Glottometrics 45
2019

RAM-Verlag

ISSN 1617-8351
e-ISSN 2625-8226

Glottometrics

Indexed in ESCI by Clarivate Analytics and SCOPUS by Elsevier

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druck-version** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de
Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 45 (2019), Lüdenscheid: RAM-Verlag, 2019. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 45 (2019)
online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)

Contents

Glottometrics 45, 2019, 1-6

Segmental and Suprasegmental Vowel Frequencies in Slovene: Statistical Modeling

Emmerich Kelih[1]

Abstract. We show that in Slovene, length, accent, and shortening of vowels represent factors influencing the frequency of vowels both in the dictionary and in texts. The results of the operation of these forces are presented by means of continuous models which are fitted to the resulting numbers.

Key words: *Slovene, vowel frequencies, accented vowels, unaccented vowels, supra segmental features*

1. Introduction

This article is devoted to the modeling of the frequency of segmental and suprasegmental properties of the Slovene vowel system. This system consists of accented and unaccented vowels, and the accented vowels can be either long or short. It will be shown that the five basic vowels can be considered as scaling property for a quantification. The empirical data are gained from a Slovene–German learner's dictionary. This allows us to differentiate the level of the dictionary (analysis of lemmas) as well as the text level when we consider the sentences which exemplify the lemmas in a natural syntactical context. First, we present the data basis, and then we propose some models (linear, exponential) and a parabola which are able to fit the retrieved frequency data of accented and unaccented Slovene vowels.

2. The data basis

For this analysis we used the Slovene–German learner's dictionary (cf. Kelih, Vučajnk 2018), although only the Slovene-to-German section is relevant for the analysis. The dictionary consists of 4,950 lemmas and 5,095 accompanying sentences where the mentioned lemmas are used in a prototypical context. For example, the lemma *imenovati se* ('to be called', 'to be named') looks as follows:

lemma	imenováti se -újem se *impf*
sentence	Kakó se imenúje vàš sodélavec?
German translation	Wie heißt Ihr Mitarbeiter?

[1] E. Kelih, Univ. Vienna (Austria), emmerich.kelih@univie@.at

The verbal form is expressed in the lexicography of a synthetic non-agglutinative language in its infinitive form with the specification of the aspect (either imperfective or perfective) and the suffix of the first person singular. The example sentence contains as already pointed out a quite typical context, which helps the learner to acquire a particular lemma more easily. The specificity of the given Slovene–German learner's dictionary is that both the lemmas and the example sentences are accented. Since in Slovene the stress position is neither fixed nor marked in the standard orthography the accent annotation gives the learner important information about the quantity, position, and quality of the accented vowels. For linguists interested in quantitative features of the Slovene vowel systems, the dictionary provides the possibility for counting the frequency of both segmental and suprasegmental properties (the distribution of long, short, and unaccented vowels) and an ongoing statistical evaluation. For the sake of simplicity, we distinguish in the following the frequency data on two different languages levels – on the lemma level (as it is given in the dictionary) and on the text level (data gained from the example sentences).

For our statistical analyses the basic specifications which are necessary for the identification and operationalization of vowel frequencies in Slovene can be found in Greenberg (2008), Herrity (2010), and Priestly (1993) and more information about Slovene suprasegmental properties in the context of Slavic languages can be found in Sussex/ Cubberley (2006: 177ff.) and Šuštaršič et al. (1995). The main features of the Slovene vowel system are:

1. The basic vowel system of Slovene consists of the five vowels /i, e, a, o, u/.

2. An important property is, however, the fact that in Slovene the accent and the vowel length and vowel shortening are inherently connected. There are five long accented vowels /í, é, á, ó, ú/, two long open vowels /ô, ê/, and five short accented vowels /ì, è, à, ò, ù/. This is the complete inventory studied here; the accented syllabic /ŕ/ is not considered in this analysis. The most outstanding property of the Slovene system are the two open vowels /e, o/, which are always marked with length and accentuation. This makes the Slovene vowel system unsymmetrical.

With regard to further operationalization it should be remarked that any annotation of the lemmas (e.g. information about parts of speech *f* (feminine), *m* (masculine), *n* (neuter), *impf* (imperfective aspect), *pf* (perfective aspect), *adj* (adjective) etc.) is excluded from the statistical counts; hence, only the "pure" Slovene material as appearing in lemmas or example sentences has been analyzed. Within inflected parts of speech (verbs, nouns, adjectives) not only the lemma, but also for example the genitive singular of nouns, the first person singular of verbs, the feminine and neuter suffixes of adjectives etc. are taken into account too. That means for example *imenováti se -újem se* ('to be named'), *Japónec –nca* ('Japanese'), *lep -a -o* ('pretty') etc. The example sentences are considered as a whole for the counts. In some rare cases, one finds two sentences with typical contexts for the given lemma. In the next section we offer a short description and discussion of the counts achieved which were obtained automatically.

2.1. Frequencies of accented and unaccented vowels

In a first step the determined vowel frequencies are presented. In Tables 1 and 2 one can find the absolute frequencies of accented and unaccented vowels of Slovene[2], based on the used

[2] As far as we know there are not many statistical analyses of the Slovene suprasegmental features. One exception where the frequency of accented and unaccented vowels of Slovene can be found is the retrograde dictionary by Hajnšek-Holz/Jakopin (1996). Further data concerning Slovene grapheme frequencies are given in Grzybek/Kelih/Stadlober (2006), but no information about accented and unaccented vowels is given there.

learner's dictionary. A (first) descriptive fact is that in all cases the unaccented vowels have the highest frequency and that the number of long accented vowels is in all cases smaller (more than half in the case of /i/ and an even greater difference for /e/ and /o/. The next interesting observation is that short, unaccented vowels do not play any role (cf. Tables 1 and 2 containing raw data) if one takes into account the quantitative rareness of these vowels. This phenomenon can be interpreted in a synergetic sense (cf. Köhler 2005). The length seems to be a constitutive feature of the accent. Seen from a synergetic background, length is obviously required for an appropriate decoding, whereas shortness of accented vowel seems to bear some unexpected infectivity during the articulation and in the decoding. The infrequent appearance of this kind of vowel is accompanied by the fact that unaccented short vowels occur in chosen positions and forms only, i.e. they are distributionally very restricted[3]. See Toporišič (2000: 60–63) for the (rare) cases in which short, accented vowels can occur – mostly in some monosyllabic nouns (especially masculine forms) and in some selected affixes.

Table 1
Frequencies of accented and unaccented vowels: Lemma

Absolute frequencies	i	e	a	o	u
unaccented	3,054	4,390	5,686	2,775	354
long, accented	1,613	999	1,704	685	428
long, open, accented		309		344	
short, accented	11	161	88	103	3
Sum	4,678	5,859	7,478	3,907	785

Table 2
Frequencies of accented and unaccented vowels: Text

Absolute frequencies	i	e	a	o	u
unaccented	8,542	10,924	9,416	10,305	1,767
long, accented	4,172	3,662	5,355	2,490	1,170
long, open, accented		749		1,264	
short, accented	84	555	570	239	21
Sum	12,798	15,890	15,341	14,298	2,958

This evident preference – short accented vowels do not play any relevant "systemic" role – can also be found in the frequencies of accented and unaccented vowels on the text level (= in the example sentences). Here, again, clearly the unaccented vowels dominate quantitatively

[3] It has to be emphasized that the opposition of long and short accented vowels holds true especially for the Slovene standard language. In the Slovene dialects, the situation is different and taking into consideration the experimental results (cf. for example Tivadar 2004, Srebot-Rejec 1988) it can be shown that obviously in the synchronic context the differentiation between long and short accented vowels seems to be obsolete. For newer references concerning the progressive loss of tones, and further outstanding characteristics of Slovene dialects, cf. Jurgec (2007). In other words, the Slovene phonological system, in particular on the suprasegmental level, is subject to ongoing changes.

(cf. Table 2) and the distance to the next category of vowels (long accented vowels) is very great. For both forms of counting (dictionary and text) it holds true that the long, open, accented vowels /ô, ê/ which in comparison to other Slavic languages represent an outstanding phonemic feature of Slovene are, as a matter of fact, characterized by a very low frequency.

In the next section, some statistical models are offered for the obtained descriptive features of Slovene.

2.2. Modeling of accented and unaccented vowels

As already mentioned in the previous chapter, there are five basic vowels in Slovene: /i, e, a, o, u/. For the modeling procedure, we start from the basic variables given as unaccented vowels, which can be modified by lengthening, shortening, openness, and accentuation. If one computes the frequency of the modified vowels, one obtains the results as given in Table 1 for the dictionary and as given in Table 2 for the texts.

Looking at the data one can see a clear gradation of frequencies. In all cases, except for /u/, the succession of frequencies is *unaccented, long accented, long open accented, short accented*. But one has to mention that of course not all vowels have the *long-open-accented* variant, but only /e/ and /o/.

Our investigation concerns the form of frequencies of these classes. Is there a common regularity followed by the frequencies? One would automatically suppose that the properties accent, length, openness, and place of articulation can be scaled, a fact that of course cannot hold for all languages of the world, but in the case of Slovene it holds true and thus for the given moment the hypotheses can only be tested for Slovene.

The sums ordered according to the place of articulation abide by a concave sequence which can be captured by a parabola, namely $y = c + a^*(x - b)^2$. The differential equation has the form of a straight line, corresponding to the theory of Wimmer and Altmann (2005). The fitting to the data where the positions are given simply by numbers yields the results presented in Table 3. Since the determination coefficient is in both cases very high one can accept the hypothesis at least preliminarily.

We obtain another function if we order the vowels according to the degree of openness (/i, e, a, o, u).

Table 3

Parabolic distribution of articulation places in Slovene

Vowel	Place	Lemma	Parabola	Text	Parabola
i	1	4,678	4,518.14	12,798	12,317.40
e	2	5,859	6,500.63	15,890	16,481.20
a	3	7,478	6,512.26	15,341	16,451.00
o	4	3,907	4,553.03	14,298	12,226.80
u	5	785	622.94	2,958	3,808.60
		a = -985.4286 b = 2.5059 c = 6,752.8343 R^2 = 0.9271		a = -2,097.0000 b = 2.4928 c = 16,990.4587 R^2 = 0.9399	

The results gained are based on the scaling of the data, which is motivated by the place of articulation. But even the individual vowels show a common course if one considers this sequence of properties: 1. unaccented, 2. long accented, 3. long-open accented, 4. short accented. We conjecture that in "normal" cases, this course is exponential (y = a*exp(-b*c)) but one can also find a straight line (SL) and the parabola (Par) as defined above. The results of the tests are presented for lemmas in Table 4 and for the text level in Table 5.

Table 4
Frequencies of vowels according to suprasegmental properties: Lemmas

Degree	e		o		i		a		u	
	Fr	F(E)	Fr	F(E)	Fr	F(SL)	Fr	F(E)	Fr	F(Par)
1	4390	4383.84	2775	2763.04	3054	3080.83	5686	5709.20	354	354
2	999	1044.90	685	767.75	1613	1559.33	1704	1531.00	428	428
3	309	249.05	344	213.33	11	37.83	88	410.56	3	3
4	161	59.3627	103	59.28						
	a = 18392.2726 b = 1.4340 R^2 = 0.9986		a = 9943.8739 b = 1.2806 R^2 = 0.9942		a = 4602.3333 b = -1521.50 R^2 = 0.9991		a = 21289.93 b = 1.3162 R^2 = 0.9919		a = -249.50 b = 1.6183 c = 458.8620 R^2 = 1.0000	

E = Exponential, SL = Straight line, Par = Parabola

Table 5
Frequencies of vowels according to suprasegmental properties: Texts

Degr.	e		o		i		a		u	
	Fr	F(E)	Fr	F(E)	Fr	F(SL)	Fr	F(SL)	Fr	F(SL)
1	10924	10948.04	10305	10266.03	8542	8495.0	9416	9536.67	1767	1859.00
2	3662	2490.85	2490	2775.62	4172	4266.0	5355	5113.67	1170	986.00
3	749	1113.08	1264	750.44	84	37.00	570	690.67	21	113.00
4	555	354.91	239	202.90						
	a = 34335.3683 b = 1.1430 R^2 = 0.9971		a = 37970.4286 b = 1.3080 R^2 = 0.9945		a = 12724.00 b = -4229.0 R^2 = 0.9993		a = 13959.67 b = -4423.0 R^2 = 0.9978		a = 2732.00 b = .873.0 R^2 = 0.9678	

In all cases one obtains satisfactory results. This shows that also within a vowel system there is a certain regularity concerning the distribution/frequency of accented and unaccented vowels.

3. Conclusions

As can be seen, each additional property (accentuation, length, shortening) applied to vowels leads to an effort with the speaker who tries to reduce it by diminishing the frequency of the vowel in the new form. That means that in this domain the Zipfian forces play a central role and in future Köhlerian (2005) synergetics could also be applied in a particular, but highly complex subsystem of the language system.

The modeling of these changes is simple. We strive for applying simple functions which can be derived from a common theory (Wimmer/Altmann 2005). When modeling, it is not relevant whether one uses a simple function or a distribution (= normalized function). One applies either discrete or continuous functions because, as we know, all models merely formally represent our concepts and can be easily formally processed by these two approaches. This is rarely possible with qualitative concepts.

REFERENCES

Greenberg, Marc L. (2008). *A short reference grammar of Slovene*. München: Lincom Europa (Lincom studies in Slavic linguistics, 30).

Grzybek, Peter; Kelih, Emmerich; Stadlober, Ernst (2006). Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen). Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik. *Anzeiger für Slavische Philologie* 34, 41–74.

Hajnšek-Holz, Milena; Jakopin, Primož (1996): *Odzadnji Slovar Slovenskega Jezika po Slovarju Slovenskega Knjižnega Jezika*. Ljubljana: ZRC, SAZU.

Herrity, Peter (2010). *Slovene. A comprehensive grammar.* London: Routledge.

Jurgec, Peter (2007). Acoustic Analysis of Tones in Contemporary Standard Slovene: Preliminary Findings. *Slovenski Jezik / Slovene Linguistic Studies* 6, 195–207.

Kelih, Emmerich; Vučajnk, Tatjana (2018). *Slovensko-nemški tematski slovar: osnovno in razširjeno besedišče. 4500 gesel, frazemov in stavčnih primerov. Grund- und Aufbauwortschatz Slowenisch-Deutsch. 4500 Lemmata, Phrasen und Satzbeispiele.* Klagenfurt/ Celovec-Ljubljana/Laibach-Wien/Dunaj: Hermagoras/Mohorjeva.

Köhler, Reinhard (2005). Synergetic Linguistics. In: Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 760-774.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Priestly, Tom (1993). Slovene. In: Bernard Comrie and Greville G. Corbett (eds.): *The Slavonic languages.* London/New York: Routledge. London, New York: Routledge, 388–451.

Srebot-Rejec, Tatjana (1988). *Word accent and vowel duration in standard Slovene. An acoustic and linguistics investigation.* München: Sagner (Slavistische Beiträge, 226).

Sussex, Roland; Cubberley, Paul V. (2006): *The Slavic languages.* Cambridge: Cambridge University Press.

Šuštaršič, Rastislav; Komar, Smiljana; Petek, Bojan (1995): Slovene: Illustrations of the IPA. *Journal of the International Phonetic Association* 25(2), 86–90.

Tivadar, Hotimir (2004). Fonetično-fonološke lastnosti samoglasnikov v sodobnem književnem jeziku. *Slavistična revija 52 (1),* 31–48.

Toporišič, Jože (2000). *Slovenska slovnica*. Maribor: Obzorja.

Wimmer, Gejza; Altmann, Gabriel (2005): Unified derivation of some linguistic laws. In: Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), 791–801.

Glottometrics 45, 2019, 7-23

Context Specificity of Lemma. Diachronic Analysis

Jan Hůla[1]
Miroslav Kubát[2]
Radek Čech[3]
Xinying Chen[4]
David Číž[5]
Kateřina Pelegrinová[6]
Jiří Milička[7]

Abstract. This study deals with the recently proposed concept of so-called Context Specificity of Lemma (*CSL*). *CSL* is based on the word embedding technique called Word2vec which enables measuring lexical context similarity between lemmas. Specifically, a recently proposed method Closest Context Specificity (*CCS*) is applied to a diachronic analysis of Czech texts. This method expresses how unique is a context within which a given lemma appears. The aim of the paper is to study what kind of semantic features can *CCS* detect and how useful could *CCS* be in a diachronic semantic analysis. The second goal is to observe the relation of *CCS* to frequencies in the corpora.

Keywords. *Word2vec, semantics, diachronic analysis, context specificity.*

1. Introduction

Generally speaking, the semantics of any linguistic unit is a very complex issue which is difficult to study in a quantitative way. Considering the number and the variation of the factors playing a role (especially pragmatic ones), it seems to be nearly impossible to express the meaning of a linguistic unit (in our case a lemma) using quantitative methods. However, very innovative methods based on neural networks approach have recently shown promising results. Namely, Word2vec technique enables measuring semantic similarities between words, where the meaning of a word is given by its context (Mikolov 2013a, 2013b). Čech et al. 2018 proposed a concept of so-called Context Specificity of a Lemma (*CLS*) which measures how unique is the context of a given lemma.

[1] Jan Hula, University of Ostrava, jan.hula21@gmail.com
[2] Miroslav Kubát, University of Ostrava: miroslav.kubat@gmail.com, https://orcid.org/0000-0002-3398-3125, corresponding author, University of Ostrava, Reální 5, Ostrava 701 03, Czech Republic
[3] Radek Čech, University of Ostrava: cechradek@gmail.com
[4] Xinying Chen, University of Ostrava, Xi'an Jiaotong University, cici13306@gmail.com, https://orcid.org/0000-0002-5052-4991
[5] David Číž, University of Ostrava, davidciz95@gmail.com
[6] Kateřina Pelegrinová, University of Ostrava, pelegrinovak@gmail.com
[7] Jiří Milička, Charles University, milicka@centrum.cz, http://orcid.org/0000-0001-8605-1199

A lemma has high context specificity when there are not many other lemmas which appear within a similar context. For instance, function words (synsemantics) like conjunctions or prepositions should have lower context specificity than content words (autosemantics). There is a limited number of function words and they have very low or no lexical meaning. Their role is to express some grammatical function. Therefore, function words should not be very tied to any context in general. Another example could be the difference between highly frequent lemmas with common usage such as *car, house, grass, money* on the one hand; and technical terms such as *atom*, *phoneme*, *molecule*, etc. on the other hand. The technical terms should have a much more specific context in general because their usage is very limited to the specific topics and style. Closest Specificity of Lemma (*CCS*) can detect the context of target lemmas and express the uniqueness of the context. This approach showed very promising preliminary results from synchronic (Kubát et al. 2018) and diachronic (Čech et al. 2018) points of view. This study follows up the recently proposed approach by the application of *CCS* to a diachronic analysis.

Context specificity can be considered as a semantic feature of lemmas which can be measured in a quantitative way and at the same time allows linguistic interpretation. This study is focused on the semantic changes of selected lemmas in Czech journalism during more than 20 years. The main goal of the paper is to discover whether *CCS* is a suitable tool for diachronic semantic analyses of lemmas and test the preliminary conclusion made by authors of this approach (Čech et al. 2018). The lemmas are selected in a qualitative way, i.e. we choose those lemmas where we intuitively expect potential changes in meaning during the analyzed time period. The following step is the linguistic interpretation of obtained data. We, therefore, cannot observe many lemmas, this study is rather focused of deeper insight into the behavior of *CCS* in individually selected cases because we want to understand what kind of semantic feature(s) (if any) the concept of measuring Content specificity can detect.

As the source of data, we use the Czech National Corpus. Specifically, we use one of the largest Czech corpora SYN_V4. This corpus consists of more than 3 billion tokens and covers the Czech language from 1990 to 2014. We can, therefore, analyze more than 20 years of development of the Czech language from the beginning of a democratic state after the so-called Velvet revolution in 1989 when the communistic regime fell.

Since many indicators from quantitative linguistic analyses such as vocabulary richness are influenced by text length (cf. Kubát 2016), we also pay attention to this problem in this study. The relation of Closest Context Specificity (*CCS*) to the relative frequencies in the corpora is tested.

2. Methods

2.1 Word Embeddings

Word Embeddings represent a set of methods which are effective for finding useful representations of textual data which are usually collected in a form that is not suitable for a task at hand. These representations are produced by taking the original representation (with dimensionality equal to the number of distinct words within the corpus) as input and transforming it through series of numerical operations to different representations (usually with much lower dimensionality) which have certain desirable properties. The exact value of the output representation is dependent on the learnable parameters which are found by maximizeing a score function on a concrete task. For word embeddings, the task is usually language modeling where we try to predict the words within the corpus conditioning on the words in its neighborhood. We can use the obtained score to update the parameters of the model in a way which tries to increase the score. By iterating this process, we are trying to maximize the score and thus to find a better representation for the task. In our case, we want the representation of a word to be a good predictor of the contexts in which the word appears (this is measured by how well it can predict the words which appear next to it within the corpus). Thus, if two words often appear in the same context, their vector representations should be close to each other.

Such word embeddings are easy to obtain with algorithms such as Word2Vec or GloVe (Mikolov et al. 2013a; Manning et al. 2014). In our work, we are focusing on the Word2Vec algorithm, concretely the Skip-Gram version of it. The algorithm aims to represent a word (in our case the lemma) as a high-dimensional (50–1000) vector which captures co-occurrence statistics between the lemma itself and other lemmas in the small window centered at this lemma. The window acts as a context for the lemma in the center. Intuitively the vector representing the lemma should contain information about the contexts where it appears. Concrete values of these vectors are found by a process which tries to maximize an objective function which measures how well can be every lemma within the window predicted based on the lemma in the center of this window. This objective function has the following form:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-m \leq j \leq m, j \neq 0} \log p(w[t+j] | w[t])$$

This function is maximized when the individual summands (log probabilities) are maximized. The first sum (indexed by *t*) iterates over all tokens within the corpus (the number of tokens is *T*). The second sum (indexed by *j*) iterates over all tokens in the small window centered at the token with an index *t*. This window is of length 2*m*+1 (there are *m* lemmas on every side of the central lemma). Intuitively we want the lemmas inside this window (*w*[*t*+*j*]) to be predictable from the central lemma (*w*[*t*]). For example, when the lemma *w*[*t*] is “funny” and the lemma *w*[*t*+1] is “joke” and such co-occurrence is frequent within the corpus, we want *p*(joke|funny) to be high so that the lemma “joke” is predictable from lemma “funny”.

This kind of predictability is measured by a function with related vectors as arguments. Concretely, the conditional probabilities in the equation above are estimated by the following function:

$$p(o|c) = \frac{exp\left(u(o)^T \cdot v(c)\right)}{\sum_{w=1}^{W} exp\left(u(w)^T \cdot v(c)\right)}$$

where *u*(*o*) and *v*(*c*) are vector representations of lemmas *o* and *c* (*o* for the outer lemma, *c* for center lemma).

The first thing to notice is that every lemma is parametrized by a set of two vectors (*u* and *v*). One vector (*v*) is used when the lemma appears in the center of the window and the second vector (*u*) is used when the lemma appears as a context lemma. For example, when the window is centered at the lemma “funny”, then the vector *v*(“funny”) is used as its representation, but when the window is centered at some other lemma and the lemma “funny” appears in this window as a context word, then we use the vector *u*(“funny”) as its representation. These two vectors are used only to simplify the optimization problem. In the end, these representations could be averaged or one of them can be discarded. After the optimization, the lemmas which appear in similar contexts will have similar vectors assigned to them. Thus, even if the exact values of these vectors are not interpretable, their closeness could be interpreted. For measuring this kind of lexical context similarity between lemmas we use the cosine similarity as suggested by Levy et al. (2015). We first normalize all vectors to unit length and then the cosine similarity is equivalent to dot product between these normalized vectors. Therefore, when the vectors point in the same direction, their similarity is 1, when they point in opposite directions their similarity is -1, and when they are orthogonal then their similarity is 0. In other words, if the similarity is close to 1, then the contexts in which these lemmas appear are positively correlated, when it is close to -1, they are negatively correlated, and when it is close to 0, then they are uncorrelated. For the concrete details about this optimization procedure see Mikolov et al. (2013b).

2.2 Context Specificity of Lemma (*CSL*)

The concept of measuring the so-called Context Specificity of Lemma (*CSL*) was recently proposed by Čech et al. (2018). This method measures how unique is the context in which the lemma appears. This approach is based on the fact that we can compute the similarity of a given lemma to all other lemmas using Word2vec technique (Mikolov et al. 2013a). Each lemma is represented by a vector. Both the size and the orientation of the vector express the position of a lemma in a contextual multi-dimensional space. Statistics of these similarities (e.g. mean value) can be used for characterizing the *CSL*. The lower the mean of similarities, the higher the *CSL*.

There are several methods of measuring the context specificity (cf. Čech et al. 2018). The most promising preliminary results in discourse analysis were obtained by Closest Context Specificity (*CCS*). This measurement is based on the average value of the similarities *S* of the 20 closest (most similar) lemmas to the target lemma. The formulas for CCS calculation is as follows:

$$CCS = 1 - \frac{\sum_{i=1}^{20} S_i}{20}$$

where S = the similarity of the lemma.

It should be mentioned that we modified a bit the originally proposed formula by Čech et al. (2018) which is as follows:

$$CCS = \frac{\sum_{i=1}^{20} S_i}{20}$$

We just use a reverse value. The reason for this modification lies in the easier interpretation. Originally, the higher the *CCS*, the less specific the context of the target lemma. After the modification the higher the *CCS*, the more specific the context of the target lemma. We consider the original version quite misleading and therefore we modified it.

For instance, we can illustrate the *CCS* calculation procedure on a lemma “banka” (a bank) based on the data from the subcorpus restricted to the year 2014. First, we need a list of the 20 closest lemmas to the target lemma “banka” (a bank) with the values of similarities S_i. The S_i values express how much similar is the context of a given lemma to the target lemma (see Table 1). Second, we apply the aforementioned formula and gain the resulting value *CCS* = 0.37 (i.e. 1 - the arithmetic mean of the *S* values).

Table 1
20 closest lemmas to the target lemma “banka” (a bank) in the subcorpus 2014

#	lemma	S
1	bankovní (bank - adjective)	0.742
2	LBBW	0.674
3	spořitelna (bank)	0.661
4	Citibank	0.660
5	Equa	0.658
6	Raiffeisenbank	0.654
7	úvěrování (crediting)	0.634
8	kreditní (credit - adjective)	0.631
9	bankéř (banker)	0.628
10	mezibankovní (interbank - adjective)	0.627
11	debetní (debit - adjective)	0.625
12	Hypoteční (mortgage - adjective)	0.625
13	Sberbank	0.622
14	bankovnictví (banking)	0.618
15	Citigroup	0.614
16	Kontokorent (overdraft)	0.613
17	mBank	0.613
18	Barclays	0.613
19	splácený (paid)	0.612
20	úročení (interest)	0.612
CCS		**0.363**

3. Data

Methods based on neural networks require large training data for producing reliable results. Since we analyze the Czech language, we decided to use the Czech National Corpus which is a suitable source for this kind of research. Namely, we work with the corpus SYN_V4. “SYN” refers to “synchronic” and every version consists of texts from all reference synchronic written corpora of the SYN series published up until the given version of the SYN corpus (Hnátková et al. 2014). This corpus is not balanced from the point of view of genres or styles. The majority of texts belong to journalism, and smaller parts consist of fiction and non-fiction texts. The structure of the corpus can be seen in Figure 1.

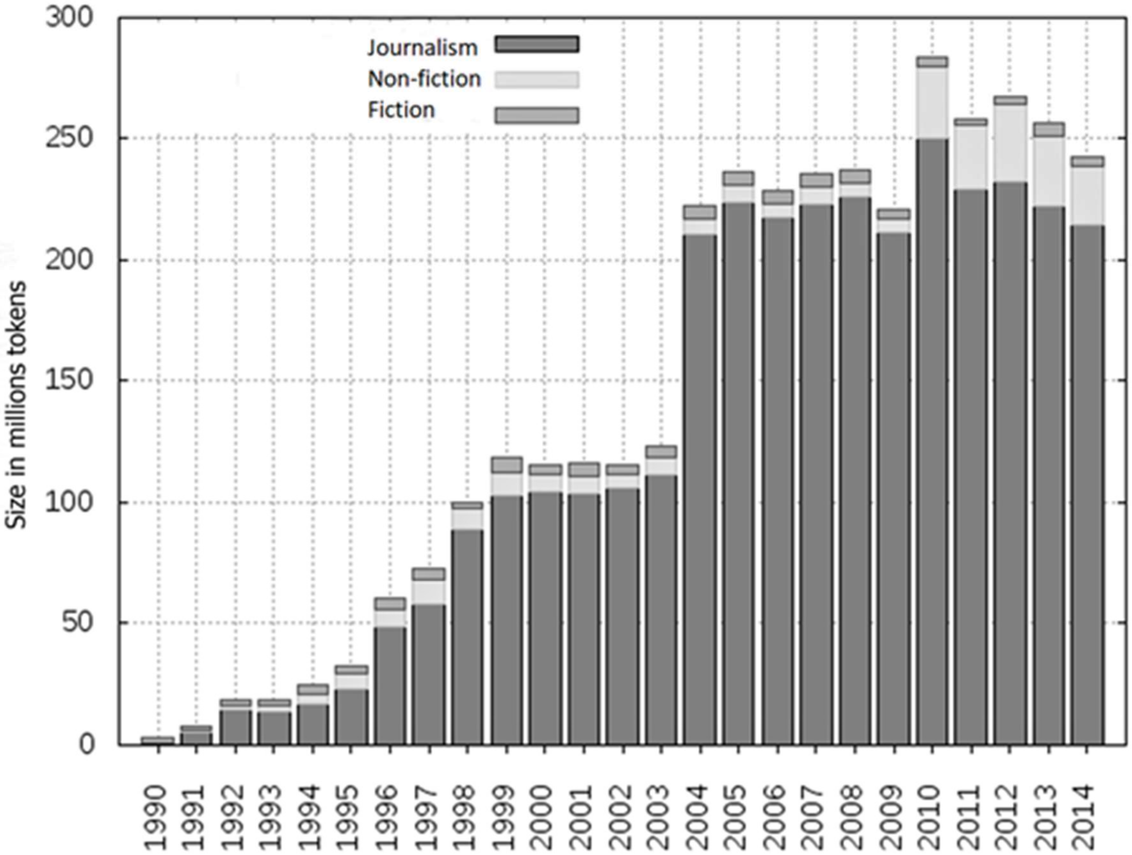


Figure 1 The composition of the corpus SYN_V4

Considering the composition of SYN_V4, we decided to use only journalistic texts due to potentially biased results. The final corpus of our study consists of more than 3 billion tokens (3,045,389,630) and more than one hundred thousand types (102,707). Since the goal is to analyze diachronic development of the *CCS*, we divide the data into 19 subcorpora where each represents one year (see Table 2). Only the subcorpus 1990-1996 consists of texts from several years because of the small data sizes (cf. Figure 1).

Table 2
The number of lemmas in each year. Years 1990-1996 are merged because of an insufficient amount of data

Year	Number of lemmas
1990-1996	37292
1997	44023
1998	40954
1999	45038
2000	45490
2001	44930
2002	44624
2003	45757
2004	64119
2005	65008

2006	64110
2007	65698
2008	66113
2009	63695
2010	69212
2011	66167
2012	66783
2013	65381
2014	64186

Czech is a highly inflected language where different endings express different grammatical categories such as case, number or gender in declension (nouns, adjectives, pronouns, numerals), and person, number or tense in conjugation (verbs). For example, the lemma *kočka* (a cat) has eleven different word forms for indicating its grammatical categories: *kočka, kočky, koček, kočce, kočkám, kočku, kočko, kočce, kočkách, kočkou, kočkami*. Since we focus on the semantic features of lexical units, lemmas are considered as the basic units in this research.

4. Diachronic Analysis

The goal of this analysis is to apply the recently proposed method called Closest Context Specificity (*CCS*) in diachronic semantic analysis. We select several lemmas from various fields where we expect some semantic changes. This study thus combines qualitative and quantitative approach. First, the lemmas are chosen qualitatively. Second, the lemmas are analyzed quantitatively. Third, the obtained results are qualitatively interpreted. We can then see what kind of semantic feature(s) (if any) could be detected by Context Specificity. It should be emphasized that this work does not have the ambition to make a final conclusion about the concept of Context Specificity of Lemma. However, we can do the first step to better understand this recently proposed method by a deeper look into several qualitatively chosen lemmas.

4.1 Political parties

The first analyzed group of lemmas is devoted to the Czech political parties. We chose traditional parties which continually existed from 1990 to 2014, namely: ODS, ČSSD, KDU-ČSL, KSČM. ODS is a right-wing conservative party. ČSSD is a left-wing labour party. KDU-ČSL is a Christian-democratic political party. KSČM is an extreme left-wing communistic party.

Looking at Figures 2-6, we can see a similar pattern of the four most traditional Czech political parties after 1989. The biggest changes can be seen during the time of the parliament election (1998, 2002, 2006, 2010, 2013). In these years the *CCS* is going down which means that the context of the names of political parties is less specific during elections. The reason for this behavior lies probably in the fact that newspapers focus more on the future agenda of the political parties and try to provide adequate information for voters for the election. The parties are mentioned in journalistic texts on various topics and that is why the context of the names of parties is less unique.

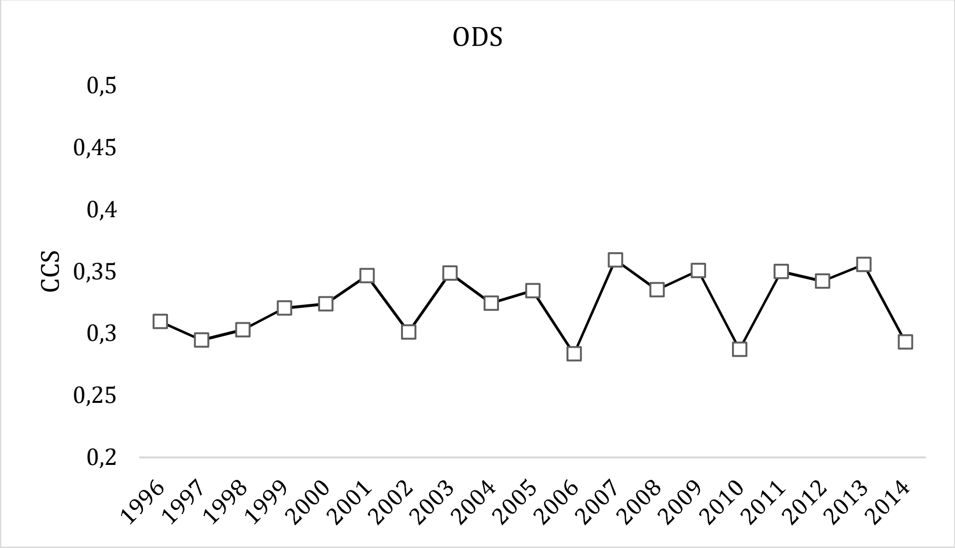


Figure 2 The *CCS* development of lemma “ODS”

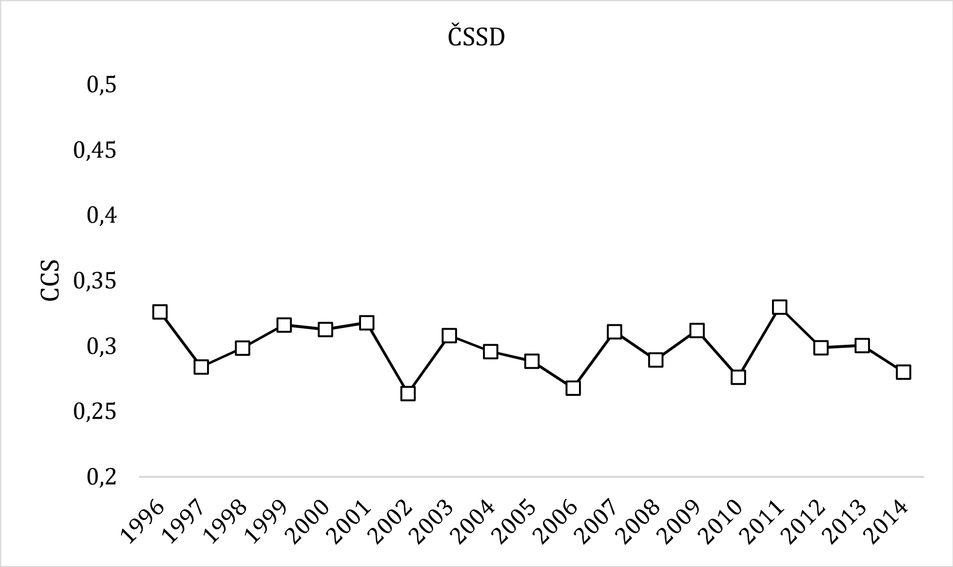


Figure 3 The *CCS* development of lemma “ČSSD”

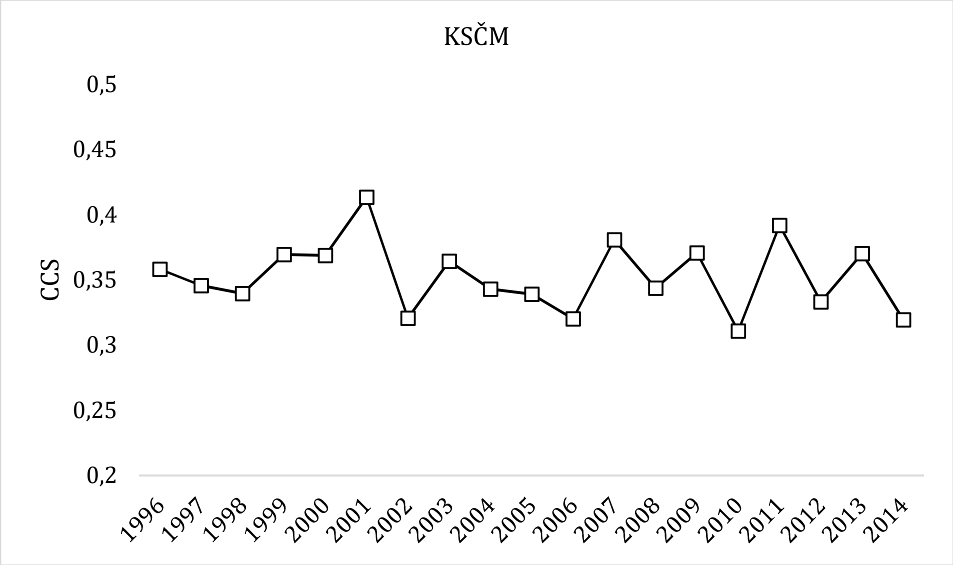


Figure 4 The *CCS* development of lemma “KSČM”

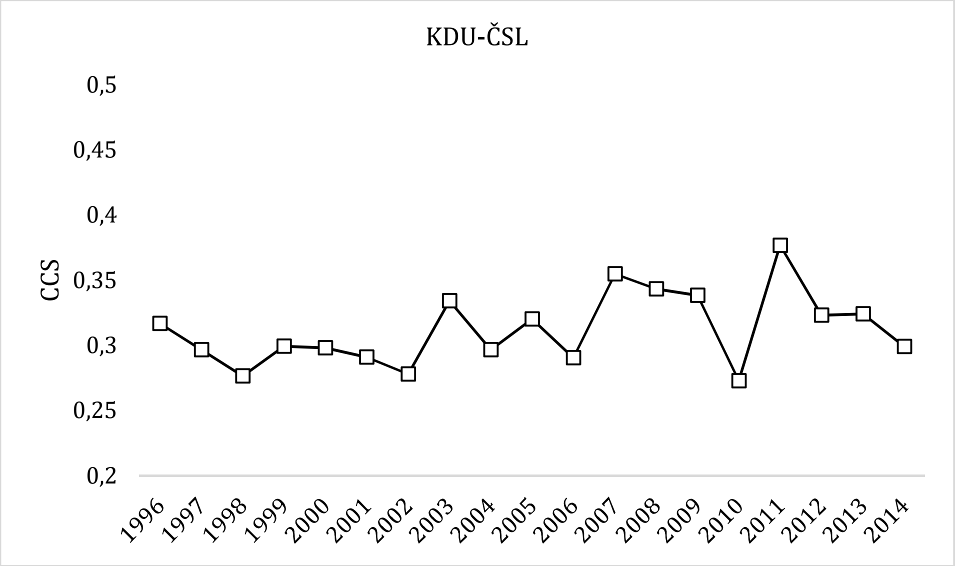


Figure 5 The *CCS* development of lemma "KDU-ČSL"

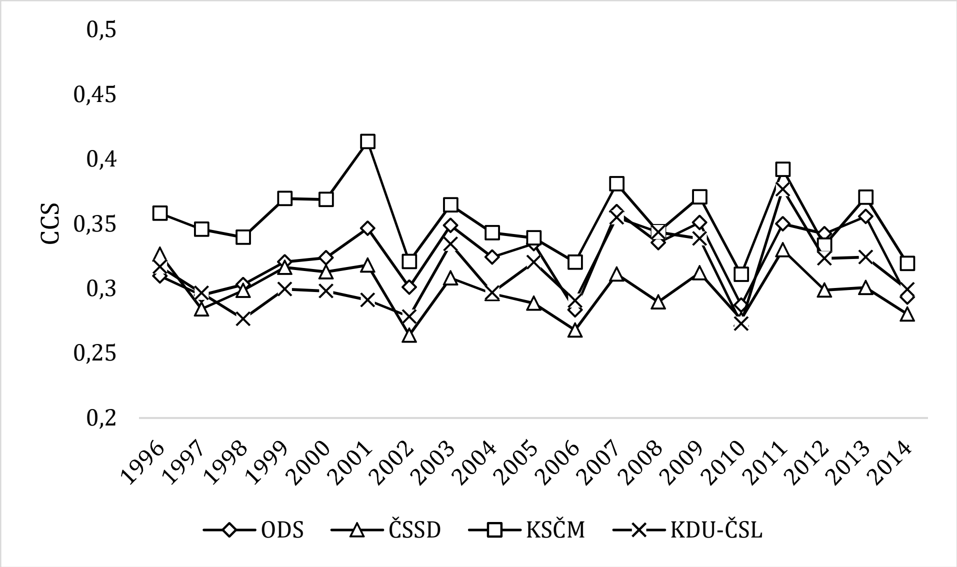


Figure 6 The *CCS* development of four traditional Czech political parties

4.2 Kraj, hejtman

In 2000, the new self-governing units were established in the Czech Republic. The name of this unit is "kraj". This word has several meanings. First, it can mean the place where something, especially surface, ends (an edge). Second, it can be used for referring to some geographical area. The last meaning is the regional unit. It should be mentioned that "kraj" also used to be a self-governing unit before 1989 with different borders and a different administration. Nowadays, the head of "kraj" is "hejtman". "Hejtman" has been used several times during the Czech history in more or less similar meanings. Thus, the usage of this lemma in newspapers in the nineties could refer to the historical meaning or to a discussion about planning new regional units. We can see in Figure 7 that the *CCS* is quite clearly reflecting the mentioned changes. The context specificity has a descending development which changes in 2000 into a rather straight curve. As we mentioned before, in the early nineties, the lemmas “kraj” and “hejtman” had very specific meaning referring to the history. Since 2000, the context of both lemmas is generally less specific because they are appearing in newspapers in a wide range of various topics.

