the most war metaphors (about 37% of all nearest neighbors). Nearest

neighbors such as 攻坚战 *tough battle*, 指挥部 *headquarter*, 打赢 *win*, and 战斗 *battle* were all war-related terminologies indicating a strong mind to confront the pandemic as an enemy. But since March, the word lost war-related neighbors and gained negative neighbors. The word 危机 *crisis* constantly stayed within top 5 of nearest neighbors from March to June, followed by other negative words like 灾难 *disaster*, 蔓延 *spread*, 严峻 *severe* and 冲击 *strike*, which also made up of 33% (13 out of 39) of all nearest neighbors in this period. This change shows that, after fighting against the outbreak of the pandemic for a period, the Chinese public was consistently paying attention to the pandemic situation and started to count the consequences of the pandemic.

Semantic change of 防控 *prevention and control*

Unlike the other two active words, the word 防控 *prevention and control* kept neutral in connotation during the six months, but the change of its nearest neighbors was also in line with the pandemic situation in China. In January and February, nearest neighbors implied an intensifying attitude towards the pandemic, such as 认真落实 *earnestly implement*, 坚决 *resolute*, 全力 *full strength* and some war-related terms like 部署 *deploy* and 抗击 *snipe*. These words indicated that the government was issuing strong orders in the fight against the pandemic. From April, such words were not close to the word 防控 *fangkong* (prevention and control). The nearest neighbors were its conventional synonyms such as 处置 *disposal* and 应对 *confront*. This is because, at this stage, the number of daily confirmed cases in China was well under control (see Figure 2). In June, as the government started to emphasize regular epidemic prevention and control, the word 常态化 *regularize* became the nearest neighbor. In general, the semantic change of 防控 *prevention and control* was in line with the orders of the government, since the word itself denotes the government's major mission during the pandemic, i.e., to mobilize the whole society to control the pandemic situation.

Table 3: Nearest neighbors of hot words in six months.

Hot word / Time	新冠 *COVID-19*	疫情 *pandemic situation*	防控 *prevention & control*	中国 *China*	美国 *US*
January 2020	新型 new type	***攻坚战 tough battle***	应急 meet emergency	湖北 Hubei	香港 Hong Kong
	无关 irrelevant	***指挥部 headquarter***	联控 joint control	全国 nationwide	增至 Increase to
	不明 unknown	全力 full strength	***抗击 snipe***	武汉 Wuhan	日本 Japan
	冠状病毒 coronavirus	新冠 COVID-19	领导小组 leading group	香港 Hong Kong	中国 China
	疫情 pandemic situation	防控 prevention and control	认真落实 earnestly implement	美国 US	知否 whether know
February 2020	本次 this time	***打赢 win***	应对 confront	日本 Japan	日本 Japan
	肺炎 pneumonia	***战役 combat***	***狙击 snipe***	美国 US	美国 US
	爆发 outbreak	***战斗 battle***	***部署 deplore***	全球 worldwide	中国 China
	疗效 curative effect	***狙击战 blocking action***	防疫 epidemic prevention	全国 nationwide	死亡 death
		战疫 fighting the pandemic	全力 full strength	武汉 Wuhan	意大利 Italy
March 2020	***爆发 outbreak***	危机 crises	应对 confront	美国 US	英规 UK
	针对 aim at	灾难 disaster	保障 guarantee	全世界 worldwide	中国 China
	死亡率 death rate	二次 second time	引发 cause	意大利 Italy	韩国 Korea
	当前 current	流行 prevalent	***部署 deploy***	日本 Japan	欧洲 Europe
	随着 along with	扩散 spread	***抗击 fight against***	韩国 Korea	日本 Japan
April 2020	肺炎 pneumonia	危机 crisis	处置 disposal	美国 UK	纽约 New York
	疫情 pandemic situation	流行 prevalent	应对 confront	海外 oversea	印度 India
	斗争 struggle	严峻 severe	应急 meet emergency	德国 Germany	法国 France
	扩散 spread	新冠 COVID-19	***抗击 fight against***	各国 all countries	意大利 Italy
	遏制 contain	冲击 strike	加强 intensify	意大利 Italy	俄罗斯 Russia
May 2020	肆虐 rage	肺炎 pneumonia	处置 disposal	武汉 Wuhan	巴西 Brazil
	最早 earliest	蔓延 spread	应对 confront	美国 US	中国 China
	何时 when	危机 crisis	抗击 fight against	人类 human	特朗普 Trump
	源头 source	流行 prevalent	控制 control	世界 world	英国 UK
	疫情 pandemic situation	冲击 strike	防疫 epidemic prevention	哪里 where	美国政府 US government
June 2020	疫情 pandemic situation	肺炎 pneumonia	应对 confront	武汉 Wuhan	印度 India
	冠状病毒 coronavirus	流行 prevalent	常态化 regularize	美国 US	全球 global
	早期 early phase	危机 crisis	当前 current	欧洲 Europe	巴西 Brazil
	或致 might cause	***爆发 outbreak***	***部署 deploy***	世界 world	英国 UK
	确诊 confirmed	流行病 disease	通告 announce	白皮书 white paper	中国 China

Notes. War-related terms are in italics, and hot words in nearest neighbors are underlined. For space-saving, half of the nearest neighbors are presented here.

Similar to Wicke and Bolognesi (2020) who found many war-related terms on Twitter during the pandemic, as mentioned in the previous subsection, we also find that the hot words fall into the War metaphor, which is frequently used in all flu-like pandemics around the world (Taylor and Kidgell 2021). Projecting the virus to an enemy is the strategy that politicians apply to encourage people from all walks of life to temporarily put contradictions aside and unite together in the "war". Huang and Hu (2021, p. 96) pointed out that the logic of war metaphor came from the shaping of the "other" and "us" in the context of dual opposition, which worked through the building and activating collective identity and unity to achieve relative social stability. Cognitively, this is a way to construct new social representations by installing the referential relation between the pandemic and war into people's episodic memory (van Dijk 1990), which primes people to be vigilant whenever there are new cases nearby. However, as Chapman and Miller (2020, p. 1109) noted, although war metaphors made the public easy to comprehend the complex social issues, they also added complexity into the public's perception of the pandemic, allowing for "the creation of discrete categories such as winner, loser, the attacked, victims, fault, blame, and enemy, all of which have implicit meanings associated with power discrepancies and blame attribution". Therefore, in March, when the pandemic in the country was gradually under control, the use of this war framing was reduced before its overuse fossilizes social cognition in the dynamic contexts, and thus no more war-related words remained the nearest neighbors of the hot words. This decrease in war framing was also detected by Wicke and Bolognesi (2021) on Twitter.

4.3 Correlation Tests

As mentioned above, both the frequency and semantic change of hot words, as well as the use of war metaphor, seemed to have some correlations with the pandemic situation. To verify this and further explore the interactions between discourse, cognition and society in the context of the COVID-19 pandemic, we did three sets of correlations tests in this subsection: between word frequency ranks and COVID-19 data, between frequency ranks of the five hot words, and between mean word frequency ranks and meaning change of hot words.

Firstly, we executed the Pearson correlation tests between hot word frequency ranks and the pandemic data of China and the US (two representative modes of pandemic development). Results are presented in Table 4. It should be noted that when the correlation value is significantly negative, it means that the case or the death number is in a positive correlation with word frequency. Clearly, US confirmed cases have the most significant ties with word frequency ranks. Its correlation between rank of 美国 *US* (r = -.584) reaches a large effect size[3]. Besides, it has a medium correlation between rank of 中国 *China* (r = -.308). US deaths have four significant correlations, and the strongest is that with 美国 *US* (r = -.426). And for pandemic data in China, there is only one significant correlation between CN

[3] Effect size: small (r = .10), medium (r = .30), large (r = .50). (Cohen, 1988: 83)

confirmed cases and rank of 美国 *US* (r = -.426) with medium effect size. Therefore, according to Table 4, it seems that the pandemic data in the US were more influential than that of China to Chinese social media trends, which is somehow out of our expectation, nor has it been explained by other scholars.

Table 4: Correlations between daily word frequency ranks and COVID-19 data.

COVID-19 data	Word frequency ranks	*r*	*p*
US confirmed cases	新冠 *COVID-19*	-.286	.000
	疫情 *pandemic situation*	-.251	.001
	防控 *prevention & control*	-.152	.042
	中国 *China*	-.308	.000
	美国 *US*	-.584	.000
US deaths	新冠 *COVID-19*	-.201	.011
	疫情 *pandemic situation*	-.184	.013
	防控 *prevention & control*	-.105	.160
	中国 *China*	-.235	.001
	美国 *US*	-.426	.000
CN confirmed cases	新冠 *COVID-19*	.055	.488
	疫情 *pandemic situation*	-.085	.259
	防控 *prevention & control*	-.139	.064
	中国 *China*	.022	.774
	美国 *US*	.410	.000
CN deaths	新冠 *COVID-19*	-.022	.780
	疫情 *pandemic situation*	-.062	.408
	防控 *prevention & control*	-.094	.210
	中国 *China*	-.030	.692
	美国 *US*	.090	.232

The significant correlations between the US pandemic data and word frequency might be attributed to the trends of public concerns. As Wang et al. (2020) observed, with COVID-19 spreading worldwide, the number of Weibo posts referring to the pandemic in other countries grew consistently. The larger number of significant correlations at or above medium effect size between the pandemic data in the US indicate that compared with the pandemic situations in China, the pandemic in the US as social situation was a larger influence over online social discourse in China. As can be seen from Figure 2, the confirmed cases in the US increased rapidly in March, while in China the number has been kept at a low speed. In late June, the daily new case number in the US was about 200 times of China. According to some Weibo contents ((2)-(4), translated), the rising number of confirmed cases in the US might arouse some social worries. Some individuals have concerns that those coming from the US would bring the virus back, leading to a new domestic outbreak, international students and their parents have uncertainties in admission, safety, and so on; and investors are more likely to be pessimistic towards the world economy, which was dominantly influenced by the US.

(2) The US **economy fell tremendously** in the first quarter, **the sharpest decline** since the 2008 financial crisis. Such news is very **shocking and worrisome**. (April 2020)

(3) COVID-19 will bring unpredictable difficulties and hardships to mankind. As the trade war between China and the US goes on, it seems that those who want to **study abroad and travel abroad** might be discouraged. (May 2020)

(4) On hearing that a Seattle COVID-19 patient received a $1.1 million bill from the hospital after he recovered, I strongly recommend that international students should buy **school insurance** in case they got infected. (June 2020)

We can conclude that the pandemic data which revealed social situations and aroused concerns during the pandemic to some extent contributed to the frequency rank change of COVID-19 Weibo hot words. But the number of significant correlations is below our anticipation. Instructed by van Dijk's (2009) discourse-cognitive-social triangle, we reckon that there might be cognitive factors in effect. Consequently, we found more through a correlation test between daily frequency ranks of hot words. Results are presented in Table 5.

According to Table 5, there are 7 significant correlations with medium or large effect size between the daily frequency ranks change of different hot words. The strongest tie is between the ranks of 新冠 *COVID-19* and 美国 *US* (r = .612, p = .000), indicating that Weibo users were concerned much about the pandemic in the US. The frequency of yiqing has the most significant correlations between that of other hot words. In both terms of effect size and number, the correlations between different word frequency ranks are stronger than that between COVID-19 data and word frequency ranks. This is understandable because these hot words are all semantically related to the pandemic. In other words, they share semantic relations, which makes the association between hot words inevitable for Weibo users when they discuss the pandemic. According to Zhu et al. (2020), in a complex network where a hot topic (word) on a microblog is a node and its semantic relations with other topics are edges, co-occurrence feature words can be regarded as subtopics of the given topics. That is probably why the frequency of 疫情 *pandemic situation* has the most correlations between other hot words. And this could be backed by the fact that there are some shared feature words (nearest neighbors in our case) facilitating the close relations between these hot words (see Table 3).

Table 5: Mutual correlations among frequency ranks of the five hot words.

Word frequency ranks		*r*	*p*
新冠 *COVID-19*	疫情 *pandemic situation*	.123	.120
	防控 *prevention & control*	-.124	.117
	中国 *China*	.463	.000
	美国 *US*	.612	.000
疫情 *pandemic situation*	防控 *prevention & control*	.520	.000
	中国 *China*	.549	.000
	美国 *US*	.307	.000
防控 *prevention & control*	中国 *China*	.243	.001
	美国 *US*	.309	.000
中国 *China*	美国 *US*	.384	.000

Yet, there remains a question: did word frequency change affect the change of word meaning, or more specifically, did they present a negative relationship as found by Englhardt et al. (2020) and Hamilton et al. (2016)? This question is worth investigating because as mentioned above, February and March did not only see mass changes in word frequency but also in word meaning. To verify whether there existed any regularities, we then conducted another correlation test between mean frequency rank and word meaning change ratio monthly from February to June. However, according to Table 6, the result is more complex than we expected. Only the word 新冠 *COVID-19* presented a strong positive relation between its frequency rank change and meaning change, i.e., a negative relation between frequency rank and meaning change, and this relation was the only one close to statistically significant ($p = 0.051$). Other two words, both show a non-significantly positive relation between frequency and meaning change.

Table 6: Correlations between mean word frequency ranks and meaning change.

Word	*r*	*p*
疫情 *pandemic situation*	-.552	.334
防控 *prevention & control*	-.250	.685
新冠 *COVID-19*	.876	.051

This seemingly strange situation may attribute to the setting of this research. Firstly, the general negative relations between word frequency and meaning change found before (Englhardt et al. 2020; Hamilton et al. 2016) were all concluded from historical corpora, in which the strangeness caused by sudden events like pandemics could be diluted. Secondly, when we take close look at these three hot words, it

is clear that while the word 新冠 *COVID-19* is a newly coined word due to the pandemic, the other two words are both existing words attached with new meanings and concerns in this pandemic. As for 新冠 *COVID-19*, both its meaning and frequency were settling down at a stable level from January, which was a trend from low frequency and unfixed meaning to stably high frequency and rather fixed meaning. Therefore, its frequency increased while the variation of its meaning decreased. In other words, it experienced a speedy process from "birth" to "maturity" which had happened to the other two words in a much longer time in history. For the other two words, they were both more frequent and semantically stable than the word 新冠 *COVID-19* at its emergence. But since the outbreak of the pandemic, new connotations were added to their meanings, while their frequencies went relatively down later compared with 新冠 *COVID-19*. This could be regarded as a fluctuation in their long and steady development. But this fluctuation was not that strong, because most changes were happening in February, and the changes in word meaning were not severe. This finding demonstrates that investigating words during social emergencies would add new understandings to the interplay between language and society.

5 Conclusion

The present study investigates the dynamics of language during the COVID-19 pandemic through changes of COVID-19-related Weibo hot words from January to June in 2020 in terms of the changing frequency and connotative meaning as well as their relationship.

Word frequency changed with both social situations and public cognition. Generally, when the pandemic was spreading rapidly, people would discuss it more on social media, thus increasing the frequency of related words. The dramatic increase of cases in the US, together with the complex relations between the two countries made the Chinese public pay much attention to its social situation, therefore the pandemic data in the US turned out to be more influential than China on hot word frequency on Weibo especially when the domestic pandemic situation eased. Moreover, as these hot words were semantically related to the main topic, i.e., the COVID-19 pandemic, people's concern of one subtopic (a hot word) would cause rise to the frequency of another, establishing the correlations between frequency rank change of related hot words.

The semantic change on the connotative level of hot words reveals the process in which the public learned the whole picture of the pandemic as well as the government strived to motivate the society by creating war metaphors. The newly coined word 新冠 *COVID-19* experienced a sharp increase in frequency and the settlement in meaning as people acquired knowledge about the virus. From nearest neighbors, it seems that language can frame people's cognition and finally affect social situations. War metaphors took their expected effect in China and were used less since March when the pandemic situation was gradually under control and the government was thinking about promoting the resumption of production.

The current study, investigating word frequency and meaning change during the COVID-19 pandemic, found acceptable "violation" to the general negative relationship between the two variables. The acceptance shall come from the interplay of social situations, public cognition and the self-adaptation of the language system. For already existing hot words like 疫情 *pandemic situation* and 防控 *prevention and control*, their connotations were added up with new feature words upon the outbreak of the pandemic while their relative frequencies experienced a limited increase, thus receiving non-significantly positive correlations between frequency and meaning change. The new word 新冠 *xin'guan*, differently, had its collection of meanings changed from a mess from a fixed one as its frequency drastically increased, thus exhibiting a quick play of birth of a new word and conforming with the law of conformity (Hamilton et al. 2016).

Investigating changes in word frequency and word meaning under the COVID-19 pandemic, this study may shed light on investigating the interplay of social and language change during the pandemic. The technology of word embeddings proved to be a capable tool for us to capture changes in the connotation of hot words. The socio-cognitive approach of discourse analysis instructed us to understand peculiarities in our results in new perspectives. Social situations and public cognition together would cause temporal language change that seemingly violates certain linguistic law, which may be observed when researchers zoom in - under certain influential social change or event.

There are some limitations of the present work. A larger size of corpora containing microblogs across countries might lead to more language changes to be found, and more distinctive or universal findings should be expected.

References

Alkandari, A., Law, J., Alhashmi, H., Alshammari, O., Bhandari, P. (2021). Staying (mentally) healthy–the impact of COVID19 on personal and professional lives. *Techniques and Innovations in Gastrointestinal Endoscopy*, 23(2), pp. 199-206.

