Glottometrics 44
2019

RAM-Verlag

ISSN 1617-8351
e-ISSN 2625-8226

Glottometrics

Indexed in ESCI by Thomson Reuters and SCOPUS by Elsevier

Glottometrics ist eine unregelmäßig er-scheinende 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.**

Herausgeber – Editors

External Academic Peers for Glottometrics

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. 44 (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 44 (2019)
online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)

Contents

Glottometrics 44, 2019, 1-15

Towards the Prepositional Meaning via Machine Learning: A Case Study of Spanish Grammar

Francesc Reina[1]
Irene Castellón[2]
Lluís Padró[3]

Abstract. Is it possible to identify or measure prepositional meaning? In our article we review a particular case of semantic universe, the verbs of movement in Spanish. In this context, we try to answer positively the initial question and validate a method. From the selection of a corpus of 71,206 prepositional phrases in Spanish, where three prepositions – *a*, *hacia* and *hasta* – are distributed, we proceed to verify the hypothesis about the semantic gradualness of the prepositions (HGSS). Applying tools of the field of machine learning, we establish a series of groupings that are compared with the hand annotation classification. The results are statistically relevant insofar as they confirm our initial hypothesis.

Keywords. *Spanish prepositions, verbs of movement, semantic gradualness, meaning, calculation, automatic learning, semantic similarities, grouping, clustering.*

1. Objectives and hypothesis

The concern, interest, and importance of the linguistic elements that express concepts and spatial meanings have been increasing in the last two decades, as can be seen in Svorou (1994), Jackendoff (1996), Saint-Dizier (2006), Levinson (2006), Ashbury (2008), Kelleher, Sloan & Mac Namee (2009), and Demonte (2011). After all, prepositions are involved in the semantics of spatial expression, which is also part of any human language.

From different perspectives and methodologies of linguistic and grammatical analyses, prepositions have been the subject of controversy, both in relation to their categorial nature and to the measurement and recognition of their semantic contribution in the multiple syntactic contexts where they are involved. Thus, they are considered a grammatical category as singular as problematic, not only in the field of descriptive and generative grammars, but also in cognitive, quantitative, and computational linguistics. The terms of this extensive debate can be found in a variety of references. We mention only some of the most outstanding ones in recent years: Cuyckens and Radden (2002), Baker (2005), Choi (2006), Baldwin, Kordoni, Valia and Villavicencio, Aline (2009), and Boleda & Herbelot (2016).

The role of prepositions in the syntactic-semantic configuration of spatial expression is crucial. Our proposal aims to shed light on both aspects: the linguistic expression of space, and the semantic perspective as the best option for the analysis of prepositions.

[1] Francesc Reina, frareina@hotmail.com.
[2] Irene Castellón, Universitat de Barcelona, icastellon@ub.edu.
[3] Lluís Padró, Universitat Politècnica de Catalunya, padro@cs.upc.edu.

In this context, our paper is a contribution that starts from the hypothesis that the semantic values of the prepositions are gradualness and progressive in any of the linguistic contexts that are found. The explicit formulation of the hypothesis of gradualness semantic similarity (HGSS) would be as follows in (1).

(1) Hypothesis of gradualness semantic similarity

Given a set of descriptive semantic values assigned to linguistics items, called prepositions, we predict that the identification of these values occurs gradually from the most functional values (with little or null semantic content) to the lexical ones (with a relevant semantic weight). In this progression, we find intermediate stages that are called semi-functional and / or semi-lexical.

The aims of our work are two, and they are inserted within a more extensive investigation on the capture and the quantification of the semantic values of all the prepositions in the Spanish language.

The first goal is to verify HGSS of certain prepositional values in a universe of restricted meanings – those understood in some real or figurative movement and actions – using empirical evidence from machine learning methods. The resulting clusters are compared with the hand-sorted groups and show, through statistically significant results, the different values predicted by and in the grammar.

The second one is to confirm the effectiveness of this quantitative method of analysis as a heuristic for the generalization of semantic evidence in the prepositional contexts of Spanish. In Mikolov, Sutskever, Chen, Corrado, Dean (2013), we are presented with some of the keys to the efficiency of these procedures, as well as in Mikolov, and Le (2014).

From the standardized or canonical grammatical descriptions of Spanish, we suggest a series of degrees or semantic phases that organize their meanings, involving three Spanish prepositions – *a, hacia*, and *hasta* – in the syntactic-argumentative contexts of a total of 92 verbs of movement and action. The most recognized and used descriptive sources come from Slager (1997), Bosque and Demonte (1999), Fernández (1999), RAE (2009), Moliner (2012), and Romo (2016).