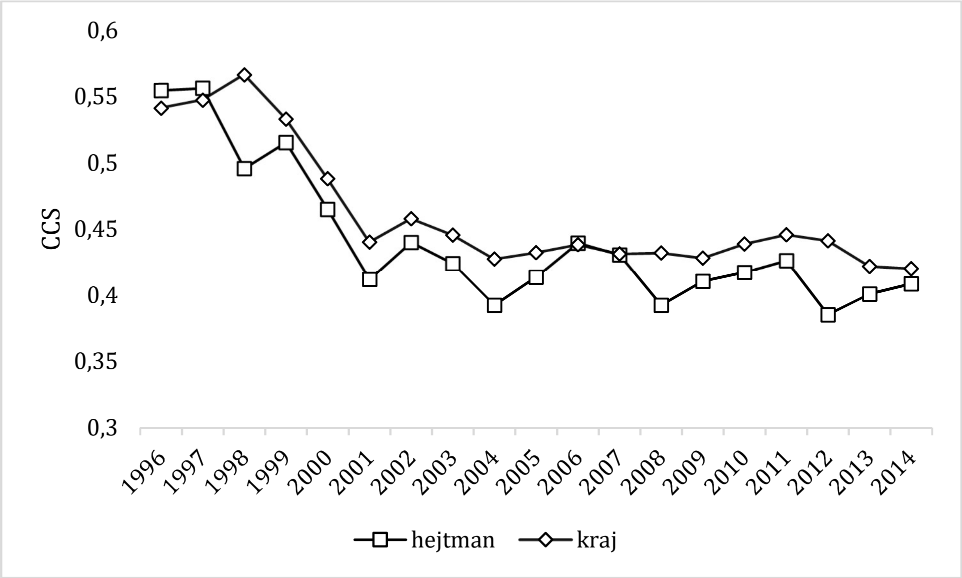


Figure 7 The *CCS* development of lemmas “hejtman” and “kraj”

The change of the meaning of the lemma “hejtman” can be also illustrated by closest lemmas at the beginning of the nineties and in 2014. In 1996, there are only those lemmas connected to the history. There are for example several lemmas referring to various administrative positions in the history of Czech lands such as “komoří“, „hofmistr“, „purkrabí“, „místodržící“, „maršálek“, „falckrabě“. Others are names of some historically important persons such as Pröll, Dietrichštejn, Radecký, Pühringer, Piccolomini. On the other hand, the majority of closest lemmas in 2014 belongs to the surnames of current hejtmans.

4.3 EU, NATO

The Czech Republic joined the North Atlantic Treaty Organization (NATO) in 1999 and European Union (EU) in 2004. These memberships, especially EU membership, has necessarily influenced the political agenda and content of newspapers as well. One could expect that the usage of the names of aforementioned institutions (EU, NATO) changed in a similar way like in the case of “kraj”.

If we look at the resulting values in Figure 8, the development is rather the opposite. In the case of both lemmas (EU, NATO) can be seen an increasing tendency of *CCS* which is contradictory to the situation of “kraj” where the new usage of this lemma caused lower context specificity. The tendency could be interpreted in the following way. Both memberships (NATO and EU) were widely discussed before the entrance to these organizations. The newspapers informed readers about all pros and cons in general. Thus, the context was rather less specific. After joining, the news about both organizations refer to some current issues. We can see in Figure 8 that NATO has generally more unique context than EU. It is quite obvious that EU is mentioned in Czech newspapers much more frequently than NATO because the European Union has a higher influence on the daily life of people. NATO is usually mentioned in the news in connection to some NATO summits or some conflicts. The range of potential topics of EU is much wider.

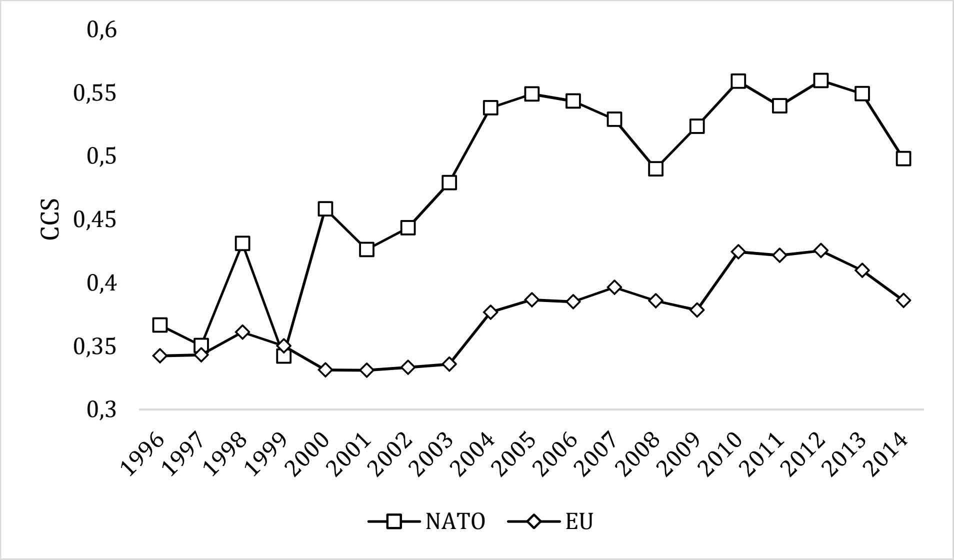


Figure 8 The *CCS* development of lemmas “NATO” and “EU”

4.4 Politicians

Another field where some semantic changes could be expected are names of famous politicians. Since we can detect changes over 20 years, we can see how *CCS* reacts to changes of politician’s carriers from a long perspective. We decided to analyze the development of *CCS* of the last three Czech presidents. These politicians can be considered the most famous and influencing Czech politics. The first one, Václav Havel, was a writer, a dissident and the first Czech democratic president from 1993 to 2003. Václav Klaus is a former economist and politician who served as the second President of the Czech Republic from 2003 to 2013, and as the first Prime Minister of the newly independent Czech Republic from 1993 to 1998. Klaus was also the principal co-founder of the Civic Democratic Party (ODS), a Czech free-market Eurosceptic political party. Miloš Zeman is the current Czech president since 2013. He is the first directly elected president in Czech history. He previously served as the Prime Minister of the Czech Republic from 1998 to 2002. For many years, Zeman was also a leader of the Czech Social Democratic Party.

We can see two clear breaking points in the development of *CCS* of Havel in 2003 and 2011 in Figure 9. In 2003, Havel left the office after his second term as Czech president. The context specificity is noticeably higher in the following years. This can be explained by the fact that Havel left politics and the range of topics he was mentioned was therefore much more narrow. Havel died in 2011 and that is why he was often mentioned in newspapers in that year.

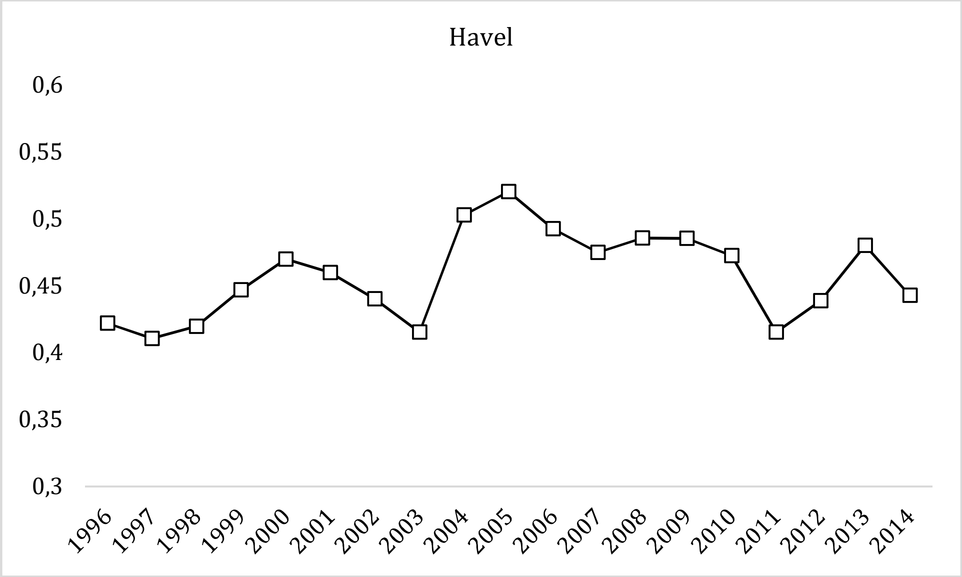


Figure 9 The *CCS* development of "Havel"

There are no such dramatic changes in *CCS* development of Klaus as in case of Havel or Zeman (see Figures 10 and 12). The reason lies in the fact that there were no big changes in his political carrier. Klaus entered Czechoslovak politics during the Velvet Revolution in 1989 and became Czechoslovakia's Minister of Finance in the same year. He served as the Prime Minister from 1992 to 1998. In 2003, he was elected as the President of the Czech Republic. Klaus has a rather stable political career where he step by step served several high positions like Minister of Finance, Prime Minister and President. Moreover, he was a leader of one of the most powerful Czech political party (ODS) from 1991-2002. He left the high politics when his presidential office ended in 2013.

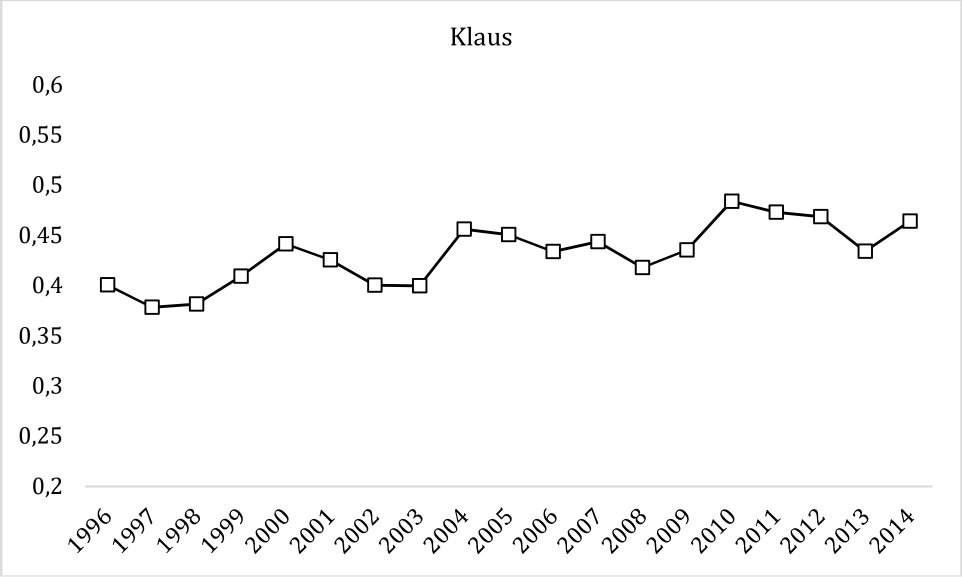


Figure 10 The *CCS* development of "Klaus"

As can be seen in Figure 11, there are two remarkable changes in the development of *CCS* values in 2003, 2013. Zeman left politics after unsuccessful presidential candidacy in 2003. He came back to politics in 2013 when he was elected as the President of the Czech Republic. We can see that the context specificity is considerably higher from 2003 until 2012 than in other years when he was an active politician.

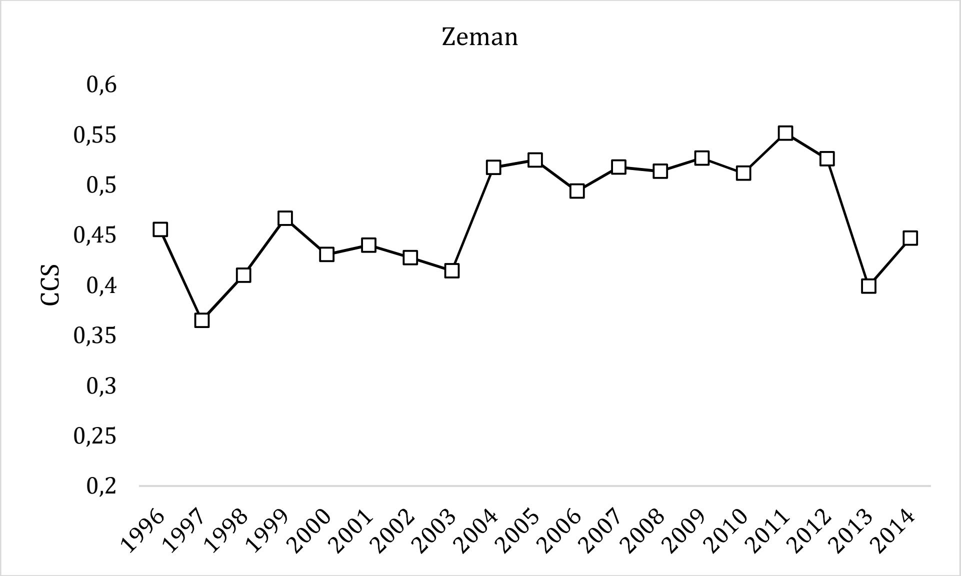


Figure 11 The *CCS* development of "Zeman"

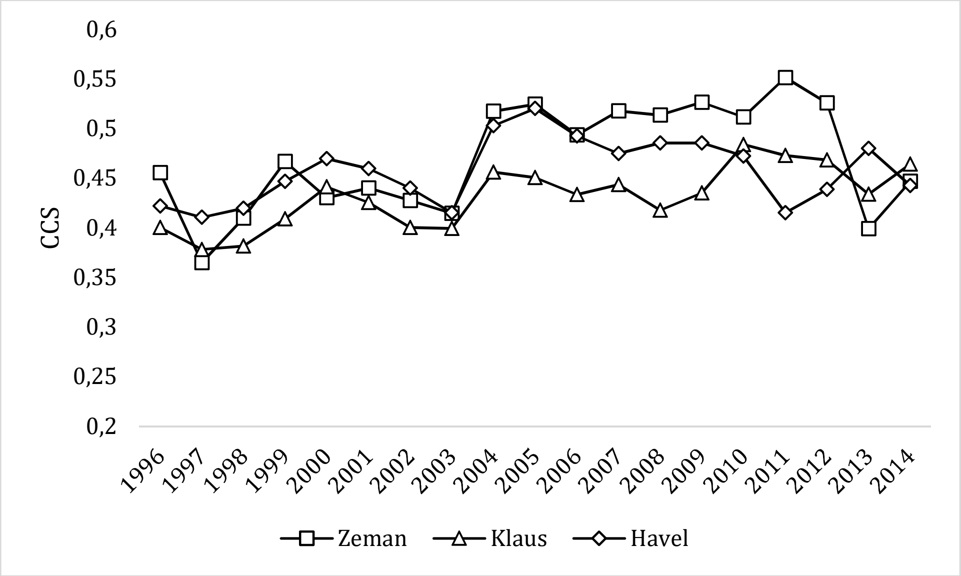


Figure 12 The *CCS* development of lemmas "Zeman", "Klaus" and "Havel"

4.5 Bird and swine flu

There were two epidemics of flu ("chřipka"), bird flu ("ptačí chřipka") and swine flu ("prasečí chřipka") in the last decade. Since these topics were widely reported in newspapers, we can expect semantic changes in the usage of lemmas "chřipka" (flu), "ptačí" (bird - adjective), and "prasečí" (swine - adjective). The years of the occurrence of these diseases are quite clearly detectable in the *CCS* development in Figures 13-16. In the Czech Republic, the bird flu emerged in 2006 and we can see that the *CCS* value drops exactly at that time. The *CCS* value has also a descendant tendency in case of the lemma "chřipka" (the flu).

The semantic changes are also very clear when we compare the closest lemmas to "ptačí" in 2006 and other years. For instance, we get following lemmas in 2000: "pták", (bird), “ptactvo“ (birds species), opeřenec (a bird), “opeřený” (adjective of "opeřenec"), hnízdící (nesting), “voliéra” (aviary), “sýkorka” (a tit), “krahujec” (a sparrowhawk), “krkavec” (a raven), “zoborožec” (a hornbill), “včelojed” (a perninae), “káně” (a buzzard), “ornitolog” (an ornithologist), “nocoviště” (a place for birds for staying overnight), “kroužkování” (bird ringing), “krmítko” (a bird feeder), “poletující” (fliting), “zobák” (a beak), “živočich” (an animal), “zob” (a bird food). We can see that all lemmas are connected to concepts connected to birds such as bird, aviary, ornithologist, etc.

In 2006, when the bird flu emerged in the Czech Republic, we get following closest lemmas to “ptačí” (bird - adjective): “chřipka” (a flu), “H5N1”, “nákaza” (an infection), “virus” (a virus), “pták” (a bird), “ptactvo” (birds - species), “vir” (a virus), “opeřenec” (a bird), “H5”, “nakažený” (infected). “drůbež” (poultry), “uhynulý” (dead), “labuť” (a swan), “nakažení” (an infection), “slintavka” (foot-and-mouth disease), “chřipkový” (flu - adjective), “pandemie” (a pandemic), “ornitolog” (an ornithologist), “H1N1”, “Nořín” (a name of a village where the bird flu emerged). We can see that most of these lemmas are connected to the emerged bird flu. Compare to the aforementioned closest lemmas in 2000, it is clear that the context substantially changed.

The epidemic of the swine flu emerged in the Czech Republic in 2009. This topic was highly reflected in newspapers and that is why the context of lemma “prasečí” (swine - adjective) changed in our corpus. This semantic change also influenced the *CCS* of the lemma “chřipka” (the flu).

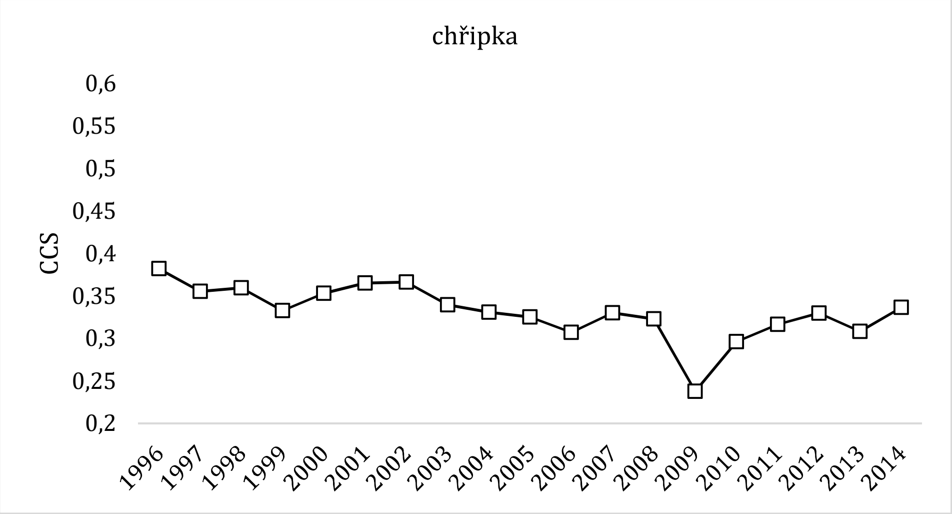


Figure 13 The *CCS* development of a lemma “chřipka” (flu)

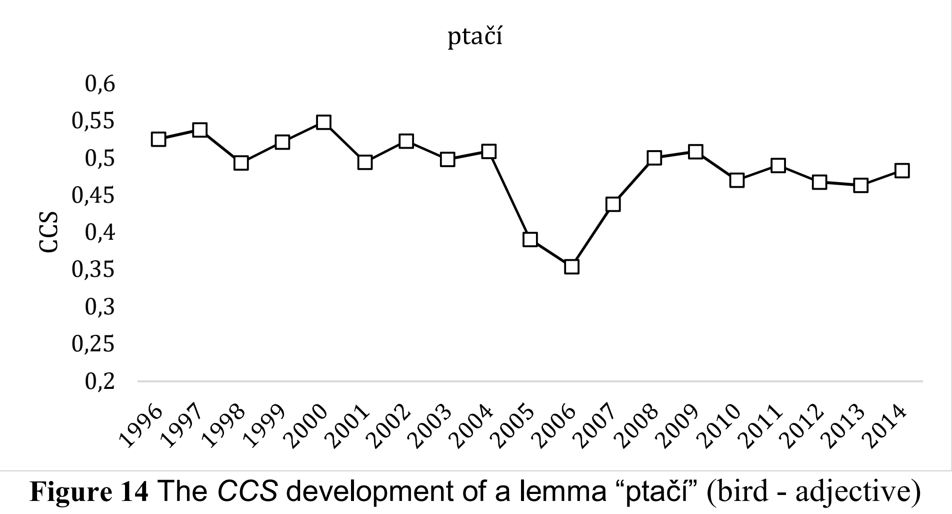


Figure 14 The *CCS* development of a lemma “ptačí” (bird - adjective)

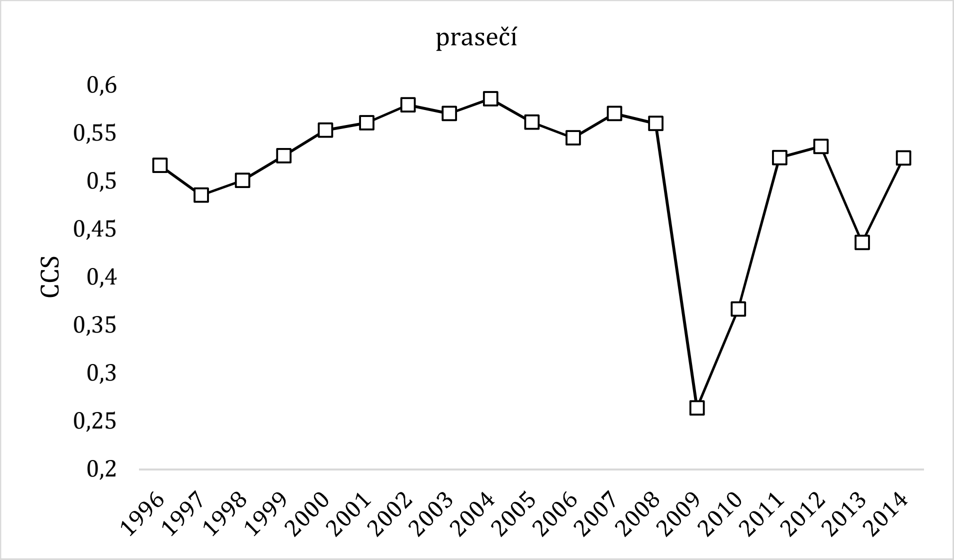


Figure 15 The *CCS* development of a lemma "prasečí" (swine - adjective)

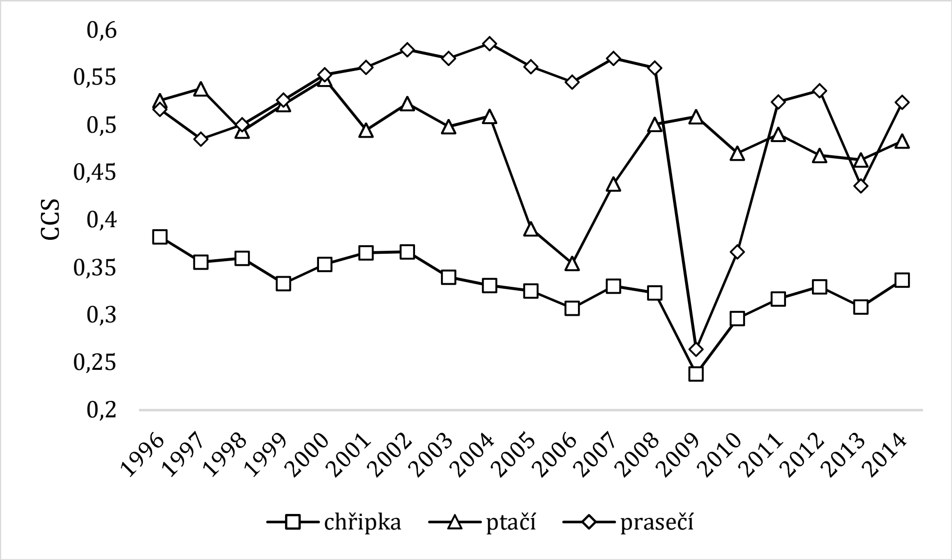


Figure 16 The *CCS* development of lemmas "chřipka", "ptačí" and "prasečí"

4.6 The relation of *CCS* to frequencies in the corpora

One of the most common obstacles of any quantitative linguistic analysis is the relation of a measured feature to frequencies in the analyzed corpus. Linguists have been dealing with this problem since they started to apply statistics to language data. The well-known case is measuring so-called vocabulary richness which is one of the common methods in quantitative linguistics, especially stylometry (cf. Kubát 2016). Given that we work with lemmas with different frequencies, we test the correlation between the obtained *CCS* values and the frequency of lemmas in the corpus. Since the analyzed subcorpora do not have the same size, the relative frequencies are used instead of the absolute frequencies. Namely, we apply the i.p.m. (instances per million) which is the average number of occurrences of the lemma in a hypothetical corpus with the size of 1 million words. We apply the Pearson correlation coefficient with the result r = -0.23. Pearson Coefficient of determination R^2 = 0.05. The correlation is visualized in Figure 17. We can see that generally *CCS* is not strongly influenced by frequencies.

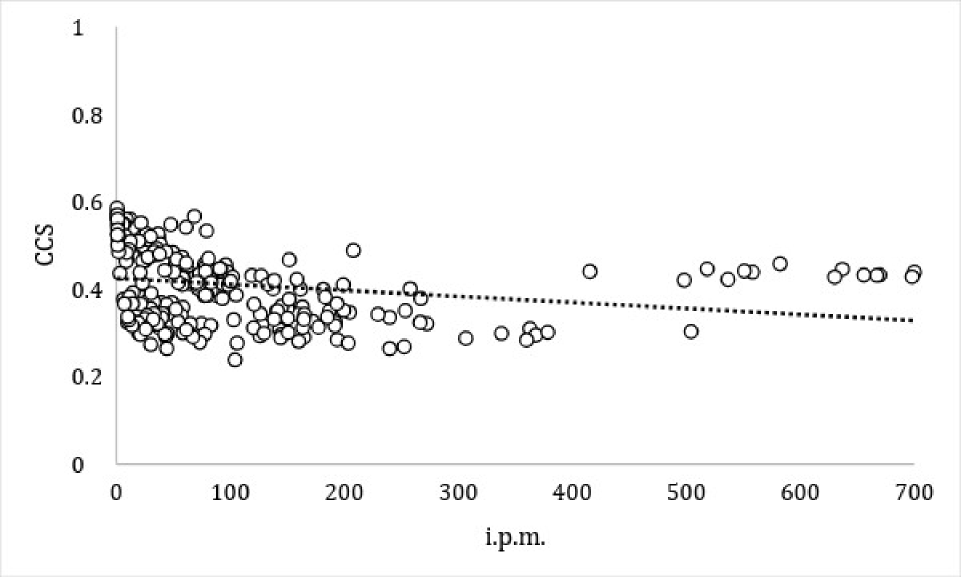


Figure 17. The correlation between *CCS* and relative frequencies (i.p.m.)

5. Conclusion

Closest Context Specificity of Lemma (*CCS*) expresses a kind of semantic feature of lemmas. The measurement is sensitive enough to study changes even in a relatively short time (several years). The behavior of the measured *CCS* development of the analyzed lemmas seems to be quite predictable and interpretable from a qualitative linguistic point of view. We tested the relation between *CCS* and frequencies of lemmas in the corpus. The results of Pearson correlation coefficient show that there is no strong correlation (r = -0.23, R^2 = 0.05).

We can state that the obtained results of this study support the preliminary conclusions given by the authors of the concept Context Specificity of Lemma (Čech et al. 2018, Kubát et al. 2018). This approach therefore seems to be promising tool for lexical semantic analyses. Since it is generally very problematic to study semantics in linguistics by quantitative methods, this method based on Word2vec technique could have a great potential in future research. The important advantage of this approach lies in the fact that even though it is based on neural networks (which are "black box" models), this concept of measuring the uniqueness of the context of the lemma allows linguistic interpretation.

Needless to say, this study is just one attempt to better understand the recently proposed method. More data must be analyzed to support or reject our conclusions based on the obtained findings in this study.

Acknowledgments

This work is supported by the University of Ostrava (No. SGS01/UVAFM/2018), the European Union & Ministry of Education of the Czech Republic (No. CZ.02.2.69/0.0/0.0/16_027/0008472), and the European Regional Development Fund (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

REFERENCES

Čech, R., Hůla, J., Kubát, M., Chen, X., Milička, J. (2018). The Development of Context Specificity of Lemma. A Word Embeddings Approach, *Journal of Quantitative Linguistics,* DOI: 10.1080/09296174.2018.1491748.

Hnátková, M., Křen, M., Procházka, P., & Skoumalová, H. (2014). The SYN-series corpora of written Czech. In Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14) (pp. 160–164). Reykjavík: ELRA

Kubát, M. (2016). *Kvantitativní analýza žánrů*. University of Ostrava.

Kubát, M., Hůla, J., Chen, X., Milička, J., Čech, R. (2018). The lexical context in a style analysis: A word embeddings approach. *Corpus Linguistics and Linguistic Theory*, DOI:10.1515/cllt-2018-0003

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics 3*. 211–225.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013b). Efficient estimation of word representations in vector space (ICLR Workshop Papers).

Mikolov, T., Chen, K., Corrado, G. S., Dean, J., & Sutskever, I. (2013a). Distributed representations of words and phrases and their compositionality. *Proceedings of Neural Information Processing Systems (NIPS 26)* (pp. 3111–3119).

Glottometrics 45, 2019, 24-57

Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis

Renkui Hou[a,b], Chu-Ren Huang[b], Mi Zhou[b], Menghan Jiang[b]
[a]College of Humanity, Guangzhou University, Guangzhou, China;
[b]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, HongKong

Corresponding author: Renkui Hou, email address: hourk0917@163.com

Abstract. This paper proposes an innovative method/index to represent the formality of a register based on the Menzerath–Altmann law and regression analysis. This index also can be used to quantify the distance between two registers. Analysis demonstrates that average word length decreases with the increase of clause length in each register and that their relationship can be fitted by the formula $y = ax^b$. It can be shown that the link between average word length and clause length abides by the Menzerath–Altmann law. Texts were represented by the fitted parameters, a and b, and their positions were plotted in 2-dimensions. Linear regression can be used to fit the functional correlation between these two parameters in each register. We show that the a-intercept of this regression line can be used as an index to represent the formality degree of the register and to compute the distance between two registers.

Keywords: *Distance between Chinese registers, The Menzerath–Altmann law, Chinese word length, Chinese clause length, Regression analysis.*

1 Introduction

Variability is inherent in human language: people use different linguistic forms on different occasions and different speakers of a language convey the same messages in different ways. Register is often considered to be the most important perspective on text varieties (Biber and Conrad 2009). The register perspective combines an analysis of the linguistic characteristics that are common in particular text varieties with an analysis of the situations of use of those varieties.

The essential features of registers involve three factors: context, linguistic materials, and fixed ways of expressing objects, the combination of which forms a discourse. We will discuss the distances between different Chinese registers based on the Menzerath–Altmann law (henceforth, the MA law), which explores the relationship between language constructs and their immediate constituents, from the perspective of quantitative linguistics.

The MA law, which is one of the best known quantitative linguistic laws, originates from the fact that the length of a construct influences the lengths of its immediate constituents in different language domains. Paul Menzerath summarized the law as “the greater the whole, the smaller its parts” after he detected the dependency of syllable length on word length (Menzerath, 1954, p.101). Altmann generalized this hypothesis to all levels of linguistic analysis, formulating it as “The longer a language construct, the shorter its components” (Altmann, 1980). Hřebíček (1992, 1995, 1997) showed that the whole hierarchy of textual levels is based on this dependency, and called this the Menzerath–Altmann law.

The theoretical derivation and corresponding differential equation of the MA law were proposed by Altmann (1980) in his seminal ‘Prolegomena to Menzerath’s Law’, as shown in Equation (1).

$$\frac{y'}{y} = -c + \frac{b}{x} \qquad \text{Equation (1)}$$

The solution to this differential equation is shown in the Formula (1):

$$y = ax^{b}e^{-cx} \qquad \text{Formula (1)}$$

where *y* is the mean size of the immediate constituents (average word length in this study), *x* is the size of the construct (clause length), and parameters *a*, *b*, and *c* depend mainly on the levels of the units under investigation, rather than on the language, the kind of text, or the author, as had previously been expected (quoted by Köhler, 2012). However, there is no convincing theoretical support for the substantiated interpretation of these parameters although it is a well-known distribution model in linguistics (Eroglu 2014). In this study, we will demonstrate that these parameter values are affected by the registers in Chinese.

It has previously been assumed that one of the two parameters, either *b* or *c*, can be neglected from the function. Then, two simplified forms are obtained:

$$y = ax^{b} \qquad \text{Formula (1a)}$$
$$y = ae^{-cx} \qquad \text{Formula (1b)}$$

A large number of observations have shown that parameter *c* is close to zero for higher levels of language whereas lower levels lead to very small values of parameter *b*; only for intermediate levels is the full formula needed (Köhler, 2012). Formula (1a) has become the most commonly used “standard form” for linguistic purposes (Grzybek, 2007).

This paper aims to establish an index to measure the formality of registers and to represent the distance between two Chinese registers based on the MA law and regression analysis.

1.1 Literature review

Generally speaking, a register is associated with a particular situation of use. It refers to the principles generated in communication and followed by speakers and listeners. Register and linguistic performance are interdependent and are not tenable without each other as register is produced and shaped by linguistic performance and, in return, its rules regulate linguistic performance once it is formulated. Except for utterances with improper register, all utterances can be categorized into a register. Biber (2012) argued strongly that reference works that describe different linguistics levels, i.e., lexical, grammatical, and lexico-grammatical, should consider register difference. For example, Cacoullos (1999) provided evidence that reductive change in grammaticalizing forms may be manifested not only as a diachronic process but also as synchronic differences between formal and informal registers. The significance of comparing different registers in studies of Chinese grammar was introduced by Lv (1992). Zhang (2012) has shown that there is much variation of linguistic properties across written Chinese registers. Consequently, we should observe the differences of manifestation of quantitative linguistic laws in different registers. For example, Hou et al. (2017) showed that the relationship between sentences and their constituting clauses abides by the MA law in written formal register texts, but not in *TV Sitcom* and *TV Conversation*. Failing to take register into account can lead to inaccurate, even incorrect, conclusions.

Biber's (1994) observation of the lack of agreement on the definitions and taxonomy of registers also applies to the study of registers in Chinese. Yuan and Li (2005) took a discrete approach and proposed seven registers: conversational, officialese, scientific, news, literary and art, lectures, and advertisements. Similar to Biber and Conrad (2009), who regard register differences as a continuum of variation, Feng (2010) thought that register is generated in interpersonal communication and that the essence of register is to adjust the psychological distance between the communicators. He held formality to be the primary element of register and proposed that register is a polarized opposite continuum, with the written formal register being the most formal, the daily informal register being the most informal, and all other registers lying in between. However, the positions of other registers in this continuum and the distances between various registers were not discussed. We adopt Biber's (1994) position to reconcile the above differences: registers are varieties in a continuum, but they are still to be analytically identified as different categories.

Köhler (2012) pointed out that the mathematical methods are worth being integrated into linguistics. Register can also be studied using such mathematical methods. Biber (1986, 1988) is generally credited with introducing quantitative methods to the linguistic study of registers. Biber (1995) restated and underlined the role of computational, statistical, and interpretive techniques using multi-dimensional analysis. He pointed out that any text characteristic that is encoded in language and can be reliably identified and counted is a candidate for inclusion. Research on register characteristics has also been undertaken from the perspective of quantitative linguistics. For example, Hou, Huang, and Liu (2017) fitted the distribution of Chinese sentence lengths using nonlinear regression and used the fitted parameters as quantitative features of the corresponding Chinese registers. In this paper, we propose an index to represent the formality of registers and quantify the distance between two registers based on the MA law and regression analysis.

As one of the best-known laws of quantitative linguistics, the MA law establishes the interrelations between successive hierarchical levels of language, providing evidence that language is a self-organizing and self-regulating system. Previous research has validated the

MA law at different language levels. For example, Köhler (1982) conducted the first empirical test of the MA law at the sentence level, analyzing short stories in German and English and philosophical texts. In his investigation, Köhler counted clause lengths in terms of the number of constituent words. Statistical tests on the data confirmed the validity of the law with high significance. Tuldava (1995) examined the dependence of average word length on clause length, finding a statistically highly significant interdependence between average word length and clause length, indicating that there are other factors that influence average word length. Motalová et al. (2014) and Ščigulinská and Schusterová (2014) verified the validity of the MA law applied to contemporary written and spoken Chinese respectively. Benešová (2016) tested the potential validity of the MA law on samples in different languages and attempted to test the concept of this language universal. Wilson (2017) used the MA law to test the hypothesis that the intonation unit is a valid language construct whose immediate constituent is the foot.