Bloomfield, L. (1933) *Language*. New York: Henry Holt and Company.

Čech, R., Hůla, J., Kubát, M., Chen, X., Milička, J. (2019). The development of context specificity of lemma. A word embeddings approach. *Journal of Quantitative Linguistics*, *26*(3), pp. 187-204.

Chapman, C. M., Miller, D. M. S. (2020). From metaphor to militarized response: the social implications of "we are at war with COVID-19"– crisis, disasters, and pandemics yet to come. *International Journal of Sociology and Social Policy*, 40(10), pp. 1107-1124.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cruse, D. (1986). *Lexical Semantics*. Cambridge, Massachusetts: Cambridge University Press.

Englhardt, A., Willkomm, J., Schäler, M., Böhm, K. (2020). Improving semantic change analysis by combining word embeddings and word frequencies. *International Journal on Digital Libraries*, 21(3), pp. 247-264.

Fairclough, N. (1989). *Language and Power*. New York: Longman.

Firth, J. R. (1957). A synopsis of linguistic theory. In *Studies in linguistic analysis*. Oxford: Philological Society.

Gao, D., Wang, Z. (2017). Research on social representation of network hot words in digital era. In: *2017 World Conference on Management Science and Human Social Development (MSHSD 2017)*, pp. 424-429. Atlantis Press.

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635-E3644.

Hafner, C. A., Sun, T. (2021). The 'team of 5 million': The joint construction of leadership discourse during the Covid-19 pandemic in New Zealand. *Discourse, Context & Media*, 44, 100523.

Hamilton, W. L., Leskovec, J. Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1489-1501. Berlin: Stroudsburg.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), pp.146-162.

Huang, Y., Yang, H. (2021). Identity from "Opposition": The Logic of Social Governance in the War Metaphor. *Journalism & Communication Review*, 74(1), pp. 96-106.

Hunt, S. (2021). COVID and the South African Family: Cyril Ramaphosa, president or father?. *Discourse, Context & Media*, 44, 100541.

Jatnika, D., Bijaksana, M. A., Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in English words. *Procedia Computer Science*, 157, pp. 160-167.

Li, S., Wang, Y., Xue, J., Zhao, N., Zhu, T. (2020). The impact of COVID-19 epidemic declaration on psychological consequences: A study on active Weibo users. *International Journal of Environmental Research and Public Health*, 17(6), 2032.

Liu, H. (2018). Language as a human-driven complex adaptive system. *Physics of Life Review*, 26–27, pp. 149–151.

Liu, W., Niu, K., He, Z., Li, Y. (2016). Trend prediction of hot words in weibo based on fuzzy time series. In: *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 354-358.

Martikainen, J., Sakki, I. (2021). Boosting nationalism through COVID-19 images: Multimodal construction of the failure of the 'dear enemy' with COVID-19 in the national press. *Discourse & Communication*, 15(4), pp. 388-414.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Mou, W., Sun, N., Zhang, J., Yang, J., Hu, J. (2015). Politicize and depoliticize: A study of semantic shifts on *People's Daily* fifty years' corpus via distributed word representation space. In: Lu Q., Gao H. (Eds.). *Chinese Lexical Semantics*, pp. 438-447. Online: Springer International Publishing Switzerland.

Rajput, N. K., Grover, B. A., Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. *arXiv preprint arXiv:2004.03925*.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one*, 16(2), e0245909.

Saeed, J. I. (2016). *Semantics (Fourth Edition)*. Chichester, West Sussex; Malden, MA: Wiley-Blackwell.

Shi, Y., Lei, L. (2019). The evolution of LGBT labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. *English Today*, 36(4), pp. 33-39.

Sindoni, M. G., Moschini, I. (2021). Discourses on discourse, shifting contexts and digital media. *Discourse, Context & Media*, 43, 100534.

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham: Packt Publishing Ltd.

Supadhiloke, B. (2012). Framing the Sino–US–Thai relations in the post-global economic crisis. *Public Relations Review*, 38(5), pp. 665–675.

Taylor, C., Kidgell, J. (2021). Flu-like pandemics and metaphor pre-covid: A corpus investigation. *Discourse, Context & Media*, 41, 100503.

Turney, P. D. Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, pp. 141-188.

Van Dijk, T. A. (1990) Social Cognition and Discourse. In: Giles, H., Robinson, W. P. (Eds.). *Handbook of Language and Social Psychology*, pp. 163-183. New York: John Wiley & Sons Ltd.

Van Dijk, T. A. (2009). Critical discourse studies: A sociocognitive approach. In: Wodak, R., Meyer, M. (Eds.). *Methods of critical discourse analysis*, pp. 62-84. London: SAGE Publications Ltd.

Wang, J., Zhou, Y., Zhang, W., Evans, R., Zhu, C. (2020). Concerns expressed by Chinese social media users during the COVID-19 pandemic: Content analysis of sina weibo microblogging data. *Journal of Medical Internet Research*, 22(11), e22152.

Wang, Y., Song, S., Zhou, F., Zheng, X. (2017). Chinese WeChat and blog hot words detection method based on chinese semantic clustering. *Intelligent Automation & Soft Computing*, 23(4), pp. 613-618.

Wicke, P., Bolognesi, M. M. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS one*, 15(9), e0240010.

Wicke, P., Bolognesi, M. M. (2021). Covid-19 Discourse on Twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time. *Frontiers in Communication*, 6, 45.

Zhu, G., Pan, Z., Wang, Q., Zhang, S., Li, K. C. (2020). Building multi-subtopic Bi-level network for micro-blog hot topic based on feature Co-Occurrence and semantic community division. *Journal of Network and Computer Applications*, 170, 102815.

A Corpus-Driven Study of the Style Variation in The Grapes of Wrath

Yiyang Hu[1], Qingshun He[1*]

[1] Sun Yat-sen University
[*] Corresponding author's email: heqsh5@mail.sysu.edu.cn

DOI: https://doi.org/10.53482/2022_52_396

ABSTRACT

The novel *The Grapes of Wrath* is distinctive in the arrangement of intercalary chapters and narrative chapters. Existing studies of the narratological distinction of this novel are primarily qualitative. This article conducted a corpus-driven study of the variation of styles in this novel from the perspectives of word cluster, type-token ratio, descriptivity and activity, keyness, and sentiment. The cluster analysis shows that the choice of words in the narrative chapters is more consistent than that in the intercalary chapters. The type-token ratio analysis testifies to the heterogeneity of the intercalary chapters in terms of lexical richness. The descriptivity and activity analysis and the keyness analysis reveal that the narrative chapters are more active than the intercalary chapters. The sentiment analysis finds that the novel is pervaded by negative sentiments and that negative sentiments are more prevalent in the narrative chapters than in the intercalary chapters. The research concludes that the corpus-driven study can provide insights into the narrative structure and the stylistic variation of the novel.

Keywords: corpus-driven; narrative structure; narratological distinction; style variation; The Grapes of Wrath

1 Introduction

1.1 Narrative Structure of The Grapes of Wrath

The Grapes of Wrath written by John Steinbeck is an epic capturing the plights of the Okies whose lives are destroyed by the Dust Bowl and the Great Depression. This novel focuses on the Joad family, who got evicted from the farm in Oklahoma, travelled along Route 66 to California and became migrant farmers in California. By consulting the spatial movement, this novel can be divided into three balanced parts.

In each of the three parts, there are some chapters not directly relevant to the storyline of the Joad family. These chapters are referred to as the intercalary chapters (Swensen, 2015) or the interchapters (Levant, 2007); they provide the narrative chapters with an epic scope. Burcar (2018) regarded this narrative

method as a dialectical montage that juxtaposes the seemingly disparate scenes to reveal the structural interconnectedness in a larger context. This setting leads the readers' attention to the cause of the systematic exploitation. Hamilton (2016) stressed the palimpsest effects of different types of chapters.

However, existing studies mainly focus on the non-verbal aspects of the narrative without considering the linguistic features. Without the description of the language, the interpretation of the narrative structure will be subjective and less convincing. The dichotomy of the intercalary chapters and the narrative chapters can be a framework for the study of style variation.

1.2 Narratology and Stylistics

Stylistic studies can provide narrative analyses with valuable insights. Most narratological studies are characterized by the analysis of text fragments and the exclusion of language, while stylists tend to focus on the linguistic features below the sentence level, such as the rhyme pattern, the word choice and the syntactic structure. It seems to exist a boundary between narratology and stylistics. Rimmon-Kenan (1989) pointed out the paradox that narratologists paid little attention to the linguistic features of the text while narratology theories were greatly influenced by linguistic theories. According to Rimmon-Kenan (1989), the reconstruction of the represented events in literary works should depend on language for its very essence. The language of the text can be described in terms of style. Stylistics is a method of describing linguistic features in the text, which provides evidence for or against the interpretation and evaluation of literature (Short, 1984). Literary criticism without the basis of stylistics lacks a convincing argument. According to Phelan (1981, p. 6), the sense of style can be developed into "those elements of a sentence or passage that would be lost in paraphrase". After paraphrasing the story, the same narrative technique can be adopted, but the linguistic features such as the dictation and the syntactic structure will be changed.

The loss of style in the paraphrase shows that a narratological study without stylistic analysis cannot fully figure out the art of the author's presentation. In the analysis of *A Farewell to Arms*, Phelan (1996) consulted the style of paratactic sentences and the use of adverbials to figure out the tension of narrative voices. He derived the hidden narrative voices from the stylistic features, thus building a bridge between narratological techniques and stylistic features. The exploration of this kind of interaction can also be found in the analysis of *Marabou Stork Nightmares* conducted by Short (1999), according to whom the viewpoint shift in the narrative structure and the movement of narrative levels are influenced by the style variation in the text. The choice of language has considerable impacts on the way texts are structured. To get a broader view of how *The Grapes of Wrath* is presented, linguistic features should be considered with organizational techniques.

1.3 Corpus Stylistics

The corpus in stylistics has gained popularity with the development of natural language processing. The relationship between linguistic description and literary appreciation can be defined as the corpus

stylistic cycle (Mahlberg, 2014). In this cycle, the researcher can use the corpus method to investigate the linguistic phenomena in quantitative ways, thus giving innovative linguistic descriptions and supporting literary appreciation. The corpus method features the quick collection and calculation of repetitions. This rough indication of the high-frequency elements can be helpful in stylistic studies.

According to Halliday (1971), the foregrounding elements can be found not only in literary deviances, i.e. the use of ungrammatical forms, but also in deflections, i.e. the departure from some expected patterns of frequency. Deviances can be seen as the law-breaking elements while deflections as law-making elements. It should be noted that these repetitions are not always foregrounding elements of stylistic relevance. Halliday (1971) emphasized the positive virtues of counting deflections under the condition that they were relevant to the thematic meaning of the text. Leech and Short (2007) maintained that there existed a dividing line between the foregrounding and the unmotivated prominences. Not all statistical deviances can be regarded as literary foregrounding. It can be seen that the statistical significance is not always relevant to the interpretive significance, but the empirical evidence can make it easier for critics to find clues of foregrounding. To improve the reliability of statistical deviance of linguistic features, researchers of corpus stylistics usually adopt the approach of matching texts against corpus. This approach is influenced by the concept of semantic prosody, referring to "a consistent aura of meaning with which a form is imbued by its collocation" (Louw, 1993, p. 157). The matching process can be achieved by comparing the extract from the individual text with a general corpus to explain the creative use of words. The comparison usually focuses more on the function of concordances rather than on the frequency of words. In the stylistic study of O'Connor's *Eyes*, Hardy (2004) investigated the frequencies of keywords in the context in comparison with the general corpus. This approach can give us evidence of the interference of negative words in the results of sentiment analysis. According to McIntyre and Walker (2019), matching texts against corpora can also be used in the calculation of keyness. The quantitative information of words can help assess whether the choice of the word can be regarded as a departure from the expected choice.

1.4 Hypothesis

The Grapes of Wrath is exquisitely weaved by the alternation of intercalary chapters and narrative chapters. In this article, we will conduct a corpus-driven study of the style variation of *The Grapes of Wrath* especially the relationship between the two types of chapters to reflect upon the role of language as a medium in the construction of the story. For this purpose, we can work on the hypothesis that the intercalary chapters and the narrative chapters of the novel have different linguistic features in terms of word cluster, type-token ratio, descriptivity and activity, keyness, and sentiment.

2 Methodology

2.1 Corpus

The novel consists of 30 chapters which can be divided into three parts according to the spatial movement. The first ten chapters talking about the eviction of the farmers are set in Oklahoma. The following ten chapters describe the events on Route 66 until the characters arrive at the destination of California. The last ten chapters concern the lives of migrant farmers in California. Of the 30 chapters, 16 are intercalary. The size of each of the intercalary chapters is smaller than 5,000 words, while 13 out of the 14 narrative chapters contain more than 5,000 words each. Most of the intercalary chapters form alternations with the narrative chapters in the narrative structure. The details of the chapters are shown in Table 1.

Table 1. An overview of the chapters in the novel

Part 1			Part 2			Part 3		
Chapter	**Intercalary/ Narrative**	**No. of words**	**Chapter**	**Intercalary/ Narrative**	**No. of words**	**Chapter**	**Intercalary/ Narrative**	**No. of words**
1	I	1,396	11	I	841	21	I	942
2	N	3,752	12	I	1,951	22	N	16,502
3	I	892	13	N	11,762	23	I	2,329
4	N	6,215	14	I	975	24	N	5,981
5	I	3,673	15	I	4,159	25	I	1,462
6	N	9,933	16	N	13,417	26	N	22,334
7	I	2,164	17	I	2,985	27	I	1,039
8	N	8,662	18	N	12,450	28	N	9,112
9	I	1,464	19	I	3,530	29	I	1,197
10	N	11,227	20	N	17,518	30	N	7,696

This research also uses the fiction sub-corpus of the Brown Corpus as the reference corpus to help calculate the keyness of the keywords. The fiction sub-corpus covers a variety of novels, including 126 sample texts totalling 253,997 words. See Table 2.

Table 2. An overview of the fiction sub-corpus of the Brown Corpus in the novel

Genre category	No. of texts	No. of words
General fiction	29	58,338
Mystery and detective Fiction	24	48,231
Science fiction	6	12,043
Adventure and Western	29	58,433
Romance and love story	29	58,675
Humor	9	18,277
Total	126	253,997

2.2 Procedure

This study adopts the top-down approach. An overview of the text can be presented by cluster analysis. Then the MATTR (Moving Average Type-Token Ratio), descriptivity, activity, nominality, keyness and

emotion shift are calculated. Finally, we zoom in on the patterns or details of interest and give an evidence-based interpretation.

Cluster analysis is based on the assumption that the works of different authors can be classified according to their distinct stylistic features. These features are the author's unique "signals" or style (Jockers, 2013, p. 63). Cluster analysis concerns the stylistic similarity and difference of texts, and hence can also be used to compare different chapters in one novel. According to Jockers (2014), high-frequency features of different texts can be compared through Euclidean Distance. The closer the distance is, the more common the feature usage habits are. The formula of Euclidean Distance is expressed as follows.

$$d\,(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_i - q_i)^2 + (p_n - q_n)^2} \quad (1)$$

In Formula 1, d represents the distance, p and q are two texts, $p_1, p_2 \ldots p_i, p_n$ are the measures of feature vectors in text p and $q_1, q_2 \ldots q_i, q_n$ are the measures of feature vectors in text q.

Type-token ratio (TTR) has been widely used in literary research to analyze lexical richness, but this indicator can be affected by the text length. For the reliability of TTR, Covington and McFall (2010) proposed the MATTR as shown in Formula 2.

$$\text{MATTR}\,(L_w) = \frac{\sum_{i=1}^{L_t - L_w} V_i}{L_w(L_t - L_w + 1)}, \quad (2)$$

In Formula 2, L_w stands for the arbitrarily chosen length of a window, L_t for text length in tokens, and V_i for the number of types in an individual window. This approach is helpful to calculate the TTR for the subset from 1 to L_w, then for the subset from 2 to $L_w + 1$, and ends when $L_w + 1$ reaches the end of the text. In this way, the indicator of lexical richness is not influenced by the segment boundaries or text length. In this study, the MATTR[1] software is used to calculate the MATTR.

Descriptivity, activity and nominality are stylometric indicators that can describe the features of a text. The verb-adjective ratio and the noun-verb ratio have been used to differentiate styles, genres and authorships (e.g., Boder, 1940; Antosch, 1969; Popescu et al., 2013; Chen and Kubát, 2021). Descriptivity is defined as the division of the number of adjectives to the sum of verbs and adjectives and activity is equivalent to 1 minus descriptivity. Nominality is defined as the ratio of nouns to the sum of nouns and verbs. The formulas are as follows.

[1] This software is designed by Covington and McFall (2010) to eliminate the effect of text length on calculating the type-token ratio.

(3) $$activity = \frac{verbs}{verbs + adjectives}$$

(4) $$descriptivity = \frac{adjectives}{verbs + adjectives}$$

(5) $$nominality = \frac{nouns}{verbs + nouns}$$

Before calculating these indicators, we need to tag the part of speech of the tokens using the TreeTagger in the R Studio[2]. Some packages[3] are used for conducting macro analyses such as word frequency study, descriptivity and activity comparison, and sentiment analysis.

The departure from the expected word choice can be assessed by keyness. According to McIntyre and Walker (2019), it is better to calculate keyness by log-likelihood because log-likelihood does not assume a normal distribution. It shows how much evidence there is for the difference between the target corpus and the reference corpus. Rayson and Garside (2000) presented the process of calculating log-likelihood. Firstly, a contingency table is constructed as follows.