Benešová & Čech (2015) proved the MA law from another perspective. They conducted that the data generated by random models does not fulfil the MA law. Consequently, they pointed out that the results can be viewed as another argument supporting the assumption considering that the MA law expresses one of important mechanisms controlling human language behavior.

In addition to applications of the MA law at different language levels, some researchers have studied the theory and formula of the law, which has been interpreted in various ways. For example, Köhler (1989) proposed that the mechanism of shortening is a consequence of memory limitations: the longer the construct, the more space must be reserved for structural information between the constituents, hence the size of the constituents must be reduced.

Hammerl and Sambor (1993) concluded that there is a negative correlation between the parameters of the MA law: the greater the value of *a*, the less the value of *b* (quoted in Kułacka, 2010). Cramer (2005) confirmed that the parameters, *a* and *b*, depend on the linguistic level of analysis and also showed that there is a functional correlation between *a* and *b*. This paper will also investigate the functional correlation between these two parameters in each register using linear regression.

1.2 Research question and methodology

This paper proposes an index to represent the formality of a register and the distance between two registers based on the MA law from the perspective of quantitative linguistics and regression analysis.

Effective register analyses are always comparative as it is virtually impossible to know what is distinctive about a particular register without comparing it to others. We have therefore selected texts from multiple registers to establish the corpus.

In contrast to Indo-European languages, it is difficult to define the terms “sentence” and “clause” in Chinese. Chinese sentences are often defined in terms of characteristics of speech (Huang and Shi, 2016; Lu, 1993). Chao (1968) and Zhu (1982) defined a sentence as an utterance with pauses and intonation changes at its boundaries. Huang and Liao (2002: P4) proposed that a sentence is a linguistic unit that has an intonation and can express a relatively complete meaning in Chinese. However, sentences are often defined using punctuation marks in corpus linguistics and quantitative linguistics. A common approach for identifying sentences in syntactically annotated corpora (e.g., Chen et al., 1996; Chen et al., 2013; Huang and Chen, 2017 for Sinica TreeBank) is to mark all segments between punctuation marks that indicate utterance pauses as sentences. Such punctuation marks include commas, semicolons, colon, periods, exclamation marks, and question marks. Wang and Qin (2014) and Chen (1994) also adopted this operational definition and called such units *sentence segments*. Chen (1994) reported that about 75% of Chinese sentences are composed of more than two sentence segments separated by commas or semicolons by corpus analysis. Wang and Qin (2014) considered the lengths of sentence segments to be relevant to language use in Chinese. In fact, sentences (as defined by Chen et al., 2003; Huang and Chen, 2017) and sentence segments (as defined by Chen, 1994; Wang and Qin, 2014) are roughly equivalent to clauses. One sentence is composed of one or more clauses, which is called simple sentence or complex sentence (Huang and Liao, 2002: P5). The structures of the simple sentences and clauses are similar in Chinese, but the latter lack a complete intonation. In complex sentences, there are generally pauses represented by commas, semicolons and colons between clauses. Pauses at the boundaries of the sentences are represented by the periods, exclamation marks, and question marks (Huang and Liao, 2002: p 159). Thus, an operational definition of Chinese clauses can also be based on the written form, and the aforementioned punctuation marks determine the boundaries of the clauses.

It has become common in quantitative linguistics to measure the length of a linguistic entity as the number of its immediate constituents. We assume that the immediate constituents of Chinese clauses are words, hence clause length can be defined as the number of words. We consider words to be the segments delineated by blank spaces in the texts segmented by a Chinese lexical analysis system. There are various perspectives to define word length, for example, from the perspectives of pronunciation, duration, and syllable number. For Chinese, we define word length as the number of Chinese characters (*Hanzi*, 汉字) in the word (Hou, Yang and Jiang, 2014; Chen and Liu, 2016).

We selected Formula (1a) to fit the function between average word length and clause length in Chinese. Formula (1a) shows that this function is nonlinear. This nonlinear function can be transformed into a linear function in order to avoid the impact of the initial parameter estimates on the fitted result.

$$y = ax^b \qquad \text{Formula (1a)}$$

Taking the logarithm of both sides of Formula (1a) gives

$$\ln(y) = \ln(a) + b\ln(x)$$

Then, defining

$$Y = \ln(y);\quad X = \ln(x)$$

The linear function stated in Formula (1a-1) is obtained:

$$Y = bX + \ln(a) \qquad \text{Formula (1a-1)}$$

If the logarithm of average word length distribution can be fitted by this linear regression, as shown in Formula (1a-1), the average word length can be fitted by the non-linear regression, as shown in Formula (1a). We will show that the fitted result using linear regression is as well as that using nonlinear regression in later section. Thus the determination coefficient (R^2) was used to validate the fitted results of this linear regression as like residual sum-of-square for the validation of nonlinear regression result; it shows the goodness-of-fit of the model to the empirically collected data. It indicates the proportion of variance in the data that can be explained by the model (Conway & White, 2013). In quantitative linguistics, a fit is generally considered good if R^2 is greater than or equal to 0.9 (Popescu et al., 2009, p.16). A fit with $0.9 > R^2 > 0.7$ is tolerable. Our study will show that the residual sum-of-squares of nonlinear regression is small if the R^2 of linear regression is large. In addition, the different settings of initial parameter values affect the fitted result. Since the aim of the paper is to obtain the parameters, *a* and *b*, to represent the texts and then calculate the distance between the different registers, an approach that does not reliably yield constant parameters is not appropriate. We adopt the linear regression approach in this study because it can be used to fit the logarithm of average word length distribution and obtain the parameters.

The function between average word length and clause length was fitted by Formula (1a-1) in each text. Then the texts from various registers were represented by the fitted parameters, *a* and *b,* using a vector space model, allowing the positions of each register texts to be displayed on a coordinate graph. The positions of the texts in each register indicate that there is a systematic link between parameters *a* and *b* in the texts from each register, which can be fitted by linear regression. The point at which the regression line intersects the *a*-axis when *b* achieves its extreme maximum value, i.e., 0, is dependent on the particular register. The value of the *a*-intercept can be used as an index to represent the position of a register in the formality continuum and to quantify the distances between various registers.

We used the open source programming language and environment R (R Core Team, 2016) to realize the fitting procedure and for the computation of both clause length and average word length. The R function *lm*() was used to fit Formula (1a-1) in order to obtain the values of parameters *a* and *b*, and to carry out regression analysis on the link between parameters *a* and *b* in texts from the same register.

2. Corpus Establishment and Preprocessing

Texts from “*News Co-Broadcasting*”, the situation comedy “*I Love My Family*”, and “*Behind the Headlines with Wentao*” were selected to represent the *News Broadcasting*, *Sitcom Conversation*, and *TV Conversation* (i.e, *TV Talkshow*) registers respectively.

The Central China TV (CCTV) program, “*News Co-Broadcasting*”, mainly consists of brief introductions of important state policies and events taking place both at home and abroad. It is characterized by formal use of language in non-interactive uni-directional speech. It is the representative of the *News Broadcasting* register.

“Behind the Headlines with Wentao” is a talk show of Phoenix Satellite TV in which the host discusses current hot issues and topics together with guests. Their dialogue is supposed to be

un-scripted with real time interaction. The speakers aim to entertain, inform, and even persuade the audience. The language use is representative of the *TV Conversation* register.

The situational comedy, "*I Love My Family*", tells the story of a family via well-constructed casual dialogues. Although the content is scripted, it is expected that the delivery should be informal and intimate. This is the representative of the *Sitcom Conversation* register.

Overall and intuitively, the *News Broadcasting* register is the most formal one, due both to its scripted nature, and the nature of being one-way communication aiming to inform. *TV Talkshow* is supposed to be less formal, due to its interactive and unscripted nature. Yet its discussion is still topical and the social inter-personal relation is only minimally expressed. Hence it is considered to be less formal. Lastly, even though *TV sitcom conversation* has to be scripted, it is scripted to reflect characteristics as well as the relation between the speaker and the addressee. And even though the conversation is meant to be heard by the audience, it doesn't need the audience to acquire information and gain information. Given that these contrasts, the register differences may be complex. We will use our result to explore whether the formality of register is dependent on one or more specific features.

The texts of *News Broadcasting* were obtained from the National Broadcast Language Resources Monitoring and Research Centre at the Communication University of China. Textual materials of "*Behind the Headlines with Wentao*" were collected from the website of Phoenix Satellite TV. The texts of "*I Love My Family*" were downloaded from the Internet. The names of speakers were deleted because they do not occur in either "*Behind the Headlines with Wentao*" or "*I Love My Family*".

The Chinese lexical analysis system created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS) was used for word segmentation. ICTCLAS has been acknowledged as having a high accuracy of 97.58%, a recall rate of over 90% for the recognition of unknown words based on role tagging, and a recall rate of approximately 98% for the recognition of Chinese names[1].

The segmented texts were screened manually. For example, words within bracket pairs in "*Behind the Headlines with Wentao*" were deleted if they were explanatory notes because explanatory notes are not considered to be parts of the texts. No special treatment was given to deal with isolated numbers and letters in the corpus.

The scales of the texts from these three registers are shown in Table 1.

Table 1
Scale of the texts from the different registers

	Number of Texts	Number of Types	Number of Tokens
News Co-Broadcasting	50	24,812	418,943
Behind the Headlines with Wentao	50	16,372	357,663
I Love My Family	60	14,107	317,661

[1] http://www.ict.ac.cn/jszy/jsxk_zlxk/mfxk/200706/t20070628_2121143.html

Having conducted preliminary research on the texts from these three registers, an index which can represent the formality degree and compute the distance between two registers was deduced. We then performed a test of validity of the index on the Lancaster Corpus of Mandarin Chinese (LCMC), which became available in 2003 (McEnery and Xiao, 2004). This corpus includes 500 texts of 2,000 word tokens each (i.e., totaling 1,000,000 words) from 15 written registers, taken from publications from mainland China between 1988 and 1992. We believe that this verification can make the conclusions that we draw here robust.

3 Experiments

3.1 Frequency distribution of clause length in terms of words

The frequency distributions of clause length in terms of words for each register were established, as shown in Figure 1. The occurrence frequency distributions and the relative occurrence frequencies of clauses with certain lengths are shown in Appendix 1 and 2 respectively. The figure demonstrates that the clause length distributions are similar in each register. In *Sitcom Conversation* texts, one-word clauses are more frequent than clauses with other lengths, reflecting the prevalence of such one-word clauses in daily conversation. The frequencies of clauses in texts from the other two registers, *News Broadcasting* and *TV Conversation*, first increase and then decrease with clause length.

The cumulative relative frequency distributions of clause lengths for each register are shown in Figure 2, from which we observe that most clauses are composed of few words. More than 98% of clauses in *TV Conversation* and *Sitcom Conversation* are composed of 1 to 15 words. About 99% of clauses in *News Broadcasting* are composed of fewer than 20 words. Figure 1 shows that the short clauses appear more frequently and longer clauses appear less frequently. Figure 2 shows that most clauses are short.

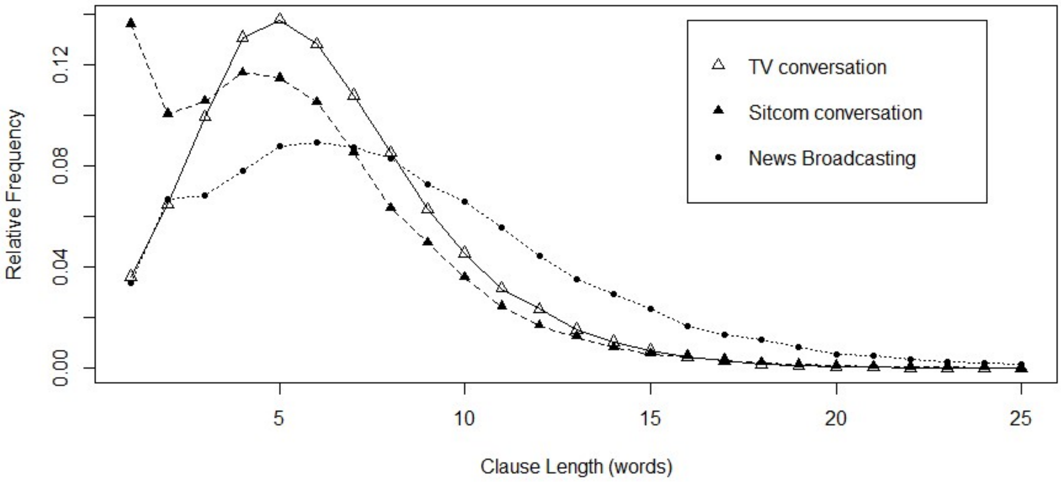


Figure 1: Frequency distributions of clause length in terms of words

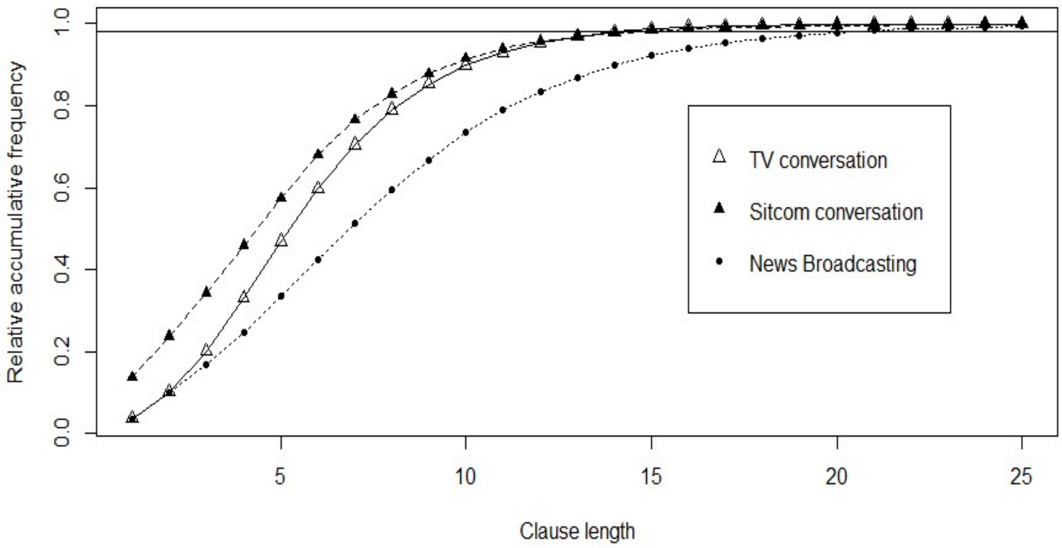


Figure 2: Cumulative relative frequency distributions of clause length in terms of words

3.2 Average word length distribution in clauses

The average word length in clauses with a certain length was calculated as the number of Chinese characters in the given clauses divided by the number of words in those clauses, which is shown in Appendix 3. As well as for texts from these three registers, we also calculated the average word length in the clauses having a certain length across texts from all registers.

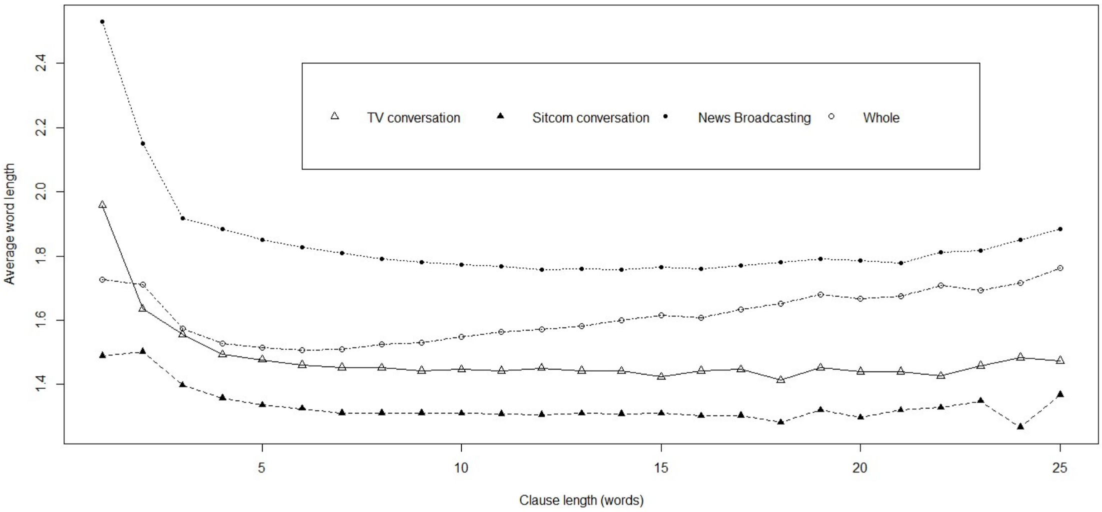


Figure 3: Average word length distributions in clauses

Figure 3 shows the negative relationship between average word length and clause length in each register. The average word length decreases with the increases of clause length in most clauses. The reason for the irregular change of average clause length in few long clauses needs to be explored in Chinese. From the figure, we observe that average word length in *News*

Broadcasting and *TV Conversation* texts decreases with clause length for most clauses. In *Sitcom Conversation*, the average word length in one-word clauses is smaller than in two-word clauses due to the large frequency of one-character words in one-word clauses, which are mostly interjections. In clauses with more than 1 word, the average word length decreases with increase of clause length. However, for all texts across registers, the average word length decreases with clause length only for short clauses of 1 to 6 words, accounting for 57.3% of all clauses. For longer clauses, the average word length increases with clause length. It is necessary to examine the distribution of average word length separately in each register in Chinese; otherwise, an incorrect conclusion would be obtained.

3.3 Regression analysis

Formula (1a-1) was selected to fit the relationship between average word length and clause length. In the fitting process, the clauses whose lengths are 15, 15 and 21 words in *TV Conversation*, *Sitcom Conversation* and *News Broadcasting* were fitted respectively. The fitted results are shown in Table 2 and Figure 4.

In Table 2, the values of determination coefficient, R^2, show that the link between the logarithm of average word length and the logarithm of clause length can be fitted by Formula (1a-1) for each of the three registers: *News Broadcasting*, *TV Conversation*, and *Sitcom Conversation*. The *p*-values, which are all smaller than 0.05, indicate the presence of a significant linear relationship between Y (the logarithm of average word length) and X (the logarithm of clause length).

The residual sum-of-squares is considered the measure to validate the result of nonlinear regression. We also calculated the residual sum-of-squares of the result of linear regression, which is the sum of squares of the difference between the predicted values and the observed values, in order to compare the results between linear regression and nonlinear regression.

Non-linear regression was used to fit the average word length distribution in *TV Conversation* text. We used the values of parameters, which obtained from the linear regression of the logarithm of average word length distribution, as the initial values of them. The residual sum-of-squares is 0.053 in the nonlinear regression result of the average word length distribution in *TV Conversation* text. In the meantime, the residual sum-of-squares is 0.054 using the fitted result of linear regression in *TV Conversation* text. The difference is 0.001 between them, which means the result of linear regression is as well as that of the nonlinear regression.

Similarly, the residual sum-of-squares is 0.009 in the nonlinear regression of the average word length distribution in *Sitcom Conversation* text. In the meantime, the residual sum-of-squares is also 0.009 when the linear regression was used to fit the logarithm of the average word length distribution in *Sitcom Conversation* text. The same values of residual sum-of-squares means the results of linear and nonlinear regressions are both well. In addition, the residual sum-of-squares in the regression result of average word length distribution in *Sitcom conversation* is less than that in *TV Conversation*. It means the regression result of the average word length distribution in *Sitcom Conversation* is better than that in *TV Conversation*. In the meantime, the R^2 of the linear regression result of average word length distribution in

Sitcom Conversation is more than that in *TV Conversation*. The linear regression result in *Sitcom Conversation* is better than that in *TV Conversation*. The conclusion is as same as that from the residual sum-of-squares.

The values of residual sum-of-squares are 0.153 in nonlinear regression of average word length distribution and 0.158 in linear regression of the logarithm of average word length distribution in *News Broadcasting*. The little difference between these two values showed that the results of linear regression is as similar as that of nonlinear regression. This residual sum-of-squares is more than that in *TV Conversation* and *Sitcom Conversation*. In the meantime, the R^2 is less than that in *TV Conversation* and *Sitcom Conversation*. They all showed that the fitted result of average word length distribution in *News Broadcasting* is not as well as that in *TV Conversation* and *Sitcom Conversation*.

We can see that the linear regression result of the logarithm of average word length distribution is similar with the nonlinear regression result of average word length from the comparison of the residual sum-of-squares. The more R^2 means the smaller residual sum-of-squares, which means that the good fitted result. The R^2 in line regression can also validate the fitted result of nonlinear regression result indirectly.

Hence we used linear regression to fit the average word length distribution because its result is similar with the nonlinear regression and the values of parameters is not set beforehand.

Table 2

Fitted results of link between average word lengths and clause length

	a	*b*	R^2	*p*-value
TV Conversation	1.784	−0.093	79.38%	8.352×10^{-6}
Sitcom Conversation	1.490	−0.055	84.74%	1.148×10^{-6}
News Broadcasting	2.240	−0.091	75.28%	3.513×10^{-7}
Whole	1.626	−0.013	6.94%	0.291

For each register, the value of parameter *b* is negative, which indicates that average word length decreases with clause length. Thus, as can be seen from Table 2 and Figure 4, the relationship between clauses and their constituent words abides by the MA law in each register. For texts across all three registers combined, $R^2 = 6.94\%$, indicating that the link between average word length and clause length cannot be fitted by Formula (1a-1), and the *p*-value, 0.291 (which is greater than 0.05), shows that there is not a linear relationship between *Y* and *X*, indicating that the relationship between clauses and their constituent words does not abide by the MA law.

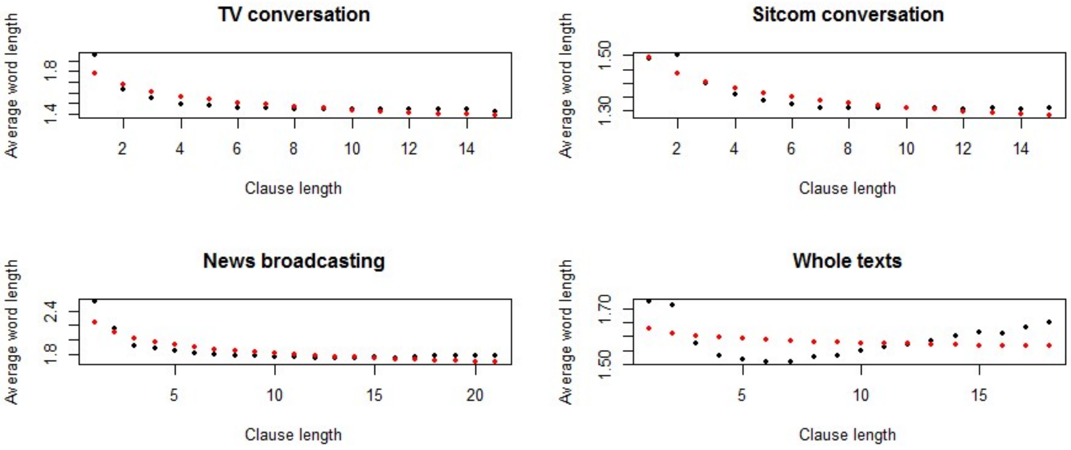


Figure 4. Fitted results of link between average word length and clause length (black dots represent the observed values of average word length; red dots represent the fitted values of average word length)

The long clauses have to be included in this experiment in order to consider as many clauses as possible, especially in the texts from *News Broadcasting*, as indicated by Figures 1 and 2. Figure 4 and Table 2 show that the link between average word length and clause length across the three registers combined cannot be fitted by Formula (1a-1) and, therefore, does not abide by the MA law. Thus, it is necessary to focus on particular registers in exploring this link based on the MA law.

3.4 Method to compute the distance between two registers

The average word length in clauses was calculated for each text in the corpus. The links between average word length and clause length were fitted by Formula (1a-1), allowing each text to be represented by its fitted parameters, *a* and *b* of the MA law (the values of these two parameters in all texts are shown in Appendix 4). The distributions of these two parameters among texts from each register are shown in Figure 5 using box plots. Box plots provide a graphical way to display median, quartiles, and extremes of a data set on a number line to summarize the distribution of the data. As can be seen from Figure 5, there are significant differences among the values of parameters *a* and *b* across the registers.

Correlation analysis examines possible correlations, such as direction and degree, between different phenomena. Pearson's correlation coefficient, the most widely used measure of dependence, was selected to compute the correlation direction and degree between parameters *a* and *b* of the MA law both within each register and across registers. Different values of the correlation coefficient indicate different directions and degrees of relevancy between the two variables. In the extreme case, a correlation coefficient value of 1 (or −1) indicates a perfectly linear positive (or negative) correlation between them. The closer the coefficient is to either −1 or 1, the stronger the correlation is between the two variables.

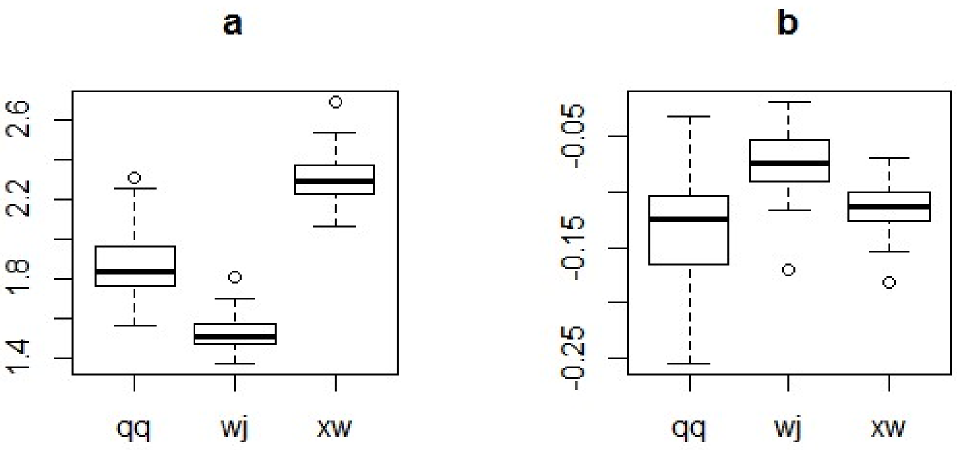


Figure 5: The distribution of fitted parameters, *a* and *b*, in the texts from different registers ("qq" refers to *TV Conversation*, "wj" refers to *Sitcom Conversation*, "xw" refers to *News Broadcasting*)

For texts across registers, the correlation coefficient between *a* and *b* is −0.634, which shows a negative correlation between them. The smooth trend line in Figure 6 shows that there is no regular functional relationship between parameters, *b* and *a*, across registers, although they are negatively correlated.

The correlation coefficients between the parameters are −0.870, −0.983, and −0.917 for texts in the *News Broadcasting*, *TV Conversation*, and *Sitcom Conversation* registers, respectively. The strong negative correlation between the parameters can be fitted by linear regression in each register. Kelih (2010) also proposed that there is a functional correlation between *a* and *b* of the MA law. On the basis of that interpretation, Köhler predicted that the borderline case forms a straight line (according to Kelih 2010).

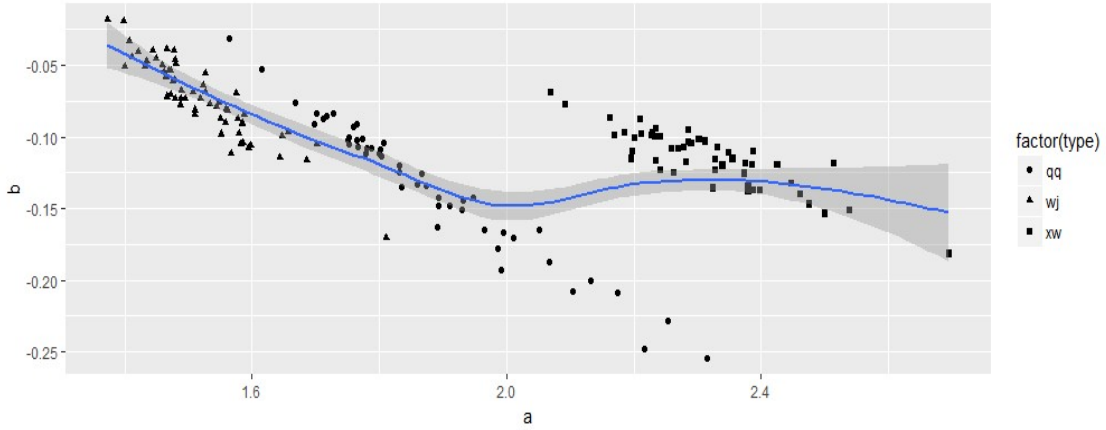


Figure 6: The negative correlation between parameters *b* and *a* across various registers ("qq", "wj", and "xw" refer to *TV Conversation*, *Sitcom Conversation*, and *News Broadcasting*, respectively)

Figure 6 shows that there are obvious boundaries among the texts from each register. In particular, the distance between the *News Broadcasting* texts and other register texts is large. The *Sitcom Conversation* and *TV Conversation* texts are close together, but far from the *News Broadcasting* texts, reflecting their different degrees of formality. From Figure 6, we also observe that parameter *b* is strongly negatively correlated with parameter *a* in each register.

Linear regression, realized by function *lm*() in R, was used to fit the functional link between these two parameters in each register. The fitted results are shown in Table 3 and Figure 7. The values of R^2 show that the fitted results are good and that there is negative linear relationship between parameters, *b* and *a*, of the MA law in each register.

Table 3

Fitted results of the relationship between parameters *b* and *a* of the MA law in each register

	Slope	*intercept*	R^2	*a*-intercept
TV Conversation	−0.288	0.405	96.53%	1.408
Sitcom Conversation	−0.304	0.389	84.01%	1.281
News Broadcasting	−0.153	0.238	75.69%	1.561

As mentioned in section 2, *News Broadcasting* is the most formal register whereas *Sitcom Conversation* is the most informal. In Table 3, for each register, the intercept is the value of the intersection of the fitted line with the *b*-axis. The *a*-axis intercept of the fitted line is obtained when *b* is equal to 0. The *a*-axis intercepts are 1.561, 1.408 and 1.281 in *News Broadcasting*, *TV Conversation*, and *Sitcom Conversation* respectively. It can be seen that the order of these values from large to small is consistent with the formality rank of the corresponding registers from formal to informal.

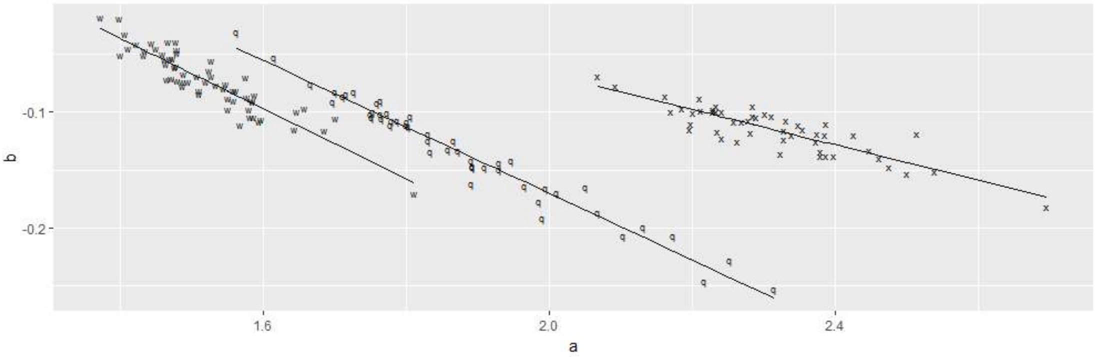


Figure 7: Regression line between fitted parameters, *b* and *a*, in each register ("*q*", "*w*", and "*x*" represent *TV Conversation*, *Sitcom Conversation*, and *News Broadcasting*, respectively)

We propose that the *a*-axis intercept can be used as an index to evaluate the formality degree of the register. For example, the formality degree of the *News Broadcasting* register is 1.561, and it is the most formal of the three registers. The distance between two registers can be quantified using the difference between their formality degrees, i.e., the *a*-axis intercepts of their fitted lines. For example, the distance between *News Broadcasting* and *TV Conversation* is 0.153, with the former register more formal than the latter.

3.5 Test of Hypothesis

We aim to test the following three hypotheses: (1) that the link between average word length and clause length abides by the MA law; (2) that there is a linear relationship between the fitted parameters, *a* and *b*, in each register; and (3) that the *a*-axis intercepts of the fitted lines can be used to represent the formality degree of Chinese registers and to quantify the distances between two registers. The Lancaster Corpus of Mandarin Chinese (LCMC) was used to verify the above conclusions. A summary of the LCMC corpus is presented in Table 5 (McEnery and Xiao)..

Table 5

Text type and number in the LCMC

Text type	Text Number	Text type	Text Number
Press reportage (A)	44	Academic prose (J)	80
Press editorial (B)	27	General fiction (K)	29
Press reviews (C)	17	Mystery/detective fiction (L)	24
Religious writing (D)	17	Science fiction (M)	6
Instructional writing (E)	38	Adventure fiction (N)	29
Popular lore (F)	44	Romantic fiction (P)	29
Biographies/essays (G)	77	Humor (R)	9
Official documents (H)	30		

We selected texts from the press reportage (A), press editorial (B), press reviews (C), official documents (H), academic prose (J), general fiction (K), science fiction (M), and adventure fiction (N) text types in LCMC. Texts from the press editorial and press reviews represent the *Press Editorials* register. Texts from general, adventure, and science fiction represent the *Fiction* register. Texts from academic prose represent the *Science* register. These registers are chosen for their variety in formality and also in terms of differences in media and modes of communication.

The cumulative relative frequencies of clause lengths, shown in Figure 8, indicate that 96% of clauses in the *Fiction* register, in the *Press Reportage* and *Press Editorials* registers, and in the *Officialese* and *Science* registers contain up to 12, 15, and 18 words, respectively.

As can be seen from Figure 9, the average word length decreases with clause length, except when the clause is very long. The average word length distributions are shown in Appendix 5. Figure 8 shows that these long clauses account for a very small proportion of clauses. We therefore infer that there is an inverse relationship between average word length and clause length.

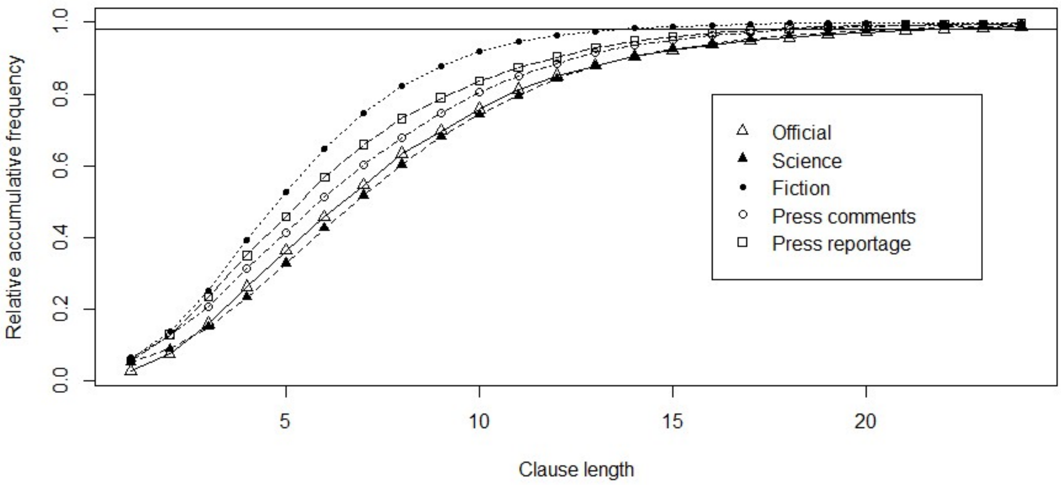


Figure 8. Cumulative relative frequencies of clause length in terms of words

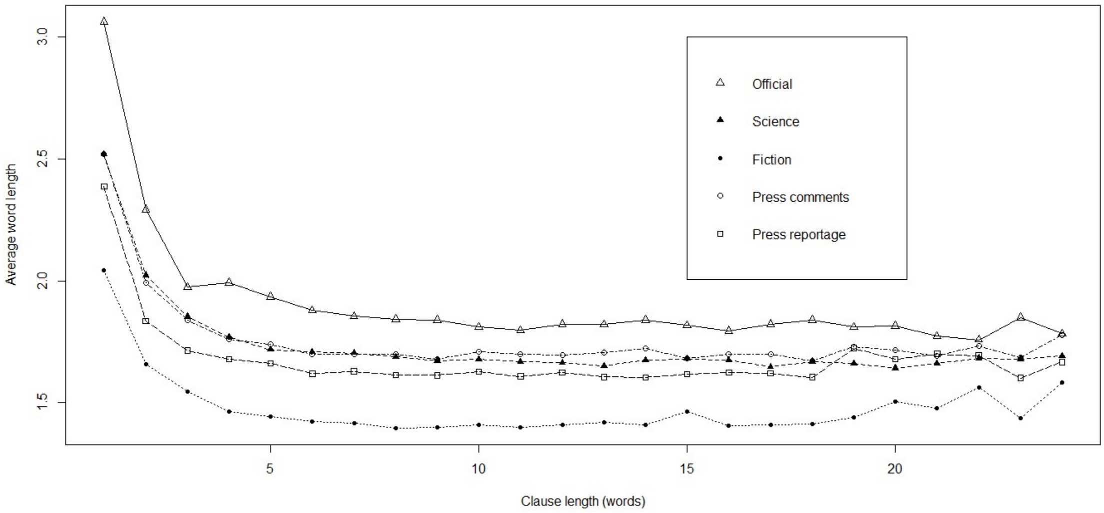


Figure 9. Distribution of average word length in clauses

Table 6

Fitted parameters of average word length distributions

	a	b	R^2	p-value
Officialese	2.697	−0.184	83.92%	2.847×10^{-5}
Science	2.295	−0.149	85.96%	1.430×10^{-5}
Fiction	1.869	−0.136	84.40%	2.437×10^{-5}
Press Editorials	2.266	−0.139	80.38%	7.825×10^{-5}
Press Reportage	2.117	−0.129	75.92%	2.228×10^{-4}

Formula (1a-1) was used to fit the average word length distribution for the texts from each of these five registers. The range of clause length was set to be 1:12. The fitted results are shown in

Table 6. The R^2 values demonstrate that the fitted results are good and the *p*-values indicate that the inverse relationships are significant. Thus, the link between average word length and clause length for the texts from each of these five registers abides by the MA law.

Next, pairs of texts in each register were merged to form a single text in the corpus — this was done because the numbers of clauses in the original texts were not enough to assess the clause frequencies of certain lengths. The average word length in clauses was calculated in this corpus. The relationships between average word length and clause length were fitted by Formula (1a-1). The texts were represented by the fitted parameters *a* and *b*, whose values are shown in Appendix 6.

Similar to section 3.3, linear regression was used to determine the systematic correlation between these two parameters, *b* and *a*, in each register. The fitted results are shown in Table 7 and the regression lines are shown in Figure 10.

Table 7
Fitted parameters of the function between parameter *b* and *a* in each register

	Slope	*b-intercept*	R^2	*a*-intercept
Officialese	−0.149	0.234	96.79%	1.570
Science	−0.189	0.281	86.62%	1.487
Fiction	−0.250	0.332	79.84%	1.328
Press Editorials	−0.238	0.401	80.36%	1.685
Press Reportage	−0.189	0.275	81.81%	1.455

The *a*-intercepts of fitted lines were calculated, which are 1.328, 1.455, 1.487, 1.570, and 1.685 in the *Fiction*, *Press Reportage*, *Science*, *Officialese*, and *Press Editorials* registers respectively, as shown in Table 7. These numbers show that the formality degree increases from *Fiction* to *Press editorials*. Hence, the *a*-intercept can be used as an index to represent the formality degree of a register and to quantify the distance between two registers. For example, the distances between *Press reportage* and *Fiction*, and between *Press reportage* and *Science* are 0.127 and -0.032 respectively. Hence, we can say that *Press reportage* is closer to *Science* than to *Fiction* in terms of formality degree and *Press reportage* is more formal than *Fiction*, while *Press reportage* is less formal than *Science*. This is consistent with our intuitive experience.

Table 8
Formality Grouping of Registers according to *a*-intercept

Formality	Register	*a*-intercept
Informal	*Sitcom Conversation*	1.281
	Fiction	1.328
	TV Conversation	1.408
Semi-formal	*Press Reportage*	1.455
	Science	1.487
High-formal	*News Broadcasting*	1.561
	Officialese	1.570
	Press Editorials	1.685

As stated in section 3.3, the *a*-axis intercepts of the regression lines are 1.281, 1.408, and 1.561 in the *Sitcom Conversation*, *TV Conversation*, and *News Broadcasting* registers respectively. Combining two studies covering eight registers from different sources, we have the following result based on *a*-intercept, as in Table 8.

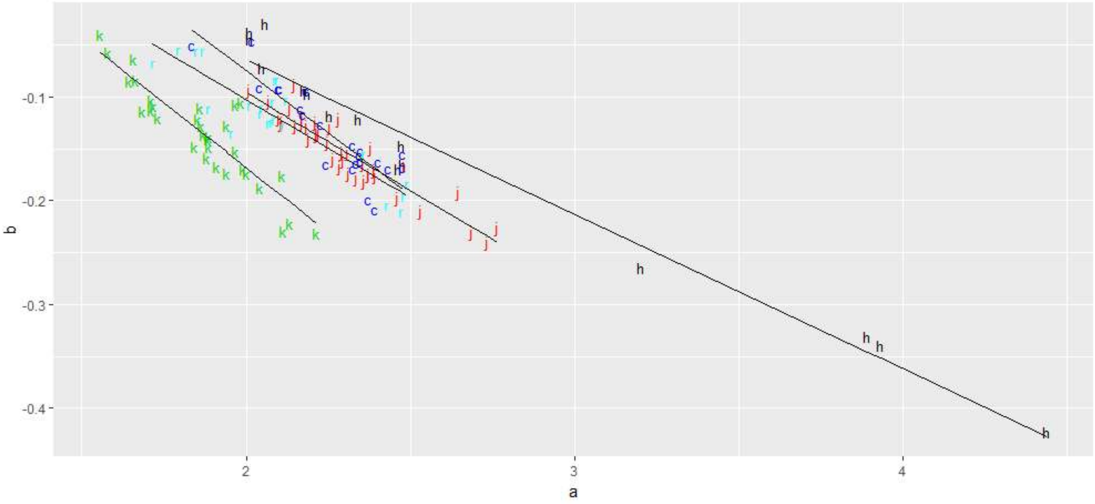


Figure 10- The regression lines for the link between *b* and *a* in each register ("h" represents *Officialese*, "j" represents *Science*, "k" represents *Fiction*, "c"represent *News Comments*, "r" represent *News Reports*)

It is interesting to observe the three clusters formed according to *a*-intercept values can be characterized by differences in degree of formality in terms of *informal*, *semi-formal* and *high-formal*. In addition, the nature of these three clusters can also be attributed to different modes of communication. The three informal registers all involve dialogue or descriptive style and could involve more than one speaker. This analysis supports the theoretical view that fictions are dialogues between the author and the reader (Bakhtin 1981). As the distributional analysis we undertake here does not consider turns and different speakers, what we capture is the planning of each text in response to and expecting responses from the other dialogue partner. This is where fiction writing is similar to the conversation and dialogue. The two semi-formal registers are conveying information with specific target audience: either to persuade (*Science*) or to inform (*Press Reportage*). In other words, although there is no direct dialogue, the speakers are aware of needs to persuade/inform when they plan their speech. The three high-formal registers involve pronouncement. I.e. the speaker is making a statement that is expected to be taken for granted. This is clear for *Officialese*, and *Press Editorials* (as newspaper editorials are considered as formal policy statement by the government in China). The somewhat surprising member of this group is *News Broadcasting*. We consider that there are two important characteristics to differentiate it from *Press Reportage*. On one hand, the person delivering *News Broadcasting* is typically different from the one who wrote it. Hence the nature of the text become strongly pronouncement. In addition, in the context where a text/speech is planned with the audience in mind, it requires time for a listener/reader to think and respond. This is not possible for *News Broadcasting* as the news broadcasting is continuous. Hence it is strictly a one-way communication with minimal influence of the addressee on the planning. This dialogic interpretation is also consistent with Biber's (1986) study showing that *Fiction* is closer to conversation than to either academic prose or planned speeches. It is also important to note that the degree of formality of register does not correspond to word length or clause/sentence averages reported earlier in this paper.

In LCMC, the number of texts in each register differs. This may affect the linear regression analysis between parameters *a* and *b*. In future studies, this factor should be considered and the number of texts from each register should be as similar as possible.

4 Conclusion

Quantitative linguistics treats languages as self-organizing and self-regulating systems. Synergetic linguistics holds that there are interrelated relationships among the various language levels (Köhler 1984, 2005). As an important law, the MA law explores the relationship between a language construct and its immediate components. This paper examined degrees of formality of register and the distance between two registers based on the MA law from the perspective of quantitative linguistics and regression analysis.

News Broadcasting, *Sitcom Conversation*, and *TV Conversation* texts were selected to form a corpus for this preliminary study. The results show that, as predicted by MA law, average word length decreases as the increase of clause length for most clauses. The logarithm of average word length distributions can be fitted by the Formula (1a-1). The fitting results shown that, for the texts from each register, the relationship between clauses and their constituent words abides by the MA law.

All the texts were represented by their corresponding fitted parameters, *a* and *b*, obtained from Formula (1a-1). There were obvious boundaries between the texts from various registers. The functional correlation between these two parameters, *a* and *b*, was fitted by linear regression in each register. Analysis indicates that the *a*-intercept can be used as an index to represent the formality degree of the register and to quantify the distances between two registers. The *News Broadcasting* register is more formal than both the *TV Conversation* and *Sitcom Conversation* registers. The same experiments were carried out on texts from 6 additional registers from LCMC, and confirmed the validity of using *a*-intercept to represent the formality degrees of registers and to quantify the distance between two registers.

In addition, by combing the results of two studies, we show that the *a*-intercept values of the 8 registers can be group into three clusters corresponding to *informal*, *semi-formal*, and *high-formal* registers. We further show that the three clusters correspond to three different modes of communication: dialogic (and informal), informative/persuasive (with targeted audience and semi-formal), and pronouncement (and high-formal). This is consistent with Hou et al.'s (under review) result showing that the average word length differences in different genres can be explained by cost of planning, where more interactive genres require more planning and hence shorter units.

In sum, we propose *a*-intercept as an effective index to represent the degrees of formality of a register and to quantify the distances between various registers based on the MA law and regression analysis. In addition, we show that the range of the *a*-intercept can be attribute to the modes of communication typical of each register. Thus our study further developed and formally realized Biber's (1994) claim that registers are varieties in a continuum which may still be analytically identified as different categories.

REFERENCES

Altmann, G. (1980). Prolegomena to Menzerath´s law. *Glottometrika 2, 1-10.*

Bakhtin, M.M. (1981). Discourse in the Novel. In: *The Dialogic Imagination: Four Essays* (Vol. 1). 262-349. Austin: University of Texas Press.

Benešová, M., & Čech, R. (2015). Menzerath-Altmann Law Versus Random Model. In: G.K. Mikros & J. Mačutek (Eds), *Sequences in Language and Text (pp. 57-69).* Berlin/Boston: de Gruyter.

Benešová, M. (2016). *Text segmentation for Menzerath-Altmann law testing*. Palacký University, Faculty of Arts.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language 62:384–414.*

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1994). An analytical framework for register studies. *Sociolinguistic perspectives on register*, 31-56.

Biber, D. (1995). On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson. *Text – Interdisciplinary Journal for the Study of Discourse*, *15*(3), 341-370

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory, 8(1), 9-37.*

Cacoullos, R. T. (1999). Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. *Language variation and change, 11(2), 143-170.*

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chen, H.H. (1994). The contextual analysis of Chinese sentences with punctuation marks. *Literary and linguistic computing, 9(4): 281-289.*

Chen, H & H Liu (2016) How to Measure Word Length in Spoken and Written Chinese, *Journal of Quantitative Linguistics, 23:1, 5-29.*

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In: B.-S. Park and J.B. Kim. (Eds.) *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation. Seoul:Kyung Hee University. pp. 167-176.*

Chen, Keh-Jiann, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, and Zhao-Ming Gao. (2003). Sinica Treebank: Design Criteria, Representational Issues and Implementation. In: Anne Abeillé (Ed.), *Treebanks: Building* *and Using Parsed Corpora (pp. 231-248).* Dordrecht; Boston: Kluwer Academic Publishers.

Conway, D., & White, J. (2013). *Machine learning for hackers*. (Chen, Kaijiang, Yizhe Liu & Xiaonan, Meng, Trans). Beijing. China: China Machine Press.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics, 12, 41–52.*

Eroglu, S. (2014). Menzerath-Altmann law: Statistical mechanical interpretation as applied to a linguistic organization. *Journal of Statistical Physics, 157(2) 392-405.*

Feng, S. (2010). On mechanisms of register system and its grammatical property. *Studies of the Chinese Language, 5, 400–412.*

Grzybek, P. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In: P. Grzybek and E. Stadlober (Eds.), *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday, 62, 205.*

Hammerl, R., & Sambor, J. (1993). *O statystycznych prawach jezykowych*. Warszawa: Zakład Semiotyki Logicznej Uniwersytetu Warszawskiego.

Hou, R., Chu-Ren Huang and Yat-mei Lee. (2018). Linguistic Characteristics of Chinese Register Based on the Menzerath – Altmann Law and Text Clustering. (Under review).

Hou, R., Yang, J., & Jiang, M. (2014). A Study on Chinese Quantitative Stylistic Features and Relation Among Different Styles Based on Text Clustering. *Journal of Quantitative Linguistics, 21(3), 246-280.*

Hou, R, Chu-Ren Huang, Hue San Do & Hongchao Liu (2017): A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law, *Journal of Quantitative Linguistics. 24(4): 350-366.*

Hou, R., Huang, C., & Liu, H. (2017). A study on Chinese register characteristics based on regression analysis and text clustering. *Corpus Linguistics and Linguistic Theory*, (Online). doi:10.1515/cllt-2016-006

Huang, B. & Liao, X. (2002). *Modern Chinese*. Beijing: High Education Press.

Huang, Chu-Ren and Shi, D. (2016). *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press.

Huang, C.-R. & K.-J. Chen. (2017). Sinica Treebank. In: N. Ide and J. Pustejovsky (eds), *Handbook of Linguistic* Annotation. Berlin & Heidelberg: Springer.

Hřebíček, L. (1992). *Text in communication: Supra-sentence structure*. Bochum, Brockmeyer.

Hřebíček, L. (1995). *Text levels: Language constructs, constituents and Menzerath-Altmann law.* Trier: WVT.

Hřebíček, L. (1997). *Lectures on text theory.* Prague: Academy of Sciences of the Czech Republic, Oriental Institute.

Kelih, E. (2010). Parameter interpretation of Menzerath's Law: Evidence from Serbian. In P. Grzybek, E. Kelih & J. Mačutek (Eds.): *Text and Language, Structures, Functions, Interrelations, Quantitative Perspectives (pp. 71–78).* Wien: Praesens.

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In: W. Lehfeldt & U. Strauss (Eds.), *Glottometrika 4 (pp. 103 – 113*). Bochum: Brockmeyer.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In: W. Lehfeldt & U. Straus (Eds.), *Glottometrika 6, 177-183*. Bochum: Brockmeyer.

Köhler, R. (1989). Das Menzerathschen Gesetz als Resultat des Sprachverarbeitungsmechanismus. In: Altmann, Schwibbe (1989): *108-112*.

Köhler, R. (2005). Synergetic Linguistics. In: R. Köhler, G. Altmann & R.G. Piotrowski (eds.). *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook*. Berlin, New York: Walter de Gruyter, 760-775.

Köhler, R. (2012). *Quantitative syntax analy*sis (Vol. 65). Berlin: Walter de Gruyter.

Kułacka, A. (2010). The coefficients in the formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, *17*(4), 257-268.

Lv, S. (1992). Studies on Chinese grammar through comparison. *Foreign Language Teaching and Research*. (2).

McEnery, A. & R. Xiao. (2004). The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In_ M. Lino, M. Xavier, F. Ferreire, R. Costa, R. Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004, pp. 1175–1178.* Lisbon, May 24 –30, 2004.

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes* (Vol. 3). F. Dümmler.

Motalová, T., Spáčilová, L., Benešová, B., Kučera, O. (2014). *An application of Menzerath-Altmann law to contemporary written Chinese*. Křížkovského, Olomouc: Univerzita Palackého v Olomouci.

Popescu, I.-I., Mačutek, J., & Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Ščigulinská, J. & Schusterová, D. (2014). *An Application of the Menzerath-Altmann Law to Contemporary Spoken Chinese.* Palacký University in Olomouc.

Tuldava, J. (1995). Informational measures of causality. *Journal of Quantitative Linguistics, 2(1), 11-14.*

Wang, K., & Qin, H. (2014). What is peculiar to translational Mandarin Chinese? A corpus-based study of Chinese constructions' load capacity. *Corpus Linguistics and Linguistic Theory, 10(1), 57-77.*

Wilson, A. (2017) Units and Constituency in Prosodic Analysis: A Quantitative Assessment, *Journal of Quantitative Linguistics, 24:2-3, 163-177.*

Yuan, H. and Li, X. (2005*). Outline of Chinese Register*. China, Beijing: The Commercial Press.

Zhang, Z. S. (2012). A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory, 8(1), 209-240.*

Zhu, D. (1982). *Lectures on Grammar.* Beijing, China: Commercial Press.

Appendix

Appendix 1:
The occurrence frequencies of clauses with certain lengths
(raw numbers)

Clause length	*TV Conversation*	*Sitcom Conversation*	*News Broadcasting*
1	*2068*	*7963*	1743
2	*3687*	*5884*	3446
3	5652	6177	3514
4	7445	6843	4020
5	7843	6704	4507
6	7294	6160	4588
7	6138	4997	4492
8	4851	3707	4260
9	3583	2907	3735
10	2593	2105	3378
11	1800	1443	2854
12	1340	993	2279
13	874	739	1821
14	594	494	1516
15	405	337	1207
16	263	281	865
17	182	183	693
18	102	137	579
19	68	90	432
20	48	65	295
21	34	52	258
22	19	38	192
23	15	37	132
24	16	25	111
25	6	21	83

Appendix 2

he relative frequency distributions of clause length (for Figure 1)

Clause length	*TV Conversation*	*Sitcom Conversation*	*News Broadcasting*
1	0.036328	0.136194	0.033945
2	0.064768	0.100636	0.067111
3	0.099287	0.105648	0.068435
4	0.130784	0.117038	0.078289
5	0.137775	0.114661	0.087774
6	0.128131	0.105357	0.089351
7	0.107824	0.085466	0.087481
8	0.085216	0.063402	0.082963
9	0.062941	0.049720	0.072739
10	0.045550	0.036003	0.065786
11	0.031620	0.024680	0.055582
12	0.023539	0.016984	0.044383
13	0.015353	0.012639	0.035464
14	0.010435	0.008449	0.029524
15	0.007114	0.005764	0.023506
16	0.004620	0.004806	0.016846
17	0.003197	0.003130	0.013496
18	0.001792	0.002343	0.011276
19	0.001195	0.001539	0.008413
20	0.000843	0.001112	0.005745
21	0.000597	0.000889	0.005025
22	0.000334	0.00065	0.003739
23	0.000263	0.000633	0.002571
24	0.000281	0.000428	0.002162
25	0.000105	0.000359	0.001616

Appendix 3

：Average word length distribution in clauses (for Figure 3)

	TV Conversation	*Sitcom Conversation*	*News Broadcasting*	*Whole*
1	1.957447	1.489263	2.530694	1.725667
2	1.635476	1.502039	2.151045	1.711646
3	1.55585	1.39728	1.916809	1.574681
4	1.493519	1.357555	1.885137	1.52869
5	1.476756	1.335561	1.850189	1.515409

6	1.460356	1.324378	1.826431	1.507021
7	1.453102	1.310958	1.810234	1.510307
8	1.452098	1.310898	1.792165	1.524282
9	1.442243	1.311318	1.782032	1.529139
10	1.446317	1.310309	1.773475	1.547709
11	1.44298	1.308574	1.767185	1.56293
12	1.451244	1.305975	1.758556	1.571824
13	1.44288	1.310086	1.759811	1.582366
14	1.441799	1.307981	1.758575	1.600834
15	1.424033	1.309397	1.764154	1.614845
16	1.44249	1.301601	1.759176	1.608809
17	1.446671	1.303439	1.769544	1.633382
18	1.412854	1.281833	1.780944	1.651453
19	1.452012	1.319883	1.790205	1.679483
20	1.439583	1.296923	1.785254	1.666789
21	1.439776	1.320513	1.777224	1.674834
22	1.425837	1.327751	1.812973	1.709383
23	1.457971	1.347826	1.816535	1.693053
24	1.484375	1.266667	1.850601	1.716009
25	1.473333	1.367619	1.883855	1.762909

Appendix 4

Fitted parameters of average word length distribution in clauses (for Figure 5, 6 and 7. “qq”, “wj”, and “xw” refer to *TV Conversation*, *Sitcom Conversation*, and *News Broadcasting*, respectively)

Files	*a*	*B*
qq01.txt	2.067773	-0.18753
qq02.txt	2.174008	-0.20846
qq03.txt	1.947793	-0.14294
qq04.txt	1.807163	-0.10454
qq05.txt	1.751832	-0.10506
qq06.txt	1.764547	-0.09116
qq07.txt	1.779792	-0.10791
qq08.txt	1.858832	-0.13347
qq09.txt	1.753004	-0.10043
qq10.txt	1.892101	-0.14266
qq11.txt	1.893699	-0.14804
qq12.txt	2.05125	-0.16539
qq13.txt	1.931095	-0.14453
qq14.txt	2.217134	-0.24779

qq15.txt	2.10442	-0.20811
qq16.txt	1.990768	-0.19279
qq17.txt	1.995214	-0.1665
qq18.txt	1.727310	-0.08338
qq19.txt	1.759958	-0.09278
qq20.txt	2.132648	-0.20043
qq21.txt	1.802140	-0.10913
qq22.txt	1.831594	-0.12511
qq23.txt	1.615169	-0.05312
qq24.txt	1.788015	-0.10808
qq25.txt	1.831414	-0.11988
qq26.txt	1.872961	-0.13428
qq27.txt	1.929761	-0.15053
qq28.txt	1.803591	-0.11334
qq29.txt	1.91029	-0.14825
qq30.txt	1.698243	-0.09146
qq31.txt	1.986195	-0.17809
qq32.txt	1.764805	-0.10245
qq33.txt	2.314057	-0.25443
qq34.txt	2.011485	-0.1705
qq35.txt	1.774028	-0.10153
qq36.txt	2.253452	-0.22888
qq37.txt	1.800376	-0.11194
qq38.txt	1.965715	-0.16464
qq39.txt	1.867041	-0.12519
qq40.txt	1.716586	-0.08507
qq41.txt	1.834335	-0.13484
qq42.txt	1.750414	-0.10254
qq43.txt	1.777919	-0.11176
qq44.txt	1.667633	-0.07651
qq45.txt	1.711596	-0.08762
qq46.txt	1.701851	-0.08325
qq47.txt	1.76669	-0.10654
qq48.txt	1.563749	-0.03122
qq49.txt	1.893359	-0.1482
qq50.txt	1.892036	-0.16296
wj01.txt	1.701541	-0.10547
wj02.txt	1.579630	-0.0974
wj03.txt	1.567663	-0.11157
wj04.txt	1.487902	-0.07347
wj05.txt	1.466492	-0.03893

wj06.txt	1.373107	-0.01848
wj07.txt	1.574886	-0.06957
wj08.txt	1.464686	-0.05789
wj09.txt	1.459430	-0.04956
wj10.txt	1.526858	-0.0552
wj11.txt	1.657220	-0.09628
wj12.txt	1.584129	-0.09103
wj13.txt	1.685830	-0.11615
wj14.txt	1.477337	-0.03961
wj15.txt	1.584944	-0.08975
wj16.txt	1.587453	-0.08484
wj17.txt	1.479296	-0.04599
wj18.txt	1.489070	-0.06742
wj19.txt	1.581871	-0.10483
wj20.txt	1.810669	-0.17034
wj21.txt	1.594398	-0.10822
wj22.txt	1.434462	-0.04705
wj23.txt	1.562341	-0.08129
wj24.txt	1.55812	-0.09029
wj25.txt	1.577619	-0.08739
wj26.txt	1.527094	-0.06899
wj27.txt	1.519326	-0.07362
wj28.txt	1.510108	-0.08433
wj29.txt	1.597706	-0.10607
wj30.txt	1.398341	-0.01865
wj31.txt	1.486941	-0.0775
wj32.txt	1.64755	-0.09942
wj33.txt	1.54406	-0.07909
wj34.txt	1.507677	-0.0689
wj35.txt	1.585655	-0.10438
wj36.txt	1.550824	-0.08769
wj37.txt	1.479014	-0.07302
wj38.txt	1.480225	-0.04912
wj39.txt	1.443864	-0.03998
wj40.txt	1.534121	-0.07684
wj41.txt	1.462054	-0.05437
wj42.txt	1.523679	-0.06365
wj43.txt	1.510244	-0.08121
wj44.txt	1.400162	-0.05061
wj45.txt	1.478317	-0.06013
wj46.txt	1.406906	-0.0327

wj47.txt	1.495283	-0.07339
wj48.txt	1.47248	-0.0704
wj49.txt	1.432348	-0.05111
wj50.txt	1.551323	-0.09809
wj51.txt	1.559035	-0.08117
wj52.txt	1.547542	-0.07581
wj53.txt	1.469425	-0.05357
wj54.txt	1.44971	-0.04541
wj55.txt	1.643353	-0.11486
wj56.txt	1.421602	-0.04071
wj57.txt	1.411729	-0.04461
wj58.txt	1.475764	-0.06114
wj59.txt	1.466146	-0.07185
wj60.txt	1.472642	-0.05403
xw01.txt	2.262991	-0.12554
xw02.txt	2.198158	-0.10987
xw03.txt	2.24177	-0.12304
xw04.txt	2.282072	-0.11802
xw05.txt	2.387058	-0.13759
xw06.txt	2.324207	-0.13598
xw07.txt	2.269689	-0.10807
xw08.txt	2.285678	-0.10362
xw09.txt	2.425591	-0.11979
xw10.txt	2.475266	-0.14716
xw11.txt	2.539164	-0.15114
xw12.txt	2.513899	-0.11853
xw13.txt	2.355283	-0.11542
xw14.txt	2.379863	-0.13813
xw15.txt	2.302483	-0.10163
xw16.txt	2.196296	-0.11534
xw17.txt	2.259619	-0.10839
xw18.txt	2.29023	-0.10474
xw19.txt	2.312217	-0.10316
xw20.txt	2.093065	-0.0775
xw21.txt	2.328397	-0.12352
xw22.txt	2.212437	-0.09836
xw23.txt	2.32851	-0.11559
xw24.txt	2.38001	-0.13449
xw25.txt	2.285232	-0.09528
xw26.txt	2.331219	-0.10743
xw27.txt	2.500296	-0.15373

xw28.txt	2.374066	-0.12564
xw29.txt	2.210489	-0.08788
xw30.txt	2.229068	-0.09742
xw31.txt	2.39812	-0.13752
xw32.txt	2.241518	-0.09986
xw33.txt	2.375414	-0.11892
xw34.txt	2.228828	-0.09917
xw35.txt	2.233510	-0.09978
xw36.txt	2.186077	-0.09676
xw37.txt	2.202082	-0.10072
xw38.txt	2.235197	-0.11707
xw39.txt	2.170009	-0.09935
xw40.txt	2.386215	-0.11978
xw41.txt	2.163245	-0.08660
xw42.txt	2.448241	-0.13281
xw43.txt	2.462103	-0.14008
xw44.txt	2.387655	-0.11001
xw45.txt	2.349125	-0.11095
xw46.txt	2.278891	-0.10759
xw47.txt	2.069112	-0.06881
xw48.txt	2.234222	-0.09467
xw49.txt	2.69523	-0.18178
xw50.txt	2.339337	-0.12004

Appendix 5

Average word length distribution in clauses (LCMC, for Figure 9, the average word length distributions in clauses whose range is 1:12 words were fitted.)

	Officialese	*Science*	*Fiction*	*Press Editorials*	*Press Reportage*
1	3.062147	2.517738	2.041815	2.520772	2.387789
2	2.291935	2.022654	1.65941	1.99269	1.834146
3	1.973881	1.851996	1.545702	1.837147	1.713834
4	1.991392	1.76871	1.46331	1.759375	1.678852
5	1.934568	1.717284	1.442179	1.74123	1.661885
6	1.878258	1.707954	1.424236	1.699459	1.61875
7	1.853913	1.703171	1.414539	1.697101	1.628486
8	1.841814	1.687843	1.395501	1.697993	1.614583
9	1.838235	1.671431	1.399111	1.678824	1.611985
10	1.810336	1.677444	1.408616	1.71008	1.627921
11	1.796671	1.666633	1.399324	1.699655	1.608276
12	1.821721	1.663522	1.408932	1.696912	1.624351

13	1.820926	1.650267	1.42096	1.705882	1.605604
14	1.838724	1.673993	1.40803	1.722084	1.602814
15	1.816798	1.679961	1.463043	1.680417	1.617687
16	1.794444	1.674213	1.407095	1.698138	1.623326
17	1.821238	1.646278	1.410256	1.700073	1.620098
18	1.838574	1.668022	1.412698	1.673127	1.60463
19	1.810729	1.660254	1.440000	1.730884	1.723977
20	1.815476	1.640761	1.504545	1.714706	1.677381
21	1.772109	1.660588	1.47619	1.690476	1.70000
22	1.758117	1.682497	1.563636	1.73445	1.693182
23	1.849275	1.678261	1.434783	1.68530	1.601449
24	1.783333	1.69086	1.583333	1.777778	1.666667

Appendix 6

The fitted parameters of average word length distribution in clauses (LCMC, “h” represents *Officialese*, “j” represents *Science*, ”k” represents *Fiction*, “c”represent *Press Editorials*, “r” represent *Press Reportage*)

	a	*B*
h01.txt	3.894010	−0.33009
h02.txt	3.931898	−0.33874
h03.txt	2.057789	−0.02896
h04.txt	2.011199	−0.04385
h05.txt	2.04932	−0.07152
h06.txt	2.478661	−0.16447
h07.txt	2.462683	−0.1686
h08.txt	2.256338	−0.1177
h09.txt	2.343932	−0.12074
h10.txt	2.011827	−0.03794
h11.txt	2.177471	−0.09293
h12.txt	2.185661	−0.09616
h13.txt	2.473934	−0.14672
h14.txt	3.203523	−0.26464
h15.txt	4.438227	−0.42259
j01.txt	2.256356	−0.12903
j02.txt	2.211942	−0.12504
j03.txt	2.21564	−0.13305
j04.txt	2.108875	−0.12448
j05.txt	2.191512	−0.14079
j06.txt	2.533191	−0.21083

j07.txt	2.392433	-0.17568
j08.txt	2.31026	-0.15318
j09.txt	2.373629	-0.17504
j10.txt	2.356452	-0.16436
j11.txt	2.151169	-0.12766
j12.txt	2.460924	-0.19736
j13.txt	2.764089	-0.22633
j14.txt	2.482294	-0.16497
j15.txt	2.390709	-0.1706
j16.txt	2.378474	-0.14962
j17.txt	2.2828	-0.12115
j18.txt	2.372927	-0.17545
j19.txt	2.360044	-0.18185
j20.txt	2.264953	-0.16002
j21.txt	2.099943	-0.12058
j22.txt	2.00819	-0.09356
j23.txt	2.169798	-0.11982
j24.txt	2.133982	-0.11114
j25.txt	2.183392	-0.12744
j26.txt	2.070529	-0.10486
j27.txt	2.146686	-0.08837
j28.txt	2.647627	-0.19237
j29.txt	2.335038	-0.17791
j30.txt	2.310879	-0.17349
j31.txt	2.294596	-0.16076
j32.txt	2.172026	-0.12477
j33.txt	2.733758	-0.2399
j34.txt	2.687748	-0.23046
j35.txt	2.285372	-0.16717
j36.txt	2.107397	-0.12615
j37.txt	2.219698	-0.13427
j38.txt	2.29143	-0.15376
j39.txt	2.214308	-0.13661
j40.txt	2.247157	-0.14454
k01.txt	1.657834	-0.06306
k02.txt	1.86132	-0.11084
k03.txt	1.851445	-0.12097
k04.txt	1.685644	-0.11368
k05.txt	1.731522	-0.11959
k06.txt	1.862992	-0.1279
k07.txt	1.880562	-0.15794

k08.txt	1.870938	-0.13575
k09.txt	1.88173	-0.14079
k10.txt	1.644948	-0.08455
k11.txt	1.557019	-0.0398
k12.txt	1.712411	-0.10302
k13.txt	1.885497	-0.14816
k14.txt	1.967994	-0.15197
k15.txt	1.94156	-0.12746
k16.txt	1.968872	-0.10686
k17.txt	1.984294	-0.10546
k18.txt	2.11039	-0.17604
k19.txt	2.131986	-0.22162
k20.txt	2.113132	-0.22887
k21.txt	1.664552	-0.08383
k22.txt	1.579961	-0.05674
k23.txt	1.908861	-0.16672
k24.txt	2.214675	-0.23127
k25.txt	1.939691	-0.17327
k26.txt	1.844373	-0.14724
k27.txt	1.991633	-0.16936
k28.txt	1.886768	-0.13761
k29.txt	1.718277	-0.10842
k30.txt	1.710238	-0.1129
k31.txt	2.043509	-0.18767
k32.txt	2.002826	-0.17315
nc01.txt	2.18501	-0.09482
nc02.txt	2.104087	-0.09234
nc03.txt	2.099482	-0.09271
nc04.txt	2.172759	-0.1178
nc05.txt	2.350765	-0.16227
nc06.txt	2.326585	-0.16922
nc07.txt	2.244471	-0.16514
nc08.txt	2.372331	-0.20013
nc09.txt	2.39383	-0.20885
nc10.txt	2.335363	-0.16448
nc11.txt	2.350964	-0.15258
nc12.txt	2.402742	-0.16292
nc13.txt	2.346136	-0.15682
nc14.txt	2.16682	-0.11142
nc15.txt	2.017009	-0.04722
nc16.txt	2.478327	-0.15619

nc17.txt	2.477328	-0.16893
nc18.txt	2.43489	-0.16924
nc19.txt	2.22708	-0.1273
nc20.txt	2.326956	-0.14678
nc21.txt	2.042853	-0.09147
nc22.txt	1.835245	-0.05065
nr01.txt	1.955519	-0.13577
nr02.txt	2.428588	-0.20546
nr03.txt	2.488269	-0.18487
nr04.txt	2.082228	-0.12232
nr05.txt	2.009029	-0.10886
nr06.txt	1.791718	-0.05526
nr07.txt	1.715746	-0.06794
nr08.txt	1.884986	-0.11167
nr09.txt	2.078765	-0.10581
nr10.txt	2.091409	-0.08449
nr11.txt	2.084471	-0.08436
nr12.txt	2.118468	-0.10367
nr13.txt	2.077391	-0.12496
nr14.txt	2.042974	-0.11603
nr15.txt	2.045316	-0.11082
nr16.txt	2.065749	-0.12572
nr17.txt	2.110759	-0.12921
nr18.txt	1.844938	-0.05597
nr19.txt	1.86599	-0.05585
nr20.txt	2.35474	-0.15559
nr21.txt	2.478857	-0.19649
nr22.txt	2.470704	-0.21101

Glottometrics 45, 2019, 57-62

The Classification of English Styles on the Basis of Lexical Parameters: A Case of Clustering Analysis

Hanna Gnatchuk[1]

Abstract: The present article is an attempt to reveal the groups of the most similar and dissimilar English styles (or genres) on the basis of three factors (variables): their average word repeat, hapax legomenas and the number of unique words. We intend here to perform a clustering analysis, which is grounded on the Euclidean distance matrix. In this research we have determined the number of clusters (= the groups) in which English styles can be divided. The results have been explained, considering Elbowplot and Dendrogram. The necessary calculations have been done in Programs R-Studio and Python.

Key words: A*gglomerative clustering analysis, styles/genres, stylistics, Euclidean distance, ward method, average silhouette means, multiscale bootstrap resampling method.*

1. Introduction: Some notes on stylistics and (functional) styles

According to Galperin (1981), stylistics refers to the branch of general linguistics, which fulfils a two-fold function. Firstly, it studies the inventory of the language media, which can have a certain impact on the audience. Secondly, it deals with certain text types (discourse), characteristic of a particular selection and organization of language means. One is able to make an analysis of the types of texts if a particular set of components is available in their interaction. In such a way, if the text types are distinguished in terms of a pragmatic aspect of the communication, they are called functional styles of the language.

Two important notions dealing with the functional styles are stylistic devices (SDs) and expressive means (EMs). They are the main objects of stylistic investigations. They are known to provide the desirable effect of the speech on a speaker. SDs and EMs deal with the following problems: the search for synonyms for designating the same notion or the same thought, a particular manner of a writer to use his/her language and the aesthetic function of the language. Moreover, the functional styles are the main objects of linguistic studies. The key issues touch upon the varieties of language – oral and written variants, the elements of texts which are higher than sentences.

It is worth mentioning that functional style has been susceptible to some changes, especially to chronological ones (from one period to another). Therefore, it is possible to refer it to a historical category. Galperin supports this statement by giving the example of emotive prose, which began to exist only in the second half of the 16th century; the newspaper style separated from the publicistic style and the oratory style faced enormous changes. These changes are often determined by social conditions, scientific progress or the development of social life in the country. As an example one can consider in the language the emotive components, which were to be found in the 18th century in the style of a scientific prose. The reason for it is a lack of scientific data which must be obtained by a thorough study. The development of science led to the compilation of the scientific data and this gave a way to

[1] Universität Klagenfurt; agnatchuk@gmail.com

arguments and evident facts. In such a way, a considerable number of English styles have been developed throughout centuries from the English language. The objectives of stylistic research can be studied in combination with other disciplines, such as theory of information, logic, psychology, statistics and literature. Nowadays, no science is isolated and borrows the necessary techniques or knowledge from other branches. This provides us with the effective study of different linguistic problems.

2. Empirical part of the research: clustering analysis

In stylistics, the lexical factors are considered to play a crucial role in the classification of styles. Different stylistic devices and means as well as stylistic differentiation of vocabulary are taken into account in the process of characterizing text properties. At the present stage it would be of great interest to reveal the groups of styles, which are similar according to 3 variables (parameters): *the average word repeat, hapax legomenas (the number of words occurring only once in a text) and the number of unique words or word types (the total words' counts without considering their repeats).*

The classification of styles (or genres) was taken in this research from the Brown Corpus, which is available in the corpus of the Python Program. "This corpus contains texts from 500 sources, and the sources have been categorized by genres" (Bird et al., 2009:42). In general, one distinguishes 15 genres in Brown Corpus: *adventure, belles-lettres, editorial, fiction, government, hobbies, humor, learned, lore, mystery, news, religion, reviews, romance and science-fiction.* In such a way, we shall analyse these 15 styles in terms of the above-mentioned three lexical factors. The values for each factor (lexical richness, the number of hapax legomenas as well as word types) and for each genre are illustrated in Table 1. All the values have been computed in the program for natural language processing – Python.

Table 1:
The values of three variables for each genre

	Genres/styles	Lexical richness	Hapax legomena	Word types
1	adventure	7.81	4933	8874
2	belles-lettres	9.39	9491	18428
3	editorial	6.22	5534	9890
4	fiction	7.36	5251	9302
5	government	8.57	3824	8181
6	hobbies	6.89	6356	11935
7	humor	4.32	3397	5017
8	learned	10.78	7982	16859
9	lore	7.6	7733	14503
10	mystery	8.18	3779	6982
11	news	6.98	7737	14394
12	religion	6.18	3635	6373
13	reviews	4.71	5339	8626
14	romance	8.28	4695	8452
15	science-fiction	4.47	2039	3233

The aim of our research is to detect the groups of the most similar genres and unite them in corresponding clusters, considering the lexical richness, the number of hapax legomenas and the number of unique words. Before tackling this task, one must be aware of the most important principles of clustering analysis.

Therefore, it would be relevant here to have a look at the aim and the procedure of the clustering analysis. According to Levshina (2015:306), the aim of the cluster analysis is "to help you discover groups of similar objects in the data". In our case, the objects are English styles. Moreover, Bortz et al. (2010:453) considers that the clustering analysis enables us to group the objects in such a way that the difference between objects in a group (=cluster) is minimal and the difference between groups (or clusters) is maximal. In our case we shall deal with the hierarchical clustering analysis. The procedure of it consists of 4 steps, described by R. Hatzinger et al. (2011: 420):

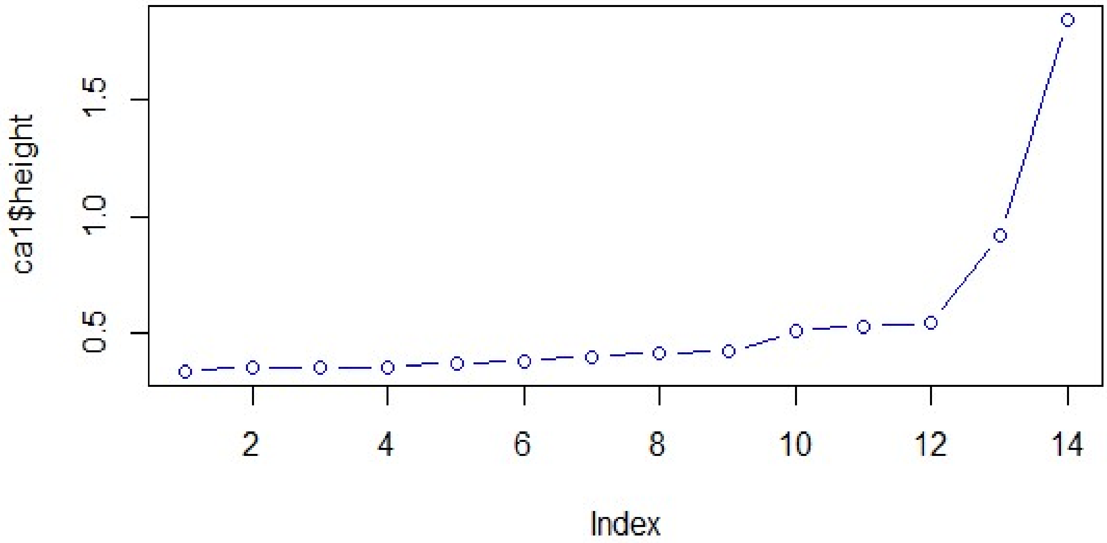
1) **Step 1**: Each observation is a cluster. Observations correspond to a style. One deals here with the distances between styles. It is worth mentioning the notion of distances. The aim of any distance is to show "how (dis)similar the constructions are with regard to the proportion of values of the variables" (Levshina, 2015 : 306). If they are similar, then the distance is small. In contrast, if the proportions of values are dissimilar, than the distance is large. In the present research we shall deal with the Euclidean distance computed according to Formula 1.1:

$$d_E(x, y) = \sqrt{\sum(x_i - y_i)^2} \qquad (1.1)$$

"The distance between two vectors (x_i and y_i) is the square root of summed squared between all pairs of numbers in the vectors"(Levshina, 2015:307)
2) **Step 2:** The fusion of the two clusters (=groups) which are the nearest/closest/the most similar;
3) **Step 3**: The calculation of the distance of a newly formed cluster to other clusters;
4) **Step 4:** One repeats step 2 and step 3 as many times till one obtains one cluster, which includes all observations (styles).