Table 3. Contingency table for word frequencies (Rayson and Garside, 2000, p. 3)

	Target corpus	**Reference corpus**	**Total**
Frequency of words	a	b	a + b
Frequency of other words	c – a	d – b	c + d – a – b
Total	c	d	c + d

Then the expected frequencies for the target corpus ($E1$) and the reference corpus ($E2$) are calculated respectively. The LL value can be calculated in Formula 8.

(6) $$E1 = c \times (a + b)/(c + d)$$

(7) $$E2 = d \times (a + b)/(c + d)$$

(8) $$LL = 2 \times \left(a \times log\left(\frac{a}{E1}\right) + b \times log\left(\frac{b}{E2}\right)\right)$$

The LL just indicates the existence of a difference between two corpora without telling the scale of difference. Thus, the log ratio is required to measure the effect size of the difference. According to

[2] A programming language like R is a more versatile tool than most ready-made software applications (Gries, 2017). R Studio can be used to calculate Euclidean Distance and draw the dendrogram by the "dist" function and the "hclust" function.

[3] The packages we used in this study are korpus.lang.en (Michalke, 2020), dplyr (Wickham et al., 2021), tidyr (Wickham, 2021), coreNLP (Arnold and Tilton, 2016), stringr (Wickham, 2019) and ggplot2 (Wickham, 2016).

Hardie (2014), log ratio is defined as the binary log of the ratio of relative frequencies, indicating the size of the difference. The calculation of log ratio (*LR*) is expressed as follows.

(9) $$LR = log_2\left(\frac{\frac{a}{c}}{\frac{b}{d}}\right)$$

The sentiment variation across the plot can be roughly evaluated by the frequency of words denoting sentiments. To identify the sentiment words, researchers usually match texts with the lexicon of sentiments, but there are some problems in the matching. Hunston (2007) recognized the problem of oversimplified exploitation of intertextuality in corpus linguistics and argued that attitudinal meanings were not always transferable from one text to another. It is necessary to check the precise phraseology in the context to avoid the uncareful classification of being positive or negative. It is possible that some negative words such as *no, not* and *without* can appear preceding the words indicating emotion. Their disruption in sentiment analysis can be checked and evaluated through *n*-grams. After gathering the 2-grams consisting of one negative word and one negated word, researchers can roughly measure the degree of their disruption in the sentiment analysis. The choice of lexicon can influence the result of sentiment analysis and hence more than one lexicon is required. This study consults the Bing (Hu and Liu, 2004) and the NRC (Mohammad and Turney, 2013) sentiment lexicons[4].

3 Results and Discussion

3.1 Unsupervised Cluster Analysis

It can be expected that Steinbeck must have controlled his selection of words to write the intercalary chapters. This selection is restricted by the location of the chapter. Each chapter is influenced by its surrounding chapters to a certain degree. However, the intercalary chapters whose contents and styles are generally different from the narrative chapters make the selection of words complicated. It is hard to evaluate the style variation just by close reading. Thus, a cluster analysis is conducted to have an overview of the styles of the chapters. See Figure 1.

[4] The ratio of negative words to positive words in the lexicons can influence the result of sentiment analysis. The Bing contains 3,324 negative words and 2,312 positive words, while the NRC includes 4,782 negative words and 2,006 positive words. The NRC also categorizes the sentiment words into anger, anticipation, disgust, joy, trust and so on.

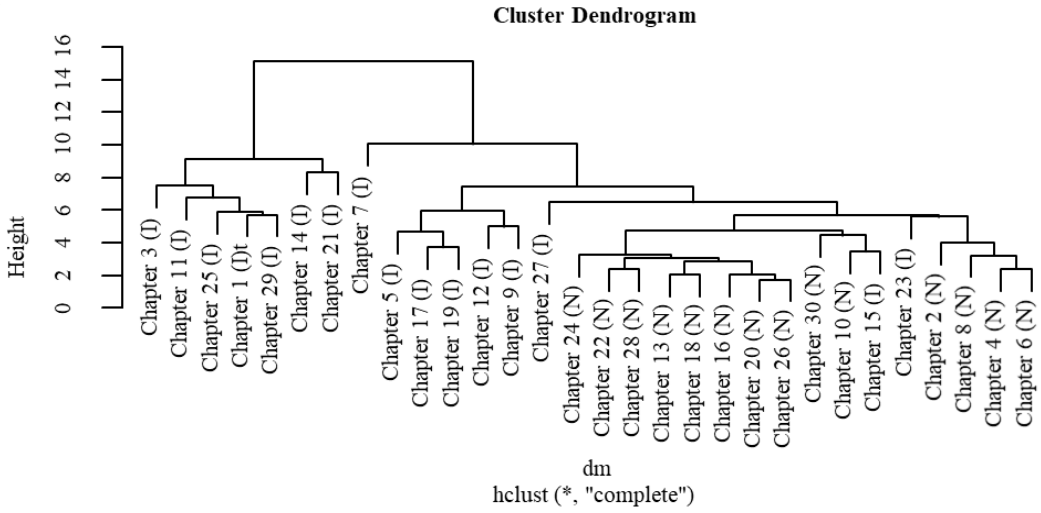

Figure 1. Cluster dendrogram of all chapters in *The Grapes of Wrath*

The dendrogram shows that most intercalary chapters are clustered on the left while the narrative chapters are on the right. The narrative chapters in Section 1 (i.e. Chapters 2, 4, 6 and 8) are merged to form their cluster, indicating that they share a high similarity. Other narrative chapters cluster nearby, the branch heights of which are below 4. The branch heights of all the narrative chapters are below 6, while the average branch height of the intercalary chapters is above 6. Even in the same section of the novel, the intercalary chapters are far away from their surrounding intercalary chapters. This attests to the heterogeneity of the intercalary chapters and the relative homogeneity of the narrative chapters.

One clade of the second-highest branch has only one leaf (i.e. Chapter 7). The isolated chunk can be interpreted as the uniqueness of Chapter 7 in the novel. The words that make Chapter 7 unique can be identified by comparing the frequency of words used in this chapter with that in the rest chapters[5]. According to Silge and Robinson (2017), the most distinct words for each file can be sorted out from the word frequency lists by consulting the log odds ratio for each word. The calculation of the log odds ratio is expressed in Formula 10.

(10) $$log\ odds\ ratio = ln(\frac{(\frac{n+1}{total+1})_{Ch-7}}{(\frac{n+1}{total+1})_{the\ rest}})$$

In formula 10, *n* is the number of times the word in question is used by each file, and *total* refers to the total words for each file. This study takes the top 15 most distinctive words for Chapter 7 and the rest chapters. The plot of these words is shown in Figure 2.

[5] With the help of the R package tidyr (Wickham, 2021), word frequencies are counted for Chapter 7 and the rest chapters.

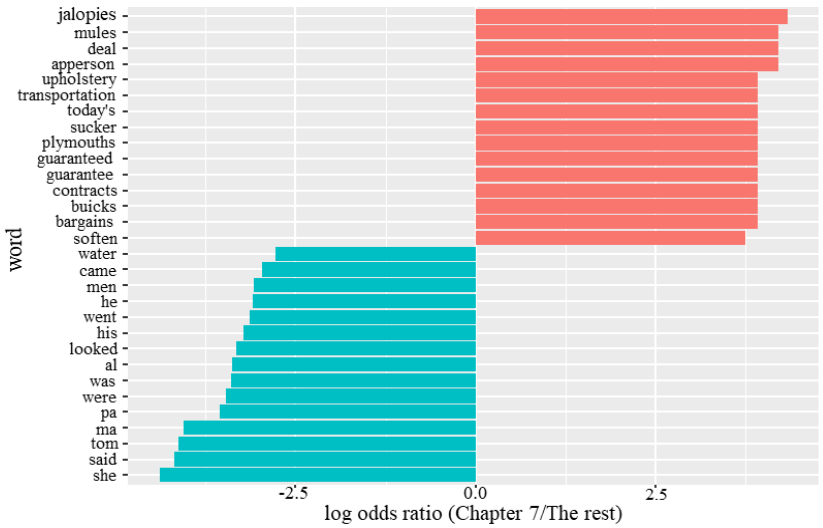

Figure 2. Comparing the odds ratio of words from Chapter 7 and that from the rest chapters

Most distinctive words in Chapter 7 form the semantic field of transportation, such as *jalopies, mules, transportation, Apperson, Buicks, Plymouths, cushions* and *upholstery*. Chapter 7 seems to be relevant to the deals and bargains in the car store. The most distinctive words in the rest chapters are characters and personal pronouns, such as *Tom, ma, pa, Al, he, his* and *she*. The lack of personal pronouns in Chapter 7 can be explained in the special speech presentation. Chapter 7 is composed of the dialogue presented in free direct speech. This special speech presentation can partly affect the lexical distinctiveness of personal pronouns in Chapter 7. In these dialogues, the speakers are mainly the owners of the car store and the salesmen involved in the dealings. As the distance between the speakers and the narrator is relatively short in the free direct speech, personal pronouns like *he* and *she* are less likely to appear in Chapter 7 than in those chapters with an omniscient narrator.

3.2 Lexical Richness

Intercalary chapters usually describe the background which is not directly relevant to the main storyline. They may provide extra information and description and are hereby expected to have a higher value of lexical richness. However, Figure 3 shows no evidence of higher MATTR values in the intercalary chapters.

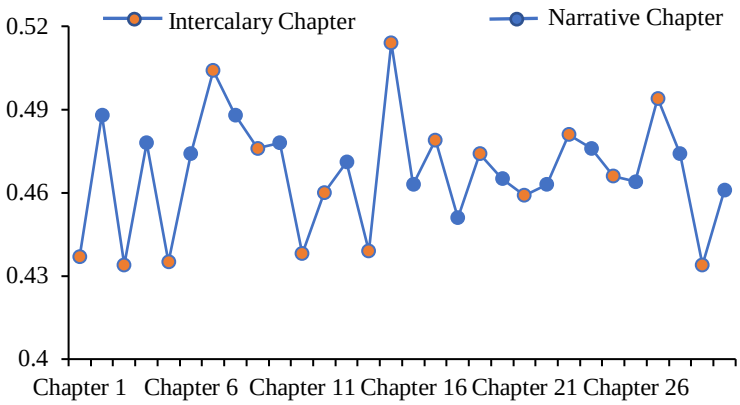

Figure 3. Variation of MATTR in *The Grapes of Wrath* (MATTR window size is 500)

According to the boxplot (see Figure 4), the average MATTR value is higher in the narrative chapters (0.471) than in the intercalary chapters (0.464). The result in the intercalary corpus varies considerably. Both the extreme values are found in the intercalary chapters. For example, the maximum value (0.514) appears in Chapter 15. The great variation testifies to the heterogeneity of the intercalary chapters.

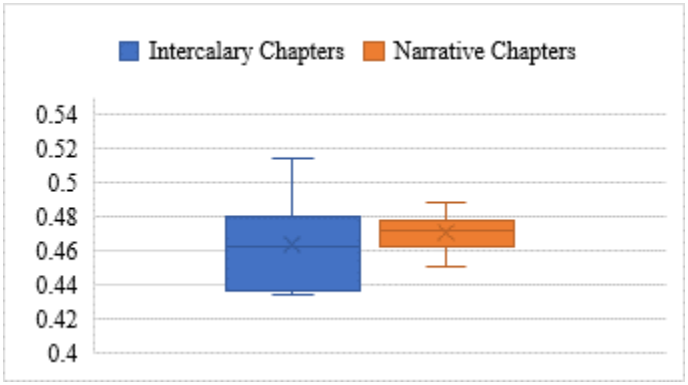

Figure 4. The boxplot of MATTR of intercalary chapters and narrative chapters

The lexical Richness of Chapter 15 can be attributed to Steinbeck's choice of words. The specific area of lexical richness in Chapter 15 can be roughly located by tracing its 3,638 subsets. As shown in Figure 5, the first half of these subsets is richer in vocabulary than the second half. In the first half, there exist three peaks of value, indicating intensive addition of new information. The intermittent upslope shows a decrease in word repetitions. The downslope following each peak suggests a more closely related extension which possibly repeats words that have appeared previously. These intense fluctuations reveal the degree of fragmentation in the text structure of Chapter 15. Moreover, the average of the second half of subsets shows a lower value of MATTR. These subsets turn out to be a conversation. Generally, people in conversations discuss an issue at length by using simple language without introducing many new words.

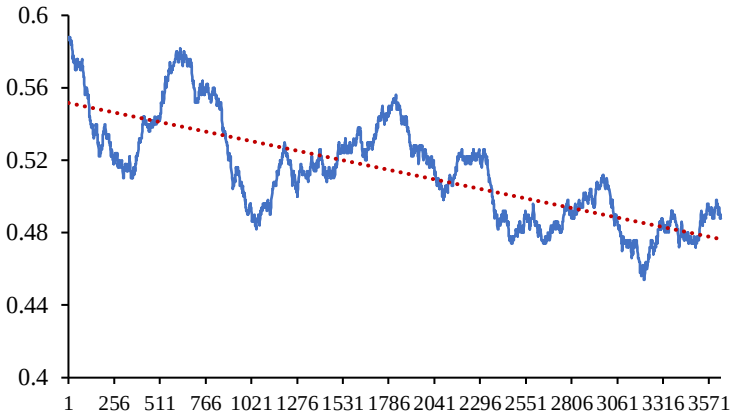

Figure 5. Variation of MATTR of subsets in Chapter 15 (MATTR window size is 500)

The second peak value (0.582) of MATTR comes from the 623rd subset. A brief extraction of this subset is shown below.

> Cars whisking by on 66. License plates. Mass., Tenn., R.I., N.Y., Vt., Ohio. Going west. Fine cars, cruising at sixty-five.
> There goes one of them Cords. Looks like a coffin on wheels.
> But, Jesus, how they travel!
> See that La Salle? Me for that. I ain't a hog. I go for a La Salle.
> 'F ya goin' big, what's a matter with a Cad'? Jus' a little bigger, little faster.
> I'd take a Zephyr myself. You ain't ridin' no fortune, but you got class an' speed. Give me a Zephyr.
> Well, sir, you may get a laugh outa this — I'll take a Buick-Puick. That's good enough.
>
> (Steinbeck, 2006, p. 154)

The number of proper nouns and the presentation of speech can expatiate lexical richness in this subset. These sentences are characterized by short and incomplete sentences that are common in spoken language. Proper nouns of cars and places account for a large proportion of words in these sentences. Moreover, free direct speech makes the talk discursive and flexible, thus reducing the possibility of repetition. These talks seem like a semi-transcription of the hubbub and impose a bunch of information on readers without introducing the identity of the speakers. If speakers can be easily identified, the content of the talk will be limited by the specific relationship between the speakers. It is the vague presentation of the scene and speakers that paves the way for a broader scope and a higher information density. Therefore, the relevance between MATTR and conversation deserves rethinking. The conversation shown in the subset has a higher value of MATTR, whereas the ordinary conversation shows a lower value of MATTR. The way of presenting conversations should be considered in the assessment of the text structure because it may impact the information density.

3.3 Descriptivity, Activity and Nominality

The descriptivity, activity and nominality of all the chapters of the novel are shown in Figure 6 and the Boxplot of descriptivity and nominality of the two types of chapters is shown in Figure 7.

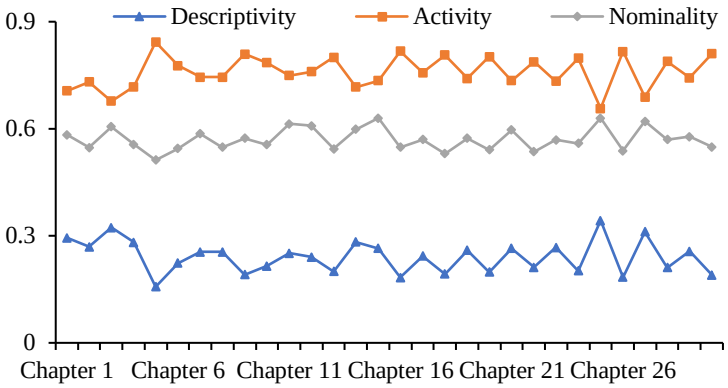

Figure 6. Descriptivity, activity and nominality of all chapters

It can be seen from Figure 6 that all the intercalary chapters have higher values of descriptivity in the last 10 chapters, i.e. Section 3. This difference is not obvious in the first two sections. This partially corroborates Levant's (2007, p. 22) criticism that "the third part…. marks an artistic decline". In the first two sections, the narrative structure does not obviously affect the style. In Section 3, the author pays much attention to the narrative structure. According to Levant (2007), style is a concomitant of structure. The high degree of manipulation seems artificial because it may lead to the overlooking of the relationship between the stories in the two types of chapters.