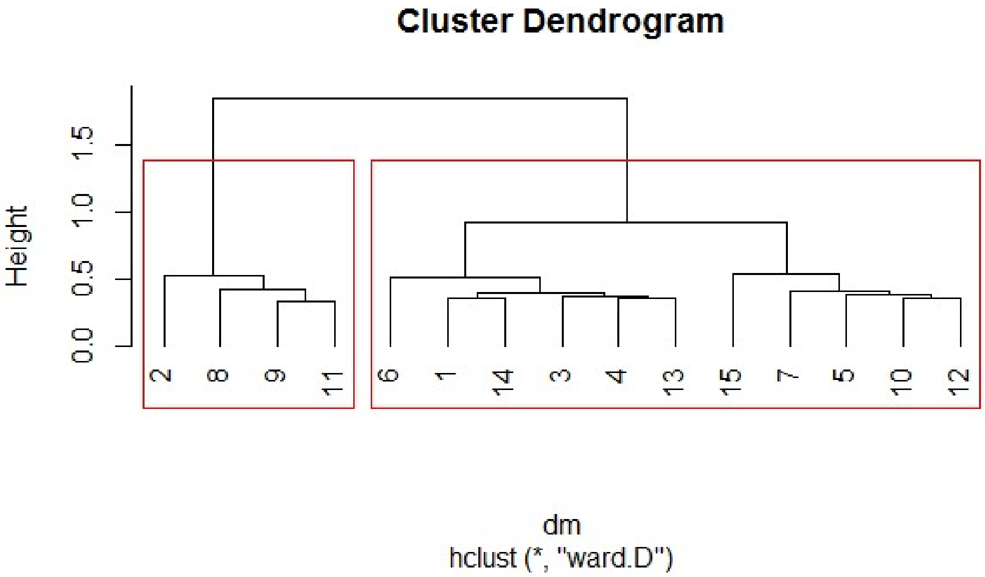
These steps are the procedures of agglomerative clustering. Graphically it shows all styles as branches of a tree (see Dendrogram 1). The clustering tree or dendrogram shows that "each object represents its own cluster, or a 'leaf'. Next, the most similar objects (the ones for which the distance between the objects is the smallest) are merged. This procedure is repeated again and again. In the end, all leaves and branches are merged into one tree." (Levshina, 2015:309). Moreover, there are a variety of methods (or algorithms) which show how the clusters are merged. In our research we have used the method according to Ward. This algorithm attempts to minimize the increase in the Variance-innen (see the y-axis of Elbowplot 1) in the distances between the members of groups (=clusters).

Before explaining the dendrograms, it is necessary at first considering Elbowplot 1. The aim of the elbowplot is to display the number of clusters (=groups), which can be distinguished.



Elbowplot 1:
Determining of the optimal number of clusters for English styles

In order to determine the optimal number of clusters, we must consider the lines in Elbowplot 1. There are the values of the total variance-innen on the y-axis, designated as height and the number of clusters on the x-axis (= index). If the variance-innen (or line) moves volatile to the next cluster (= index on our elbowplot), it means that two dissimilar clusters are fusioned. This demands choosing the largest of the two clusters, situated on the x-axis – 13. The Variance-innen at 13 clusters is 1.0. This plays an important role for determining the optimal number-cluster solution considering Dendrogramm 1:



Dendrogramm 1.

One can see all observations (styles) on the x-axis of cluster dendrogram, which are united by strokes. This is a sign of a fusion of certain styles. The height of the horizontal uniting lines corresponds to the variance-innen of a fusion. In our case, Dendrogram 1 suggests canceling the clustering process at about 1.0 (see Elbowplot 1, y-axis). This favours a two-cluster solution. In particular, this leaves 2 clusters, one of which contains 4 styles, the rest cluster – 11 styles. In particular, the styles (genres) **belles-lettres, learned, lore** and **news** are combined to one cluster on the basis of average word repeats, the number of hapax legomenas and the number of unique words; the second cluster consists of **hobbies, adventure, romance, editorial, fiction, reviews, science-fiction, humor, government, mystery and religion.**

The solution to the optimal number of clusters can be made with the help of the **average silhouette width**. This measurement can vary from 0 to 1: 0 means no cluster structure and 1 denotes excellent separation of clusters. According to Kaufman$Rousseeuw (1990) the average silhouette width below 0.2 means a lack of cluster structure. Levshina (2015) considers that average silhouette means show well-formedness of certain clusters for a solution. This means that the objects of one cluster are near or close to each other and far from the objects of the other clusters.

At the present stage it would be of great interest to reveal which average silhouette width values the different number-cluster solutions can have. This also helps us to reveal which number-cluster solutions are the most effective, considering the values of the average silhouette widths computed in R-Studio. The values are given in Table 1:

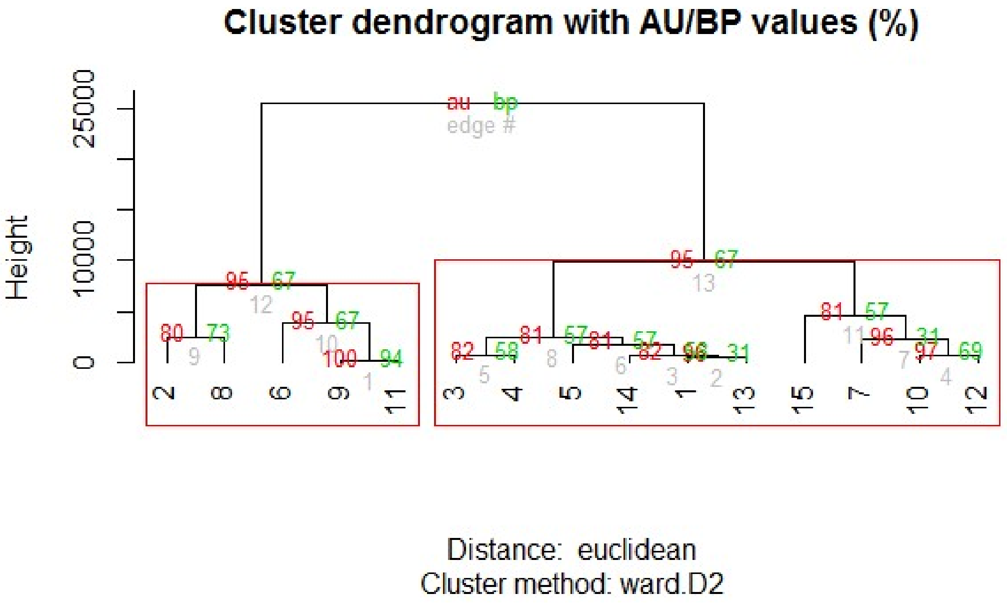
Table 1:
The values of average silhouette width for n-cluster solutions

The number of clusters	Average silhouette width
2	**0.33**
3	**0.20**
4	0.17
5	0.13
6	0.11
7	0.10
8	0.07
9	0.04
10	0.04
11	0.03
12	0.027
13	0.02

As one can see from Table 1, the perfect separation can be found for the two-number solutions. The greatest silhouette width is 0.33 which belongs to a two-cluster solution.

3. Diagnostics of a two-cluster solution

With the help of the average silhouette width we have determined the optimal number of clusters for our research. At this stage we must be certain of how reliable our results are when one repeats this research using another sample. This task can be done by means of multiscale bootstrap resampling in the package *pvclust* of R-Studio Program. This algorithm deals with a random sample considering the replacement from the original sample and calculates the necessary statistics. This is repeated for many times (i.e. 1000). The result of this resampling is given in Plot 2:



Plot 2.

The values on the plot correspond to the cluster probabilities. The probabilities to the left are Approximately Unbiased (AU) p-values and BPs to the right are bootstrap probabilities. If the p-value is closer to 1, the more reliable and stable support the cluster receives. The AV is considered here to be exacter measure. It is possible to notice here that the first cluster (belles_lettres(2) + learned(8) + hobbies (6) + lore (9) + news(11)) is supported by the data at 0.95 as well as the second cluster (editorial (3) + fiction (4) + government (5) + romance (14) + adventure (1) + reviews (13) + science-fiction (15) + humor (7) + mystery (10) + religion(12)). Within the first cluster the styles lore and news are supported at the level of 100. Within the second cluster one can see that adventure (1) and reviews (13) obtain the highest support at the level 0.96 as well as humor (7), mystery (10) and religion at the levels of 0.96, 0.97.We may conclude here that these clusters can be revealed in other research if one uses another sample.

REFERENCES

Bird, S., Klein, E and Loper, E. (2009). *Natural Language Processing with Python.* O'REILLY.

Bortz., J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaften.* (7 Aufl.). Heidelberg: Springer-Verlag

Galperin, I.R. (1981). *Stylistics*. Moscow Vyssaja Shkola

Hatzinger, R., Hornik, K., Nagel, H. (2011). *Einführung durch angewandte Statistik.* Pearson Studium.

Kaufman, L & Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley-Interscience

Levshina, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis.* John Benjamins Publishing Company.

Glottometrics 45, 2019, 63-88

Quantitative Analysis of Queen Elizabeth II's and American Presidents' Christmas Messages over 50 Years (1967–2018)

Zheyuan Dai[1], *Haitao Liu* [2*]

Abstract. Over the past century, the UK and the US have evolved new Christmas traditions, namely Queen's Christmas Broadcasts for the UK and Lighting the National Christmas Tree for the US. Queen Elizabeth II and American Presidents deliver their Christmas felicitations as accompaniments to new celebrations. This study intends to evaluate stylistic features – both synchronically and diachronically, and especially at the lexical level – of Queen Elizabeth II and American Presidents' Christmas messages based on the material from over 50 years. The results exhibit that overall, Queen Elizabeth II has a higher level of vocabulary richness along the half century. Detailed indicators, big words and hapax legomena**,** further show that Queen Elizabeth II's vocabulary is more complex and diversified than the lexis of American Presidents. Nevertheless, American Presidents surpass Queen Elizabeth II in thematic concentration. Discourse analysis discovers that Queen Elizabeth II concentrates on many smaller-scale themes, ignoring political ones, and cares for accuracy of words. On the contrary, in addition to conveying good wishes, American Presidents take Christmas messages as a good opportunity to publicize political opinions, which leads to their overall higher thematic concentration level.

Keywords: *Stylistic analysis, Christmas messages, quantitative analysis, Queen Elizabeth II, American Presidents*

1. Introduction

In order to express secular emotions and sincere wishes better, Christmas has developed new traditions over the recent hundred years. Different countries have their own distinct characteristics (Miles, 1976). Christmas has also become an important opportunity for politicians, royals, and other political figures to express their affinity to the people and communicate with them. In the UK, since 1952, every year on Christmas Day at Buckingham

[1] Department of Linguistics, Zhejiang University, Hangzhou, China.

[2] Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing, China; Department of Linguistics, Zhejiang University, Hangzhou, China; Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com, ORCID-No.: https://orcid.org/0000-0003-1724-4418

Palace, Queen Elizabeth II delivers her annual message to the Commonwealth in a tradition started by her grandfather George V in 1932 (Mount, 2015). Queen Elizabeth II's Christmas Broadcasts, the drafts of which she has written herself, is one of a few occasions where the Head of the Commonwealth is entitled to speak freely about her views. In the US, the President lights the National Christmas Tree in a national park every year. The remarks of Presidents at the lighting ceremonies are blessings from the White House for the beginnings of Christmas seasons.

In the 20th century, both British royal, and American presidential functions have changed. For American Presidents, "speaking is power" (Caesar et al., 1981) – it means that they must take advantage of their political speeches to win the Congress and the chosen citizens' supports. Political texts – such as inaugural speeches, campaign debates, State of the Union Addresses, etc. – have attracted many linguists (Hoffman & Howard, 2006; Kubát & Čech, 2016a; Lim, 2002; Savoy, 2010 & 2016; Wang & Liu, 2017). The modern British monarch is not the figure of political power (Billig, 2003). Representing the image of the Commonwealth and maintaining national unification in the spiritual perspective have become their primary responsibility. Due to the political particularity of the British Royal family, the Queen seldom expresses her independent political views. Some attention was paid to her political speeches (Jennings & John, 2009; Kelso, 2017). Queen's English has always been regarded as the most standard, accurate, and elegant. As to taking royal Christmas messages as the study objective, Queen Elizabeth II's pronunciation – Received Pronunciation – becomes a hotspot in linguistic research (Harrington, 2000 & 2006). There are also some qualitative studies concerning grammatical elements in the texts (Kredátusová, 2009; Li, 2014). However, qualitative methods emphasize description, and then turn to viewpoints, feelings, and experiences. Quantitative approaches complement qualitative analyses to make them more scientific and accurate, helping to draw extensive and in-depth conclusions.

Quantitative research is characterized by logical rigour and reliability. Therefore, quantitative approaches are widely employed to analyze individual stylistic features. A speaker's language style and characteristics can be grasped on the basis of the fundamental component of the article – the lexicon. Traditionally, to evaluate the richness of the textual vocabulary, type-token ratio (TTR) has been verified to be a reliable indicator (Herdan, 1960; Kubát et al., 2014). However, TTR is strongly length-dependent; its usage in the Christmas addresses should thus be justified – for example, by the fact that they are of approximately the same length. Its application as a metric to capture the vocabulary richness in a text is extensively exhibited in political speech analyses (Kubát & Čech, 2016a; Savoy, 2010 & 2016; Wang & Liu, 2017). To explore the complexity or diversity of the text at the lexical level further, more specific indicators – such as big words (BW), Hapax Legomena (HL), Lexical Density (LD), and Average Word Length (AWL), etc. – are employed (Fan et al., 2014a; Popescu & Altmann, 2008; Savoy, 2017). Language and ideology are closely linked (Van Dijk, 2006). The degree of how close the relationship between them is can be measured as the thematic concentration (TC) in the stylistic research. TC indicates the speakers' intention to focus on certain themes more intensively than on others (Čech et al., 2015). This indicator has been applied to investigate the speakers' stylistic characteristics widely, especially in political speeches or debates (Čech, 2014; Kubát & Čech, 2016a; Wang & Liu, 2017).

Qualitative analysis should also be adopted because it is the premise of quantitative analysis. Only when the two methods are combined flexibly, the best results can be achieved. According to critical discourse analysis (CDA), a text creates its sense only when the knowledge of the text and the world is related (Van Dijk, 2003). Christmas messages delivered by Queen Elizabeth II and American Presidents summarize the past year and expect the next. Therefore, the study of Christmas messages should include both a characterization of the text in particular as well as the systematic description of its context (Fairclough, 1995). What's more, CDA also proposes that all texts are interrelated both diachronically and synchronically (Wodak & Krzyżanowski, 2008). The two sets of Christmas messages have covered a period spanning over 50 years, which allows us to analyze the evolution of their stylistic features diachronically.

Based on previous studies, the paper pays attention to quantitative analysis of stylistic features and to the diachronic evolution of Queen Elizabeth II's Christmas Broadcasts (QCB) as well as of American Presidents' remarks upon lighting the National Community Christmas Tree (RLNCT). For holiday felicitations, stylistic studies have already been conducted to describe the distinguished styles of political characters (Čech, 2014; Jičínský & Marek, 2017; Rovenchak & Rovenchak, 2018). Being stripped of the political framework, holiday felicitations accurately reveal the characteristics of textual messages and speakers' delivery styles. In this paper, two specific questions are answered:

Question 1. What are the differences in vocabulary richness between Queen Elizabeth II and American Presidents' Christmas messages, and what causes them?

Question 2. What are the reasons behind the choices and expressions of thematic words in their Christmas messages?

The arrangement of the paper goes as follows. The second section introduces the basic information about the selection of our corpus and main methods. In the third section, a set of analyses describes and compares the stylistic features and evolutions of British QCB and American RLNCT over 50 years, being based on overall measurements with computational tools. Discourse analyses for main thematic words are exhibited as well, for better comprehension. The last section summarizes the paper briefly.

2. Selection of the Christmas messages and methods

2.1 Text selection

The texts of QCB were collected from the official website of British Monarchy[3]. The texts of RLNCT were collected from American Presidency Project[4]. Detailed information can be

[3] This can be accessed at https://www.royal.uk.

checked in Appendix.For American Presidents' RLNCT, the entire set includes 50 remarks delivered by 10 presidents, from Lyndon B. Johnson (Dec 15, 1967) to Donald J. Trump (Nov 29, 2018). For the record, president Richard Nixon was absent from the lighting ceremony in 1971 and 1972. Vice president Spiro Agnew lit the Tree[5]. Therefore, the year to select the material was pushed back to 1967 to ensure that the total number of texts tested is 50. For British QCB, this paper excerpted 50 texts from 1967 to 2018, except for texts of the years 1971 and 1972 to keep the material balanced with the US.

Many people may take broadcasts or remarks as a one-way communication, which may not correspond to written scripts. However, a person reading a written text aloud will produce a speech that has the linguistic characteristics of the written text (Biber & Conrad, 2009). In other words, under the processing of memory mechanism, both written texts, and oral speeches can be converted to each other equally.

2.2 Methods

As we have mentioned above, three quantitative indicators (MATTR, BW, HL) were exploited for studying lexical richness.

First, type-token ratio (TTR) – distinct types of words divided by the text length (Baayen, 2008) – is an indicator of lexical richness. What should be emphasized is that this index relies on textual length greatly. Solutions – such as standardized TTR (STTR), Lambda (Λ), measure of textual lexical density (MTLD), Moving-Average Type-Token Ratio (MATTR) – have been proposed to fix it (Covington & McFall, 2010; McCarthy & Jarvis, 2010; Popescu et al., 2011). This paper adopted MATTR to calculate TTR through a moving window to avoid the impact of text length. This method has already been proved feasible and reliable (Kubát, 2014). The algorithm of MATTR goes as follows.

With the window – a randomly chosen size *W*, moving one step at a time –, the text of length *N* is divided into several overlapped subtexts of the same length. Each move produces a sub-TTR. The average mean of all sub-TTRs is MATTR. Here comes the formula:

$$\text{(1)} \qquad \mathrm{MATTR} = \frac{\sum_{i=1}^{N-W} V_i}{W(N-W+1)}$$

In (1), W signifies window size, N the total text length ($W < N$), and V_i the numbers of types in the text. For this paper, taking some edge cases – such as Queen Elizabeth II's

[4] This can be accessed at https://www.presidency.ucsb.edu. So far, this website has not updated the latest President's remarks. The latest remarks are available on the official White House website: https://www.whitehouse.gov/articles/christmas-tree-lighting-president-trump-revives-traditions-religious-spirit/.

[5] Richard Nixon was in Key Biscayne, Florida, in 1971, and absent from the ceremony. In 1972, the tree was lighted by the vice president, too. The information can be accessed at https://potus-geeks.livejournal.com/1038940.html.

message in 1969 (263 tokens) and Clinton's remark in 1997 (139 tokens) into consideration –, the suggested window size of 500 words for stylometric analysis (Covington & McFall, 2010) is adjusted to 100 words. The value is measured by the software MATTR[6], based on word forms of our corpus.

Word length is a globally recognized measurement for lexical complexity. The longer the word is, the more complex the text is. A text which has a higher percentage of big words (BW) – with six letters or more – can be considered semantically complex (Savoy, 2016). For lexical diversity, hapax legomena (words that appear only once in the text) is a measure to reveal the degree of synthesis of the language in texts (Lardilleux & Lepage, 2007). The higher the hapax percentage is, the lower the repetition rate of words, and the higher the diversity of the vocabulary. These two indicators are measured in word forms by WordSmith Tools[7] and QUITA[8], respectively.

To find thematic words, analysis of thematic concentration lays the foundation. TC was first introduced by Popescu (2007), elaborated by Popescu et al. (2009), and further developed by Popescu and Altmann (2011). To measure TC, h-point – calculated on the basis of word frequency – should be counted first. H-point first entered linguistics with Popescu's work (2007). It marks the moment that the rank of a certain word equals to its occurrence if we rank word frequencies of a text in descending order. The computation of h-point can be expressed as follows:

$$h = \begin{cases} \mathrm{r}_i, & \mathrm{r}_i = f(r_i) \\ \dfrac{f(r_i)\mathrm{r}_{i+1} - f(\mathrm{r}_{i+1})\mathrm{r}_i}{\mathrm{r}_{i+1} - \mathrm{r}_i + f(r_i) - f(\mathrm{r}_{i+1})}, & \mathrm{r}_i \neq f(r_i) \end{cases} \tag{2}$$

H-point fuzzily functions as the cut-off boundary of so-called frequent synsemantics (i.e., pronouns, participles, prepositions, and articles), and autosemantics (i.e., nouns, adjectives, and adverbs) [Popescu et al., 2009]. Autosemantic words appearing before the h-point always play the roles of bearers of textual themes. Therefore, in this paper, only autosemantics (in the form of lemmata) are taken into consideration. Based on the value of the h-point, TC can be calculated through (3), i.e. –

$$\mathrm{TC} = 2\sum_{r'}^{T} \frac{(h-r')f(r')}{h(h-1)f(1)} \tag{3}$$

[6] MATTR can be available at http://ai1.ai.uga.edu/caspr/.

[7] WordSmith Tools is available at https://lexically.net/wordsmith/. It can help to count the number of words with different letters in the text. The proportion of BW is calculated on the basis of the data provided by the software.

[8] QUITA (Quantitative Index Text Analyzer) is available at http://oltk.upol.cz/software.

where f (1) is the frequency of the first rank. T is the number of autosemantics before the h-point, r' is the average rank of lemma sharing the same frequency with others ($r' < h$), and $f(r')$ denotes the frequency of the lemma which ranks r'. Let's take American incumbent President Trump's Christmas messages in the past two years for examples.

Table 1

18 most frequent lemmas in American incumbent President Trump's RLNCT

Rank	Average Rank	Lemma	Frequency	Rank	Average Rank	Lemma	Frequency
1	1	the	69	10	6.2	we	23
2	2	be	62	11	11.5	for	22
3	3	and	57	12	11.5	you	22
4	4	of	47	13	13.3	all	20
5	5	a	25	14	13.3	I	20
6	6.2	in	23	15	13.3	that	20
7	6.2	our	23	**16**	**16.5**	**Christmas**	**19**
8	**6.2**	**thank**	**23**	17	16.5	very	19
9	6.2	to	23	18	18	have	17

Note. Thematic words and related information are highlighted in bold.

In this text, $r_{17} < f(r_{17})$, while $r_{18} > f(r_{18})$. According to (3), h-point is 17.6.

$$h = \frac{f(r17)r18 - f(r18)r17}{r18 - r17 + f(r17) - f(r18)} = \frac{19 \times 18 - 17 \times 16.5}{18 - 16.5 + 19 - 17} \approx 17.571 \approx 17.6$$

Two thematic words, *thank* and *Christmas*, are before the h-point. Therefore, the value of TC can be computed as follows:

$$\text{TC}_{\text{American President Trump}} = 2 \times \left(\frac{(17.6 - 6.2) \times 23}{17.6 \times (17.6 - 1) \times 69} + \frac{(17.6 - 16.5) \times 19}{17.6 \times (17.6 - 1) \times 69} \right) \approx 0.02808665$$

Selected texts were lemmatized by TreeTagger (Schmid, 1994). Kubát & Čech (2016a) also suggested that TC should be independent of the text length roughly within the range of 200 – 6,500 tokens. Mostly, our texts fall into this interval. However, as we mentioned before, edge cases – such as Clinton's remark in 1997 (139 tokens) – are too short to be analyzed. Instead of looking at 50 years' values of TC, we might as well evaluate individual stylistic features of American Presidents. Hence, in terms of discussing TC, our texts (both British QCB, and American RLNCT) will be categorized according to the terms of office of American Presidents.

In the process of calculating TC, autosemantics with relative high frequencies were selected, which enabled us to make qualitative analyses of the themes. The concordances or collocations of the main thematic words were tracked through AntConc 3.2.4w (Anthony, 2011).

3. Discussion on stylistic features

3.1 Vocabulary selections – rich, or not?

As is shown in Table 2, British QCB's average MATTR value is higher than American RLNCT's in the past 50 years.

Moreover, a non-parametric test was conducted on two countries' political figures' MATTR values of traditional Christmas messages (a Mann-Whitney U-test was employed since the set of data of British QCB violates the normal distribution). The results have shown that average MATTR value of British QCB (M = .717, SD = .0197) is significantly distinguished from that of American RLNCT (M = .675, SD = .0285, U = 245*,* p = .000). From a diachronic perspective (see Figure 1), the overall trends are based on fluctuations around the average level. It is much clearer that over the last 50 years, in terms of vocabulary richness, Queen Elizabeth II has maintained a relative high level than American presidents. Only in two years (1982 & 2010), MATTR values are lower than American presidents'.

Generally speaking, British QCB have had a richer vocabulary than American RLNCT in the past 50 years. Formally, although both are live speeches, QCB are televised in the Buckingham Palace – Queen Elizabeth II speaks to the camera alone, monologue-like, without any response –, while American Presidents deliver Christmas remarks to the audience in front of the White House, which require a certain degree of interaction[9]. Compared with Queen Elizabeth II's live broadcasts, this form is more like a two-way communication where one party has got ready in advance, and the other party responds by applause, cheer, or other non-language forms. Garrod & Pickering (2004) have discovered that two-way communication is easy because of the automatic links between perception and behaviour in social interaction. Therefore, fuelled by the Christmas atmosphere and the anticipation of the audience, American Presidents' remarks need to be more concise and vivid. Too complex and diversified vocabulary is not conducive to interaction with the audience. As for Queen Elizabeth II, she pays no attention to the response of the audience. She just needs to make her speech (or we can call it a monologue) well. The cognitive load of monologue makes her tend to perfect her language, leading to the increase of vocabulary richness, to a certain extent.

[9] The scene of the lighting ceremony can be seen on https://thenationaltree.org. Some Presidents had simple interactions with the audience or the host during delivering his Christmas messages. This study captures Presidents' words of the interactions.

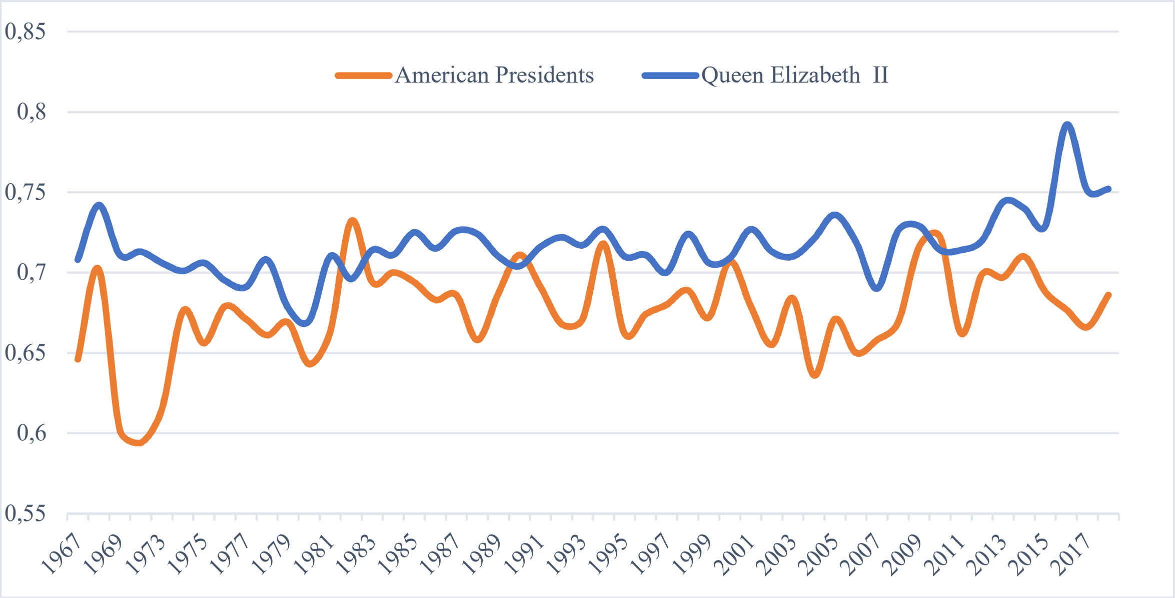


Figure 1. MATTR values of British QCB & American RLNCT

To compare vocabulary richness of the Christmas messages further, BW and HL, namely the lexical complexity and diversity of the texts, were investigated respectively. Independent-Samples T Tests were conducted, and it was discovered that the differences were significant. For BW, Levene's test shows that with the equal variances assumed ($F(1, 98) = .002$, $p = .964 > .05$), t-test (2-tailed) proves that in terms of BW, there were significant differences in the two sets of data ($t = 5.362$, $p = .000$). For HL, Levene's test shows that with the equal variances assumed ($F(1, 98) = 3.269$, $p = .074 > .05$), t-test (2-tailed) testifies that in terms of HL, there were significant differences in the two sets of data ($t = 2.845$, $p = .005 < .05$) as well.

Table 2
Mean relative frequencies of BW & HL in British QCB & American RLNCT

	Queen Elizabeth II	American Presidents
BW	0.277	0.25
HL	0.33	0.303

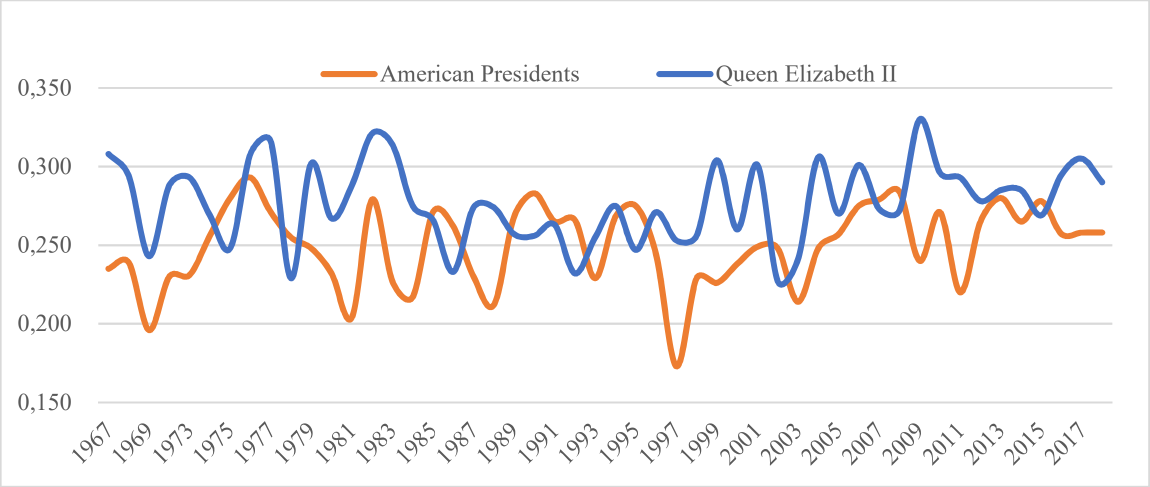


Figure 2. Relative frequency of BW in British QCB & American RLNCT

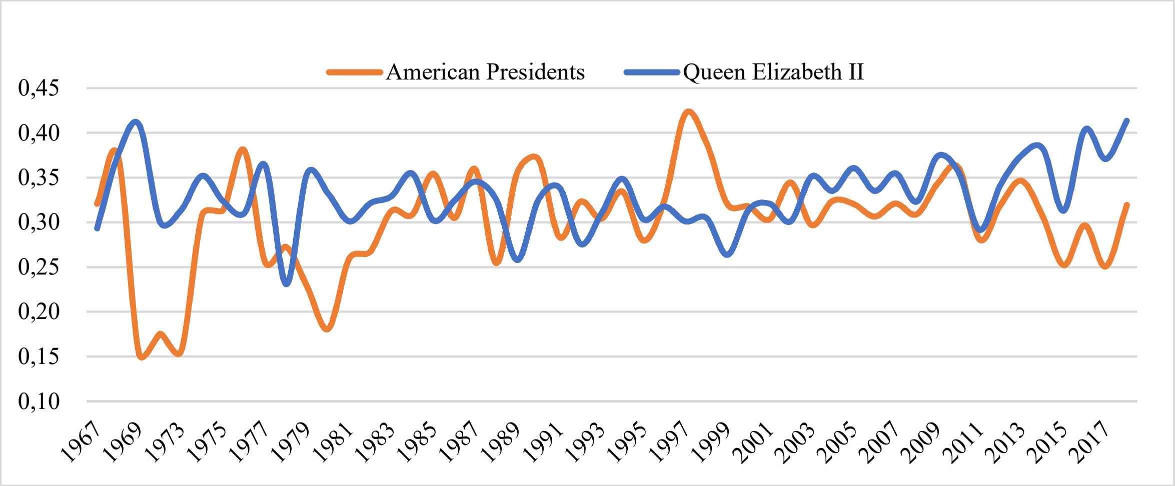


Figure 3. Relative frequency of HL in British QCB & American RLNCT

As is exhibited in Figure 2 and Figure 3, overall, the four sets of data fluctuated above and below their own mean values respectively. Mean values (see Table 2) show that the two sets of data are basically on the same level, while t-test demonstrates that they are significantly different. Queen Elizabeth II should not only speak for her "Queen's English", but also maintain the image of the whole country, which can be reflected by her words and deeds. More precisely, Queen Elizabeth II represents the image of the British Royal family, which is the most famous noble house in British history. Speech can convey people's temperament and image, and reflect their social status (Cuerie, 1952; Ellis, 1967). Therefore, besides conveying Christmas greetings to the world and expressing the kinship of the Royal family, it is still necessary to maintain the pride and identity of the nobility. As we have mentioned in the introduction, Queen's English has always been regarded as the most accurate and elegant English. This standardized language needs to maintain a high level of writing, especially the accuracy of expression, first and foremost on such occasions.

Since 10 presidents have been in office in the past 50 years in US, their personal stylistical features should be considered. Correspondingly, several data groups with significant differences can be seen in Figure 2 and Figure 3 (e.g., 1969–1973; 1979; 1980; 1997; 2016–2018). According to American Presidents' respective terms of office, the texts are divided into 10 parts respectively.

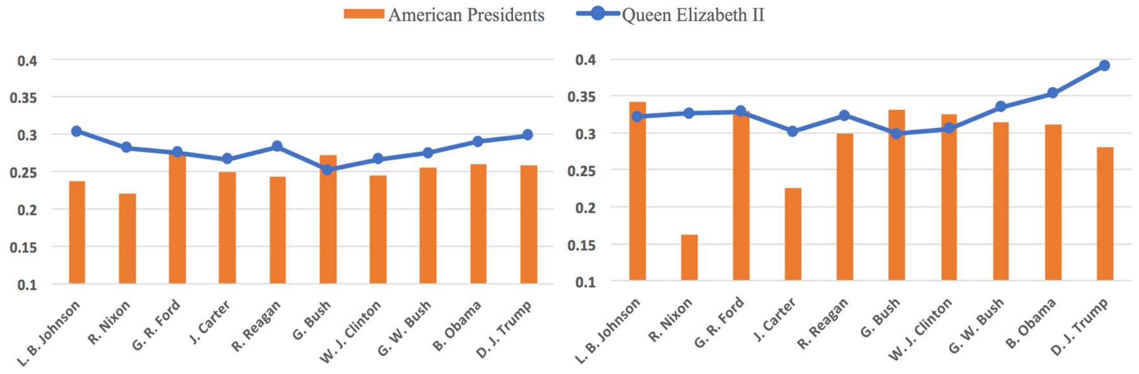


Figure 4. Relative frequency of BW & HL in British QCB & American RLNCT

Note. The left chart denotes the data of BW, and the right denotes the data of HL.

Two Presidents caught our attention – G. Bush and R. Nixon. G. Bush is the only President in the 50 years that has a higher level of lexical complexity and diversity than Queen Elizabeth II. However, his MATTR value fails to surpass Queen Elizabeth II's. As to President R. Nixon, not only relative frequencies of BW and HL (especially HL) in his Christmas remarks, but also the integral vocabulary richness indicator – MATTR – are much lower. Arguments exist that Christmas remarks need no complex expressions which may make the felicitation too formal and rigorous. Nevertheless, studies have already found that on more formal political occasions – for inaugural speeches as well as annual SOTU –, Nixon always ranks at the bottom level in terms of vocabulary richness among all American Presidents (Kubát & Čech, 2016a; Savoy, 2016), although his level of second thematic concentration is much higher than the average. Repetitions of thematic words in the texts lead to an increase in the degree of theme concentration. These discoveries coincide with our findings about Nixon – with less BW, there will be more simple and popular expressions. With less HL, there will be a higher repetition of words. His vocabulary richness obviously lags behind other presidents, leading to the decline of their overall vocabulary richness level, compared with Queen Elizabeth II's.

3.2 Thematic concerns – monotonous, or not?

3.2.1 Comparison of TC levels

Comparison between TC levels comes first to give a general introduction to investigate Queen Elizabeth II's and American Presidents' stylistic features from the perspective of content.

American Presidents have a relatively higher mean value of thematic concentration (*M* = .0174, *SD* = .0080) than Queen Elizabeth II (*M* = .0036, *SD* = .0033) when delivering Christmas messages, signifying American Presidents' efforts to express certain themes more intensively. At the other end of the spectrum, Queen Elizabeth II's lower TC level indicates the diversity of her themes in Christmas broadcasts. Specifically, during two periods (1967–1969; 2017–2018), Queen Elizabeth II's Christmas broadcasts show zero TC value, which means decentralization of themes (see Figure 5 and Table 3). Čech (2014) suggests that low TC values can be viewed as a reflection of the speaker's attempt to reflect the complexity and diversity of the real world where we live. So, what topics does Queen Elizabeth II care about in her Christmas broadcasts? – This exploration is presented in the next section.

An Independent-Samples T Test was carried out. Through Levene's test, the variances are assumed to be not equal ($F(1,11.998) = 7.921$, $p = .011 < .05$), and adjusted t-test (2-tailed) exhibited that there were significant differences in the two sets of data ($t = -5.031$, $p = .000$).

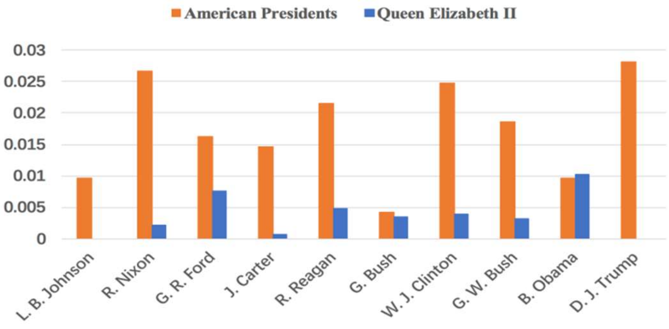


Figure 5. TC of British QCB & American RLNCT in Presidents' respective tenures

As we have mentioned in the introduction, language has close connections to ideology (Van Dijk, 2006). TC level can mirror a tendency of ideology, namely the higher the TC value is, the more totalitarian the speaker may be. By contrast, the lower the TC value is, the more democratic one may be (Čech, 2014). Howbeit, American Presidents such as Nixon, Clinton, Reagan, and G. W. Bush, etc., cannot be casually regarded as totalitarian leaders because of their higher TC values, compared with Queen Elizabeth II (see Figure 5). The United States is a federalist country with a presidential regime as its organizational form of political power. Although American Presidents are checked and balanced by the system of separation of powers, they still hold real power in national political affairs. Take Trump as an example – studies have proven that his high TC value in campaign speeches (Wang & Liu, 2017) does not mean his high totalitarian tendency, but portray his supporters as authoritarians on the other hand (Morgan & Shanahan, 2017). Wang & Liu (2017) reckoned that people's interests in having a leader with an authoritarian style may be aroused by Trump's concentration on certain themes.

Totalitarianism was the great mobilizing and unifying concept of the Cold War (Gleason, 1995); some materials we chose originated in the Cold War period. American Presidents used to regard totalitarianism as their enemy (Brooks, 2006). Thus, we may speculate boldly that Presidents present high TC values in RLNCT because of political reasons. With the help of Christmas messages, American Presidents have strengthened the focus of the theme, demonstrated tough and vigorous leadership, and gave people confidence in the government. The main central themes of American Presidents over the past 50 years in their RLNCT are further discussed in the following section.