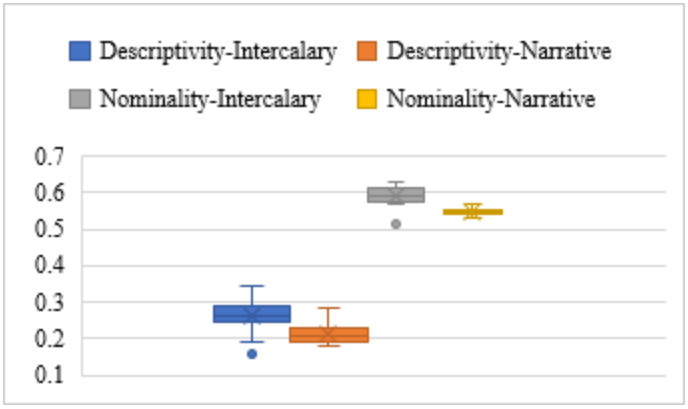

Figure 7. The boxplot of descriptivity and nominality of the intercalary and narrative chapters

The average values in Figure 7 indicate that the narrative chapters are more dynamic than most intercalary chapters. However, there exist two deviant points in descriptivity and nominality of the intercalary chapters respectively, both in Chapter 5. It is reasonable to infer that Chapter 5 is more dynamic than the other chapters. The deviant abundance of verbs signals the departure from the usual diction in the novel. Therefore, the verbs in Chapter 5 deserve in-depth exploration.

3.4 Keyness

In this section, we count, sort and lemmatize[6] the verbs in Chapter 5. It is interesting that the lemma *be* accounts for nearly 65% and *do* for almost 24.8% of all the verbs in this chapter. These prominent statistics spur the comparison of the frequencies of these two verbs in Chapter 5 and the fiction sub-corpus of the Brown Corpus. The results of the comparison will be used in the log-likelihood test. See Table 4.

Table 4. Contingency table for the keyness of *be* and *do* using log-likelihood

	Chapter 5		**Brown Corpus**	
	be	***do***	***be***	***do***
Frequency of words	134	51	268	154
Frequency of other words	3,390	3,473	253,729	253,843
Total	3,524	3,524	253,997	253,997

The *LL* and *LR* of *be* and *do* can be calculated respectively. The *LL* for the lemma *be* reaches 280.45, a value well above the log-likelihood critical value of 15.13. As the degree of freedom for the data is 1, the value 15.13 is associated with the *p*-value of 0.0001. Likewise, the *LL* for the lemma *do* is about 102.30. Therefore, there is a 99.99 per cent probability that both the results of the lemma *be* and *do* are significant and are not due to chance. The *LR* for the lemma *be* is about 5.17, indicating that the lemma *be* is 16 times more frequent in Chapter 5 than in the Brown Corpus. The *LR* of the lemma *do* is about 4.58, indicating that the lemma *do* is 12 times more frequent in Chapter 5 than in the Brown Corpus.

Lemma *be* will be taken as an example in the following stylistic analysis. The dispersion of the lemma *be* is shown in Figure 8.

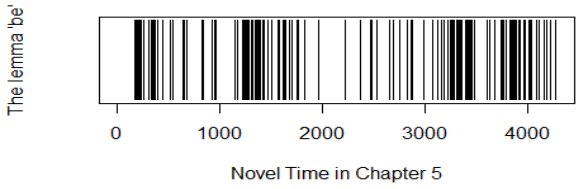

Figure 8. The dispersion plot of the lemma *be* in Chapter 5

It can be seen that there are three clusters of *be* in the plot. The extraction of the second cluster (1250-1450 in the novel time of Chapter 5) is shown as follows.

[6] This process consults the package coreNLP (Arnold and Tilton, 2016).

> We know that – all that. **It's** not us, **it's** the bank. A bank **isn't** like a man. Or an owner with fifty thousand acres, he **isn't** like a man either. **That's** the monster.
> Sure, cried the tenant men, but **it's** our land. We measured it…. Even if **it's** no good, **it's** still ours. **That's** what makes it ours – being born on it, working it, dying on it….
> **We're** sorry. **It's** not us. **It's** the monster. The bank **isn't** like a man.
> Yes, but the bank **is** only made of men.
> No, **you're** wrong there – quite wrong there. The bank **is** something else than men…. The bank **is** something more than men, I tell you. **It's** the monster. Men made it, but they can't control it.

(Steinbeck, 2006, p. 33)

This subset is the talk between one tenant and the spokesman for the farm owner. Different forms of the lemma *be* are scattered throughout the sentences. Most instances of *be* are found in the collocation *it's*. The linking verbs are the backbones of these direct statements. It is through the lemma *be* that the different meanings of *it* directly convey the conflict between the spokesman and the tenant. The pronoun *it* in the first paragraph refers to the perpetrator who drives the farmers away. In the second paragraph, *it* refers to the land in the tenants' minds. The last *it* means the bank. It can be seen that the spokesman wants to inform the tenant of the one who is to blame for the eviction, but the tenant is only concerned about the land before his eyes.

The talk is not just about the question of who is to blame; but rather it presents different ways of perceiving the world. The perception is embodied clearly in their use of the lemma *be*. The argument of the real attribution forms a tension between different perceptions in the conversation. The tenant is not informed of the bank system. What he cares about are concrete objects, such as the land and people, but the spokesman is conscious of his location in the system. To convey the abstract social system to the tenant, the spokesman directly defines the bank as a monster. The tenant, however, cannot understand why the monster is made of people. The spokesman can only say, "Men made it, but they cannot control it". The tenant faces the fate that he has to abandon his perception of concrete lands and people and accept the destiny of being evicted by the abstract bank.

3.5 Sentiment

We matched each chapter of the novel with the Bing and the NRC lexicons respectively[7]. The value of sentiment was calculated by the value of positive words minus the value of negative words. The result is presented in Figure 9.

[7] The matching process is conducted in R Studio with the help of the dplyr package (Wickham et al., 2021) and tidyr package (Wickham, 2021).

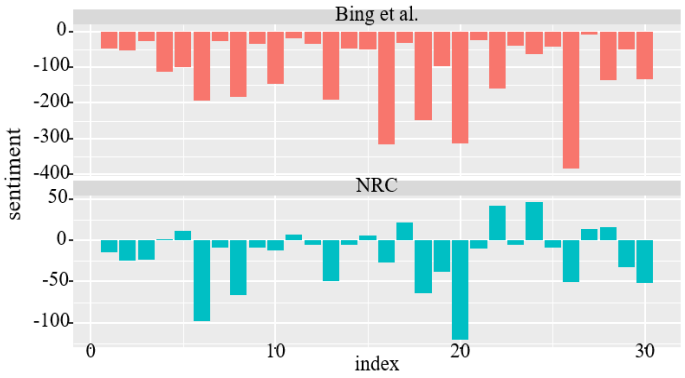

Figure 9. Sentiment through the narrative of *The Grapes of Wrath*

If the novel is matched with the Bing, all chapters contain the value indicating negative sentiments. If the novel is matched with the NRC, only a small number of chapters have positive values and other chapters still show negative sentiments. The existence of a block of negative values shows that negative sentiments permeate the novel. Moreover, the absolute value of the intercalary chapters tends to be lower than that of the narrative chapters. The narrative chapters are more abundant in negative sentiments, whereas the intercalary chapters seem to serve as the subsidiary texts to build up momentum for displaying more sentiments in the narrative chapters.

The result of sentiment analysis in Figure 9 can be roughly checked by calculating the negated words in 2-grams extracted from the novel. Figure 10 shows the numbers of negated words from the Bing lexicon.

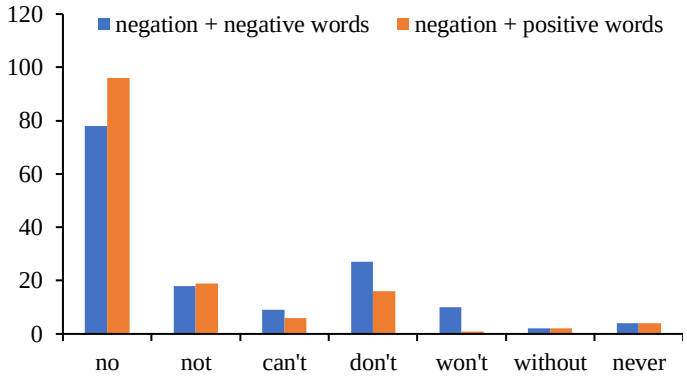

Figure 10. The number of negated words from the Bing lexicon in the sentiment analysis

In Figure 10, most negated words follow the negation *no*. Moreover, the number of negated negative words (144) is nearly as large as that of negated positive words (148). The distribution of negated words supports the result of previous sentiment analysis that negative sentiments generally permeate the novel.

4 Conclusion

The Grapes of Wrath is characterized by intercalary chapters and narrative chapters. This study adopted the narratological distinction of chapters as a framework and conducted a corpus-driven stylistic analysis in terms of word cluster, lexical richness, descriptivity, activity, nominality, keyness and sentiment. Starting from these indicators, the analysis then drills down to the specific value of statistical significance and interprets the linguistic features.

The cluster analysis shows that the narrative chapters are much more homogenous than the intercalary chapters. Of the intercalary chapters, Chapter 7 is unique in the word choice. This uniqueness can be interpreted from the semantic field of the words and the presentation of speech. The MATTR analysis shows that the intercalary chapters vary more than the narrative chapters in terms of lexical richness. Chapter 15 possesses the highest MATTR value, and its lexical richness is relevant to the presentation of speech. The study of descriptivity, activity and nominality finds that the author manipulates the style in Section 3 of the novel. Although the narrative chapters are more active than most intercalary chapters, the intercalary Chapter 5 is the most active in the novel. The keyness analysis shows that the activity of Chapter 5 features the frequent use of the lemma *be*, which helps create the tension between different perceptions of characters. The sentiment analysis finds that negative sentiments generally permeate the whole novel, and the narrative chapters are more abundant in negative sentiments than the intercalary chapters.

Acknowledgements

This research is supported by the National Philosophy and Social Sciences Fund of China (21BYY043).

References

Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio. In: Doležel, L., Bailey, R. W. (Eds.). *Statistics and Style*, pp. 57-65. New York: American Elsevier.

Arnold, T., Tilton, L. (2016). coreNLP: Wrappers around Stanford coreNLP tools. R package version 0.4-2. https://CRAN.R-project.org/package=coreNLP

Boder, D. P. (1940). The adjective-verb quotient: A contribution to the psychology of language. *Psychological Record*, 3, pp. 310-343. https://doi.org/10.1007/BF03393230

Burcar, L. (2018). The (Forgotten) significance of interchapters in John Steinbeck's The Grapes of Wrath: From tenancy to seasonal migrant farm labor. *Arcadia*, 53(2), pp. 360-378. https://doi.org/10.1515/arcadia-2018-0027

Chen, X., Kubát, M. (2021). Rural versus urban fiction in contemporary Chinese literature: Quantitative approach case study. *Digital Scholarship in the Humanities*. (Online) https://doi.org/10.1093/llc/fqab094

Covington, M. A., McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), pp. 94-100. https://doi.org/10.1080/09296171003643098

Gries, S. T. (2017). *Quantitative Corpus Linguistics with R: A practical introduction (2nd ed.)*. New York: Routledge.

Halliday, M. A. K. (1971). Linguistic function and literary style: An inquiry into the language of William Golding's The Inheritors. In: Chatman, S. (Ed.). *Literary Style: A symposium*, pp. 330-368. Oxford: Oxford University Press.

Hamilton, S. (2016). The legacy of Steinbeck's interchapters: The effects of palimpsest on group consciousness and universality. *Steinbeck Review*, 13(2), pp. 169-178. https://doi.org/10.5325/steinbeckreview.13.2.0169

Hardie, A. (2014) 'Log ratio: An informal introduction', Blog post. ESRC Centre for Corpus Approaches to Social Science (CASS). <http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/> (last accessed 29 December 2021).

Hardy, D. E. (2004). Collocational analysis as a stylistic discovery procedure: The case of Flannery O'Connor's Eyes. *Style*, 38(4), pp. 410-427. https://www.jstor.org/stable/10.5325/style.38.4.410

Hu, M., Liu, B. (2004). Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177. New York: Association for Computing Machinery.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics,* 12, pp. 249-268. https://doi.org/10.1075/ijcl.12.2.09hun

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. Champaign, IL: University of Illinois Press.

Jockers, M. L., Thalken, R. (2014). *Text Analysis with R for Students of Literature*. New York: Springer.

Leech, G. N., Short, M. (2007). *Style in Fiction: A linguistic introduction to English fictional prose (2nd ed.)*. Harlow: Pearson Education.

Levant, H. (2007). The fully matured art: The Grapes of Wrath. In: Bloom, H. (Ed.). *Bloom's Modern Critical Interpretations: The Grapes of Wrath*, pp. 7-37. New York: Infobase Publishing.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Baker, M., Francis G., Tognini-Bonelli, E. (Eds.). *Text and Technology: In honour of John Sinclair*, pp. 157-176). Amsterdam: John Benjamins. https://doi.org/10.1075/z.64.11lou

Mahlberg, M. (2014). Corpus stylistics. In: M. Burke (Ed.). *The Routledge Handbook of Stylistics*, pp. 378-393). New York: Routledge.

McIntyre, D., Walker, B. (2019). *Corpus Stylistics: Theory and practice*. Edinburgh: Edinburgh University Press.

Michalke, M. (2020). koRpus.lang.en: Language support for 'koRpus' package: English (Version 0.1-4). Available from https://reaktanz.de/?c=hacking&s=koRpus.

Mohammad, S. M., Turney, P. D. (2013). *NRC Emotion Lexicon.* National Research Council, Canada, 2.

Phelan, J. (1981). *Worlds from Words: A theory of language in fiction*. Chicago: University of Chicago Press.

Phelan, J. (1996). Voices, distance, temporal perspective, and the dynamics of A Farewell to Arms. In: *Narrative as Rhetoric: Techniques, audiences, ethics and ideology*, pp. 59-84). Columbus, OH: Ohio State University Press.

Popescu, I. I., Čech, R.,, Altmann, G. (2013). Descriptivity in Slovak lyrics. *Glottotheory*, 4(1), pp. 92-104. https://doi.org/10.1524/glot.2013.0007

Rayson, P., Garside, R. (2000). Comparing corpora using frequency profiling. In: *The Workshop on Comparing Corpora* (Volume 9), pp. 1-6. Stroudsburg: Association for Computational Linguistics. http://dx.doi.org/10.3115/1117729.1117730

Rimmon-Kenan, S. (1989). How the model neglects the medium: Linguistics, language, and the crisis of narratology. *The Journal of Narrative Technique*, 19(1), pp. 157-166. https://www.jstor.org/stable/30225242

Short, M. H. (1984). Who is Stylistics. *Journal of Foreign Languages*, 5, pp. 14-21.

Short, M. (1999). Graphological deviation, style variation and point of view in Marabou Stork Nightmares by Irvine Welsh. *Journal of Literary Studies*, 15(3-4), pp. 305-323. https://doi.org/10.1080/02564719908530234

Silge, J., Robinson, D. (2017). *Text Mining with R: A tidy approach*. Sebastopol, CA: O'Reilly Media, Inc.

Steinbeck, J. (2006). *The Grapes of Wrath*. New York: Penguin Books.

Swensen, J. R. (2015). *Picturing migrants: The Grapes of Wrath and new deal documentary photography (Vol. 18)*. Oklahoma: University of Oklahoma Press.

Wickham, H. (2016). *ggplot2 - Elegant graphics for data analysis*. Cham: Springer International Publishing.

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr

Wickham, H. (2021). tidyr: Tidy messy data. R package version 1.1.4. https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., Müller, K. (2021). dplyr: A grammar of data manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

Memory limitations are hidden in grammar

Carlos Gómez-Rodríguez[1], Morten H. Christiansen[2,3], Ramon Ferrer-i-Cancho[4*]

[1] Universidade da Coruña, CITIC, FASTPARSE Lab, LyS Research Group, Depto. de Ciencias de la Computación y Tecnologías de la Información, A Coruña, Spain.
[2] Department of Psychology, Cornell University, Ithaca, NY, USA.
[3] Interacting Minds Centre and School of Communication and Culture, Nobelparken, Aarhus University, Denmark.
[4] Complexity and Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain.
[*] Corresponding author's email: rferrericancho@cs.upc.edu

DOI: https://doi.org/10.53482/2022_52_397

ABSTRACT

The ability to produce and understand an unlimited number of different sentences is a hallmark of human language. Linguists have sought to define the essence of this generative capacity using formal grammars that describe the syntactic dependencies between constituents, independent of the computational limitations of the human brain. Here, we evaluate this independence assumption by sampling sentences uniformly from the space of possible syntactic structures. We find that the average dependency distance between syntactically related words, a proxy for memory limitations, is less than expected by chance in a collection of state-of-the-art classes of dependency grammars. Our findings indicate that memory limitations have permeated grammatical descriptions, suggesting that it may be impossible to build a parsimonious theory of human linguistic productivity independent of non-linguistic cognitive constraints.

Keywords: dependency syntax, dependency distance minimization, memory, grammar, network science

1 Introduction

An often celebrated aspect of human language is its capacity to produce an unbounded number of different sentences (Chomsky, 1965; Miller, 2000). For many centuries, the goal of linguistics has been to capture this capacity by a formal description—a grammar—consisting of a systematic set of rules and/or principles that determine which sentences are part of a given language and which are not (Bod, 2013). Over the years, these formal grammars have taken many forms but common to them all is the assumption that they capture the idealized linguistic competence of a native speaker/hearer, independent of any memory limitations or other non-linguistic cognitive constraints (Chomsky, 1965; Miller, 2000). These abstract formal descriptions have come to play a foundational role in the language sciences, from linguistics, psycholinguistics, and neurolinguistics (Hauser et al., 2002; Pinker, 2003) to

computer science, engineering, and machine learning (Dyer et al., 2016; Gómez-Rodríguez et al., 2018; Klein and Manning, 2003). Despite evidence that processing difficulty underpins the unacceptability of certain sentences (Hawkins, 2004; Morrill, 2010), the cognitive independence assumption that is a defining feature of linguistic competence has not been examined in a systematic way using the tools of formal grammar. It is therefore unclear whether these supposedly idealized descriptions of language are free of non-linguistic cognitive constraints, such as memory limitations.

If the cognitive independence assumption should turn out not to hold, then it would have wide-spread theoretical and practical implications for our understanding of human linguistic productivity. It would require a reappraisal of key parts of linguistic theory that hitherto have posed formidable challenges for explanations of language processing, acquisition and evolution (Gold, 1967; Hauser et al., 2002; Pinker, 2003)—pointing to new ways of thinking about language that may simplify the problem space considerably by making it possible to explain apparently arbitrary aspects of linguistic structure in terms of general learning and processing biases (Christiansen and Chater, 2008; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). In terms of practical ramifications, engineers may benefit from building human cognitive limitations directly into their natural language processing systems, so as to better mimic human language skills and thereby improve performance. Here, we therefore evaluate the cognitive independence assumption using a state-of-the-art grammatical framework, dependency grammar (Nivre, 2005), to search for possible hidden memory constraints in these formal, idealized descriptions of natural language.

In dependency grammar, the syntactic structure of a sentence is defined by two components. First, a directed graph where vertices are words and arcs indicate syntactic dependencies between a head and its dependent. Such a graph has a root (a vertex that receives no edges) and edges are oriented away from the root (Figure 1). Second, the linear arrangement of the vertices of the graph (defined by the sequential order of the words in a sentence). Thus, syntactic dependency structures constitute a particular kind of spatial network where the graph is embedded in one dimension (Barthélemy, 2018), a correspondence that has led to the development of syntactic theory from a network theory standpoint (Gómez-Rodríguez and Ferrer-i-Cancho, 2017).

Dependency grammar is an important framework for various reasons. First, categorial grammar defines the syntactic structure of a sentence as dependency grammar (Morrill, 2010). Second, equivalences exist between certain formalisms of dependency grammar and constituency grammar (Gaifman, 1965; Kahane and Mazziotta, 2015). Third, there has been an evolution of minimalism towards dependency grammar (Osborne et al., 2011). Finally, dependency grammar has become a *de facto* standard in computational

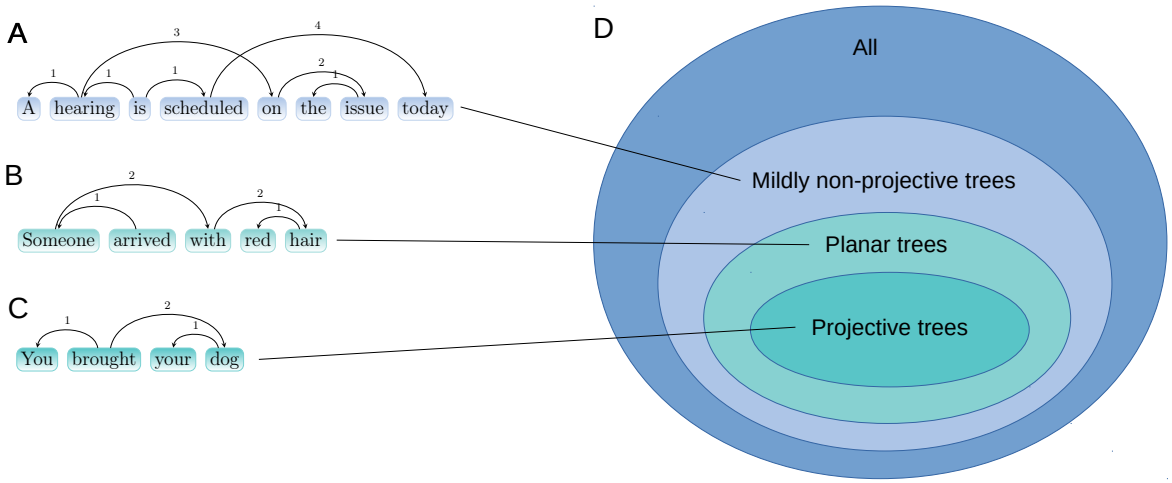

Figure 1: Examples of syntactic dependency structures. Arcs indicate syntactic dependencies from a head to its dependent and are labelled with the distance between them (distance is measured in words; consecutive words are at distance 1). n is the number of words of the sentence, D is the sum of dependency distances and $\langle d \rangle = D/(n-1)$ is the average dependency distance. A. A mildly non-projective tree from the classes $1EC$ and MH_4 (adapted from Nivre, 2009) where $n = 8$ and $\langle d \rangle = 13/7 \approx 1.85$. B. A planar but non-projective tree where $n = 5$ and $\langle d \rangle = 3/2$ (adapted from Groß and Osborne, 2009). C. A projective tree (adapted from Groß and Osborne, 2009) where $n = 4$ and $\langle d \rangle = 4/3$. D. A diagram of the superset relationships between projective, planar, mildly non-projective and unrestricted (all) syntactic dependency structures.

linguistics (Kübler et al., 2009).

To delimit the set of possible grammatical descriptions, various classes or sets of syntactic dependency structures have been proposed. These classes can be seen as filters on the possible linear arrangements of a given tree. Here, we consider four main classes. First, consider planar structures, where edges do not cross when drawn above the words of the sentence. The structure in Figure 1 B-C are planar while that of Figure 1 A is not. Second, we have projective structures, the most well-known class. A dependency tree is projective if, and only if, it is planar and its root is not covered by any dependency (Figure 1 C). Third, there are mildly non-projective structures, comprising the union of planar structures and additional structures with further (but slight) deviations from projectivity, e.g., by having a low number of edge crossings (Figure 1 A). Finally, the class of all structures, that has no constraints on the possible structures.

Figure 1 D shows the inclusion relationships among these classes. However, the whole picture, encompassing state-of-the-art classes is more complex. Mildly non-projective structures are not actually a class but a family of classes. We have selected three representative classes: MH_k, WG_1 and $1EC$ structures, that are supersets of projective structures but whose definition is more complex (see Methods).

Here we validate the assumption of independence between grammatical constraints and cognitive limitations in these classes of grammar using the distance between syntactically related words in a dependency tree as a proxy for memory constraints (Liu et al., 2017; Temperley and Gildea, 2018). Such a distance is defined as the number of intermediate words plus one. Thus, if the linked words are consecutive they are at distance 1, if they are separated by an intermediate word they are at distance two, and so on, as shown in Figure 1. Dependency distance minimization is a pressure to reduce the distance between syntactically related words that is supported statistically by large-scale analyses of syntactic dependency structures in languages (Ferrer-i-Cancho et al., 2022; Futrell et al., 2020; Futrell et al., 2015; Jing et al., 2021; Liu, 2008). As such, dependency distance minimization is a type of memory constraint, believed to result from pressure against decay of activation or interference during the processing of sentences (Liu et al., 2017; Temperley and Gildea, 2018). Dependency distances tax memory and cognition in general. Dependency distances reduce in case of cognitive impairment (Aronsson et al., 2021; Roark et al., 2011). There is an association between the level of cognitive impairment and dependency distance: as the severity of the impairment increases, dependency distances tend to be reduced (Aronsson et al., 2021). Moreover, an association between the level of competence of L2 learners and dependency distance has also been found: as learners of a second language become more competent in the new language, dependency distances increase (Ouyang and Jiang, 2018; Yuan et al., 2021).

The article is written so that reading the next section, *Materials and methods* (Section 2) is not essential

to understand the *Results* section (Section 3). Therefore, it is up to reader to decide whether to proceed with Section 2 or to skip to Section 3, reading Section 2 later on.

2 Material and Methodology

2.1 Control for Sentence Length

In our study, we do not investigate the average dependency distance over a whole ensemble of dependency structures but instead we condition on sentence length (Ferrer-i-Cancho and Liu, 2014; Futrell et al., 2015). Then for a given n, we calculate $\langle d \rangle_{AS}$, the average dependency length for an ensemble of artificial syntactic dependency structures (AS), and also $\langle d \rangle_{RS}$, the average dependency length for an ensemble of attested syntactic dependency structures (RS). By doing that, we are controlling for sentence length, getting rid of the possible influence of the distribution of sentence length in the calculation of $\langle d \rangle_{RS}$ or $\langle d \rangle_{AS}$ (Ferrer-i-Cancho and Liu, 2014).

2.2 Attested Syntactic Dependency Structures

We estimated the average dependency distances in attested sentences using collections of syntactic dependency treebanks from different languages. A syntactic dependency treebank is a database of sentences and their syntactic dependency trees.

To provide results on a wide range of languages while controlling for the effects of different syntactic annotation theories, we use two collections of treebanks:

- Universal Dependencies (UD), version 2.4 (Nivre et al., 2019). This is the largest available collection of syntactic dependency treebanks, featuring 146 treebanks from 83 distinct languages. All of these treebanks are annotated following the common Universal Dependencies annotation criteria, which are a variant of the Stanford Dependencies for English (de Marneffe and Manning, 2008), based on lexical-functional grammar (Bresnan, 2000), adapting them to be able to represent syntactic phenomena in diverse languages under a common framework. This collection of treebanks can be freely downloaded[1] and is available under free licenses.

- HamleDT 2.0 (Rosa et al., 2014). This collection is smaller than UD, featuring 30 languages, all of which (except for one: Bengali) are also available in UD, often with overlapping source material. Thus, using this collection does not meaningfully extend the diversity of languages covered beyond using only UD. However, the interest of HamleDT 2.0 lies in that each of the 30 treebanks is annotated with not one, but two different sets of annotation criteria: Universal Stanford dependencies (de Marneffe et al., 2014) and Prague Dependencies (Hajič et al., 2006). We abbreviate these two

[1] https://universaldependencies.org/. Last accessed 17 February 2022.

subsets of the HamleDT 2.0 collection as "Stanford" and "Prague", respectively. While Universal Stanford dependencies are closely related to UD, Prague dependencies provide a significantly different view of syntax, as they are based on the functional generative description (Sgall, 1969) of the Praguian linguistic tradition (Hajicova, 1995), which differs from Stanford dependencies in substantial ways, like the annotation of conjunctions or adpositions (Passarotti, 2016). Thus, using this version of HamleDT[2] makes our analysis more robust, as we can draw conclusions without being tied to a single linguistic tradition. The HamleDT 2.0 treebanks are available online.[3] While not all of the treebanks are made fully available to the public under free licenses, to reproduce our analysis it is sufficient to use a stripped version where the words have been removed from the sentences for licensing reasons, but the bare trees are available. This version is distributed freely.[4]

A preprocessed file with the minimal information needed to reproduce our measurements on attested syntactic structures (Figure 6 A) is available. [5]

To preprocess the treebanks for our analysis, we removed punctuation, following common practice in statistical research of dependency structures (Gómez-Rodríguez and Ferrer-i-Cancho, 2017). We also removed tree nodes that do not correspond to actual words, such as the null elements in the Bengali, Hindi and Telugu HamleDT corpora and the empty nodes in several UD treebanks. To ensure that the dependency structures are still valid trees after these removals, we reattached nodes whose head has been deleted as dependents of their nearest non-deleted ancestor. Finally, in our analysis we disregarded syntactic trees with less than three nodes, as their statistical properties are trivial and provide no useful information (a single-node dependency tree has no dependencies at all, and a 2-node tree always has a single dependency of distance 1). Tables 1 and 2 summarize the languages in each collection of treebanks.

2.3 Artificial Syntactic Dependency Structures

Apart from the attested trees, we used a collection of over 16 billion randomly-generated trees. For values of n (the length or number of nodes) from 3 to $n^* = 10$, we exhaustively obtained all possible trees. The number of possible dependency trees for a given length n is given by n^{n-1}, ranging from 9 possible trees for $n = 3$ to 10^9 for $n = n^*$. From $n > n^*$ onwards, the number of trees grows too large to be manageable, so we resort to uniformly random sampling of 10^9 trees for $n^* < n \leq 25$. For each tree

[2]While there is a later version (HamleDT 3.0), it abandoned the dual annotation and adopted Universal Dependencies instead, thus making it less useful for our purposes.

[3]https://ufal.mff.cuni.cz/hamledt/hamledt-treebanks-20. Last accessed 17 February 2022.

[4]https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-9551-4?show=full. Last accessed 17 February 2022.

[5]https://doi.org/10.7910/DVN/XHRIYX

Table 1: The languages in the UD collection grouped by family. The counts attached to the collection name indicate the number of different families and the number of different languages. The counts attached to family names indicate the number of different languages.

Collection	Family	Languages
UD (19, 83)	Afro-Asiatic (7)	Akkadian, Amharic, Arabic, Assyrian, Coptic, Hebrew, Maltese
	Turkik (3)	Kazakh, Turkish, Uyghur
	Austro-Asiatic (1)	Vietnamese
	Austronesian (2)	Indonesian, Tagalog
	Basque (1)	Basque
	Dravidian (2)	Tamil, Telugu
	Indo-European (46)	Afrikaans, Ancient Greek, Armenian, Belarusian, Breton, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Faroese, French, Galician, German, Gothic, Greek, Hindi, Hindi-English, Irish, Italian, Kurmanji, Latin, Latvian, Lithuanian, Marathi, Norwegian, Old Church Slavonic, Old French, Old Russian, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, Upper Sorbian, Urdu, Welsh
	Japanese (1)	Japanese
	Korean (1)	Korean
	Mande (1)	Bambara
	Mongolic (1)	Buryat
	Niger-Congo (2)	Wolof, Yoruba
	Other (1)	Naija
	Pama-Nyungan (1)	Warlpiri
	Sign Language (1)	Swedish Sign Language
	Sino-Tibetan (3)	Cantonese, Chinese, Classical Chinese
	Tai-Kadai (1)	Thai
	Tupian (1)	Mbya Guarani
	Uralic (7)	Erzya, Estonian, Finnish, Hungarian, Karelian, Komi Zyrian, North Sami

Table 2: The languages in the HamleDT collections (Stanford and Prague) grouped by family. The counts attached to the collection names indicate the number of different families and the number of different languages. The counts attached to family names indicate the number of different languages.

Collection	Family	Languages
Stanford (7, 30)	Afro-Asiatic (1)	Arabic
	Turkik (1)	Turkish
	Basque (1)	Basque
	Dravidian (2)	Tamil, Telugu
	Indo-European (21)	Ancient Greek, Bengali, Bulgarian, Catalan, Czech, Danish, Dutch, English, German, Greek, Hindi, Italian, Latin, Persian, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish
	Japanese (1)	Japanese
	Uralic (3)	Estonian, Finnish, Hungarian
Prague (7, 30)	Afro-Asiatic (1)	Arabic
	Turkik (1)	Turkish
	Basque (1)	Basque
	Dravidian (2)	Tamil, Telugu
	Indo-European (21)	Ancient Greek, Bengali, Bulgarian, Catalan, Czech, Danish, Dutch, English, German, Greek, Hindi, Italian, Latin, Persian, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish
	Japanese (1)	Japanese
	Uralic (3)	Estonian, Finnish, Hungarian

in the collection, the classes it belongs to are indicated in the dataset[6].

The reason why we do not go beyond length 25 is that, for larger lengths, trees that belong to our classes under analysis are very scarce (Figure 2 A). For example, even sampling 10^9 random trees for each length, no projective trees are found for $n > 18$. The same can be said of planar trees for $n > 19$, $1EC$ trees for $n > 22$, MH_4 trees for $n > 23$, and WG_1 trees for $n > 24$. For the MH_5 class, some trees can still be found in the sample for length 25, but only 69 out of 10^9 belong to the class. Due to undersampling, the plot on artificial structures in the results section only shows points represented by at least 30 structures for $n > n^*$. 30 is considered a rule of thumb for the minimum sample size that is needed to estimate the mean of random variables that follow short tailed distributions (Hogg and Tanis, 1997). Figure 2 B shows average dependency distances not excluding any point.

For $n \leq n^*$, the ensemble of AS used to calculate $\langle d \rangle_{AS}$ contains all possible syntactic dependency structures for all classes. For $n > n^*$, it contains a random sample of them. Within a given ensemble, each structure is generated from a labelled directed tree whose vertex labels are interpreted as vertex positions in the linear arrangement. The values of $\langle d \rangle_{AS}$ for each class are exact (the mean over all possible syntactic dependency structures) for $n \leq n^*$ and random sampling estimates for $n > n^*$. A detailed explanation follows.

For a given n, an ensemble of syntactic dependency structures is generated with a procedure that is a generalization of the procedure used to generate random structures formed by an undirected tree and a linear arrangement (Esteban et al., 2016). The procedure has two versions: the exhaustive version, that was used for $n \leq n^*$, and the random sampling version, that was used for $n > n^*$. The exhaustive version consists of

1. Generating all the $T(n)$ labelled (undirected) trees of n vertices using Prüfer codes (Prüfer, 1918). It is known that $T(n) = n^{n-2}$ (Cayley, 1889).

2. Converting each of these random trees into labelled directed trees (i.e., dependency trees) by rooting it in all possible ways. A rooting consists in choosing one node of the tree as the root, and making all edges point away from the root via a depth-first traversal. This produces $nT(n) = n^{n-1}$ syntactic dependency structures.

3. Producing a syntactic dependency structure from every directed tree using vertex labels (integers from 1 to n) as vertex positions in a linear arrangement (Esteban et al., 2016).