Contrastly, Queen Elizabeth II was completely overpowered in her political life because of the Constitutional monarchy, a system of government derived from Britain's imperial history. Monarchy prefers "peace and order", the guiding principle of government is its authority over its "subjects". While in republican democracies, which prefer liberty, the guiding principle is unity, or whether it works in a beneficial sense for the citizens (Kennedy, 2005). In the light of that, to realize republican ideal on the premise of retaining monarchy, the position of the Head of the Commonwealth should not be an office, but rather an expression of a symbolic character without any separate constitutional standing or capacity (Bogdanor, 1997). Queen Elizabeth II's lower TC values just reflect her support for absolute democracy. She talks about many small topics and avoids to participate in politics excessively, exercising her formal powers and authorities of the Head of the Commonwealth prescribed within an established legal framework, namely acting as a visible symbol of national unity.

Table 3 and Table 4 demonstrate pre-h autosemantics in British QCB and American RLNCT respectively. Since nouns occupy large proportions of the autosemantics and reflect the theme of texts effectively, this paper concentrates on thematic nouns. A simple question comes out quickly – What is the common theme of their Christmas messages? – The answer is obvious – *Christmas*.

Table 3
Relevant information of Thematic Concentration of British QCB

Year	Speaker	h-point	f (1)	Autosemantics (average rank *r'*/frequency *f(r')*)	TC
1967–1968	Queen Elizabeth II	14.5	98	/	/
1969, 1970, 1973	Queen Elizabeth II	14	77	year (13/15);	0.0021
1974–1976	Queen Elizabeth II	16	101	people (12/19); good (15/17)	0.0077
1977–1980	Queen Elizabeth II	19.5	221	Christmas (18/20)	0.0008
1981–1988	Queen Elizabeth II	30.5	399	year (17.5/43); Christmas (25/35); Commonwealth (27/33)	0.0048
1989–1992	Queen	23	190	year (17/28)	0.0035

	Elizabeth II				
1993–2000	Queen Elizabeth II	28	345	year (16.5/45)	0.0040
2001–2008	Queen Elizabeth II	26	274	Christmas (20/33); people (24.5/28); year (24.5/28)	0.0032

2009–2016	Queen Elizabeth II	28	309	Christmas (16/41); year (17/40);
				people (20/33);
					0.0102

2017–2018	Queen Elizabeth II	12.5	72	/	/

Note. Autosemantics calculated in the form of lemma.

Table 4

Relevant information of Thematic Concentration of American RLNCT

Year	Speaker	h-point	f (1)	Autosemantics (average rank *r'*/frequency *f(r')*)	TC
1967–1968	L. B. Johnson	12.75	61	life (10/16)	0.0096
1969, 1970, 1973	R. Nixon	27	227	peace (11/50); Christmas (16.5/39); year (18/35); tree (19.5/34); America (21.5/32); world (23.5/31); light (25/28)	0.0266
1974–1976	G. R. Ford	15	67	Christmas (9/19)	0.0162
1977–1980	J. Carter	24.5	218	Christmas (12/61); nation (19/29)	0.0147
1981–1988	R. Reagan	27.5	269	Christmas (9/79); light (19.5/41); tree (21/40); time (25.5/30)	0.0215
1989–1992	G. Bush	17.5	127	Christmas (14/22)	0.0042
1993–2000	W. J. Clinton	25	214	Christmas (12/53); peace (13/51); thank (20/31); year (21.5/30); light (24/25)	0.0247
2001–2008	G. W. Bush	26	278	Christmas (13/60); thank (15/57); national (19/31); peace (24/29)	0.0186

2009–2016	B. Obama	28	275	Christmas (20/44); holiday (21/39);
				tree (22/38); national (25.5/30);
				year (25.5/30)
					0.0096

2017–2018	D. J. Trump	17.6	69	thank (6.2/23); Christmas (16.5/ 19)	0.0281

Note. Autosemantics calculated in the form of lemmata.

In terms of TC, Queen Elizabeth II shows relative lower values than American Presidents. However, the main thematic word – *Christmas* – runs through the whole 50 years' messages in both two parties. Interestingly, there are obvious differences between the two parties in expressing Christmas greetings (see Table 5). Instead of employing the popular collocation *Merry Christmas* as American Presidents did, Queen Elizabeth II preferred a different, "strange" expression – *Happy Christmas*. this collocation appears in different syntactic structures, such as declarative sentence, emphatic sentence, imperative sentence, etc. According to etymology[10], the word *Merry* had much wider senses in Middle English, among which a low slang *Merry-bout*, meaning an incident of sexual intercourse was widely used, making *merry* linked to the meaning of *lust*. William Shakespeare, the greatest English writer at that time, wrote a famous comedy – *Merry Wives of Windsor*. In this play, *Merry* denotes the decadent concept of women in the old society – a symbol of lust and a source of evil. Coincidentally, the surname of Queen Elizabeth II is *Windsor*. Given the popularity of this comedy, Queen Elizabeth II is prone to avoid this embarrassment even more. Anyhow, since the word used to have a negative meaning, in terms of vocabulary, Queen Elizabeth II pays more attention to the dignity of nobility and turns to a relatively plain, but safer choice – *happy*.

Table 5

Occurrences of *Happy Christmas* or *Merry Christmas* in British QCB & American RLNCT

	Happy Christmas	Merry Christmas
Queen Elizabeth II	47	3
American Presidents	2	58

3.2.2 Analyses of the main thematic words

To obtain a more comprehensive understanding of the stylistic features, discourse analyses of thematic words of British QCB and American RLNCT are discussed in this section.

[10] An online etymology dictionary (https://www.etymonline.com) can track the wheel-ruts of modern English.

According to Figure 6, thematic nouns are within a small scale as to frequency, and centralized together. Besides, the distribution of thematic words also conforms to Zipf's law, a power function relation.

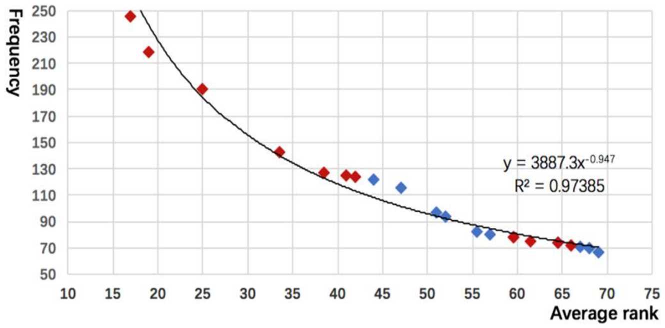


Figure 6. Rank frequency distribution of pre-h thematic words of British QCB over 50 years. Red diamonds denote thematic nouns.

Although some of thematic words seem to be politicized (i.e., world, Commonwealth, country), Queen Elizabeth II's expressions, unlike those of politicians, still show great affinity to the people, which is an integral part of Queen Elizabeth II's Christmas broadcasts. Thematic nouns in Table 6 – such as *family*, *life*, *child* – roused our interests.

Table 6
Thematic words in Queen Elizabeth II's Christmas broadcasts

Average rank	Lemma	Frequency	Average rank	Lemma	Frequency
17	**year**	246	55.5	come	82
19	**Christmas**	218	57	see	80
25	**people**	190	59.5	**child**	78
33.5	**world**	143	59.5	give	78
38.5	**family**	127	61.5	**day**	75
41	**time**	125	64.5	**country**	74
42	**Commonwealth**	124	64.5	own	74
44	**life**	122	66	**hope**	72
47	good	116	67	**help**	71
51	make	97	68	bring	70
52	great	94	69	happy	67

Note. H-point is 68.5. Thematic nouns are highlighted in bold. Words like *hope*, which can be used as nouns and verbs alike, are also given in bold.

Typical sentences containing thematic words – *family*, *life*, *child* – are selected as examples:

Family

1. We are trying to create a wider *family* of Nations and it is particularly at Christmas that this *family* should feel closest together.

2. Christmas is for most of us a time for a break from work, for *family* and friends, for presents, turkey and crackers.

3. I first came here for Christmas as a grandchild. Nowadays, my grandchildren come here for the same *family* festival.

4. Like many other *families*, we have lived through some difficult days this year.

Life

1. The responsibility for the way we live *life* with all its challenges, sadness and joy is ours alone.

2. The very act of living a decent and upright *life* is in itself a positive factor in maintaining civilised standards.

3. Success in industry and commerce, for instance, creates the wealth that provides so many of the things that make *life* happier and more comfortable.

4. (…) [Jesus Christ] managed to live an outgoing, unselfish and sacrificial *life*. Countless millions of people around the world continue to celebrate his birthday at Christmas…

Child

1. Everything we do now is helping to shape the world in which our *children* are going to live.

2. They never lost hope and they never lacked confidence in themselves or in their *children*.

3. The sight of the happy faces of *children* and young people in Russia, in South Africa…

4. There are some *children* who are much less fortunate than others, for they come from countries where nature makes life very hard…

Examples reveal that the word *family* is used mostly to depict three situations: the big family in political sense, common families, and Queen Elizabeth II's royal family. As we mentioned before, Queen Elizabeth II represents the image of the Commonwealth and symbolizes the national unity. On the one hand, Queen Elizabeth II has no real power and stays away from real political life; on the other hand, as the Head of the Commonwealth and a member of the Royal family, she is far from the masses, making her out of reach. *Family*, a warm and cohesive word, functions as a bridge connecting Queen Elizabeth II's ordinary emotions with her political missions.

Queen Elizabeth II cares about the well-being of the people. On Christmas Day, Queen Elizabeth II acts like an elder of an ordinary family. She talks about her life experiences and feelings by the fireside, making everything warm and touching. Besides, Queen Elizabeth II's vocabulary richness is well concentrated around this ordinary word – *life* (*with challenges*, *sadness and joy*; *decent and upright*; *happier and more comfortable*; *outgoing*, *unselfish and sacrificial…*).

What's more, Queen Elizabeth II emphasizes the quality of *life* in an ordinary manner. She attaches great importance to the future of the country – *children* as well. Taking *children* as a carrier, Queen Elizabeth II expresses her concerns about some hot topics, such as environmental protection, education, etc. Unexpectedly, Queen Elizabeth II have made special mentions of children from all over the world, hoping that they could be taken care of, showing her great affinity to the people as well as her sympathy.

As for American RLNCT, apart from thematic words or their collocations – such as *Christmas*, *national Christmas tree*, etc. –, another autosemantic noun *peace* caught our attention (see Table 4). The concordance plot of *peace* over the 50 years goes as follows.

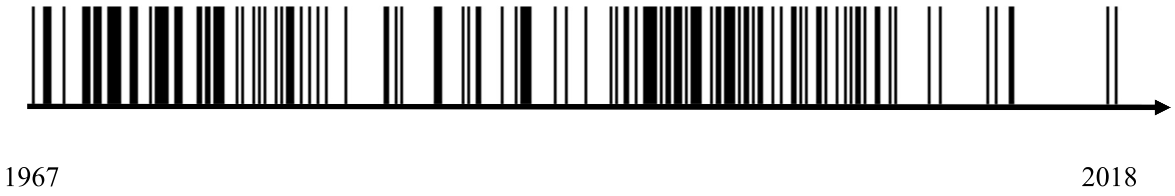


Figure 7. Concordance plot of *peace* in American RLNCT

Note. A black line indicates that the word appears once, and the more the word appears in the same or adjacent periods, the thicker the black line becomes.

According to classification on Wikipedia[11] – "Military History of the United States", which can be supported by the latest American Congressional Research Service (Torreon, 2018) [updated on Dec 14, 2018] –, America has gone through three big war or conflict periods in the past half century – Vietnam Era (1964–1975), Post-Cold War Era (1990–2001) and War on Terrorism (2001–present). Table 7 shows the hits of *peace* in American RLNCT. Three Presidents who mentioned *peace* most frequently (highlighted in bold) served the tenures basically coinciding with the three war periods. The three relatively dense lines on the bar chart are prominent as well (see Figure 7).

[11] https://en.wikipedia.org/wiki/Military_history_of_the_United_States.

Table 7
Concordance hits of *peace* in American Presidents' RLNCT

Presidents	Terms of office	Hits
Lyndon B. Johnson	1967–1968	6
Richard Nixon	**1969–1973**	**50**
Gerald R. Ford	1974–1976	12
Jimmy Carter	1977–1980	20
Ronald Reagan	1981–1988	23
George Bush	1989–1992	6
William J. Clinton	**1993–2000**	**51**
George W. Bush	**2001–2008**	**29**
Barack Obama	2009–2016	6
Donald J. Trump	2017–2018	2

We gleaned expressions about *peace* from the following sources:

Richard Nixon

1. Seventy years ago, America was at *peace*. Today, America is not at *peace*. And what we want for this Nation is not only *peace* now but peace in the years to come, *peace* for all people in the years to come.

2. Our wish, our prayer, is for *peace*, the kind of *peace* that we can live with, the kind of *peace* that we can be proud of, the kind of *peace* that exists not just for now but that gives a chance for our children also to live in *peace*.

3. (…) *peace* in the world, *peace* in our homes, and *peace* in our hearts.

4. (…) for the fact that this is the first Christmas in 12 years that a President has stood here at a time when America was at *peace* with every nation in the world.

5. And our greatest hope in this Christmas season and in all seasons is, of course, ***peace*** in the whole world. We can be grateful in this Christmas season that already we have been able to bring 200,000 men back from Vietnam, more coming home.

William J. Clinton

1. At this holiday season also, my fellow Americans, let us extend our special gratitude and prayers for the men and women of our Armed Forces who protect the *peace* and stand sentry for our freedom.

2. They see a nation graced by *peace* and prosperity, a land of freedom and fairness.

3. Let us be grateful that our Nation is at *peace* and rejoice in the progress we have made to bring about *peace* on Earth. And let us not forget the work still to be done, from Bosnia to the Middle East, to the Korean Peninsula.

4. I hope that we can finish the business of *peace* there and help, again, America to give a gift to the rest of the world.

5. Our Nation is at *peace*, and all around the world we are privileged to make *peace*, from Bosnia to Northern Ireland, to the Middle East, the land where a homeless child grew up to be the Prince of *Peace*.

George W. Bush

1. America seeks *peace* and believes in justice. We fight only when necessary. We fight so that oppression may cease, and even in the midst of war, we pray for *peace* on Earth and good will to men.

2. They [the American military forces] serve in the cause of *peace* and freedom. They wear the uniform proudly, and we are proud of them.

3. We have service men and women celebrating the holidays at bases from Europe to East Asia and on many fronts in the war on terror. Especially for those deployed in Afghanistan and Iraq, the work is dangerous and the mission is urgent. American service men and women are bringing freedom to many and *peace* to future generations.

4. America's military men and women stand for freedom, and they serve the cause of *peace*. Many of them are serving in distant lands tonight, but they are close to our hearts.

5. We rejoice in the Christmas promise of *peace* to men of good will.

From above expressions, obviously, Nixon adopted a large number of parallel sentences to emphasize his point. The same components appear denser, and thereof his vocabulary is relatively simple. With respect to the content, the three Presidents have repeatedly reiterated their pursuit of peace. At the traditional beginning ceremony of the Christmas season, the emphasis on *peace* not only conforms people's wishes, but also receives the resonance of the world. It can soothe people's hearts injured by the war, making them full of expectations for the peaceful life or reunion in the new year. Also, as we mentioned before, to fight against so-called enemy – totalitarianism (Brooks, 2006) –, American Presidents need highly centralized political power and national cohesion, which projects, to a certain extent, a tendency to totalitarianism. Correspondingly, in special war periods, the country also needs a strong leader. Therefore, Nixon gradually brought America out of the quagmire of Vietnam War, while Clinton borrowed the name of *peace* to start the wars in Somalia and other several wars / conflicts. In order to fight terrorism and achieve so-called *peace*, Bush launched a series of Wars on Terrorism. In words, the politicians took advantages of Christmas felicitations not only to convey their desires for the world peace, but also to publicize their political views to the people further, win their hearts, and seek political supports. In contrast to being a member of British Royal family, an American president is more concerned about whether or not he has made outstanding achievements during his term of office.

4. Conclusion

In this paper, we conducted a stylistic analysis of Queen Elizabeth II's Christmas Broadcasts and American Presidents' Remarks on Lighting National Christmas Tree over the past 50 years. Mainly lexis-measuring quantitative methods were adopted. We further compare the themantic concentration of the two parties respectively. Next, discourse analyses of the main thematic words selected from Christmas messages were carried out to discuss the possible factors.

For vocabulary richness, Queen Elizabeth II's vocabulary is richer than American Presidents'. Specifically, in terms of vocabulary complexity and diversity, American President Nixon drags the whole team back. Contrastly, the values of Queen Elizabeth II and American Presidents' thematic concerntration vary greatly. Topics of Queen Elizabeth II's concern are wide and unconcentrated; they involve no political opinions and tend to show a strong affinity to the people. Higher indexes for vocabulary and lower values for theme concentration – namely formal and elegant expressions without any substantive contents – showing that the Queen's image has little political significance. Moreover, Queen Elizabeth II cares for words selection to represent nobility's dignity. American Presidents with high TC values circle around a limited number of themes. Discourse analyses reflect that it mirrors their ambitions to firmly seize every opportunity to speak as a means of propaganda for political positions.

Nevertheless, there is still a lack of further systematic analysis. Factors such as social and political backgrounds are not fully considered. Besides, American Presidents' own changes within their tenures were not investigated. All Christmas messages during their terms of office were classified into wholes, in order to be studied in comparisons. Last but not least, due to the limited time and space, this paper does not compare syntactical complexity between Queen Elizabeth II and American Presidents.

Acknowledgements

This work is partly supported by the National Social Science Foundation of China (Grant No. 17AYY021) and the Special Project for Humanities and Social Science, Zhejiang University (*Quantitative Analysis of Sentence Hierarchical Structure in English*).

REFERENCES

Anthony, L. (2011). AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Billig, M. (2003). *Talking of the royal family*. London and New York: Routledge.

Bogdanor, V. (1997). The sovereign and the commonwealth. *Monarchy & the Constitution*, 240–298.

Brooks, J. (2006). Totalitarianism Revisited. *The Review of Politics*, 68, 318–328.

Caesar, J. W., Thurow, G. E., Tulis, J., & Bessette, J. M. (1981). The Rise of Rhetorical Presidency. *Presidential Studies Quarterly*, 11(2), 158–171.

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). *Quality & Quantity*, 48(2), 899–910.

Čech, R., Garabík, R., & Altmann, G. (2015). Testing the thematic concentration of text. *Journal of Quantitative Linguistics*, 22(3), 215–232.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.

Cuerie, H. C. (1952). A projection of socio-linguistics: The relationship of speech to social status. *The Southern Speech Journal*, 18(1), 28–37.

Ellis, D.S. (1967). Speech and Social Status in America. *Social Forces*, 45(3), 431.

Fairclough, N. (1995). *Critical Discourse Analysis: The Critical Study of Language* (Language in Social Life series). London: Routledge.

Fan, F., Wang, Y., & Gao, Z. (2014a). Some macro-quantitative features of low-frequency word classes. *Glottometrics*, 28, 1–12.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *TRENDS in Cognitive Sciences*, 8(1), 8–11.

Gleason, A. (1995). *Totalitarianism: The Inner History of The Cold War*, New York: Oxford University Press.

Harrington, J. (2000). Does the Queen speak the Queen's English? *Nature*, 408, 927–928.

Harrington, J. (2006). An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics*, 34, 439–457.

Herdan, G. (1960). *Type-token mathematics. A textbook of mathematical linguistics*. The Hague: Mouton and Co.

Hoffman, D. R., & Howard, A. D. (2006). *Addressing the* State of the Union*. The Evolution and Impact of the President's Big Speech*. Boulder: Lynne Rienner.

Jennings, W., & John, P. (2009). The Dynamics of Political Attention: Public Opinion and the Queen's Speech in the United Kingdom. *American Journal of Political Science*, 53(4), 838–854.

Jičínský, M., & Marek, J. (2017). New Year's Day speeches of Czech presidents: phonetic analysis and text analysis. In: Saeed K., Homenda W., Chaki R. (eds.). *Computer Information Systems and Industrial Management.* CISIM 2017. Lecture Notes in Computer Science, vol. 10244.

Kelso, A. (2017). The politics of parliamentary procedure: An analysis of Queen's speech debates in the House of Commons. *British Politics*, 12(2), 267–288.

Kennedy, G. (2005). Constitutional Monarchy. *In: Adam Smith's Lost Legacy.* London: Palgrave Macmillan, 78–80.

Kredátusová, M. (2009). Queen's Christmas Speeches 1952–2007: Discourse Analysis. Brno: Masaryk University.

Kubát, M. (2014). Moving window type-token ratio and text length. In: Altmann G., Čech R., Mačutek J. et al. (eds.). *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM, 105–113.

Kubát, M., & Čech, R. (2016a). Quantitative analysis of US presidential inaugural addresses. *Glottometrics*, 34, 14–27.

Kubát, M., Matlach, V., & Čech, R. (2014). *QUITA. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Lardilleux, A., & Lepage, Y. (2007). Hapax Legomena: Their Contribution in Number and Efficiency to Word Alignment. Conference: *Human Language Technology. Challenges of the Information Society, Third Language and Technology Conference*. Poznań, Poland.

Li, X. (2014). *A Quantitative Study of the Grammatical Features of the Sixty Christmas Speeches Broadcast by Queen Elizabeth II.* Shanghai: Shanghai Normal University.

Lim, E.T. (2002). Five Trends in Presidential Rhetoric: An Analysis of Rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, 32(2), 328–348.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392.

Miles, C.A. (1976). *Christmas customs and traditions: their history and significance*. New York: Courier Dover Publications.

Mount, H. (2015). *The Queen's speech: a Christmas tradition worth keeping*. The Telegraph. Retrieved at http://www.telegraph.co.uk/culture/tvandradio/bbc/120 57926/The-Queens-speech-a-Christmas-tradition-worth-keeping.html.

Morgan, M., & Shanahan, J. (2017). Television and the cultivation of authoritarianism: a return visit from an unexpected friend. *Journal of Communication*, 67, 424–444.

Popesecu, I.-I., Čech, R., & Altmann, G. (2011). *The Lambda-structure of Texts*. Lüdenscheid: RAM-Verlag.

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P. (eds.). *Exact Methods in the Study of Language and Text*. Berlin: Mouton de Gruyter, 555–566.

Popescu I.-I., & Altmann, G. (2011). Thematic concentration in texts. In: Kelih, E., Levickij, V., and Matskulyak, Y. (eds.). *Issues in Quantitative Linguistics 2*. Lüdenscheid: RAM, 110–116.

Popescu I.-I., & Altmann, G. (2008). Hapax legomena and language typology. *Journal of Quantitative* Linguistics, 15(4), 370–378.

Rovenchak, A., & Rovenchak, O. (2018). Quantifying Comprehensibility of Christmas and Easter Addresses from the Ukrainian Greek Catholic Church Hierarchs. *Glottometrics*, 41, 57–66.

Savoy, J. (2010). Lexical Analysis of US Political Speeches. *Journal of Quantitative Linguistics*, 17(2), 123–141.

Savoy, J. (2016). Analysis of the Style and the Rhetoric of the American Presidents Over Two Centuries. *Glottometrics*, 38, 55–76.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. vol.12.

Scott, M. (1996). *WordSmith Tools Manual*. Oxford: Oxford University Press.

Torreon, B. S. (2018). U.S. Periods of War and Dates of Recent Conflicts. Congressional Research Service, Retrieved at *https://crsreports.congress.gov*.

Van Dijk, T. A. (2003). The discourse-knowledge interface. In: Weiss, G. and Wodak, R. (eds.). *Critical Discourse Analysis*. Basingstoke: Palgrave Macmillan, 85–109.

Van Dijk, T. A. (2006). Ideology and discourse analysis. *Journal of Political Ideology*, 11, 115–140.

Wang, Y., & Liu, H. (2018). Is Trump always rambling like a fourth-grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election. *Discourse & Society*, 29(3), 299–323.

Wodak, R., & Krzyżanowski, M. (eds.) (2008). *Qualitative Discourse Analysis in the Social Sciences*. Basingstoke: Palgrave Macmillan.

Appendix: Texts Information

Text	Speaker	Date	Words
1	Queen Elizabeth II	Dec 25, 1967	953
2	Queen Elizabeth II	Dec 25, 1968	507
3	Queen Elizabeth II	Dec 25, 1969	263
4	Queen Elizabeth II	Dec 25, 1970	625
5	Queen Elizabeth II	Dec 25, 1973	491
6	Queen Elizabeth II	Dec 25, 1974	628
7	Queen Elizabeth II	Dec 25, 1975	573
8	Queen Elizabeth II	Dec 25, 1976	630
9	Queen Elizabeth II	Dec 25, 1977	433
10	Queen Elizabeth II	Dec 25, 1978	1101
11	Queen Elizabeth II	Dec 25, 1979	549
12	Queen Elizabeth II	Dec 25, 1980	711
13	Queen Elizabeth II	Dec 25, 1981	868
14	Queen Elizabeth II	Dec 25, 1982	938
15	Queen Elizabeth II	Dec 25, 1983	774
16	Queen Elizabeth II	Dec 25, 1984	567
17	Queen Elizabeth II	Dec 25, 1985	873
18	Queen Elizabeth II	Dec 25, 1986	604
19	Queen Elizabeth II	Dec 25, 1987	603
20	Queen Elizabeth II	Dec 25, 1988	1035
21	Queen Elizabeth II	Dec 25, 1989	917
22	Queen Elizabeth II	Dec 25, 1990	767
23	Queen Elizabeth II	Dec 25, 1991	845

24	Queen Elizabeth II	Dec 25, 1992	783
25	Queen Elizabeth II	Dec 25, 1993	764
26	Queen Elizabeth II	Dec 25, 1994	739
27	Queen Elizabeth II	Dec 25, 1995	734
28	Queen Elizabeth II	Dec 25, 1996	678
29	Queen Elizabeth II	Dec 25, 1997	786
30	Queen Elizabeth II	Dec 25, 1998	833
31	Queen Elizabeth II	Dec 25, 1999	989
32	Queen Elizabeth II	Dec 25, 2000	607
33	Queen Elizabeth II	Dec 25, 2001	662
34	Queen Elizabeth II	Dec 25, 2002	578
35	Queen Elizabeth II	Dec 25, 2003	577
36	Queen Elizabeth II	Dec 25, 2004	582
37	Queen Elizabeth II	Dec 25, 2005	548
38	Queen Elizabeth II	Dec 25, 2006	594
39	Queen Elizabeth II	Dec 25, 2007	593
40	Queen Elizabeth II	Dec 25, 2008	680
41	Queen Elizabeth II	Dec 25, 2009	521
42	Queen Elizabeth II	Dec 25, 2010	625
43	Queen Elizabeth II	Dec 25, 2011	736
44	Queen Elizabeth II	Dec 25, 2012	641
45	Queen Elizabeth II	Dec 25, 2013	648
46	Queen Elizabeth II	Dec 25, 2014	667
47	Queen Elizabeth II	Dec 25, 2015	680
48	Queen Elizabeth II	Dec 25, 2016	614
49	Queen Elizabeth II	Dec 25, 2017	679
50	Queen Elizabeth II	Dec 25, 2018	569

Text	Speaker	Date	Words
1	Lyndon B. Johnson	Dec 15, 1967	609
2	Lyndon B. Johnson	Dec 16, 1968	422
3	Richard Nixon	Dec 16, 1969	1028
4	Richard Nixon	Dec 16, 1970	1199
5	Richard Nixon	Dec 14, 1973	1230
6	Gerald R. Ford	Dec 17, 1974	464
7	Gerald R. Ford	Dec 18, 1975	443
8	Gerald R. Ford	Dec 16, 1976	341
9	Jimmy Carter	Dec 15, 1977	960
10	Jimmy Carter	Dec 14, 1978	807
11	Jimmy Carter	Dec 13, 1979	887
12	Jimmy Carter	Dec 18, 1980	1588

13	Ronald Reagan	Dec 17, 1981	494
14	Ronald Reagan	Dec 16, 1982	998
15	Ronald Reagan	Dec 15, 1983	781
16	Ronald Reagan	Dec 13, 1984	649
17	Ronald Reagan	Dec 12, 1985	609
18	Ronald Reagan	Dec 11, 1986	507
19	Ronald Reagan	Dec 07, 1987	244
20	Ronald Reagan	Dec 15, 1988	453
21	George Bush	Dec 14, 1989	447
22	George Bush	Dec 13, 1990	640
23	George Bush	Dec 12, 1991	687
24	George Bush	Dec 10, 1992	319
25	William J. Clinton	Dec 09, 1993	471
26	William J. Clinton	Dec 07, 1994	585
27	William J. Clinton	Dec 06, 1995	621
28	William J. Clinton	Dec 05, 1996	524
29	William J. Clinton	Dec 04, 1997	139
30	William J. Clinton	Dec 09, 1998	428
31	William J. Clinton	Dec 08, 1999	518
32	William J. Clinton	Dec 11, 2000	622
33	George W. Bush	Dec 06, 2001	511
34	George W. Bush	Dec 05, 2002	443
35	George W. Bush	Dec 04, 2003	765
36	George W. Bush	Dec 02, 2004	644
37	George W. Bush	Dec 01, 2005	517
38	George W. Bush	Dec 07, 2006	491
39	George W. Bush	Dec 06, 2007	498
40	George W. Bush	Dec 04, 2008	577
41	Barack Obama	Dec 03, 2009	583
42	Barack Obama	Dec 09, 2010	468
43	Barack Obama	Dec 01, 2011	651
44	Barack Obama	Dec 06, 2012	669
45	Barack Obama	Dec 06, 2013	600
46	Barack Obama	Dec 04, 2014	577
47	Barack Obama	Dec 03, 2015	611
48	Barack Obama	Dec 01, 2016	824
49	Donald J. Trump	Nov 30, 2017	689
50	Donald J. Trump	Nov 28, 2018	550

Glottometrics 45, 2019; 89-113

The Effects of Source Languages on Syntactic Structures of Target Languages in the Simultaneous Interpretation: A Quantitative Investigation Based on Dependency Syntactic Treebanks

Yawen Wang[1], *Haitao Liu*[2*]

Abstract. Dependency distance (DD), as the distance between two linked words in one sentence is widely used to explore the cognitive demands and cross-linguistic syntactic features in language processing. The purpose of simultaneous interpreting is to enable smooth communication between two languages, though it imposes a large burden on interpreters. However, previous studies have not yet investigated the impact of source languages on target languages in the simultaneous interpreting process between different language pairs from a typological perspective quantitatively. It is still indispensable to examine carefully how essential the role is played by different source languages in simultaneous interpreting. With recourse to quantitative methods, the current study explores English simultaneous interpretations from distinct source languages. From the cognitive perspective, results via mean dependency distance demonstrate that the structures of English interpretations are interfered marginally significantly by diverse source languages in simultaneous interpreting. Meanwhile, language typology of source languages has moderately small impact on English interpretations with resort to dependency direction. This research firstly investigates the effect of diverse source languages on the same target language in simultaneous interpreting, suggesting the overwhelming impact of mean dependency distance minimization on language processing.

Key words: *Dependency distance, source languages, simultaneous interpreting, dependency direction, quantitative linguistics*

1. Introduction

Simultaneous interpretation, as a type of interpreting, is a very difficult time-limited cross-language communication activity. In simultaneous interpreting, with resort to professional equipment, interpreters communicate the content with the audience in one language without interrupting the speaker of another language through the synchronization of listening and speaking. The delay between the speaker and the interpreter is no more than a few

[1] Department of Linguistics, Zhejiang University, China.

[2] Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing, China; Department of Linguistics, Zhejiang University, Hangzhou, China; Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com, ORCID-No.: https://orcid.org/0000-0003-1724-4418

seconds during the simultaneous process. It is the simultaneity of language comprehension and production that requires enormous demands on the interpreters' cognitive capabilities (Mizuno, 2005; Padilla, Bajo, & Macizo, 2005). Interpreters simultaneously focus on and comprehend a new unit of meaning or chunk in one source language while simultaneously translating and producing the previous content in another target language.

Interpretation as the bridge between two languages may be interfered by source languages and impact on target texts. Systemic differences between source and target language have traditionally been viewed as a significant source of difficulty, and available empirical research has primarily focused on contrastive analysis of specific syntactic construction taken from a SI corpus. For example, Uchiyama (1991) analyzed Japanese and English; Setton (1998) Mandarin and English; Seeber (2005) German and English. Moreover, the academic interest in corpus-based source language research and its effect on the target language focused on language contact in translation in respect of English as the source language (Baumgarten, 2007, 2008; Fischer, 2007; House, 2011a, 2011b; Kranich, 2011; Kranich, House, & Becher, 2012; Malamatidou, 2013). In addition, for decades, researches in this area have been limited to the comparison between related European languages. It is of vital importance to find evidence from genetically distinct language pairs such as English and Chinese. A gap remains to locate the variation of cognitive difficulty in processing distinct source languages during simultaneous interpretation with recourse to treebanks.

A previous study resorted to the dependency grammar and dependency distance to measure the cognitive difficulty and found that consecutive interpreting entails smaller dependency distance (DD) and bears heavier cognitive demands than simultaneous interpreting (Liang, Fang, Lv, & Liu, 2017). Inspired by this former research, this study aims to investigate the cognitive burden caused by source languages in the simultaneous interpreting process via dependency distance.

Dependency Grammar is a grammar based on the dependency relations, proposed by Lucien Tesnière (1965). One important property of dependency relations is "dependency distance (DD)", which was created by Heringer, Strecker, and Wimmer (1980) and introduced by Hudson (1995). Its definition is "the distance between words and their parents, measured in terms of intervening words." (Liu, Hudson, & Feng, 2009). Measuring DD is useful for predicting syntactic difficulty (Liu, Hudson, et al., 2009). The close relationship between linguistic complexity, working memory, and sentence length has attracted a lot of attention in the linguistic community. Numerous psycholinguists have developed many theories, such as the Depth Hypothesis (Yngve, 1960), Early Immediate Constituents (EIC) (Hawkins, 1994), the Dependency Locality Theory (Gibson, 1998, 2000), and Minimize Domain (MiD) (Hawkins, 2004). All these theories found that linear distance between words in one sentence exerts a significant impact on the syntactic difficulty. The longer the sentence, the larger the dependency distance, the more difficult is language processing. Though Eppler (2010) and Hiranuma (1999) calculated the distance in terms of the number of intervening words, this study follows Liu's measurement of distance in terms of the difference between the words' position numbers, namely the mean dependency distance (MDD) (Liu, 2008, 2010; Liu, Hudson, et al., 2009).

As an effective predictor of syntactic difficulty, MDD is widely applied in numerous researches of language processing (Eppler, 2013; Jiang & Liu, 2015; Liang et al., 2017; Liu, 2008; Y. Wang & Liu, 2017). MDD also facilitates the discovery of a language universal preference for dependency distance minimization so as to reduce the memory burden (Liu, Xu, & Liang, 2017). Therefore, with the benefit of MDD, this study endeavors to illustrate the relationship between source and target languages in the simultaneous interpreting process.

Since dependency distance as one feature of dependency relations reflects the complexity of language processing, another property of dependency relation is dependency direction which is closely related with word-order language typology (Liu, 2010). Dependency direction reveals the unique syntactic structures of different languages, especially the linear order between a dependent and its governor. It is well-known that word order is essential in distinguishing the typological features of languages (Dryer, 1992; Greenberg, 1963; Liu, 2010). Dependency direction suggests whether the dependency relation is head-initial or head-final. Hudson (2003) assumed that languages are inclined to be consistently head-initial like Welsh or head-final like Japanese, or consistently mixed like English. Liu (2010) confirmed this assumption with the aid of a 20-language treebank. Dependency direction may enable this study to advance people's understandings about whether the typology of source languages may result in dissimilar syntactic features of the English interpretations.

Based on the previous studies, the current study aims to quantitatively investigate the syntactic features of English simultaneous interpretations by means of MDD and dependency direction. Research findings would put some insights on the simultaneous interpretation processes and language processing. A treebank of simultaneous interpretation from five different source languages to English is established to measure their MDDs and dependency directions, while another treebank of native English speeches is also developed for comparison. Dependency relations hold a considerable potential for measuring and calculating the cognitive difficulty of processing different languages in simultaneous interpreting, so as to provide a new perspective into the study of language processing. By virtue of dependency relations, the study will address the following questions:

(1) From the cognitive perspective, does handling distinct source languages impose different cognitive demand in the simultaneous interpretation process and then influence the syntactic structures of English interpretations?

(2) In regard to the language typology, are the syntactic structures in English interpreted texts interfered by their source language?

The first question aims to explore whether processing assorted source languages imposes different cognitive demands in simultaneous interpreting and then further influences their English interpretations. If so, how big is the influence? The second question investigates the variation among English interpretation texts caused by source languages in view of language typology. These questions endeavor to illustrate the substantial role played by source languages in the simultaneous interpretation.

Language materials and research methods are introduced in the next section. The results and detailed discussions are provided in the third section. Conclusions are described in the last section.

2. Materials and Methods

2.1 Methods

This study resorts to two quantitative indexes — mean dependency distance and dependency direction — to investigate the impacts of different source languages on simultaneous interpretation. These measurements rely on the dependency relation between two linked words within a sentence (Hudson, 2007; Liu, Hudson, et al., 2009; Tesnière, 1965). A dependency relation has three widely-accepted core qualities: (i) a binary relation between two linguistic elements, (ii) an asymmetric relation in which one element is a governor whereas the other serves as a dependent, (iii) a label on the top of an arc linking two elements (Liu, 2009). To present three qualities more transparently, a syntactic dependency tree or a directed dependency graph is built. Figure 1 clearly displays the syntactic structure of the sentence *The student reads a novel.* via a directed dependency graph.

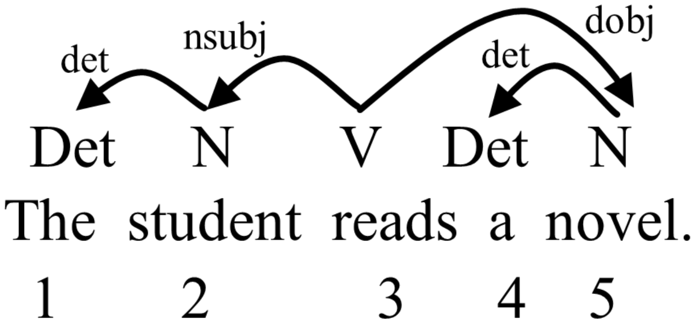


Figure 1. Dependency structure of the sample sentence *The student reads a novel.*.

Such dependency relations are labeled based on the Penn Treebank part-of-speech tags and phrasal labels (De Marneffe & Manning, 2008). The numbers below the sentence are the linear word order, which are used to compute the mean dependency distance of sentences and texts, developed by Liu, Hudson, et al. (2009).

Firstly, the sentence is labeled in linear word order as W_1, W_2, W_3, W_i… and W_n. If there is a dependency relation between a governor W_a and its dependent W_b, the dependency distance (i.e. DD) between W_a and W_b is defined as the difference between a and b (i.e. “a-b”). Thus, the DD of two adjacent words is 1 or -1, which is also known as the adjacent dependency. The DD value is positive if the dependent is before the governor, while a negative number shows up if the governor is before. More notably, just the absolute value of the DD is adopted for the calculation in this context.

The mean dependency distance (MDD) of one sentence is measured as follows:

$$\mathrm{MDD(the\ sentence)} = \frac{1}{n-1}\sum_{i=1}^{n-1}|\mathrm{DD}_i| \quad (1)$$

In the above formula, “*n*” is the total number of words in one sentence and “DD_i” refers to the dependency distance of the *i*-th syntactic link within one sentence. Specifically, root in the sentence has no governor and thereby a zero DD. So, it is eliminated in the calculation. It is feasible to employ this formula to compute the MDD of a text or even a treebank via the following formula.

$$\mathrm{MDD(the\ text)} = \frac{1}{n-s}\sum_{i=1}^{n-s}|\mathrm{DD}_i| \qquad (2)$$

Similarly, in (2), “*n*” represents the total number of words in the text and “DD_i” refers to the dependency distance of the *i*-th syntactic link within one text. Therefore, all the absolute DD values within the sample sentence *The student reads a novel.* are |2-1|+|3-2|+|3-5|+|5-4| by subtracting the number of the dependent word from that of its governor. To this end, the MDD of the sample sentence is 5/4 = 1.25 according to the formula (1).