[6]The trees are freely available from https://doi.org/10.7910/DVN/XHRIYX

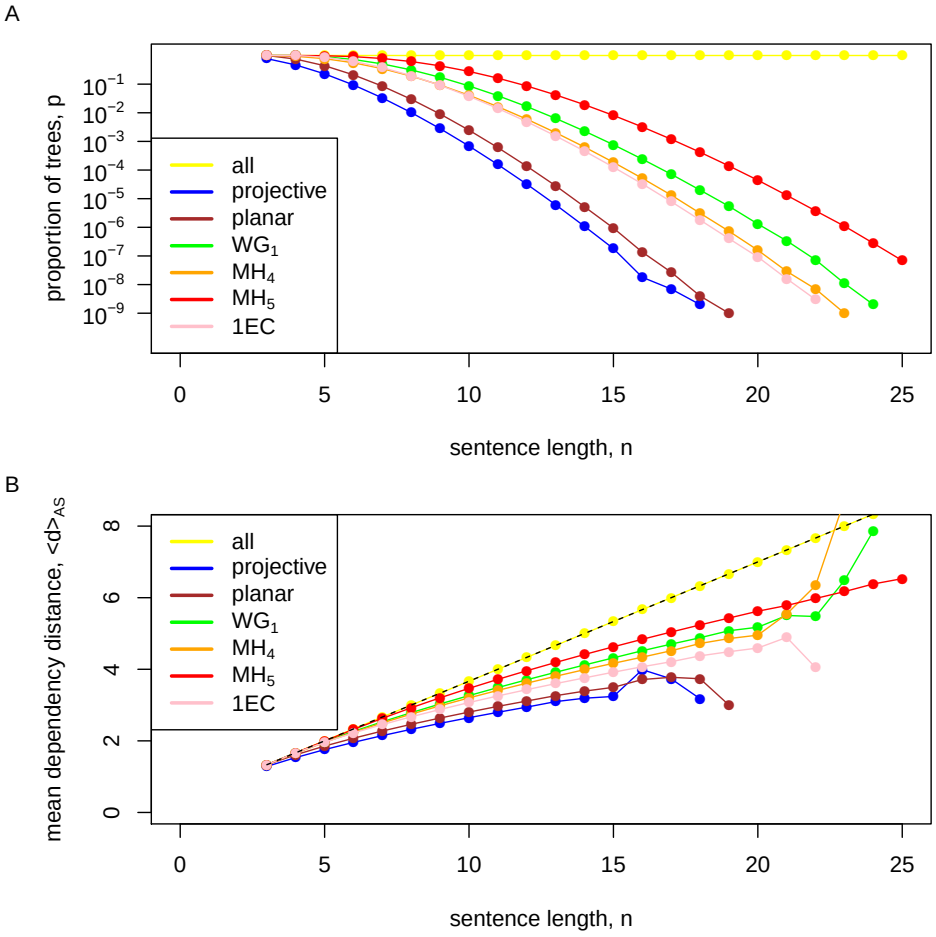

Figure 2: Undersampling in artificial syntactic dependency structures (AS). A. p, the proportion of artificial structures of a certain class in the sample. B. The average dependency length (in words), $\langle d \rangle_{AS}$, as a function of n, the sentence length (in words). For reference, the base line defined by a random linear arrangement of the words of the sentence, $\langle d \rangle_{rla}$ (Eq. 3) is also shown (dashed line).

4. Discarding the trees that do not belong to the target class.

The random sampling version consists of

1. Generating S uniformly random labelled (undirected) trees of n vertices, via uniformly random Prüfer codes (Prüfer, 1918).
2. Converting these uniformly random labelled trees to uniformly random labelled directed trees (i.e., dependency trees) by randomly choosing one node of each tree as the root, and making all edges point away from the root via a depth-first traversal. This produces S syntactic dependency structures.
3. Same as exhaustive version.
4. Same as exhaustive version.

Note that Step 2 warrants that labelled directed trees in the ensemble are uniformly random: if we call K_n the probability of generating each undirected tree of n vertices with a random Prüfer code, we can observe that each possible directed tree corresponds to exactly one undirected tree (obtained by ignoring arc directions), and each undirected tree corresponds to exactly n distinct directed trees (resulting from picking each of its n nodes as the root). Thus, the method of generating a random Prüfer code and then choosing a root generates each possible directed tree with a uniform probability K_n/n (as the probability of choosing the underlying undirected tree is K_n, and the probability of choosing the relevant root is $1/n$).

After each procedure, the average dependency length $\langle d \rangle$ for a given n and a given class is calculated. While the exhaustive procedure allows one to calculate the true average dependency length over a certain class, the random sampling algorithm only allows one to estimate the true average. Put differently, the exhaustive procedure allows one to calculate exactly the expected dependency length in a class assuming that all labelled directed trees are equally likely whereas the random sampling procedure only allows one to obtain an approximation.

We explore all values of n within the interval $[n_{min}, n_{max}]$ with $n_{min} = 3$ and $n_{max} = 25$ and $n^* = 10$ and $S = 10^9$. The total number of syntactic dependency structures generated for our study is

$$U = (n_{max} - n^*)S + \sum_{n=n_{min}}^{n^*} nT(n) = (n_{max} - n^*)S \sum_{n=n_{min}}^{n^*} n^{n-1}.$$

Applying the parameters above, one obtains

$$U \approx 1.6 \cdot 10^{10} \quad (1)$$

2.4 The Random Baseline

Although the random baseline

$$\langle d \rangle_{rla} = (n+1)/3 \tag{2}$$

follows from Jaynes' maximum entropy principle in the absence of any constraint (Kesavan, 2009), it may be objected that our baseline is too unconstrained from a linguistic perspective. In previous research, random baselines that assume projectivity or consistent branching, whereby languages tend to grow parse trees either to the right (as in English) or to the left (as in Japanese), have been considered (Futrell et al., 2015; Gildea and Temperley, 2010; Liu, 2008). However, it has been argued that these linguistic constraints could be a reflection of memory limitations (Christiansen and Chater, 1999; Ferrer-i-Cancho and Gómez-Rodríguez, 2016b). Therefore, incorporating these linguistic constraints into the baseline for evaluating dependency distances would not provide an adequate test of the cognitive independence assumption because they could mask the effect of dependency distance minimization (DDm). Consistently, the planarity assumptions reduces the statistical power of a test of DDm (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). In addition, these additional constraints compromise the parsimony of a general theory of language for neglecting the predictive power of DDm (Ferrer-i-Cancho and Gómez-Rodríguez, 2016b).

A priori, $\langle d \rangle_{AS}$ could be below the random baseline as it occurs typically in human languages (Ferrer-i-Cancho, 2004; Ferrer-i-Cancho and Liu, 2014) but it could also be above. As for the latter situation, empirical research in short sentences has shown that there are languages where dependency lengths are larger than expected by chance (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). In addition, there exist syntactic dependency structures where $\langle d \rangle > \langle d \rangle_{rla}$ from a network theoretical standpoint. For instance, among planar syntactic structures, the maximum average dependency distance is $\langle d \rangle_{max} = n/2$ (Ferrer-i-Cancho, 2013).

$\langle d \rangle_{AS}$ never exceeds $\langle d \rangle_{rla}$ and it deviates from $\langle d \rangle_{rla}$ when $n = 3$ for projective trees, $n = 4$ for planar trees and MH_4 and $n = 5$ for $1EC$, MH_5 and WG_1. For the class of all syntactic dependency structures (Figure 1 D), we find that $\langle d \rangle_{AS}$ matches Eq. 2 as expected from previous research (Esteban et al., 2016).

2.5 The Classes of Dependency Structures

Planar trees: A dependency tree is said to be *planar* (or *noncrossing*) if its dependency arcs do not cross when drawn above the words. Planar trees have been used in syntactic parsing algorithms (Gómez-Rodríguez and Nivre, 2010), and their generalization to noncrossing graphs has been widely studied both for its formal properties (Yli-Jyrä and Gómez-Rodríguez, 2017) and for parsing (Kuhlmann and Jonsson, 2015).

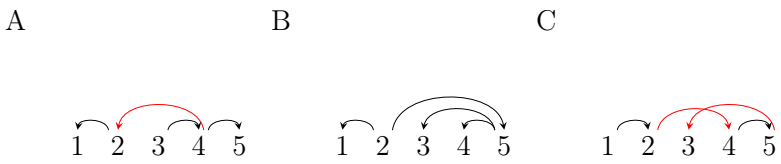

Figure 3: Planarity and projectivity. A. A tree that is planar (dependencies do not cross) but not projective (the root node, 3, is covered by the dependency in red). B. A tree that is planar and projective. C. A tree that is not planar (the dependencies in red cross), and thus not projective.

Projective trees: A dependency tree is said to be projective if it is planar and its root is not covered by any dependency (Figure 3). Projectivity facilitates the design of simple and efficient parsers (Nivre, 2003, 2004), whereas extending them to support non-projective trees increases their computational cost (Covington, 2001; Nivre, 2009). For this reason, and because treebanks of some languages (like English or Japanese) have traditionally had few or no non-projective analyses, many practical implementations of parsers assume projectivity (Chen and Manning, 2014; Dyer et al., 2015).

However, non-projective parsing is needed to deal with sentences exhibiting non-projective phenomena such as extraposition, scrambling or topicalization. Non-projectivity is particularly common in flexible word order languages, but generally present in a wide range of languages. However, non-projectivity in natural languages tends to be *mild* in the sense that the actually occurring non-projective trees are very close to projective trees, as they have much fewer crossing dependencies than would be expected by chance (Ferrer-i-Cancho et al., 2018).

For this reason, there has been research interest in finding a restriction that would be a better fit for the phenomena observed in human languages. From a linguistic standpoint, the goal is to describe the syntax of human language better than with the overly restrictive projective trees or the arguably excessive permissiveness of admitting any tree without restriction, disregarding the observed scarcity of crossing dependencies. From an engineering standpoint, the goal is to strike a balance between the efficiency provided by more restrictive parsers with a smaller search space and the coverage of the non-projective phenomena that can be found in attested sentences. In this line, various sets of dependency structures that have been proposed are supersets of projective trees allowing only a limited degree of non-projectivity. These sets are called mildly non-projective classes of dependency trees (Kuhlmann and Nivre, 2006).

Here, we focus on three of the best known such sets, which have interesting formal properties and/or have been shown to be practical for parsing due to providing a good efficiency-coverage trade-off. We briefly outline them here, and refer the reader to Gómez-Rodríguez, 2016 for detailed definitions and coverage

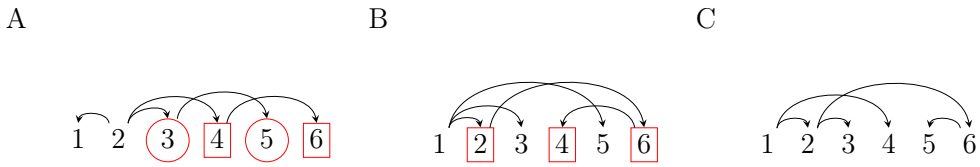

Figure 4: Well-nestedness and gap degree. A. An ill-nested tree (the yields of node 3—circled—and node 4–squared—form an interleaving pattern). B. A tree with gap degree 2 (the yield of node 2, squared, has two discontinuities, at nodes 3 and 5). C. A tree that is well-nested and has gap degree 1, and thus is in WG_1.

statistics of these and other mildly non-projective classes of trees.

Well-nested trees with Gap degree 1 (WG_1): A dependency tree is well-nested (Bodirsky et al., 2005) if it does not contain two nodes with disjoint, interleaving yields. Given two disjoint yields $a_1 \dots a_p$ and $b_1 \dots b_q$, they are said to interleave if there exist i, j, k, l such that $a_i < b_j < a_k < b_l$. On the other hand, the gap degree of a tree is the maximum number of discontinuities present in the yield of a node, i.e., a dependency tree has gap degree 1 if every yield is either a contiguous substring, or the union of two contiguous substrings of the input sentence. Figure 4 provides graphical examples of these properties. WG_1 trees have drawn interest mainly from the formal standpoint, for their connections to constituency grammar (Kuhlmann, 2010), but they also have been investigated in dependency parsing (Corro et al., 2016; Gómez-Rodríguez et al., 2011; Gómez-Rodríguez et al., 2009).

Multi-Headed with at most k heads per item (MH_k): Given $k \geq 3$, the set of MH_k trees contains the trees that can be parsed by an algorithm called MH_k (Gómez-Rodríguez et al., 2011). k is a parameter of the class, such that for $k = 3$ the class coincides with projective trees, but for $k > 3$ it covers increasingly larger sets of non-projective structures (but the parser becomes slower). A recent neural implementation of the MH_4 parser has obtained competitive accuracy on UD (Gómez-Rodríguez et al., 2018). For $k > 4$, the MH_k sets have been shown to be Pareto optimal (among known mildly non-projective classes) in terms of balance between efficiency and practical coverage (Gómez-Rodríguez, 2016). In this paper, we will consider the MH_4 and MH_5 sets.

1-Endpoint-Crossing trees (1EC): A dependency tree has the property of being 1-Endpoint-Crossing if, given a dependency, all other dependencies crossing it are incident to a common node (Pitler et al., 2013). This property is illustrated in Figure 5. 1EC trees were the first mildly non-projective class of dependency trees to have a practical exact-inference parser (Pitler, 2014), which was reimplemented with a neural architecture in (Gómez-Rodríguez et al., 2018). They are also in the Pareto frontier with respect to coverage and efficiency, according to Gómez-Rodríguez, 2016.

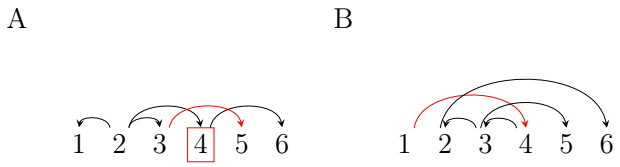

Figure 5: 1-Endpoint-Crossing property. A. A 1-Endpoint-Crossing tree (given any dependency, dependencies crossing it are incident to a common node—for example, here the dependencies crossing the one marked in red are incident to node 4). B. A tree that is not 1-Endpoint-Crossing. The dependency arc in red has two crossing dependencies which are not incident to any common node.

3 Results

3.1 Short Dependency Distances in Attested Structures Revisited

Assuming that all the linear arrangements are equally likely, $\langle d \rangle$, the average of dependency distances in a sentence of n words, is expected to be (Ferrer-i-Cancho, 2004)

$$\langle d \rangle_{rla} = (n+1)/3. \tag{3}$$

Figure 6 A shows that $\langle d \rangle_{RS}$, the average dependency distance in attested syntactic dependency structures (RS), is below the random baseline defined by $\langle d \rangle_{rla}$ (see Methods for a justification of this baseline). This is in line with previous statistical analyses (Ferrer-i-Cancho, 2004; Futrell et al., 2015; Liu, 2008; Park and Levy, 2009) (see Liu et al., 2017; Temperley and Gildea, 2018 for a broader review of previous work) and the expected influence of performance constraints on attested trees.

The fact that $\langle d \rangle_{RS}$ is below 4 has been interpreted as a sign that dependency lengths are constrained by working memory limitations (Liu, 2008). For this reason, we test whether memory effects have permeated the classes of grammar by determining if $\langle d \rangle_{AS}$, the average dependency distance in a collection of artificial syntactic dependency structures (AS) from a certain class, is also below $\langle d \rangle_{rla}$ (Eq. 3). The purpose of Figure 6 A is merely to provide the reader with a baseline derived from attested dependency structures in natural language as a backdrop for the main contribution of the article, which is based on artificial syntactic dependency structures.