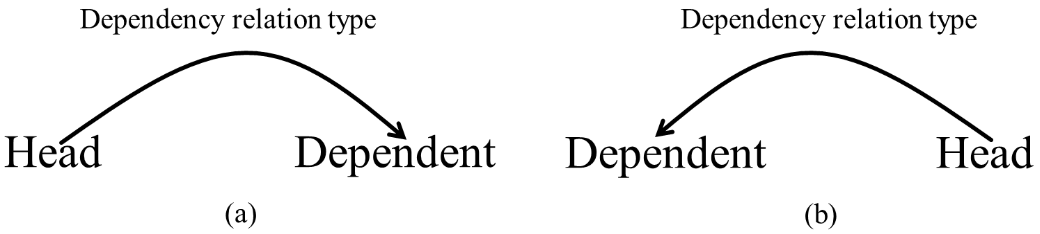


Figure 2. Two asymmetric dependency relations between a head and its dependent.

Dependency direction is concerned with the asymmetric relationship within a dependency relation, as illustrated in Figure 2. If the head precedes the dependent, this is a head--initial dependency relation like Figure 2(a), while a head-final dependency relation is obtained if the head follows the dependent as Figure 2(b). Many scholars have paid heed to the intimate connection between dependency direction and the classification of languages (Hudson, 2003; Liu, 2010; Tesnière, 1965). Some languages prefer head-initial structures, while others have more head-final ones.

In practice, there is no need to calculate both percentages of head-initial dependency relations and head-final ones, due to the fact that their sum is always 1. Here, we just measure the proportion of head-initial dependency relations. Its calculation is via dependency distance. If the dependency distance of one dependency relation is a positive number, its dependency direction is head-final, whereas a negative number presents a head-initial dependency direction. Take Figure 1 as an example. There are three (75%) head-final dependency relations and one (25%) head-initial. Thus, its percentage of head-initial dependency relations is 25%.

2.2 Materials

As treebank is an essential resource to quantitatively measure and analyze the common syntactic features of texts (Liu & Huang, 2006; Liu, Hudson, et al., 2009), a small-sized treebank is established based on speeches made by diplomats with their own official and native language from Arabic-speaking countries, China, France, Spain, and the Russian Federation at the 71st session of General Assembly of the United Nations (UN). These speeches are simultaneously interpreted into English by professional interpreters of the UN.

The United Nations as a large international organization, has a huge demand for interpreters and owns specialized interpreting branches, representing the highest level in the industry. In most cases, the requirements of the United Nations are to carry out accurate and complete literal translation within a limited time. Other factors influencing the output texts are excluded beforehand such as the individual interpreting styles (van Besien & Meuleman, 2008) and interpreting strategies (Kajzer-Wietrzny, 2012). Since the interpretation in the United Nations is well organized, all the interpreters are all highly professional and highly experienced to ensure the accuracy and consistence. Meanwhile, another small-sized treebank, namely treebank 2, is built for comparison, with recourse to accumulating English speeches made by American diplomats at the same session. All these texts are collected from the United Nations official website. Table 1 displays an overview of the two treebanks.

Table 1

An overview of the two treebanks

Languages	Language Family	Size (Words)	Sentence Numbers
Treebank 1			
Arabic	the Afro-Asiatic family	9246	324
Chinese	the Sino-Tibetan family	8459	308
French	a Romance language of the Indo-European family	8796	312
Spanish	a Western Romance language of the Indo-European family	8257	262
Russian	an East Slavic language of the Indo-European family	8884	338
Treebank 2			
English	a Germanic language of the Indo-European family	8354	276

All the texts in the treebanks are analyzed by the Stanford Parser version 3.9.1, a natural language parser that figures out the grammatical structures of sentences designed by the Natural Language Processing Group of Stanford University. It directly provides the dependency relations and parts of speech of words (De Marneffe & Manning, 2008). After careful manual check and correction, the Stanford Parser's parsed outputs are transferred to EXCEL formats for further analysis.

Table 2 provides an example of the format, which enables us to calculate the DD easily. As there is no governor of the main verb *is*, its dependency relation *root* is irrelevant in the calculation and therefore ignored. Moreover, the *punct* dependency relation is also deleted in the syntactic analysis because it is useless to this regard. According to the formula (1), the mean dependency distance of the sample sentence is (1+3+1+1+1)/5 = 1.4. Meanwhile, with regard to dependency direction, the percentage of head-initial dependency relations is 60%. The corresponding results will be discussed in detail in the next section.

Table 2

Dependency relations of the sample sentence

Word order	Word	Part of Speech	Word Order of Governor	Dependency Relation	Dependency Distance
1	Our	PRP$	2	poss	1
2	course	NN	5	nsubj	3
3	of	IN	2	prep	-1
4	action	NN	3	pobj	-1
5	is	VBZ	0	root	-5
6	clear	JJ	5	acomp	-1
7	.	.	5	punct	-2

3. Results and Discussion

3.1 Source languages, mean dependency distance, and simultaneous interpreting

To begin with, mean dependency distances of all dependency relations in two treebanks are chosen as our first indicator so as to reveal the diverse cognitive difficulty in processing different source languages in simultaneous interpretation. However, mean dependency distance is liable to be interfered by many factors, such as sentence length (Ramon Ferrer-i-Cancho & Liu, 2014; Jiang & Liu, 2015; Oya, 2011), genre (Liu, Zhao, & Li, 2009; Oya, 2013; Y. Wang & Liu, 2017), language types (Eppler, 2010; Hiranuma, 1999; Liu & Xu, 2012), and grammar (Gildea & Temperley, 2010; Liu, 2008). Thus, before the direct analysis of mean dependency distance in the two treebanks, we need to obtain a general picture of dependency distances with recourse to dependency distance distribution and the adjacent dependencies.

3.1.1. The probability distribution of dependency distance and sentence length

First and foremost, it is rational to check the distribution regularities of the dependency distance values for the two treebanks because some regularities have been found in English, Chinese, and different genres of one language (Ramon Ferrer-i-Cancho & Liu, 2014; Jiang & Liu, 2015; Liu, 2007; Y. Wang & Liu, 2017).

To begin with, same numbers of sentences of each sentence length are selected. In two treebanks, 10 to 30 sentence lengths account for the majority of sentences, which is consistent with that obtained in a previous study (Jiang & Liu, 2015). Yet due to the special genre of texts in our treebanks as political speeches, the 27-word sentence length appears most frequently. For each sentence length from 25 to 29 words, 6 sentences are randomly selected. There are all together 30 sentences.

Altmann-Fitter is a quantitative program for the iterative fitting of univariate discrete probability distributions to frequency data. By virtue of Altmann-Fitter software (2013), the distributions of dependency distances of English interpretations from different source languages and sentence lengths are investigated based on previously selected 30 sentences. The frequency of the dependency distance in the 30 sentences is fitted well by the probability distribution models: Right truncated modified Zipf-Alekseev (a, b; n = x-max, α fixed), Right truncated Waring (b, n), and Right truncated zeta (a, R = x-max). All the formulae of these distributions are presented in Appendix A. Table 3 illustrates R^2 (the coefficient of determination) values of the 30 sentences in each distribution.

Table 3

R^2 of dependency distances fitted by several distributions

Treebank	Source Language	Right truncated modified Zipf-Alekseev	Right truncated Waring	Right truncated zeta
Treebank 1	Arabic	0.995	0.991	0.949
	Chinese	0.991	0.984	0.940
	French	0.988	0.992	0.962
	Russian	0.995	0.996	0.959
	Spanish	0.996	0.990	0.958
Treebank 2	English	0.991	0.993	0.953

The model fittings of the Right truncated modified Zipf-Alekseev and Right truncated Waring are both excellent, with most R^2 values over 0.99. The model fitting of Right truncated zeta is not as good as the former two models, yet it is still acceptable with the coefficient of determination (R^2) above 0.9. The results are similar to that of Wang and Liu's (2017).

To further advance understanding of the influences of different source languages, Figure 3 takes the Right truncated zeta distribution of the 30-sentences' dependency distances as an example. More impressively, all the distributions have similar long tails. It indicates that all the investigated languages have a similar probability distribution, with the shortest dependency distance accounting for the largest proportion. The longer the dependency distance is, the fewer its amount is.

Table 3 and Figure 3 both show that sentence lengths of English simultaneous interpretations from diverse source languages make almost no difference in the probability distribution of dependency distances. It suggests that all language users tend to minimize the dependency distance and lessen the cognitive demands in language processing, mainly due to the limited working memory (Liu et al., 2017).

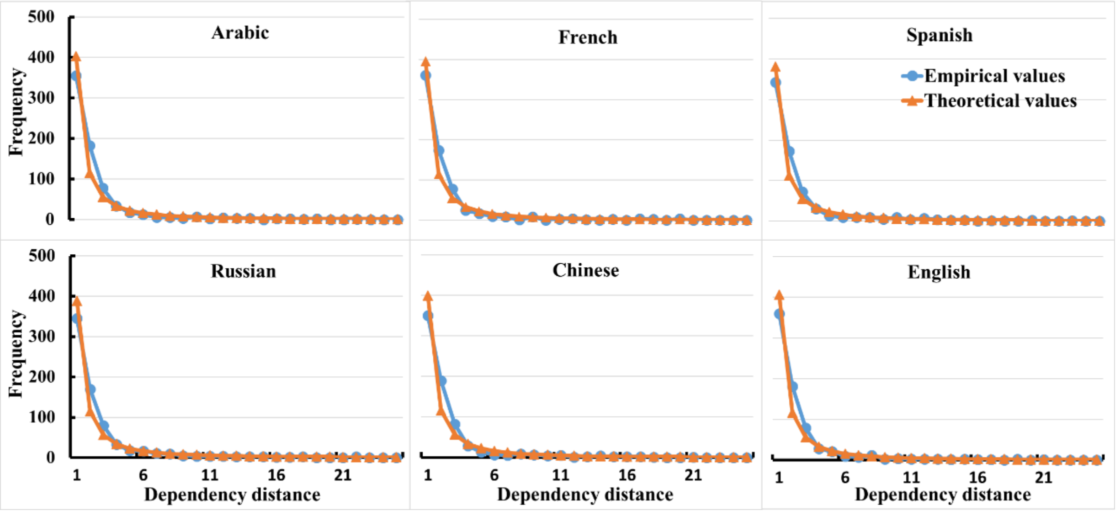


Figure 3. Fitting the Right truncated zeta distribution to dependency distances of 30 sentences from two treebanks. The blue dotted line represents the empirical values, the orange triangle-dotted line the theoretical values. Appendix B provides the raw data.

As Figure 3 shows, the adjacent dependency accounts for almost a half of the total amounts of DD and plays a critical role in language processing. Large numbers of the adjacent dependency may exert certain influences on the mean dependency distance. It is essential to have a deeper insight at the adjacent dependency before the analysis of mean dependency distance. Next section would explain how the adjacent dependencies would vary over different source languages.

3.1.2. Source language, adjacent dependency, and sentence length

Another factor that may also influence mean dependency distance in one language is the adjacent dependency, namely the dependency link between adjacent words (Liu, 2008). Based on this previous study of 20 languages, almost half of the dependency relations belong to the adjacent ones. However, the ratios of adjacent dependencies in English interpretations are much smaller when compared to those of previous studies, i.e. 74.2 in Collins (1996), 61.7 in Jiang and Liu (2015) and 51.3 in Liu (2008), thanks to different annotation schemes and treebank types (Y. Wang & Liu, 2017).

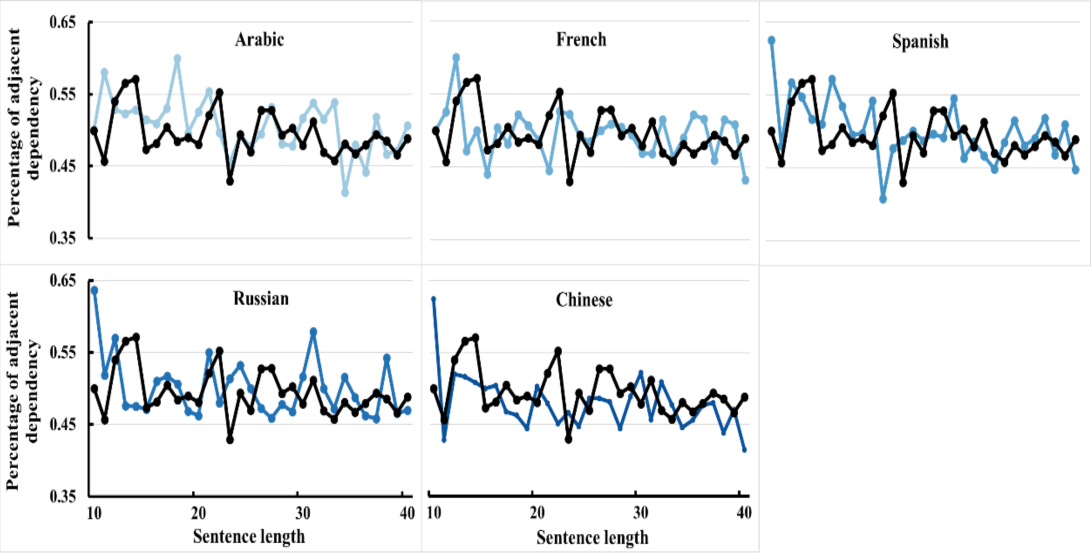


Figure 4. Percentages of adjacent dependency of diverse sentence lengths with different source languages. The black lines represent the correlated English values. Raw data are presented in Appendix C.

The percentages of adjacent dependency in our two treebanks are slightly lower than those of the previous studies with relatively stable values from 46% to 50%. This tendency is in good agreement with Liu's work that a lower MDD is available if a language includes more adjacent dependencies (Liu, 2008). Meanwhile, the much lower percentages of the adjacent dependency in the two treebanks also correspond well with the higher MDD in this research, as also confirmed by the previous studies.

Figure 4 reveals a general tendency of adjacent dependencies to decline with the increase of sentence lengths. Specifically, Arabic-English interpretations have higher percentages of adjacent dependencies as compared with those of the English native texts. In contrast, Chinese-English interpreted texts tend to have fewer adjacent dependencies. Besides, the differences are non-significant between French-, Russian-, and Spanish-English interpretations and English native texts based on Figure 4. In order to further probe the differences between two treebanks, a likelihood ratio test was employed.

A logistic regression presents a significant but weak correlation between dependency adjacency (adjacent or not-adjacent) and sentence length (G = 14.90; df = 1; $p < 0.001$; $R^2 = 0.001$; C = 0.512). Among them, R^2 and C-value work as indicators for the classification quality of the model. R^2 usually emerges in the range from 0 to 1 and a C-value appears from 0.5 to 1. If R^2 and C-value are above 0.8, they are considered good (Gries, 2013). In addition, a second logistic regression model is fitted, predicting the adjacent dependency with sentence length and different source languages. Adjacent dependency has a significant relationship but very low correlation with sentence length and different source languages (G = 24.46; df = 6; $p < 0.001$; $R^2 = 0.001$; C = 0.516). Considering the two important indexes - R^2 and C-value, the relationship between different source language and adjacent dependency is quite small. The likelihood ratio test is marginally significant ($p = 0.089$) between the two models. Nevertheless, the values of two correlation indicators, i.e. R^2 and C-value, expose limited effects on the relationship between different source languages and the adjacent dependency. Appendix D illustrates the effect plot. With the increase of sentence length from 10 to 40, interpretations from different source languages have similar predicted probabilities of the adjacent dependency.

The minor differences among source languages can be explained by the language typology, with Arabic from the Afro-Asiatic family, Chinese from the Sino-Tibetan family, and French, Spanish, and Russian all from the Indo-European family. The effect of source languages in simultaneous interpretation is in good accordance with Liu's study of native texts of these languages (2008). More interestingly, Arabic's and Chinese's interpretations to English have the similar tendency, since Arabic has the largest ratio of adjacent dependency and the smallest is found in Chinese. Although, due to English as the target language, the overall percentages of adjacent dependency lie within the scope of English, it is obvious that source languages play a minor rather than decisive role in the simultaneous interpretation.

With regard to the close relationship between adjacent dependencies and mean dependency distance, what is the impact of different source languages and sentence lengths on mean dependency distance?

3.1.3. Source language, mean dependency distance, and sentence length

Table 4 provides a general feature of mean dependency distances in two treebanks. Particularly, except Arabic and Chinese, mean dependency distances of all other three source languages to English interpretations are larger than that of native English. The French-English interpretation texts yield the highest MDD, whereas the lowest MDD is obtained from the Arabic-English interpretation texts. This finding is in line with previous studies that mean dependency distances differ cross-linguistically, although former investigations hardly exclude the influence of genre (Temperley, 2007; L. Wang & Liu, 2013). Herein, the interpretations in two treebanks belong to the same genre. The MDD values of Chinese- English SI texts correspond well to Liang et al.'s previous study (2017). Generally speaking, the highest mean dependency distance (2.78) in the two treebanks is below the threshold limited by cognitive capacity of human beings: 4 (Cowan, 2001). In the past, certain agreement has been reached about which languages have the shortest and which ones have the longer DDs. Namely, English has the shortest MDD, followed by Arabic, Spanish, and Chinese, as supported by previous study (Liu, 2008), which is different from the order of MDDs in two treebanks.

Table 4

An overview of mean dependency distance in two treebanks

		Mean Dependency Distance
Treebank 1	Arabic	2.55
	French	2.78
	Spanish	2.73
	Russian	2.75
	Chinese	2.69
Treebank 2	English	2.72

The main reasons behind diverse MDDs of texts interpreted from six source languages lie in the variation of the syntactic structures of source languages. In other words, the closeness between source and target languages may exert certain impact on this process. Take Chinese as an example. As an isolating language, Chinese uses free morpheme to mark tense, number, and aspect, whereas the Indo-European languages resort to numerous inflections of words. Such a difference may have a significant impact on the interpretation processes. In contrast, Spanish, French, and Russian belong to the same Indo-European language family of English. Then, their influence to English in the simultaneous interpreting tends to be contrary to Arabic and Chinese.

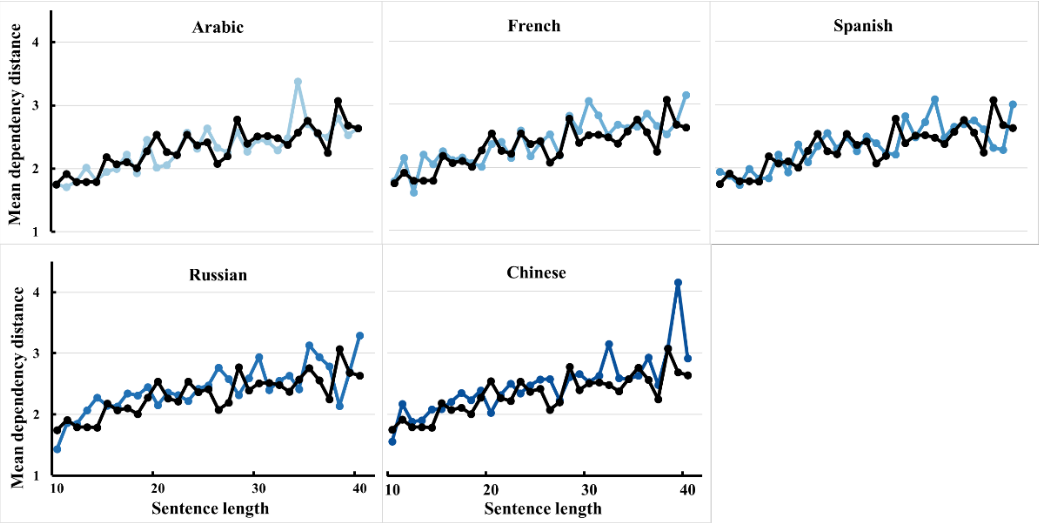


Figure 5. Percentages of mean dependency distance of different sentence lengths with variant source languages. The dark blue lines represent the English values. Raw data are provided in Appendix E.

Moreover, the largest percentages of adjacent dependencies of Arabic interpretation may partly bring about the smallest MDDs, because more adjacent dependencies in a language would produce a lower MDD (Liu, 2008).

Besides, numerous previous studies have found a close relationship between mean dependency distance and sentence lengths (R Ferrer-i-Cancho & Arias, 2013; Jiang & Liu, 2015; Oya, 2011; Y. Wang & Liu, 2017), the interference of sentence length has to be examined beforehand. As Figure 5 reveals, with the increase of sentence lengths, mean dependency distances of different source languages have a rising inclination. Compared with Figure 4, mean dependency distances are climbing up with an increment of sentence lengths, while adjacent dependencies are falling. In other words, longer sentences are inclined to have fewer adjacent dependencies and larger mean dependency distances, which is in good accordance with Liu's study (2008).

It is therefore indispensable to investigate how significant the difference is, with regard to the sentence lengths and source languages, with the obvious variation in MDDs among English interpretations from different source languages.

To answer the question, this study establishes a linear regression model, with an aim of predicating mean dependency distance with sentence length. The results reveal a highly significant relationship yet a minor correlation (F = 298.4, df_l = 1, df_2 = 184, p < 2.2e-16, Adjusted R^2 = 0.6165). Then, another model is fitted between mean dependency distance and sentence length with an interaction of different source language. This model is also significant (F = 53.93, df_l = 6, df_2 = 179, p < 2.2e-16, Adjusted R^2 = 0.6319). The result of the likelihood ratio test between the two models is noteworthy (p < 2.2e-16), indicating that mean dependency distances are closely correlated with source languages. The effect plot (Appendix F) presents an intimate relation among mean dependency distances, sentence length and source languages, since MDDs change with the increase of sentence length among different source languages. However, the effects of the correlation among mean dependency distances with sentence lengths and source languages are marginally significant (R^2 = 0.6319).

The results suggest that processing different source languages in simultaneous interpretation does not impose dramatically variant burden on interpreters' cognitive demands. These findings are consistent with the universal tendency to dependency distance minimizeation in language production. For the sake of reliable, coherent, and effective communication, dependency distance minimization has to be obeyed in accordance with the syntactic structures of the same target language - English. Although diverse source languages have quite minor influence, it will make no difference to the English interpretations. The reason behind lies in the nature of language as a complex system (Liu et al., 2017). Language is capable of self-organizing and self-adapting, implying that language would use some strategies to relieve the heavy memory demands made by unique linguistic patterns from different source languages and strive to be as close as possible with the target language. Thus, this study is in good consistence with the tendency of dependency distance minimization of human languages. It is possible to assume that target languages in the interpretation process may play a more critical and decisive role than source languages.

After examining the cognitive factors related to the dependency relations, it is essential to analyze the impact of source languages from the stance on linguistic typology via dependency direction. Dependency directions can be applied to classify language types (Liu, 2010). In the next section, we strive to explore the effect of different source language types on English interpretations through dependency directions.

3.2. Source language, dependency direction, and simultaneous interpreting

Table 5 shows the overall dependency direction via percentages of head-initial dependency in both treebanks. Arabic-English interpretation has the largest percentages of head-initial dependencies, immediately followed by native English texts, whereas interpretations from other source languages have subtly smaller percentages. This variation obviously attributes to source languages. As we all know, many languages have a dominant dependency direction (Eppler, 2013; Liu, 2010). For example, Arabic is predominantly head initial. Other languages such as English and Chinese are more or less mixed. Previous studies have found that Chinese has a moderately larger proportion of head-initial dependency than that of English (Jiang & Liu, 2015; Liu, Zhao, et al., 2009).

Table 5

An overview of dependency direction in two treebanks

		Percentage of Head-initial Dependency
Treebank 1	Arabic	55.45%
	French	51.88%
	Spanish	51.48%
	Russian	52.86%
	Chinese	52.91%
Treebank 2	English	54.12%

Moreover, the percentages of dependency directions fluctuate within a small range from 45% to 55%, indicating that the target language significantly determines the head-initial percentages of interpretations. Since in English, head-initial dependencies account for half of all dependency relations based on previous findings (Hudson, 2003; Liu, 2010; Y. Wang & Liu, 2017). This can be accounted from the perspective of cognitive capacity. Although each language has its own prevailing dependency direction, some short dependents are placed closer to the head and branch in the opposite direction rather than crowding consistently in the same dominant direction (Dryer, 1992; Liu, 2010). In such a way, shorter dependencies are obtained and cognitive burdens are reduced.

Generally speaking, all the interpretations are English and thus their percentages of head-initial dependencies are similar to native English, while the unique features of different source languages may exert certain influences on the simultaneous interpretation process. Arabic as a source language reveals such an obvious tendency, with its ratio larger than that of native English.

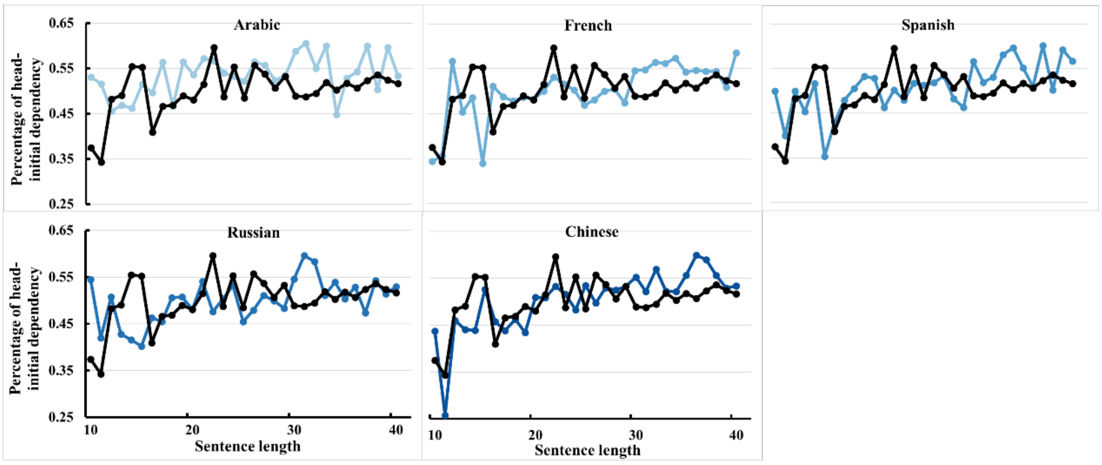


Figure 6. Percentages of head-initial dependencies of different sentence lengths with different source languages. The dark blue lines represent the relevant English values. Appendix G provides the raw data.

As Figure 6 displays, with the interaction of sentence lengths, Arabic-English interpretation texts tend to have more head-initial dependencies, whereas interpretations from other source languages do not show an obvious tendency compared with native English texts.

To explore whether simultaneous interpretations are liable to be influenced by diverse source languages and sentence lengths, a logistic regression model is fitted to predict the dependency direction with the sentence length. The model is highly significant with a weak correlation ($G = 60.27$; $df = 1$; $p < 0.0001$; $R^2 = 0.003$; $C = 0.525$). Taken different source languages into consideration, another model is further fitted. The results remain significant, yet the correlation is quite low ($G = 86.02$; $df = 6$; $p < 0.0001$; $R^2 = 0.004$; $C = 0.531$). The result of the likelihood ratio test between the two models is significant ($p < 0.0001$). However, the R^2 and C values indicate that the effect of source languages to dependency direction is quite small and their percentages only change within a limited range, as shown in the effect plot (Appendix H). This finding demonstrates the overwhelming power of cognitive capacity in language processing. According to Gibson (1998, 2000), it is a great burden to keep track of long incomplete dependencies on memory load, and impose cognitive demand on linking a new word into the existing sentence structure which seems to be influenced by dependency direction. Especially in such a highly demanding simultaneous interpreting process, the interpreters may strive to reduce cognitive burden as much as possible. Therefore, a subtler transformation of syntactic structures between the two languages makes language processing easier.

All in all, the source language only has a marginally significant effect on the interpretation process. The typology of different source languages exerts limited influence on the English interpretations due to the stronger role of cognitive factors in these simultaneous interpreting processes. In other words, source languages make little difference to the target interpretations during simultaneous interpretation.

4. Conclusion

Based on the two treebanks, our study suggests that different source languages have limited impact on English interpretations in the simultaneous interpreting processes. The effect of source languages is examined from the following two perspectives: one is cognitive factors via mean dependency distance; and the other is linguistic typology by means of dependency direction.

First and foremost, due to the complexity of the index - mean dependency distance, the distribution of dependency distances and percentages of adjacent dependency in interpretations are examined beforehand. All interpretations present similar regularities of the dependency distance distributions as the native English. Meanwhile, the percentages of adjacent dependencies fluctuate within a minor limited range. A logistic regression model is fitted to predict adjacent dependencies due to different source languages and sentence length. The model is quite significant but has a weak correlation, indicating that different source languages have little effect to the variability of percentages of adjacent dependencies. Then, the mean dependency distance is investigated thoroughly. The MDDs of different source languages have a similar rising tendency with the increase of sentence length. The Arabic to English interpretation has the smallest MDD while French the largest. The Indo-European languages-English interpretations all have similar larger MDDs than the native English texts, revealing the close relationship between mean dependency distance and human cognition. A linear regression model predicting mean dependency distance from sentence length with an interaction of different source languages is highly significant yet has a quite weak correlation. These findings demonstrate that the effect of source languages is closely correlated with human cognition constraints in a small scale. In other words, these findings coincide with the universal tendency of dependency distance minimization.

Next, the relationship between dependency relations in the English interpretations and linguistic typology is investigated via dependency direction. When it comes to the dependency direction, this study resorts to the percentages of head-initial dependencies. Their ratio also fluctuates within a limited range. A likelihood ratio test shows that a binary logistic regression model predicting percentages of head-initial dependencies with an interaction of source languages is significantly different from those without such an interaction. Yet, the correlation of the model is small. This presents source languages make a marginally significant variance to dependency directions under the limitation of human cognition.

To put it into a nutshell, this study investigates two essential properties of dependency relations, namely the cognitive part and the linguistic part, aiming to reveal the influence of source languages in the simultaneous interpretation process. Results indicate that the effects of source languages on dependency distances and dependency directions are modest, because of the well-acknowledged dependency distance minimization. Quantitative methods used in this study provide some insights to other researches. Further specific studies on interpretations would enable us to better understand what is happening in the simultaneous interpreting processes.

Acknowledgements

This work is partly supported by the National Social Science Foundation of China (Grant No. 17AYY021) and the Special Project for Humanities and Social Science, Zhejiang University (Quantitative Analysis of Sentence Hierarchical Structure in English).

REFERENCES

Altmann, G. (2013). Altmann-Fitter User Guide.

Baumgarten, N. (2007). Converging conventions? Macrosyntactic conjunction with English and and German und. *Text & Talk-An interdisciplinary journal of language, discourse communication studies, 27*(2), 139-170.

Baumgarten, N. (2008). Writer construction in English and German popularized academic discourse: The uses of we and wir: Walter de Gruyter GmbH & Co. KG.

Collins, M. J. (1996). *A new statistical parser based on bigram lexical dependencies.* Paper presented at the Proceedings of the 34th annual meeting on Association for Computational Linguistics.

Cowan, N. (2001). The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity. *The Behavioral and brain sciences, 24*, 87-114; discussion 114. doi: 10.1017/S0140525X01003922

De Marneffe, M.-C., & Manning, C. D. (2008). Stanford typed dependencies manual:

Technical report, Stanford University.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 81-138.

Eppler, E. (2010). *Emigranto: The syntax of german-english code-switching* (Vol. 99): Braumüller.

Eppler, E. (2013). *Dependency distance and bilingual language use: evidence from German/English and Chinese/English data.* Paper presented at the Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013).

Ferrer-i-Cancho, R., & Arias, M. (2013). Non-linear Regression on Dependency Trees. Lecture on Complex and Social Networks (2013–2014).

Ferrer-i-Cancho, R., & Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory, 5*(2), 143-155.

Fischer, K. (2007). Komplexität und semantische Tranparenz im Deutschen und Englischen. *Sprachwissenschaft, 32*(4), 355-405.

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 95-126.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science, 34*(2), 286-310.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language, 2*, 73-113.

Gries, S. T. (2013). *Statistics for Linguistics with R: A Practical Introduction (Mouton Textbook)*: De Gruyter Mouton.

Hawkins, J. A. (1994). *A performance theory of order and constituency* (Vol. 73): Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*: Oxford University Press on Demand.

Heringer, H. J., Strecker, B., & Wimmer, R. (1980). *Syntax* (Vol. 251): W. Fink.

Hiranuma, S. (1999). Syntactic difficulty in English and Japanese: A textual study. *UCL Work. Pap. Linguist, 11*, 309-322.

House, J. (2011a). Linking constructions in English and German translated and original texts. *Multilingual discourse production: Diachronic and synchronic perspectives*, 163-182.

House, J. (2011b). Using translation and parallel text corpora to investigate the influence of global English on textual norms in other languages. *Corpus-based translation studies: Research and applications*, 187-208.

Hudson, R. (1995). Measuring syntactic difficulty. *Manuscript, University College, London*.

Hudson, R. (2003). The psychological reality of syntactic dependency relations. *MTT2003, Paris*.

Hudson, R. (2007). *Language networks: The new word grammar*: Oxford University Press.

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*, 93-104.
doi: 10.1016/j.langsci.2015.04.002

Kajzer-Wietrzny, M. (2012). Interpreting universals and interpreting style. *Unpublished PhD*

Dissertation, Adam Mickiewicz University, Poznan, Poland. Available at: https://repozytorium. amu. edu. pl/jspui/bitstream/10593/2425/1/Paca% 20doktorska% 20Marty% 20Kajzer-Wietrzny. pdf [last accessed 11 August 2013].

Kranich, S. (2011). To hedge or not to hedge: the use of epistemic modal expressions in popular science in English texts, English–German translations, and German original texts. *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies, 31*(1), 77-99.

Kranich, S., House, I., & Becher, V. (2012). Changing conventions in English-German translations of popular scientific texts. *Multilingual individuals and multilingual societies, 13*, 315.

Liang, J., Fang, Y., Lv, Q., & Liu, H. (2017). Dependency Distance Differences across Interpreting Types: Implications for Cognitive Demand. *Front Psychol, 8*, 2132. doi: 10.3389/fpsyg.2017.02132

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics, 15*, 1-12.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159-191.

Liu, H. (2009). Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics, 16*(3), 256-273.

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120*(6), 1567-1578.

Liu, H., & Huang, W. (2006). *A Chinese Dependency Syntax for Treebanking.* Paper presented at the Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation.

Liu, H., Hudson, R., & Feng, Z. (2009). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory, 5*(2), 161-174.

Liu, H., & Xu, C. (2012). Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics, 48*. doi: 10.1515/psicl-2012-0027

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Phys Life Rev, 21*, 171-193.

Liu, H., Zhao, Y., & Li, W. (2009). Chinese Syntactic and Typological Properties Based on Dependency Syntactic Treebanks. *Poznań Studies in Contemporary Linguistics, 45*(4). doi: 10.2478/v10010-009-0025-3

Malamatidou, S. (2013). Passive voice and the language of translation: A comparable corpus-based study of modern Greek popular science articles. *Meta: Journal des traducteurs/Meta: Translators' Journal, 58*(2), 411-429.

Mizuno, A. (2005). Process model for simultaneous interpreting and working memory. *Meta: Journal des Traducteurs/Meta: Translators' Journal, 50*(2), 739-752.

Oya, M. (2011). *Syntactic dependency distance as sentence complexity measure.* Paper presented at the Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics.

Oya, M. (2013). *Degree centralities, closeness centralities, and dependency distances of different genres of texts.* Paper presented at the Selected Papers from the 17th International Conference of Pan-Pacific Association of Applied Linguistics.

Padilla, F., Bajo, M. T., & Macizo, P. (2005). Articulatory suppression in language interpretation: Working memory capacity, dual tasking and word knowledge. *Bilingualism: Language and Cognition, 8*(3), 207-219.

Seeber, K. G. (2005). Temporale Aspekte der Antizipation beim Simultandolmetschen von SOV-Strukturen aus dem Deutschen. *Bulletin VALS-ASLA, 81*, 123-140.

Setton, R. (1998). Meaning assembly in simultaneous interpretation. *Interpreting, 3*(2), 163-199.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition, 105*(2), 300-333.

Tesnière, L. (1965). Eléments de syntaxe structurale.

Uchiyama, H. (1991). Problems caused by word order when interpreting/translating from English into Japanese: The effect of the use of inanimate subjects in English. *Meta: Journal des Traducteurs/Meta: Translators' Journal, 36*(2-3), 404-413.

vAn Besien, F., & Meuleman, C. (2008). Style differences among simultaneous interpreters: A pilot study. *The Translator, 14*(1), 135-155.

Wang, L., & Liu, H. (2013). Syntactic variations in Chinese–English code-switching. *Lingua, 123*, 58-73.

Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences, 59*, 135-147.
doi: 10.1016/j.langsci.2016.09.006

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society, 104*(5), 444-466.

Appendix

Appendix A. Formulae of these distributions

(1) The formula for the Right truncated modified Zipf-Alekseev distribution:

$$P_x = \begin{cases} \alpha, & x = 1 \\ \dfrac{(1-\alpha)x^{-(a+b\ln x)}}{T}, & x = 2,3,\ldots,n \end{cases}$$

where

$$T = \sum_{j=2}^{n} j^{-(a+b\ln j)}, a,b \in \Re, 0 < \alpha < 1$$

(2) The formula for the Right truncated Waring distribution:

$$P_x = c\frac{a^{(x)}}{(a+b+1)^{(x)}}, x = 0,1,2,\ldots,n$$

(3) The formula for the Right truncated zeta distribution:

$$P_x = \frac{1}{x^a[\Phi(1,0,a) - \Phi(1,R,a)]}, \quad x = 1,2,\dots,R$$

Appendix B. Raw data of Figure 3.