3.2 Short Dependency Distances in Artificial Structures

For a given n, we generate an ensemble of artificial syntactic dependency structures by exhaustive sampling for $n \leq n^* = 10$ and random sampling for $n > n^*$ (Methods). These artificial syntactic dependency trees are only constrained by the definition of the different classes. They are thus free

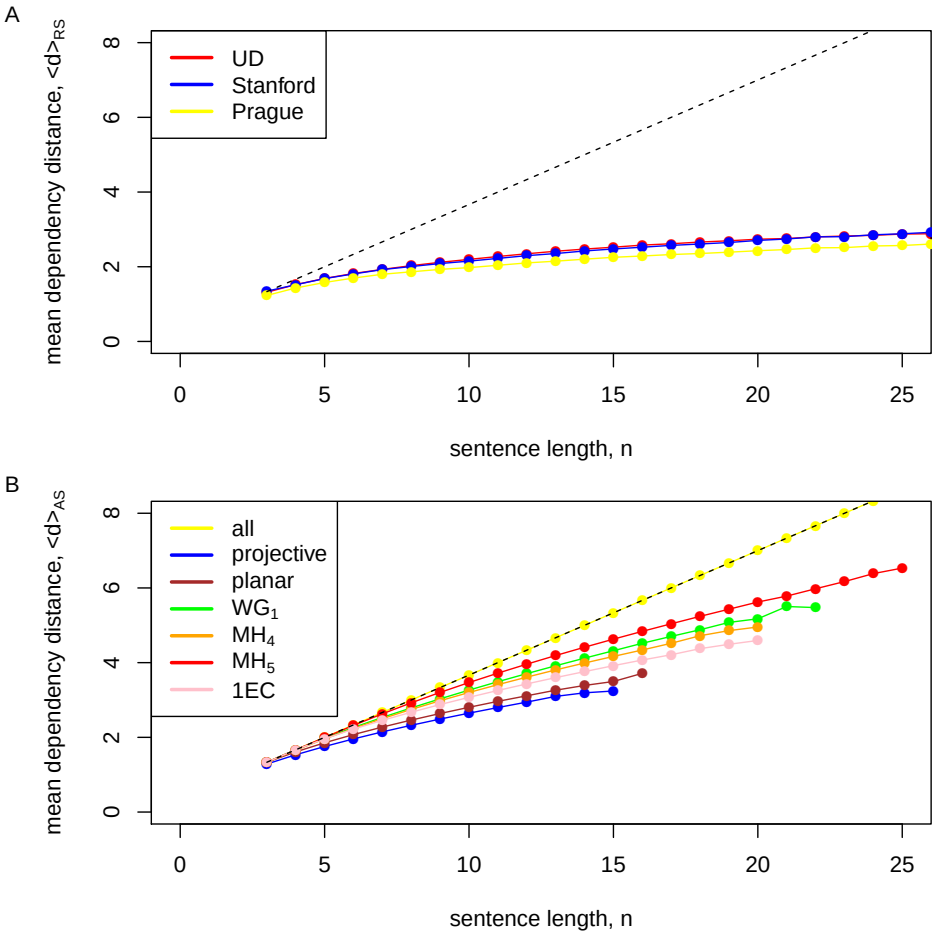

Figure 6: The average dependency length (in words), $\langle d \rangle$, as a function of n, the sentence length (in words). For reference, the baseline defined by a random linear arrangement of the words of the sentence, $\langle d \rangle_{rla}$ (Eq. 3) is also shown (dashed line). A. Attested syntactic dependency trees (RS) following three different annotation criteria: UD, Prague and Stanford dependencies. B. Artificial syntactic dependency structures (AS) belonging to different classes of grammars. Due to undersampling, only points represented by at least 30 structures are shown for $n > n^*$.

from any memory constraint other than the ones the different classes of grammars may, perhaps, impose indirectly. Still, these artificial syntactic structures have dependency lengths that are below the chance level (Figure 6 B), indicating that memory constraints are hidden in the definition of these classes. Interestingly $\langle d \rangle_{AS}$ is below chance for sufficiently large n in all classes of grammars although $\langle d \rangle_{AS}$ could be above $\langle d \rangle_{rla}$ (Eq. 3) in principle (see Methods). In general, the largest reduction of $\langle d \rangle_{AS}$ with respect to the random baseline is achieved by the projective class, followed by the planar class.

It is worth noting that a reduction of $\langle d \rangle_{AS}$ with respect to our random baseline has been observed for the projective class in past work, but with some important caveats: Liu, 2008 did not control for sentence length as in Figure 6 B; and whereas Park and Levy, 2009 did implement this control and considered another class of marginal interest (2-component structures) in addition to projective trees, their use of attested dependency trees instead of artificial control trees suggests that memory limitations might have influenced the results.

4 Discussion

The reduction of $\langle d \rangle$ with respect to the random baseline in artificial trees from a wide range of state-of-the art classes is consistent with the hypothesis that the scarcity of crossing dependencies is a side-effect of pressure to reduce the distance between syntactically related words (Gómez-Rodríguez and Ferrer-i-Cancho, 2017). The smaller reduction of dependency distances with respect to the random baseline in artificial dependency structures can be explained by the fact that the curves in Figure 6 B derive from uniform sampling of the space of all possible trees. In contrast, real speakers may not only choose linear arrangements that reduce dependency distance, but also sample the space of possible structures with a bias towards structures that facilitate that such reduction or that satisfy other cognitive constraints (Ferrer-i-Cancho and Gómez-Rodríguez, 2021).

Our findings complete our understanding of the relationship between projectivity or mildly non-projectivity and dependency distance minimization. It has been shown that such minimization leads to a number of edge crossings that is practically zero (Ferrer-i-Cancho, 2006), and to not covering the root, one of the conditions for projectivity, in addition to planarity (Ferrer-i-Cancho, 2008). Here, we have demonstrated a complementary effect, i.e., that dependency distance reduces below chance when edge crossings are minimized (planarity) or projectivity is imposed. Whereas a recent study of similar classes of grammars suggested that crossing dependencies are constrained by either grammar or cognitive pressures rather than occurring naturally at some rate (Yadav et al., 2019), our findings strongly demonstrate that it is not grammar but rather non-linguistic cognitive constraints, that limit the occurrence of crossing dependencies in languages. Since we released the first version of this article in August 2019, https://arxiv.org/abs/1908.06629, other researchers have confirmed that dependency distance minimiza-

tion contributes significantly to the emergence of formal constraints on crossing dependencies (Yadav et al., 2021, 2022). Yadav et al., 2021 have also confirmed the findings of previous research indicating that the effect of dependency distances alone leads to overestimate the actual number of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017); a critical point is that Gómez-Rodríguez and Ferrer-i-Cancho (2017) use a normalized score leading to the conclusion that such overestimation implies a small relative error.

We sampled about 16 billion syntactic dependency structures, that differed in length and syntactic complexity, to determine whether linguistic grammars are free of non-linguistic cognitive constraints, as is typically assumed. Strikingly, while previous work on natural languages has shown that dependency lengths are considerably below what would be expected by a random baseline without memory constraints (Ferrer-i-Cancho, 2004; Ferrer-i-Cancho and Liu, 2014; Liu, 2008; Park and Levy, 2009), we still observe a drop in dependency lengths for randomly generated, mildly non-projective structures that supposedly abstract away from cognitive limitations. Our interpretation of these results is that memory constraints, in the form of dependency distance minimization, have become inherent to formal linguistic grammars. We have demonstrated that distinct formal classes of mild non-projectivity manifest the sort of burden of dependency distances for memory and cognition that is observed in psychological experiments (Liu et al., 2017, Section 2) and that has been observed to become more marked in case of cognitive impairment (Aronsson et al., 2021) or second language learning (Ouyang and Jiang, 2018; Yuan et al., 2021).

It may be objected that our argument that memory limitations have permeated grammars is based on artificially generated syntactic structures instead of real ones. However, it is all but impossible to study real dependency structures without possible contamination from linguistic or non-linguistic cognitive constraints other than the formal mild non-projectivity classes. For that reason, here and in previous research (Ferrer-i-Cancho, 2014), we have focused on artificially generated syntactic structures. Notice this research is part of a larger research program were we have already used real syntactic dependency structures, but minimizing assumptions to argue that the scarcity of crossing dependencies can be explained to a large extent by dependency distance minimization (Gómez-Rodríguez and Ferrer-i-Cancho, 2017). Nonetheless, further research is needed with real syntactic dependency structures and the current study is a key, necessary step in this direction.

It may also be objected that our conclusions are limited by the sample of classes that we have considered and that we cannot exclude the possibility that, in the future, researchers might adopt a new class of mildly non-projective structures whose dependency distances cannot be distinguished from the random baseline. However, we believe that this is very unlikely for the following reasons: (1) our current sample of classes is representative of the state of the art (Gómez-Rodríguez, 2016), and spans classes that

originated with different goals and motivations (from purely theoretical to parsing efficiency), with all sharing the drop in dependency lengths, (2) while one could conceivably engineer a class to have lengths in line with the baseline while still having high coverage of linguistic phenomena, this would mean forwarding more responsibility for dependency distance reduction to other parts of the linguistic theory in order to warrant that dependency distances are reduced to a realistic degree (Figure 6) and hence would preclude a parsimonious approach to language, and (3) given the positive correlation between crossings and dependency lengths (Alemany-Puig, 2019; Ferrer-i-Cancho and Gómez-Rodríguez, 2016a), such a class would be likely to have many dependency crossings, so it would be, at the least, questionable to call it mildly non-projective.

Beyond upending longheld assumptions about the nature of human linguistic productivity, our findings also have key implications for debates on how children learn language, how language evolved, and how computers might best master language. Whereas a common assumption in the acquisition literature is that children come to the task of language learning with built-in linguistic constraints on what they learn (Gold, 1967; Pinker, 2003), our results suggest that language-specific constraints may not be needed and instead be replaced by general cognitive constraints (Tomasello, 2005). The strong effects of memory on dependency distance minimization provide further support for the notion that language evolved through processes of cultural evolution shaped by the human brain (Christiansen and Chater, 2008), rather than the biological evolution of language-specific constraints (Pinker, 2003). Finally, our results raise the intriguing possibility that if we want to develop computer systems that target human linguistic ability in the context of human-computer interaction, we may paradoxically need to constraint the power of such systems to be in line with human cognitive limitations, rather than giving them the super-human computational capacity of AlphaGo. Memory limitations in the form of dependency minimization have already been applied to machine learning methods, but imposing planarity as if planarity and memory limitations were unrelated constraints (Eisner and Smith, 2010; Smith and Eisner, 2006, for instance). This suggests that considering planarity and other formal constraints as the effect of dependency minimization could boost machine learning methods

Our study was conducted using the framework of dependency grammar, but because of the close relationship between this framework and other ways of characterizing the human unbounded capacity to produce different sentences (Chomsky, 1965; Miller, 2000), such as categorial grammar (Morrill, 2010), phrase structure grammar (Gaifman, 1965; Kahane and Mazziotta, 2015), and minimalist grammar (Osborne et al., 2011), our results suggest that any parsimonious grammatical framework will incorporate memory constraints. Notice that, as a result of our study, we cannot refute the cognitive independence assumption. Our point is that the independence assumption leads to a less parsimonious theory of syntax. We are simply invoking Occam's razor so that formal constraints and the cognitive burden of

dependency distances are not treated as separate entities. Moreover, given that dependency grammars constitute a special case of a graph that is embedded in one dimension, the physics toolbox associated with statistical mechanics and network theory may be applied to provide further insight into the nature of human linguistic productivity (Barthélemy, 2018; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). However, these future explorations notwithstanding, our current findings show that memory limitations have permeated current linguistic conceptions of grammar, suggesting that it may not be possible to adequately capture our unbounded capacity for language, at least in the context of a parsimonious theory compatible with the idea of mild non-projectivity, without incorporating non-linguistic cognitive constraints into the grammar formalism.

Acknowledgments

This article is dedicated to the memory of G. Altmann, 1931-2020 (Köhler et al., 2021). We are grateful to L. Alemany-Puig, A. Hernandez-Fernandez and M. Vitevitch for helpful commments. CGR has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), from ERDF/MICINN-AEI (ANSWER-ASAP, TIN2017-85160-C2-1-R; SCANNER-UDC, PID2020-113230RB-C21), from Xunta de Galicia (ED431C 2020/11), and from Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (ERDF - Galicia 2014-2020 Program), by grant ED431G 2019/01. RFC is supported by the grant TIN2017-89244-R from MINECO and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya).

References

Alemany-Puig, L. (2019). Edge crossings in linear arrangements: From theory to algorithms and applications (Master thesis). Barcelona School of Informatics.

Aronsson, F. S., Kuhlmann, M., Jelic, V., Östberg, P. (2021). Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology*, *35*(7), 900–913. https://doi.org/10.1080/02687038.2020.1742282

Barthélemy, M. (2018). *Morphogenesis of spatial networks*. Springer. https://doi.org/10.1007/978-3-319-20565-6

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199665211.001.0001

Bodirsky, M., Kuhlmann, M., Möhl, M. (2005). Well-nested drawings as models of syntactic structure. *10th Conference on Formal Grammar and 9th Meeting on Mathematics of Language*, 195–203.

Bresnan, J. (2000). *Lexical-functional syntax*. Blackwell.

Cayley, A. (1889). A theorem on trees. *Quart. J. Math*, *23*, 376–378. https://doi.org/10.1017/CBO9780511703799.010

Chen, D., Manning, C. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 740–750. https://doi.org/10.3115/v1/D14-1082

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Christiansen, M. H., Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205. https://doi.org/10.1016/S0364-0213(99)00003-8

Christiansen, M. H., Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558. https://doi.org/10.1017/S0140525X08004998

Corro, C., Le Roux, J., Lacroix, M., Rozenknop, A., Wolfler Calvo, R. (2016). Dependency parsing with bounded block degree and well-nestedness via Lagrangian relaxation and branch-and-bound. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 355–366. https://doi.org/10.18653/v1/P16-1034

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. *Proceedings of the 39th Annual ACM Southeast Conference*, 95–102.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 4585–4592). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf

de Marneffe, M.-C., Manning, C. D. (2008). The Stanford typed dependencies representation. *COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8. http://aclweb.org/anthology/W08-1301

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 334–343. https://doi.org/10.3115/v1/P15-1033

Dyer, C., Kuncoro, A., Ballesteros, M., Smith, N. A. (2016). Recurrent neural network grammars. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 199–209. https://doi.org/10.18653/v1/N16-1024

Eisner, J., Smith, N. A. (2010). Favor short dependencies: Parsing with soft and hard constraints on dependency length. In H. Bunt, P. Merlo, J. Nivre (Eds.), *Trends in parsing technology: Dependency parsing, domain adaptation, and deep parsing* (pp. 121–150). Springer. http://cs.jhu.edu/~jason/papers/#eisner-smith-2010-iwptbook

Esteban, J. L., Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2016). The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics*, 063401. https://doi.org/10.1088/1742-5468/2016/06/063401

Ferrer-i-Cancho, R., C. Gómez-Rodríguez, J. L. E., Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, *105*(1), 014308. https://doi.org/10.1103/PhysRevE.105.014308

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*, 056135. https://doi.org/10.1103/PhysRevE.70.056135

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters*, *76*(6), 1228–1235. https://doi.org/10.1209/epl/i2006-10406-0

Ferrer-i-Cancho, R. (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, *11*(3), 393–414. https://doi.org/10.1142/S0219525908001702

Ferrer-i-Cancho, R. (2013). Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics*, *25*, 1–21.

Ferrer-i-Cancho, R. (2014). A stronger null hypothesis for crossing dependencies. *Europhysics Letters*, *108*(5), 58003. https://doi.org/10.1209/0295-5075/108/58003

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L. (2018). Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, *493*, 311–329. https://doi.org/10.1016/j.physa.2017.10.048

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2016a). Crossings as a side effect of dependency lengths. *Complexity*, *21*, 320–328. https://doi.org/10.1002/cplx.21810

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2016b). Liberating language research from dogmas of the 20th century. *Glottometrics*, *33*, 33–34.

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2021). Anti dependency distance minimization in short sequences. A graph theoretic approach. *Journal of Quantitative Linguistics*, *28*(1), 50–76. https://doi.org/10.1080/09296174.2019.1645547

Ferrer-i-Cancho, R., Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, *5*, 143–155. https://doi.org/10.1515/glot-2014-0014

Futrell, R., Levy, R. P., Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412. https://doi.org/10.1353/lan.2020.0024

Futrell, R., Mahowald, K., Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and Control*, *8*, 304–337. https://doi.org/10.1016/S0019-9958(65)90232-9

Gildea, D., Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447–474. https://doi.org/10.1016/S0019-9958(67)91165-5

Gómez-Rodríguez, C. (2016). Restricted non-projectivity: Coverage vs. efficiency. *Computational Linguistics*, *42*(4), 809–817. https://doi.org/10.1162/COLI_a_00267

Gómez-Rodríguez, C., Carroll, J., Weir, D. (2011). Dependency parsing schemata and mildly non-projective dependency parsing. *Computational Linguistics*, *37*(3), 541–586. https://doi.org/10.1162/COLI_a_00060

Gómez-Rodríguez, C., Ferrer-i-Cancho, R. (2017). Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, *96*, 062304. https://doi.org/10.1103/PhysRevE.96.062304

Gómez-Rodríguez, C., Nivre, J. (2010). A transition-based parser for 2-planar dependency structures. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1492–1501. http://portal.acm.org/citation.cfm?id=1858681.1858832

Gómez-Rodríguez, C., Shi, T., Lee, L. (2018). Global transition-based non-projective dependency parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2663–2674. https://doi.org/10.18653/v1/P18-1248

Gómez-Rodríguez, C., Weir, D., Carroll, J. (2009). Parsing mildly non-projective dependency structures. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, 291–299. https://doi.org/10.3115/1609067.1609099

Groß, T., Osborne, T. (2009). Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics*, *22*, 43–90.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., Urešová, Z. (2006). Prague dependency treebank 2.0.

Hajicova, E. (1995). Prague school syntax and semantics. In E. Koerner R. Asher (Eds.), *Concise history of the language sciences* (pp. 253–262). Pergamon. https://doi.org/10.1016/B978-0-08-042580-1.50045-3

Hauser, M. D., Chomsky, N., Fitch, W. T. (2002). The faculty of language: What is it, who has it and how did it evolve? *Science*, *298*, 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

Hogg, R. V., Tanis, E. A. (1997). *Probability and statistical inference* (7th). Prentice Hall.

Jing, Y., Blasi, D. E., Bickel, B. (2021). Dependency length minimization and its limits: A possible role for a probabilistic version of the Final-Over-Final Condition. *Language*, in press.

Kahane, S., Mazziotta, N. (2015). Syntactic polygraphs. a formalism extending both constituency and dependency. *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, 152–164. https://doi.org/10.3115/v1/W15-2313

Kesavan, H. K. (2009). Jaynes' maximum entropy principle. In C. A. Floudas P. M. Pardalos (Eds.), *Encyclopedia of optimization* (pp. 1779–1782). Springer US. https://doi.org/10.1007/978-0-387-74759-0_312

Klein, D., Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 423–430. https://doi.org/10.3115/1075096.1075150

Köhler, R., Kelih, E., Goebl, H. (2021). Gabriel Altmann (1931–2020). *Journal of Quantitative Linguistics*, *28*(2), 187–193. https://doi.org/10.1080/09296174.2021.1902057

Kübler, S., McDonald, R., Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, *1*(1), 1–127.

Kuhlmann, M. (2010). *Dependency structures and lexicalized grammars. an algebraic approach* (Vol. 6270). Springer. https://doi.org/10.1007/978-3-642-14568-1

Kuhlmann, M., Jonsson, P. (2015). Parsing to noncrossing dependency graphs. *Transactions of the Association for Computational Linguistics*, *3*, 559–570. https://doi.org/10.1162/tacl_a_00158

Kuhlmann, M., Nivre, J. (2006). Mildly non-projective dependency structures. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 507–514.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*, 159–191.

Liu, H., Xu, C., Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

Miller, P. (2000). *Strong generative capacity: The semantics of linguistic formalism*. Cambridge University Press.

Morrill, G. (2010). *Categorial grammar: Logical syntax, semantics, and processing*. Oxford University Press.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, 149–160.

Nivre, J. (2004). Incrementality in deterministic dependency parsing. *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together (ACL)*, 50–57. https://aclanthology.org/W04-0308/

Nivre, J. (2005). *Dependency grammar and dependency parsing* (tech. rep. MSI 05133). Växjö University, School of Mathematics and Systems Engineering. http://stp.lingfil.uu.se/~nivre/docs/05133.pdf

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, 351–359. https://aclanthology.org/P09-1040

Nivre, J., Abrams, M., Agić, Ž., et al. (2019). Universal dependencies 2.4 [LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-2988

Osborne, T., Putnam, M., Gross, T. (2011). Bare phrase structure, label-less trees, and specifier-less syntax: Is minimalism becoming a dependency grammar? *The Linguistic Review*, *28*, 315–364. https://doi.org/10.1515/tlir.2011.009

Ouyang, J., Jiang, J. (2018). Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics*, *25*(4), 295–313. https://doi.org/10.1080/09296174.2017.1373991

Park, Y. A., Levy, R. (2009). Minimal-length linearizations for mildly context-sensitive dependency trees. *Proceedings of the 10th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) conference*, 335–343. https://aclanthology.org/N09-1038

Passarotti, M. C. (2016). How far is Stanford from Prague (and vice versa)? Comparing two dependency-based annotation schemes by network analysis. *L'analisi Linguistica e Letteraria*, *1*, 21–46.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. In M. H. Christiansen S. Kirby (Eds.), *Language evolution* (pp. 16–37). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199244843.003.0002

Pitler, E. (2014). A crossing-sensitive third-order factorization for dependency parsing. *Transactions of the Association for Computational Linguistics*, *2*, 41–54. https://doi.org/10.1162/tacl_a_00164

Pitler, E., Kannan, S., Marcus, M. (2013). Finding optimal 1-endpoint-crossing trees. *Transactions of the Association for Computational Linguistics*, *1*, 13–24. https://doi.org/10.1162/tacl_a_00206

Prüfer, H. (1918). Neuer Beweis eines Satzes über Permutationen. *Arch. Math. Phys*, *27*, 742–744.

Roark, B., Mitchell, M., Hosom, J., Hollingshead, K., Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2081–2090. https://doi.org/10.1109/TASL.2011.2112351

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2334–2341). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/915_Paper.pdf

Sgall, P. (1969). *A functional approach to syntax in generative description of language*. Elsevier.

Smith, N. A., Eisner, J. (2006). Annealing structural bias in multilingual weighted grammar induction. *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, 569–576. https://doi.org/10.3115/1220175.1220247

Temperley, D., Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/Cognitive universal? *Annual Review of Linguistics*, *4*(1), 67–80. https://doi.org/10.1146/annurev-linguistics-011817-045617

Tomasello, M. (2005). *Constructing a language. A usage-based theory of language acquisition*. Harvard University Press.

Yadav, H., Husain, S., Futrell, R. (2019). Are formal restrictions on crossing dependencies epiphenominal? *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 2–12. https://doi.org/10.18653/v1/W19-7802

Yadav, H., Husain, S., Futrell, R. (2021). Do dependency lengths explain constraints on crossing dependencies? *Linguistics Vanguard*, *7*(s3), 20190070. https://doi.org/10.1515/lingvan-2019-0070

Yadav, H., Husain, S., Futrell, R. (2022). Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity. *Computational Linguistics*, 1–27. https://doi.org/10.1162/coli_a_00437

Yli-Jyrä, A., Gómez-Rodríguez, C. (2017). Generic axiomatization of families of noncrossing graphs in dependency parsing. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1745–1755. https://doi.org/10.18653/v1/P17-1160

Yuan, J., Lin, Q., Lee, J. S. Y. (2021). Discourse tree structure and dependency distance in EFL writing. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, 105–115. https://aclanthology.org/2021.tlt-1.10

Book review
On Invisible Language in Modern English: A Corpus-based Approach to Ellipsis. By Evelyn Gandón-Chapela. London: Bloomsbury Academic. 2020.

Zheyuan Dai[1*]

[1] Linguistics Department, Zhejiang University
[*] Corresponding author's email: zydai@zju.edu.cn

DOI: https://doi.org/10.53482/2022_52_398

Instead of directly addressing "ellipsis" in the main title, Evelyn Gandón-Chapela describes this linguistic phenomenon as "invisible language" colloquially and vividly. However, readers may be misled by the title of the book "On Invisible Language in Modern English – A Corpus-based Approach to Ellipsis", which in fact only concentrates on the post-auxiliary ellipsis (PAE) rather than the overall ellipsis in Modern English. PAE covers ellipsis cases in which a verb phrase, prepositional phrase, noun phrase, adjective phrase, or adverbial phrase is omitted after one of the following licensors: modal auxiliaries, auxiliaries *be*, *have*, and *do*, and infinitival marker *to*. Concretely, the book discusses two main sub-types of PAE, namely verb-phrase ellipsis (VP-Ellipsis) and pseudo-gapping (PG). See examples (1) and (2) as follows for further illustration, respectively:

(1) I have [eaten an apple] this morning, but Mary *hasn't* ~~eaten an apple~~.

(2) Peter [kissed] Daisy, and Paul *did* ~~kiss~~ Nancy.

Example (1) of VP-Ellipsis shows the omission of the VP triggered by the licensor *have*. The elided VP antecedent (eaten an apple) is highlighted by square brackets. Though example (2) is close to the structure of example (1), in PG, there would be a complement left after the auxiliary, as the directly object Nancy after the licensor *did*.

In all, the book stands out among other ellipsis-related volumes for its qualitative discussions accompanied by quantitative analysis. With this work, Dr. Gandón-Chapela conducted the first sustained diachronic corpus investigation towards PAE. Aiming to provide an empirical account for PAE, this book not only reports the descriptive overview of PAE in the Penn Parsed Corpus of Modern English (PPCME) (1700-1914) but also compares its results with former corpus studies. What is noteworthy is that the book presents a new series of algorithms for automatically detecting and retrieving PAE from the Penn Parsed Corpus of Modern English. Retrieving a considerable amount of PAE accurately has always been a premise for quantitative research on the ellipsis phenomenon. Nonetheless, it is not easy

due to their highly-liberalized structures. The methodology Dr. Gandón-Chapela updated was stated in a detailed and specific style, making it quite applicable to a wide range of parsed corpora, contributing to the more efficient realization of this premise.

The book is very much written in the traditional academic style. With four main chapters, this book presents literature review in the first Chapter, methodology in the second, and then they are followed by descriptive results and the analyses. In the end, it goes with concise and clear conclusions. One of the most striking advantages is that the illustrative instances are widely scattered around the book for clarifying some concepts and the author's opinions. For this reason, this book is well suited for those who have relatively weak theoretical backgrounds as well as linguists with an intense interest in English ellipsis, corpus linguistics, and language evolution.

The beginning of the introduction sets the scene for the whole book by showing the necessity for the research with vivid explanations of illustrative sentences. The methodology used in the book is briefly mentioned, along with the employed corpus and the variables to be analyzed. Ellipsis is a unique linguistics phenomenon at the semantics-syntax-pragmatics interfaces (Merchant 2010). The mismatches between the invisible information (the elided structure with intended meaning) and the visible elements (what is actually pronounced) can cause an ambiguity. As the consideration of ellipsis is a complex linguistic phenomenon, in Section 1.2, along with the three most influential theoretical accounts, that is, Comprehensive Grammar of English, Systematic Functional Grammar, and Transformational Generative Grammar (TGG), the book helps the readers to have a broad understanding of ellipsis. Take the TGG part as an example. TGG mainly focuses on the formal characteristics of ellipsis with the endeavor of answering three relevant questions about the structure, the identity, and the licensing. The book follows Bîlbîie's (2011: 129) clear-cut criteria for the identification of elliptical structure. However, the criteria proposed by Bîlbîie concentrate mainly on the syntactic and semantic sides. Discourse-related influences such as context and thematic structure matter as well. To supplement the theoretical analyses, the empirical findings of ellipsis from the psycholinguistic perspective were also provided.

As stated in Liu (2018), language can be considered as a human-driven complex adaptive system. The syntactic structure of PAE is highly-liberalized owing to both grammatical and interactional reasons. To detect and retrieve such a linguistic phenomenon as accurately and completely as possible in a corpus, the query should be designed with careful consideration for all possible PAE scenarios. In light of that, Chapter 2 further primes the reader with the literature review on studying English PAE with corpus-based approaches. Section 2.1 introduces the research target, namely, the data source, PPCME (1700-1914). For two main branches of English PAE, VP-Ellipsis and PG, a number of reviews with empirical methods are revisited. For academic disputes over some structural details, the author posits herself clearly. For instance, the book follows Sag's (1976: 53) correction on the misnomer of so-called "VP-Ellipsis", and reclassifies it into a subtype of PAE. VP-Ellipsis together with PG, these two main subtypes of PAE, is dissected in terms of syntactic characteristics. Because of that, all types of PAE cases

are listed in Section 2.2.3.2. The difficulty with retrieving PAE structures is that, as the author states, the omission words are *invisible* in a sentence. Fortunately, the problem can be solved through retrieving algorithms designed in concise and basic formal logical running with the Corpus Search 2, a java-based software assorted with PPCMBE. In the annotation scheme of the PPCMBE, the asterisk * thus could correspondingly represent those invisible elements. This symbol is a wild card for any combinations of ellipsis resources and targets. With the assistance of the symbol, the invisible part of the sentence can be visualized. Take the query ((MD* iPrecedes HV*) AND (HV* iPrecedes [.,])) as an instance, the query returns to examples where any modal auxiliary immediately precedes the auxiliary *have*, and in turn, immediately precedes any punctuation mark. For the technical operability, corresponding algorithms for detecting PAE in the corpus are outlined one by one. From my perspective, one of the most outstanding merits of the monograph is that with as many PAE structural scenarios considered, the recall rate has been raised from 0.89 to 0.97. The increase in recall rate surely would optimize the automatic corpus retrieving work, guaranteeing the quantity as well as the quality of the database for further analysis. Appendix 1 shows the basic query language of Corpus Search 2, making the retrieving process accessible for linguists as well as beginners.

In Chapter 3, based on the retrieved corpus, 32 variables are divided into three main groups for different research purposes, "Core defining variables", "Usage variables" and "Processing variables". Each group is accompanied by illustrative sentences for definition interpretation, statistical analysis, comparison with results of Present-Day English, and a succinct summary. What is noteworthy is that some variables, such as type of anaphora, sloppy identity, remnants, etc., have been studied empirically in Present-Day English (e.g., Bos and Spenader 2011; Miller 2014), which paves the way for the analysis of the diachronic evolution of PAE.

Defining variables focuses on grammatical and discursive aspects. The present monograph focuses on the licensors of PAE in PPCME. The licensor, a grammatical element that triggers the appearance of ellipsis, is fairly useful in helping language learners or researchers quickly detect the ellipsis. It discovers that modal auxiliaries, such as can/could, will/would, may/might, etc., are the most frequent licensors of PAE in Late Modern English. This conclusion remains valid after comparing with that of Present-Day English (Bos and Spenader 2011). Moreover, through the diachronic exploration, the book discovered that some licensors employed in Late Modern English such as *shouldest*, *shalt*, *durst*, *dost*, and *ought*, are no longer used in the Present-Day English. As for the discursive perspective, for example, the frequencies of four clausal types under the framework of discourse conditions were calculated for comparing the types of clause of the source of ellipsis versus the target ellipsis. Usage variables concentrate on the dynamic description of PAE, especially for their diachronic evolutions and genre distributions. To explore the possible diachronic variations of PAE, examples were sorted arbitrarily into 5 groups with every 50 years as nodes for classification. As for the genre part, 18 genres were equally divided into speech-related and writing-related genres (Fiction is treated as a mixed type). Processing

variables, as the name implies, process the connection between the resource and the target of ellipsis, primarily concerning the co-textual aspects of PAE. Two types of processing distances were estimated. The lexical distance measures the words between the resource and the target of ellipsis, while the syntactic distance is measured in the number of IPs. Take the syntactic distance for illustration, in the vast majority (around 76.88 %) of the VP-Ellipsis examples, there are no clauses intervening between the source and target of ellipsis. The concept of lexical distance is close to that of dependency distance between two words in a sentence (Heringer et al. 1980; Hudson 1995; Liu 2008) in Dependency Grammar. Dependency distance is a measurement of syntactic complexity as well as human cognitive load (Hudson 1995; Liu 2008). Coincidentally, according to the Dependency Locality Theory proposed by Gibson (1998 & 2000), the longer the syntactic dependency is, the higher the possibility for a sentence to bear larger syntactic complexity. Therefore, the calculation of the lexical distance variable may be helpful empirical assistance to further analysis for measuring the syntactic complexity of PAE. Moreover, Popescu et al. (2014) proposed that the length distribution of all types of language units would conform to the Zipf-Alekseev distribution. Following their steps, Jiang & Liu (2015) discovered that the DD distribution of natural human languages also conforms to the right-truncated Zipf-Alekseev distribution. With the empirical contribution of the present book, the regularity of the PAE phenomenon can also be captured based on the processing variables.

Chapter 4, the conclusion, summarizes the main research results with a brief presentation of the research targets, methods and steps once again. With all the quantitative descriptions of PAE, the book verifies many hypotheses and theoretical claims raised by former research. For example, VP-Ellipsis can be licensed by more than one auxiliary, whereas, as a general rule, PG cannot. Also, the research has established the syntactic linking between the antecedent and the ellipsis site in PAE. All the discoveries may assist both the researchers and language learners to get acquainted with PAE more.

There are, however, a few points that could be possibly improved, though the book is generally well-organized and informative. First, this book quantitatively describes the PAE of Late Modern English but less consideration has been given to the linguistic interpretation behind these statistics. Second, for the diachronic analysis of the PAE, the monograph mainly takes Bos and Spenader (2011)'s work on PAE of Present-Day English for reference, which is based on the data collected from the Wall Street Journal sections of the Penn Treebank. As mentioned by Dr. Gandón-Chapela herself, Bos and Spenader's data may be biased due to their genre or theme monotonicity in the database. In all, this book is an excellent contribution to the corpus-based study on English PAE and also a valuable empirical addition to theoretical research. It provides a complete set of quantitative research methodology that can be applied to traditional research on a language phenomenon for reference, namely, the construction of theoretical background, the selection of appropriate corpus, the retrieval of the target structures, and the analysis of quantitative results. Besides, readers may also find that it can be referred to as an encyclopedia for PAE in Late Modern English. As a pioneering study of English PAE with corpus-

based approaches, this book has laid a solid foundation for the follow-up diachronic and synchronic research. Moreover, with the process of detecting and retrieving highly-liberalized structures in a corpus, the book has demonstrated good operability of employing corpus for its readers, from amateur to professional. The valuable experience has great reference significance for the quantitative research of specific linguistic structures.

Acknowledgements

This review is supported by the National Social Science Fund of China for Distinguished Young Scholars (20CYY030).

References

Bîlbîie, G. (2011). *Grammaire des Constructions Elliptiques. Une étude Comparativedes Phrases sans Verbe en Roumain et en Français*. Université Paris Diderot-Paris 7, Paris. (PhD thesis)

Bos, J., Spenader, J. (2011). An Annotated Corpus for the Analysis of VP Ellipsis. *Language Resources and Evaluation*, 45(4), pp. 463-94.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, pp. 1-76.

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. In: Marantz, A., Miyashita, Y., O'Neil, W. (Eds.). *Image, language, brain*. Cambridge MA: MIT Press, pp. 95-126.

Heringer, H. J., Strecker, B., Wimmer, R. (1980). *Syntax: Fragen-Lösungen- Alternativen*. München: Wilhelm Fink Verlag.

Hudson, R. A. (1995). *Measuring Syntactic Difficulty*. Manuscript. London: University College London.

Jiang, J., Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. *Language Sciences* 50, pp. 93-104.

Liu, H. (2008). *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.

Merchant, J. (2001). *The Syntax of Silence: Sluicing, Islands, and the Theory of Ellipsis*. Oxford: Oxford University Press.

Miller, P. (2014). A Corpus Study of Pseudogapping and its Theoretical Consequences. In: Piñón, C. (ed.). *Empirical Issues in Syntax and Semantics 10*, pp. 73–90. Available online: http://www.cssp.cnrs.fr/eiss10/ (accessed 29 November 2018).

Popescu, I.-I., Best, K. H., Altmann, G. (2014). *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Sag, I. A. (1976). *Deletion and Logical Form.* MIT: Cambridge, MA. (PhD thesis)