	Arabic		French		Spanish		Russian		Chinese		English	
DD	EV	TV	EV	TV	EV	TV	EV	TV	EV	TV	EV	TV
1	355.00	402.46	361.00	395.95	343.00	381.98	345.00	388.49	350.00	399.58	360.00	406.20
2	182.00	115.05	175.00	114.78	173.00	112.02	170.00	114.85	190.00	116.63	181.00	115.81
3	78.00	55.31	78.00	55.62	72.00	54.66	80.00	56.30	83.00	56.76	79.00	55.58
4	33.00	32.89	25.00	33.27	30.00	32.85	33.00	33.95	29.00	34.04	28.00	33.02
5	17.00	21.98	17.00	22.33	13.00	22.14	18.00	22.93	14.00	22.90	21.00	22.05
6	12.00	15.81	10.00	16.12	9.00	16.03	16.00	16.64	7.00	16.57	9.00	15.85
7	5.00	11.97	9.00	12.24	10.00	12.20	11.00	12.69	6.00	12.60	7.00	11.99
8	5.00	9.40	2.00	9.64	9.00	9.64	10.00	10.04	10.00	9.94	12.00	9.41
9	3.00	7.60	9.00	7.81	5.00	7.82	5.00	8.16	7.00	8.06	3.00	7.61
10	7.00	6.28	1.00	6.47	9.00	6.49	4.00	6.78	5.00	6.69	4.00	6.29
11	2.00	5.29	3.00	5.46	4.00	5.48	4.00	5.73	6.00	5.64	3.00	5.29
12	4.00	4.52	4.00	4.67	7.00	4.70	4.00	4.92	1.00	4.84	3.00	4.52
13	3.00	3.91	2.00	4.05	4.00	4.08	2.00	4.27	3.00	4.19	3.00	3.91
14	3.00	3.42	1.00	3.55	2.00	3.58	2.00	3.75	5.00	3.68	1.00	3.42
15	0.00	3.02	3.00	3.14	2.00	3.17	2.00	3.32	3.00	3.25	3.00	3.02
16	2.00	2.69	1.00	2.80	0.00	2.83	1.00	2.97	2.00	2.90	1.00	2.68
17	2.00	2.41	4.00	2.51	1.00	2.54	1.00	2.67	2.00	2.60	1.00	2.41
18	1.00	2.17	3.00	2.27	0.00	2.29	3.00	2.41	1.00	2.35	0.00	2.17
19	2.00	1.97	0.00	2.06	0.00	2.08	0.00	2.19	0.00	2.14	2.00	1.97
20	0.00	1.80	4.00	1.88	2.00	1.90	0.00	2.00	0.00	1.95	0.00	1.79
21	0.00	1.64	1.00	1.72	0.00	1.75	0.00	1.84	1.00	1.79	1.00	1.64
22	1.00	1.51	0.00	1.58	0.00	1.61	2.00	1.70	0.00	1.65	0.00	1.51
23	0.00	1.40	0.00	1.46	1.00	1.49	0.00	1.57	0.00	1.52	0.00	1.39
24	0.00	1.29	0.00	1.35	0.00	1.38	0.00	1.45	0.00	1.41	0.00	1.29
25	0.00	1.20	1.00	1.26	0.00	1.28	0.00	1.35	0.00	1.31	0.00	1.20

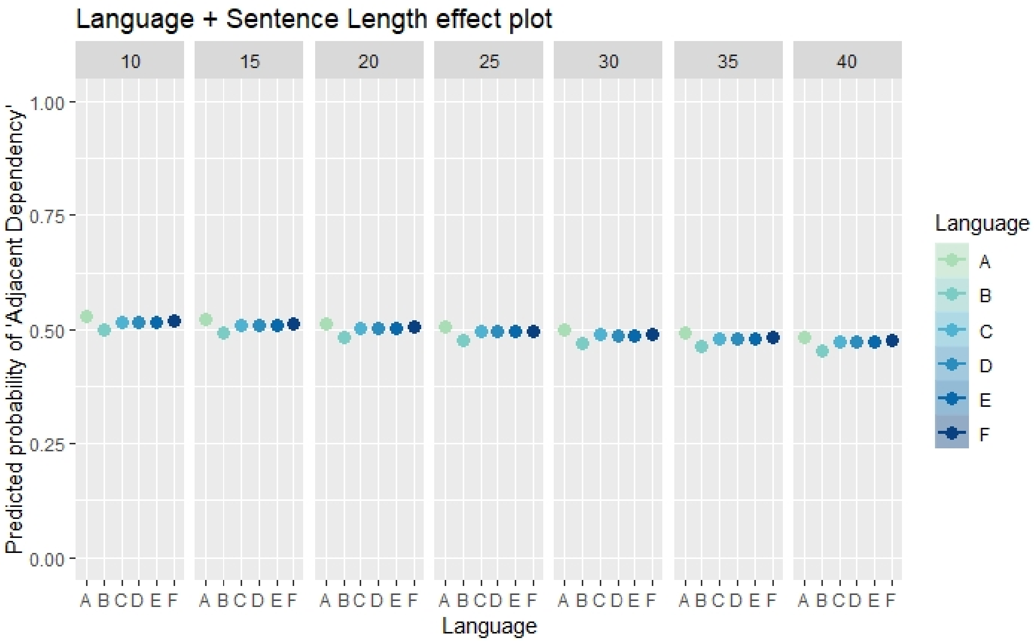
Note: DD refers to absolute value of dependency distance. Empirical values and theoretical values are abbreviated to EV and TV.

Appendix C. Raw data of Figure 4.

Sentence Length	Arabic	French	Russian	Spanish	Chinese	English
10	50.0%	50.0%	63.6%	62.5%	62.5%	50.0%
11	58.1%	52.5%	51.9%	48.0%	42.9%	45.7%
12	52.9%	60.0%	57.0%	56.7%	52.0%	54.0%
13	52.3%	47.2%	47.6%	54.7%	51.7%	56.6%

14	52.8%	50.0%	47.5%	51.7%	50.9%	57.1%
15	51.5%	44.0%	47.1%	51.0%	50.0%	47.4%
16	50.9%	50.4%	51.0%	57.1%	50.5%	48.2%
17	53.1%	48.1%	51.7%	53.4%	46.8%	50.5%
18	60.0%	52.1%	50.6%	49.5%	46.3%	48.4%
19	49.6%	50.7%	46.9%	49.7%	44.4%	49.0%
20	52.6%	48.7%	46.2%	54.1%	50.3%	48.1%
21	55.4%	44.5%	55.0%	40.7%	48.0%	52.1%
22	49.7%	52.6%	48.0%	47.6%	45.1%	55.3%
23	45.5%	52.2%	51.4%	48.7%	46.6%	43.0%
24	49.3%	49.1%	53.2%	50.0%	44.7%	49.4%
25	47.8%	48.5%	50.0%	48.7%	48.6%	47.0%
26	49.5%	50.0%	47.3%	49.6%	48.6%	52.8%
27	53.2%	50.9%	45.9%	49.1%	48.1%	52.8%
28	48.1%	50.5%	47.8%	54.5%	44.4%	49.3%
29	47.9%	49.4%	46.8%	46.3%	48.9%	50.3%
30	51.7%	46.9%	51.7%	48.5%	52.2%	47.9%
31	53.8%	46.8%	57.9%	46.7%	45.7%	51.2%
32	51.5%	51.5%	50.0%	44.8%	50.9%	47.0%
33	53.9%	46.4%	47.2%	48.4%	47.8%	45.8%
34	41.4%	48.9%	51.6%	51.5%	44.6%	48.1%
35	48.0%	52.1%	48.8%	48.0%	45.6%	46.8%
36	44.2%	51.5%	46.2%	49.0%	47.7%	48.0%
37	51.9%	45.9%	45.8%	51.8%	48.0%	49.4%
38	46.7%	51.5%	54.3%	46.8%	43.8%	48.6%
39	47.2%	50.8%	46.6%	50.9%	47.1%	46.6%
40	50.7%	43.2%	47.0%	44.8%	41.5%	48.9%

Appendix D. The effect plot of the binary logistic regression of adjacent dependency predictions in regard to the sentence length and different source languages.



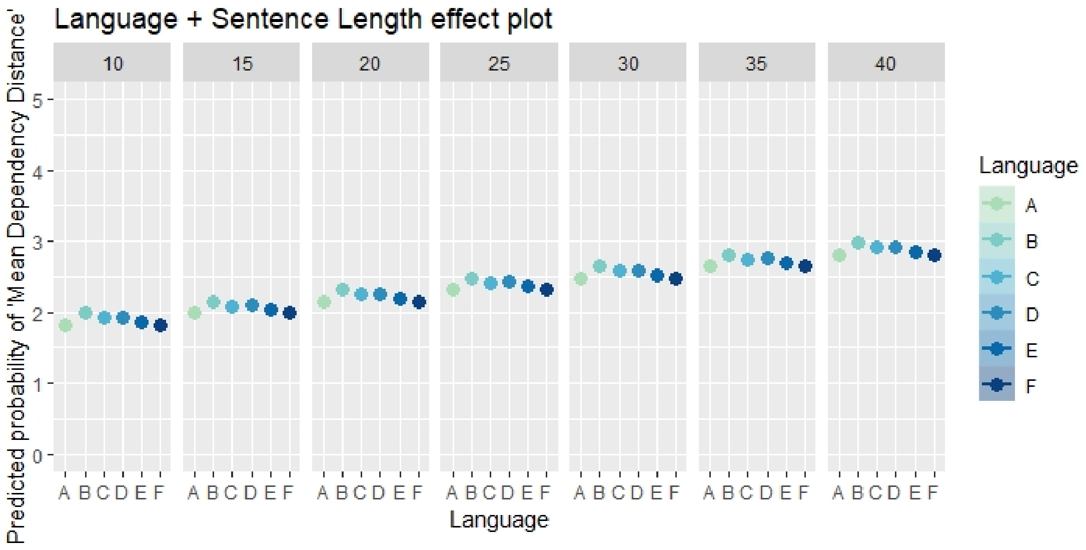
In each panel, the x-axis stands for texts from two treebanks. A is Arabic; B Chinese; C French; D Russian; E Spanish; F English (Appendix F and H have the same x-axis). The y-axis represents the predicted probability of adjacent dependency.

Appendix E. Raw data of Figure 5.

Sentence Length	Arabic	French	Russian	Spanish	Chinese	English
10	1.750	1.781	1.436	1.938	1.563	1.750
11	1.710	2.153	1.864	1.880	2.171	1.914
12	1.794	1.600	1.852	1.733	1.880	1.793
13	2.015	2.204	2.071	1.987	1.907	1.792
14	1.811	2.049	2.277	1.833	2.079	1.790
15	1.950	2.260	2.149	1.843	2.092	2.184
16	1.994	2.117	2.135	2.214	2.206	2.072
17	2.225	2.160	2.345	1.932	2.348	2.107
18	1.933	2.071	2.312	2.376	2.231	2.008
19	2.458	2.014	2.453	2.091	2.394	2.275
20	2.015	2.373	2.151	2.344	2.026	2.538
21	2.059	2.404	2.358	2.556	2.315	2.267
22	2.222	2.151	2.319	2.304	2.497	2.219
23	2.571	2.588	2.222	2.487	2.342	2.537
24	2.314	2.181	2.420	2.265	2.475	2.369
25	2.640	2.385	2.481	2.503	2.568	2.418
26	2.333	2.525	2.768	2.391	2.578	2.075

27	2.255	2.190	2.583	2.231	2.235	2.196
28	2.574	2.824	2.320	2.214	2.611	2.777
29	2.274	2.577	2.595	2.823	2.659	2.395
30	2.459	3.053	2.942	2.485	2.526	2.511
31	2.430	2.830	2.404	2.729	2.630	2.518
32	2.288	2.515	2.556	3.089	3.145	2.482
33	2.487	2.678	2.639	2.478	2.589	2.377
34	3.379	2.624	2.415	2.655	2.576	2.574
35	2.702	2.644	3.134	2.692	2.628	2.759
36	2.553	2.856	2.938	2.751	2.923	2.561
37	2.496	2.662	2.792	2.614	2.480	2.250
38	2.800	2.529	2.143	2.323	3.079	3.072
39	2.528	2.698	2.699	2.287	4.147	2.683
40	2.635	3.148	3.291	3.005	2.909	2.635

Appendix F. The effect plot of the linear regression of mean dependency distance predictions in regard to the sentence length and different source languages.

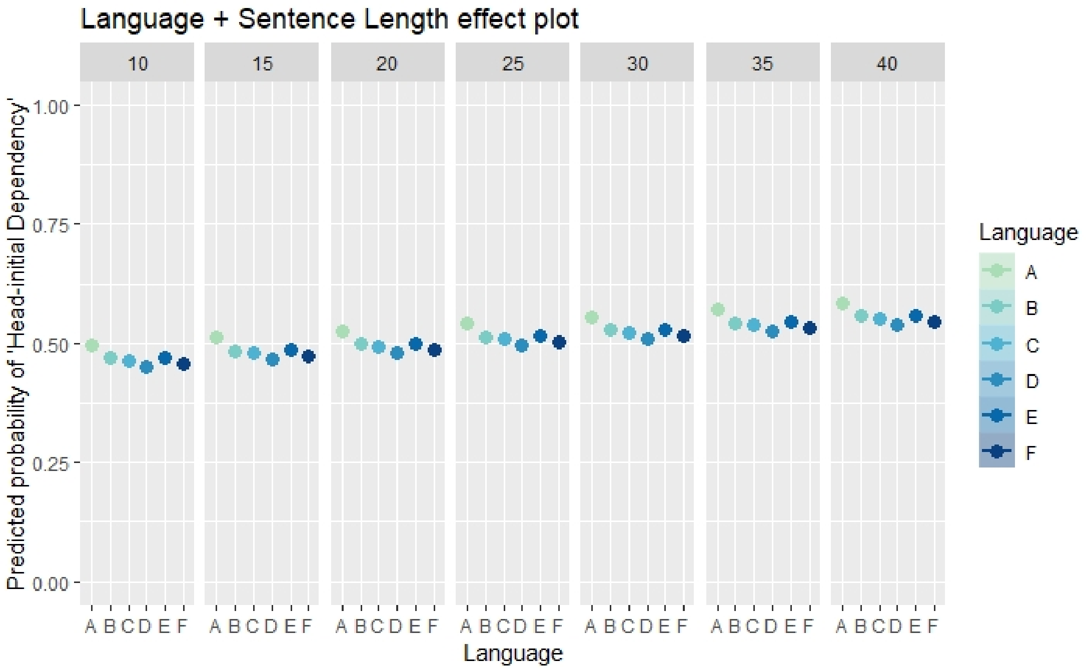


In each panel, the x-axis stands for texts from two treebanks, similar to the Appendix D. The y-axis represents the predicted probability of mean dependency distance.

Appendix G. Raw data of Figure 6.

Sentence Length	Arabic	French	Russian	Spanish	Chinese	English
10	53.1%	34.4%	54.5%	50.0%	43.8%	37.5%
11	51.6%	35.6%	42.0%	40.0%	25.7%	34.3%
12	45.6%	56.7%	50.8%	50.0%	46.0%	48.3%
13	46.9%	45.4%	42.9%	45.3%	44.1%	49.1%
14	46.2%	48.6%	41.6%	51.7%	43.9%	55.5%
15	51.5%	34.0%	40.2%	35.3%	52.6%	55.3%
16	49.7%	51.1%	46.4%	42.9%	45.8%	41.0%
17	56.3%	48.8%	45.5%	47.9%	43.8%	46.6%
18	46.7%	47.9%	50.6%	50.5%	46.3%	46.9%
19	56.5%	48.6%	50.8%	53.3%	43.4%	49.0%
20	53.6%	48.2%	48.1%	52.9%	51.0%	48.1%
21	57.2%	50.0%	54.1%	46.3%	50.9%	51.5%
22	56.7%	53.1%	47.6%	50.3%	53.3%	59.6%
23	54.0%	51.6%	50.4%	47.9%	51.6%	48.8%
24	53.3%	50.3%	53.2%	51.8%	48.2%	55.4%
25	52.2%	46.9%	45.5%	51.3%	53.5%	48.5%
26	56.5%	48.1%	47.9%	51.9%	49.7%	55.7%
27	55.7%	50.0%	51.1%	53.4%	53.0%	53.7%
28	52.3%	50.5%	50.0%	48.3%	52.5%	50.7%
29	53.4%	47.4%	48.3%	46.3%	53.3%	53.3%
30	58.8%	54.6%	54.6%	56.6%	55.2%	48.9%
31	60.5%	54.8%	59.6%	51.9%	52.2%	48.8%
32	55.0%	56.4%	58.3%	53.1%	57.0%	49.5%
33	60.0%	56.2%	51.1%	58.1%	52.2%	51.9%
34	44.8%	57.4%	53.9%	59.7%	52.2%	50.3%
35	52.8%	54.3%	50.4%	55.2%	55.7%	51.8%
36	54.3%	54.6%	52.9%	51.0%	60.0%	50.7%
37	60.0%	54.4%	47.4%	60.2%	59.0%	52.4%
38	50.4%	54.4%	54.3%	50.2%	55.7%	53.6%
39	59.7%	50.8%	51.5%	59.3%	52.9%	52.4%
40	53.4%	58.5%	53.0%	56.7%	53.4%	51.7%

Appendix H. The effect plot of the binary logistic regression of head-initial dependency predictions in regard to the sentence length and different source languages.



In each panel, the x-axis stands for texts from two treebanks, similar to the Appendix D. The y-axis represents the predicted probability of head-initial dependency.

Glottometrics 45, 2019, 114-123

Frequency and Length of Syllables in Serbian

Marija Radojičić[1]
Biljana Lazić[2]
Sebastijan Kaplar[3]
Ranka Stanković[4]
Ivan Obradović[5]
Ján Mačutek[6]
Lívia Leššová[7]

Abstract. Basic analyses of several properties of syllables (the rank-frequency distribution, the distribution of length, and the relation between length and frequency) in Serbian is presented. The syllabification algorithm used combines the maximum onset principle and the sonority hierarchy. Results indicate that syllables behave similarly to words as far as mathematical models are concerned, but values of parameters in models for syllables are quite different from those for words.

Keywords: *syllable frequency, syllable length, Serbian*

1. Introduction

Syllable is a language unit which „has become a stepchild in linguistic description“ (Haugen, 1956, p. 213) because of the lack of its precise definition[8] (cf. also Crystal, 2008, pp. 467-468;

[1] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia, and Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: marija.radojicic@uns.ac.rs
[2] Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: biljana.lazic@rgf.bg.ac.rs
[3] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21101 Novi Sad, Serbia
e-mail: kaplar@uns.ac.rs
[4] Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: ranka@rgf.rs
[5] Faculty of Mining and Geology, University of Belgrade, Đušina 7, 11120 Beograd, Serbia
e-mail: ivan.obradovic@ rgf.bg.ac.rs
[6] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 84248 Bratislava, Slovakia
e-mail: jmacutek@yahoo.com
[7] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 84248 Bratislava, Slovakia
e-mail: livia.lessova@fmph.uniba.sk

[8] It is quite common that there are several definitions of a linguistic unit (cf. e.g. Crystal, 2008, pp. 521-523 for word, p. 367 for phrase, and pp. 432-433 for sentence). However, syllable seems to be more problematic than other units – here we do not face the problem of having to choose from among several established definitions (introduced by different linguistic schools), but the lack of a proper definition as such.

Cairns & Raimy, 2011, p. 1; Ladefoged & Johnson, 2011, p. 310). Consequently, it is very difficult to conduct a systematic study of syllable properties, as different definitions – which are to be expected if there is no established approach – inevitably lead to results which are not comparable (at the very least not directly). Quantitative linguistics also suffers from this problem. Investigations on the level of syllables appear relatively rarely.[9] In the situation described above, with a general syllable definition lacking, a scientist can apply language-specific rules for syllabification (e.g. using morpheme borders as one of the criteria for syllable borders). While the application of language-specific rules is not bad per se, if one wants to compare models, parameter values etc., a general approach to all languages under investigation is indispensable.

If a language allows only open syllables (such as Old Slavonic, cf. Rottmann, 1999), the syllabification is straightforward (provided that diphthongs – if the language under investigation contains any – can be reliably distinguished from sequences of two adjacent monophthongs). Consonant clusters (especially in intervocalic positions) are the most problematic aspect of syllabification. The problem can be solved either empirically, with the help of native speakers (or, in a psycholinguistic research, relying fully on them), or by following syllabification rules prescribed by an authority, or theoretically, establishing rules for syllable borders. Experiments were carried out e.g. by Rubach & Booij (1990) for Polish, by Schiller et al. (1996, 1997) for Dutch, and by Eddington et al. (2013a,b) for American English[10]. Rottmann (2002) acknowledges consultations with native speakers of some Slavic languages in cases of more complicated consonant clusters. The second approach was chosen e.g. by Best (2011, 2013), who refers to a prestigious German pronunciation dictionary (which suggests also syllabification rules).

The approach according to which only those syllable onsets exist that are observable word-initially, and those syllable codas that occur word-finally (cf. e.g. Kelih, 2012), is perhaps the best known theoretical framework. A more detailed description can be found in Pulgram (1970). However, this approach requires a comprehensive dictionary that contains practically all words used in a language. Lehfeldt (1971) presented a modification, distinguishing between marginal (rarely occurring and considered to be exceptions) and non-marginal (found with a high frequency) consonant clusters at beginnings and ends of words; only those which are not marginal are allowed to form syllable onsets and codas. If one follows his modification, a large enough corpus is needed to perform statistical tests, based on which a decision on the (non-) marginality of a particular consonant cluster is made. Finding or creating such a corpus can be problematic for minor languages (such as e.g. Lower and Upper Sorbian among Slavic languages). In addition, the rules derived from Pulgram's approach can change relatively quickly, as lexicon is one of the more dynamic language features. Therefore, we follow another approach, namely, a combination of the maximum onset principle and the sonority sequencing principle.

The paper is organized as follows. The syllabification algorithm is described in Section 2. Section 3 presents some properties of Serbian phonology that are relevant for syllabification, and the Serbian alphabets (both Latin and Cyrillic). Then, the language material used is introduced. In Section 4, mathematical models for syllable properties under study (the rank-frequency distribution, the distribution of length) are suggested, together with parameter estimations and

[9] See e.g. the bibliography by Karl-Heinz Best at http://wwwuser.gwdg.de/~kbest/litlist.htm and compare the number of entries for syllables and for words.

[10] Needless to say, the lists of works mentioned here as examples is by no means exhaustive.

goodness-of-fit evaluation. The relation between length and frequency of syllables is also discussed. Section 5 contains concluding remarks.

2. Methodology

The maximum onset principle (Pulgram, 1970) requires that the sylable onset be the longest allowed[11] (i.e., as many consonants in intervocalic positions as possible are attached to the onset). Allowed onsets are determined by the sonority sequencing principle, according to which „[b]etween any member of a syllable and the syllable peak, a sonority rise or plateau[12] must occur“ in the onset (Blevins, 1995, p. 210).

A sonority hierarchy must be established based on which the behaviour of phoneme sequences with respect to the sonority sequencing principle can be evaluated. Several sonority scales were suggested (Blevins, 1995, p. 210: „[s]uch scales come in a variety of types ... fine-grained vs. not-so-fine-grained“), see e.g. Clements (1990) or Zec (1995). We chose perhaps the simplest one – we distinguish only sonorant and obstruent consonants, with approximants and nasals being sonorants. Admittedly, this scale puts many consonants with different phonological characteristics into one category (e.g. stops and fricatives); however, according to Zec (1995, p.86), it „is not nearly as elaborate as some of the scales proposed in the literature, but is sufficient to capture the most common subdivisions of segments with respect to sonority”.

To sum up, in this paper we divide words into syllables using the following algorithm:

1. In the first step, all syllables end after their nuclei (i.e., after a vowel or a syllabic consonant). The maximum onset principle is „blindly“ respected in this step, and thus, preliminarily, all syllables are kept open.
2. If, after Step 1, consonant clusters occur in intervocalic positions, the borders between syllables are reconsidered taking into account the sonority sequencing principle.

If some irregularities which contradict these two principles occur at the beginning of a word (i.e. if a word begins with a consonant cluster in which sonority decreases; examples from different languages are presented in Clements, 1990, p. 288), we take these onsets as they are.

It must be noted that our choice of syllable definition is motivated purely by pragmatic reasons, as it is easy to implement automatically and it is applicable to (almost) all languages.[13] We do not have the ambition to introduce a definition which would be better than other options, e.g. the ones mentioned in Section 1.

We divide words into syllables, hence the definition of word we use deserves a mention. We define words orthographically, as sequences of letters between spaces. We are aware of problems related to this definition, but it facilitates easy automatic text processing (see e.g. a discussion on this topic in Antić et al., 2006, pp. 118-121). The text under analysis (see Section 3) is pre-processed, so that it does not contain any zero-syllable words.

[11] The maximum onset principle implies the minimal codas.

[12] Many authors (e.g. Clements, 1990) speak about a strict increase of sonority.

[13] E.g. Berber languages can be problematic, see Ridouane (2008).

3. Language material

Serbian is a South Slavic language. It has the official status in Serbia (exclusively) and in Bosnia and Herzegovina (as one of three languages, together with Bosnian and Croatian), and the status of a minority language in several other countries. Given the scope of our research, we briefly mention the Serbian phonology and orthography; more information on the language can be found e.g. in Browne (1993).

The Serbian phonological system consists of 30 phonemes - 5 vowels and 25 consonants, out of which 8 are sonorants (Stanojčić & Popović, 1999, or Piper & Klajn, 2013). By manner of articulation, phonemes are classified as plosives (their graphemic representations are b, p, d, t, g, k), affricates (c, č, ć, dž, đ), fricatives (f, z, s, ž, š, h), nasals (m, n, nj), laterals (l, lj), a vibrant (r) and semivowels (v, j). The Serbian language uses two alphabets: Latin and Cyrillic. Serbian graphemes are presented in Table 1, first Latin ones, then, in brackets, their Cyrillic equivalents[14]. Every phoneme in Serbian can be presented by a grapheme or by a digraph, in accordance with the principle "write as you speak". In Cyrillic script, every grapheme represents one sound. In Latin script, there are three digraphs – dž, nj, and lj (Cyrillic equivalents: џ, њ, љ), which are pronounced as one sound.

Table 1.
Graphemic representation of phonemes in Serbian language

vowels		a(а), e(е), i(и), o(о), u(у)
consonants	sonorants	j(ј), l(л), lj(љ), m(м), n(н), nj(њ), r(р), v(в)
	obstruents	b(б), c(ц), č(ч), ć(ћ), d(д), dž(џ), đ(ђ), f(ф), g(г), h(х), k(к), p(п), s(с), š(ш), t(т), z(з), ž(ж)

Two further aspects of Serbian must be taken into consideration. First, the consonant r is syllabic (i.e. it forms a syllable nucleus) if it is surrounded by two other consonants; e.g *srce* (heart) is a two-syllabic word (syllabified *sr-ce*). Second, there are two zero-syllable words in Serbian, both prepositions – *k* and *s* –, which are, following the approach from Antić et al. (2006), attached to the words which they precede.

As an example we present an application of the algorithm described in Section 2 to the first sentence from the Universal Declaration of Human Rights (in English: All human beings are born free and equal in dignity and right):

Sva ljudska bića rađaju se slobodna i jednaka u dostojanstvu i pravima.
Sva lju-dska bi-ća ra-đa-ju se slo-bo-dna i je-dna-ka u do-sto-jan-stvu i pra-vi-ma.

We apply the algorithm to the complete Serbian translation of the Russian socialist realist novel "*Kak zakalyalas' stal*"" (How the Steel Was Tempered) by N. Ostrovsky. The choice is motivated by the fact that a parallel corpus consisting of the first ten chapters of the novel and their translations to all standard Slavic languages (except for Lower Sorbian) is available (Kelih, 2009), which will make possible to conduct typological studies on the level of syllable when

[14] The Cyrillic alphabet follows a different order of letters, see e.g. Comrie (1996), p. 704.

automatic tools for syllabification of other Slavic languages are prepared.[15] The output of the automatic syllabification was manually checked and several mistakes (caused most probably by OCR deficiencies) were corrected or deleted (e.g. abbreviations).

4. Results

The syllabified text provides a valuable source of data (word forms: 114348 tokens, 21378 types; syllables: 239219 tokens, 2417 types) which can be used to investigate many properties of syllables. In this paper we limit ourselves to analyses of three aspects: 1) the rank-frequency distribution, 2) the distribution of length, and 3) the relation between length and frequency. The goodness-of-fit of a model is evaluated in terms of the discrepancy coefficient $C = \chi^2/N$, where χ^2 is the value of the test statistic from the Pearson χ^2 goodness-of-fit test and N is the sample size. As a rule of thumb, the fit is considered satisfactory if $C < 0.02$ (Mačutek & Wimmer, 2013).

Strauss et al. (2008, p. 11) formulated the hypothesis that „[t]he rank-frequency distribution of syllables behaves like the rank-frequency distribution of words". Word frequencies mostly follow Zipf-like distributions (Köhler, 2005; Popescu et al., 2009, pp. 127-142); according to the abovementioned hypothesis, the rank-frequency distribution of Serbian syllables (see Table 2, full data can be found at rgf.rs/projekti/bil/sk/results/KakoSeKalioCelik_2019_01_14.xlsx) can be modelled by one of these distributions as well.

Table 2.
Rank-frequency distribution of syllables in Serbian

rank	frequency	syllable
1	10103	o
2	6970	je
3	5778	u
4	5291	na
5	5248	da
6	4827	i
7	4436	se
8	4278	po
9	4252	ko
10	4062	ne
⋮	⋮	⋮
2417	1	ut

The Zipf-Mandelbrot distribution (Wimmer & Altmann, 1999, p. 666),

[15] In addition to works by Rottmann (1999, 2002) already mentioned in Section 1, syllables in Slavic languages were studied within the framework of quantitative linguistics in several other papers. However, borders between syllables were determined either using language-specific rules (Obradović et al., 2010, for Serbian; Meštrović et al, 2015, for Croatian), or using the approach suggested by Pulgram (1970) and modified by Lehfeldt (1971), with its drawback of needing a sufficiently large corpus (Kelih & Mačutek, 2013, for Russian and Slovene), or not at all (because the mean syllable length in words was sufficient for the purposes of the research, as in Mačutek & Rovenchak, 2011, for Ukrainian).

$$P_x = \frac{k}{(x+b)^a}, \qquad x = 1,2,\ldots,n,$$

achieves a good fit ($C = 0.0177$) for parameter values $a = 1.87$, $b = 30.12$ (we remind that the distribution has two parameters; k is a normalization constant and not an independent parameter, i.e. its value depends on parameters a and b). These parameter values are out of the range of values for rank-frequency distribution for word forms[16] (cf. Popescu et al. 2009, pp. 137-138; the highest value of a is 1.6543 for a Hawaiian text, i.e. for a text written in a very analytical language). It can be a consequence of the fact that the inventory of syllables is, at least for Slavic languages, much more restricted that the one of words. The trend of the empirical repeat rate ($RR = (\sum_{i=1}^{K} f_i^2)/N^2$, with K being the inventory size, N the sample size, and f_i, $i = 1,\ldots,K$ the frequencies) to decrease with the increasing inventory size is presented e.g. by Kelih (2013) for graphemes; it can be presumed that, in general, the less different units are available to the language user, the more often they will be repeated. In our text we have $RR = 0.0098$ for syllables (2417 types) and $RR = 0.0059$ for words (21378 types). The repeat rate is one of the characteristics of an empirical distribution; its values are reflected also in the parameter values.

An analogy in the behaviour of syllables and words can be observed also with respect to their length (frequencies of syllable length can be found in Table 2). Word length is usually modelled by the Poisson distribution or by one of its generalizations or modifications, see e.g. Best (2005) and Popescu et al. (2013).

Table 3.
Distribution of syllable length in Serbian

length	frequency
1	23505
2	135938
3	54556
4	6982
5	236
6	2

The data can be fitted e.g. by the hyper-Poisson distribution[17] (Wimmer & Altmann, 1999, pp. 281-282),

$$P_x = k\,\frac{a^{x-1}}{b^{(x-1)}}, x = 1,2,\ldots,$$

[16] Parameters values are not directly comparable, as the Zipf-Mandelbrot distribution is not a good model for word frequencies in the language material we used ($C = 0.0880$). Given that we work with a complete novel consisting of 110104 words, it is necessarily a text mixture rather than a homogeneous text (Popescu et al., 2009, set an upper limit - admittedly an arbitrary one - of 10000 words for a homogeneous text, see p.3). Lower language units, such as graphemes, phonemes, or syllables, which do not bear a meaning (at least not in the full sense of the word) can behave regularly even in text mixtures.

[17] This distribution is usually defined for $x = 0,1,2,\ldots$, i.e. it is shifted here to the right by 1.

with $a = 0.3410$, $b = 0.0521$, and $C = 0.0050$ (k is, again, a normalization constant). As several other Poisson-like distributions also fit the data very well, we postpone any attempts to formulate conclusions that could be deduced from the model and parameter values until data for more languages are available.

Stretching the analogy between words and syllables even further, one can suppose that more frequent syllables are shorter.[18] Indeed, the value of the Spearman correlation coefficient between syllable frequency and length in the text under analysis is -0.397. It is quite clearly statistically significant, with p-value < 0.001. The negative correlation between frequency and length of syllables seems to be stronger that the one for words[19], for which the Spearman correlation attains value -0.267 if word length is measured in syllables, and -0.299 if word length is measured in letters[20] (statistically significant also in both of these cases).

The tendency to favour shorter syllables is obvious also from Table 4. Data from Table 2 were pooled so that each group contained at least 20000 syllables (tokens), and the weighted mean of syllable length (with frequencies serving as the weights; differences between the weighted means and the means computed without the weights are negligible) was calculated in each group. Syllables with higher ranks (i.e., with higher frequencies) are, on average, shorter.

Table 4.
Mean syllables length for pooled data

ranks	mean length
1-3	1.31
4-8	1.80
9-14	2.00
15-23	2.00
24-34	1.91
35-47	2.08
48-66	2.16
67-97	2.28
98-155	2.50
156-309	2.91
310-2417	3.18

[18] This hypothesis (now known as the law of brevity) was first formulated for words by Zipf (1935).

[19] Ferrer-i-Cancho & Hernández-Fernández (2013) provide Spearman correlations between word frequency and length (measured in the number of letters) in seven languages. The correlation is -0.269 for Croatian, a language which is close to Serbian. No other language in their study achieves a stronger correlation. This fact tempts us to formulate a conjecture (which, of course, must be corroborated on many other languages) that the correlation between frequency and length of syllables is stronger than the one for words.

[20] We prefer to measure word length in syllables, as they are direct constituents of words (the more immediate the constituents, the stronger the dependency, see e.g. Altmann, 1983; our translation from German); however, in order to be able to compare the correlation with that from Ferrer-i-Cancho & Hernández-Fernández (2013), word length in letters was considered as well. Given that the correlation is stronger if word length is measured in letters, one could perhaps hypothesize that not only shorter words are used more frequently, but also that short words consisting of short syllables are favored over short words which contain longer syllables.

5. Conclusion

This paper can be considered a pilot study as far as a systematic quantitative approach to syllables in Slavic languages is concerned. The syllabification algorithm used here can be easily applied to all of them (and also to many other languages).

Our data support the hypothesis suggested by Strauss et al. (2008), according to which syllables, as far as models are concerned, behave like words. Syllables in the Serbian text under analysis „mimic“ the behaviour of words with respect to their frequencies, length, and the relation between these two properties. The models used belong to a very general family of distributions and functions introduced by Wimmer & Altmann (2005), which is a generalization of many linguistic laws (and thus can be considered to be a linguistic theory). Hence regularities in the syllable behaviour follow the same pattern as other linguistic units.

However, there are important differences if not only models, but also parameters are considered. Their values in the model for the rank-frequency distribution of syllables exceed those for words, and in the model for the distribution of syllable length they are out of the range of values observed for word length (Popescu et al., 2013, pp. 229-233).[21] The different strengths of correlations (for the relations between word frequencies and word length, and syllable frequencies and syllable length) were shortly addressed in Section 4.

It must be emphasized that we have preliminary results only, as the analyses were so far performed on one language only. In future, other Slavic languages and other aspects of syllables will be investigated. As there is a parallel corpus of Slavic languages available, properties of syllables can be used to construct a data-based typology of Slavic languages and to compare it with other approaches (see e.g. Koščová et al., 2016).

Acknowledgements

Supported by research projects APVV SK-SRB-2016-0021 (M. Radojičić, B. Lazić, R. Stanković, J. Mačutek, L. Leššová) and VEGA 2/0054/18 (J. Mačutek, L. Leššová).

REFERENCES

Altmann, G. (1983). H. Arens' "Verborgene Ordnung" und das Menzerathsche Gesetz. In: Faust, M., Harweg, R., Lehfeldt, W., Wienold, G. (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik: 31-39*. Tübingen: Narr.

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language: 117-156*. Dordrecht: Springer.

Best, K.-H. (2005). Wortlänge. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 260-273.* Berlin, New York: de Gruyter.

[21] For the time being, it remains an open question whether this property is generally valid or whether it is specific for the hyper-Poisson distribution.

Best, K.-H. (2011). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics 21, 1-13.*

Best, K.-H. (2013). Silbenlängen im Deutschen. *Glottotheory 4, 36-44.*

Blevins, J. (1995). The syllable in the phonological theory. In: Goldsmith, J. (ed.), *The Handbook of Phonological Theory: 206-244*. Oxford: Blackwell.

Browne, W. (1993). Serbo-Croat. In: Comrie, B., Corbett, G.G. (eds.), *The Slavonic Languages: 306-387*. London: Routledge.

Cairns, C., Raimy, E. (2011). Introduction. In: Cairns, C.E., Raimy, E. (eds.), *The Handbook of the Syllable: 1-30*. Leiden, Boston: Brill.

Clements, G.N. (1990). The role of the sonority cycle in core syllabification. In: Kingston, J., Beckman, M.E. (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech: 283-333*. Cambridge: Cambridge University Press.

Comrie, B. (1996). Adaptations of the Cyrillic alphabet. In: Daniels, P.T., Bright, W. (eds.), *The World's Writing Systems: 700-726*. Oxford: Oxford University Press.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Eddington, D., Treiman, R., Elzinga, D. (2013a). Syllabification of American English: Evidence from a large-scale experiment. Part I. *Journal of Quantitative Linguistics 20, 45-67.*

Eddington, D., Treiman, R., Elzinga, D. (2013b). Syllabification of American English: Evidence from a large-scale experiment. Part II. *Journal of Quantitative Linguistics 20, 75-93.*

Ferrer-i-Cancho, R., Hernández-Fernández, A. (2013). The failure of the law of brevity in two New World primates. Statistical caveats. *Glottotheory 4(1), 45-55.*

Haugen, E. (1956). The syllable in linguistic description. In: Halle, M., Lunt, H., MacLean, H., van Schooneveld, C. (eds.), *For Roman Jakobson: 213-221.* The Hague: Mouton.

Kelih, E. (2009). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis: 106-124.* Chernivtsi: ČNU.

Kelih, E. (2012). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation*. München: Otto Sagner.

Kelih, E. (2013). Grapheme inventory size and the repeat rate in Slavic languages. *Glottotheory 4(1), 56-71.*

Kelih, E., Mačutek, J. (2013). Number of canonical syllable types: A continuous bivariate model. *Journal of Quantitative Linguistics 20, 241-251.*

Koščová, M., Mačutek, J., Kelih, E. (2016). A data-based classification of Slavic languages: Indices of qualitative variation applied to grapheme frequencies. *Journal of Quantitative Linguistics 23, 177-190.*

Köhler, R. (2005). Properties of lexical units and systems. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 305-312.* Berlin, New York: de Gruyter.

Ladefoged, P., Johnson, K. (2011). *A Course in Phonetics*. Boston: Wadsworth / Cengage Learning.

Lehfeldt, W. (1971). Ein Algorithmus zur automatischen Silbentrennung. *Phonetica, 24, 212-237.*

Mačutek, J., Rovenchak, A. (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.), *Issues in Quantitative Linguistics 2: 136-147*. Lüdenscheid: RAM-Verlag.

Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics 20, 227-240.*

Meštrović, A., Martinšić-Ipšić, S., Matešić, M. (2015). Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik. *Govor 32, 3-34.*

Obradović, I., Obuljen, A., Vitas, D., Krstev, C., Radulović, V. (2010). Canonical syllable types in Serbian. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives: 145-157.* Wien: Praesens.

Piper P., Klajn I. (2013). *Normativna gramatika srpskog jezika*. Novi Sad: Matica srpska.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word Frequency Studies*. Berlin, New York: de Gruyter.

Popescu, I.-I., Naumann, S., Kelih, E., Rovenchak, A., Sanada, H., Overbeck, A., Smith, R., Čech, R., Mohanty, P., Wilson, A., Altmann, G. (2013). Word length: aspects and languages. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 3: 224-281.* Lüdenscheid: RAM-Verlag.

Pulgram, E. (1970). *Syllable, Word, Nexus, Cursus*. The Hague, Paris: Mouton.

Ridouane, R. (2008). Syllables without vowels: Phonetic and phonological evidence from Tashlhiyt Berber. *Phonology 25, 321-359.*

Rottmann, O.A. (1999). Word and syllable lengths in East Slavonic. *Journal of Quantitative Linguistics 6, 235-238.*

Rottmann, O.A. (2002). Syllable lengths in Russian, Bulgarian, Old Church Slavonic and Slovene. *Glottometrics 2, 87-94.*

Rubach, J., Booij, G. (1990). Syllable structure assignment in Polish. *Phonology 7, 121-158.*

Schiller, N.O., Meyer, A.S., Baayen, R.H., Levelt, W.J.M. (1996). A comparison of lexeme and speech syllables in Dutch. *Journal of Quantitative Linguistics 3, 8-28.*

Schiller, N.O., Meyer, A.S., Levelt, W.J.M. (1997). The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants. *Language and Speech 40, 103-140.*

Stanojčić, Ž., Popović L. (1999). *Gramatika srpskoga jezika.* Beograd: Zavod za udžbenike i nastavna sredstva.

Strauss, U., Fan, F., Altmann, G. (2008). *Problems in Quantitative Linguistics 1.* Lüdenscheid: RAM-Verlag.

Wimmer, G, Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions.* Essen: Stamm.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807.* Berlin, New York: de Gruyter.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology 12, 85-129.*

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Boston: Houghton Mifflin.

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4.* 2016, 287 pp.
24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.
26.M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme.* 2017, VI+125 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6.* 2018, IX+118 pp.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, 129 pp.