2. Semantic values of prepositions (*a, hacia,* and *hasta*)

We shall now describe the semantic values of the prepositions for the selected movement verbs and exemplify them with real statements of the corpus. These types, abbreviated as F1, SF1, SL1, SL2, L1, and L2 (Tables I, II, and III), constitute the various groups of the manually classified models that are used for comparison with the classes resulting from automatic clustering. They correspond to functional, semi-functional, semi-lexical (1 indicates modality and 2 events), and lexical (1 for spatial values and 2 for temporal values) classes, according to their load or semantic content. These values are carefully characterized in the following tables.

The preposition *a* is split in six values (those indicated in the previous parenthesis and collected in Table I), the *hacia* in three (SL1, L1, and L2 – Table II) and the *hasta* in three, too (SL1, L1 and L2 – Table III).

Table 1
Description and examples of the semantic values of *a*

Preposition *a*	
Type of semantic value and description	**Examples**
Functional (F1) – it is a grammatical mark of verbal periphrasis or infinitive complement.	Se lanzó a criticar (He started criticizing) Va a empezar (He will start) Pasa a exigir (He happens to demand) Vuelve a arremeter (He returns to attack)
Semifunctional (SF1) – expresses or indicate the beneficiary of the action or mandatory verb pattern.	Envían a Pedro (They send Pedro) Empuja a decidir (He pushes to decide)
Semilexical 1 (SL1) – indicates the mode or instrument in which the action is performed.	Fueron al rescate (They went to the rescue) Andan a gritos (They scream) Continúan a nado (They continue to swim)
Semilexical 2 (L2) – indicates the event or the figurative place where the action takes place.	Parte a la reunión (He party to the meeting) Acude a la convocatoria (He attends the call)
Lexical 1 (L1) – indicates the real place (determined point or physical space) where the action is directed.	Van a Madrid (They go to Madrid) Se dirigen al límite (They go to the limit)
Lexical 2 (L2) – indicates the point, time, or period of time in which the action takes place.	Vuelven a la semana (They come back a week) Recorren al atardecer (They walk at sunset) Regresa a las tres (He returns at three o'clock)

Table2
Description and examples of the semantic values of *hacia*

Preposition *hacia*	
Type of semantic value and description	**Examples**
Semilexical (SL) – indicates the beneficiary, or the figurative concept that receives the verbal action.	Desplazan hacia los afectados (They move towards the affected) Aparta hacia la democracia (he moves towards democracy)
Lexical 1 (L1) – expresses location, direction, or physical or figurative orientation	Fluye hacia el oeste (He flows westward) Acompañan hacia adelante (They accompany forward)
Lexical 2 (L2) – indicates a temporary location	Andan hacia la fase (They walk towards the phase) Asientan hacia siglos (They settle centuries ago)

Table 3
Description and examples of the semantic values of HASTA

Preposition *hasta*	
Type of semantic value and description	**Examples**
Semilexical (SL) – introduces complements of infinitive or abstract states.	Pasan hasta comprender (They go on to understand) Llegan hasta la locura (They get to madness)
Lexical (L1) – expresses term, limit, or place where an action ends.	Escalan hasta la cima (They climb to the top) Transportan hasta Amsterdam (They transported to Amsterdam)
Lexical (L2) – expresses the time limit in which an action begins.	No se retiran hasta las tres (They do not retire until three o'clock) Van hasta el domingo (They go until Sunday)

3. Methodology

We start with a corpus of analysis of 71,206 prepositional phrases (triplets) from two corpora: *AnCora* and *Wikicorpus*. Its distribution is 57,815 for *a*, 6,389 for *hacia*, and 7,102 for *hasta*.

Wikicorpus is a trilingual corpus (Catalan, Spanish, and English) that contains huge amounts of words from Wikipedia, and has been enriched automatically with linguistic information. In its current version, it houses more than 750 million words. For a description of how this corpus is made, Reese, Boleda, Cuadros, Padró, Rigau (2010) can be consulted.

Regarding *AnCora*, it is a corpus of Catalan (*AnCora-CA*) and Spanish (*AnCora-ES*) with different levels of annotation, (cf. Taulé, Mariona, M. Antònia Martí, Marta Recasens, 2008).The corpus of each language contains 500,000 words and mainly consists of journalistic texts.

The analysis tool is the CLUTO / GLUTO application, Karypis (2003). CLUTO is a software package for grouping data sets of low and large dimensions and to analysing the characteristics of the different clusters. CLUTO is well suited to store datasets that arise in various application areas, including information retrieval, customer purchase transactions, web, GIS, science, and biology.

We have used the *Word embedding* technique to build the CLUTO vectors, like word2vec models (cf. Mikolov, Chen, Corrado and Dean 2013). The objective of this procedure is to quantify and categorise semantic properties among linguistic elements from the contexts where they cooperate. These *vector space models* (VSMs) represent embedded-words in a continuous vector space in which semantically similar words are assigned to nearby points ("embed each other").

The clustering algorithm parameter used is called 'direct'. This method has been implemented due to its ability to identify the optimum set of groups based on our hypothesis. In this method, the desired *k*-way cluster solution is calculated by finding all the *k* groups simultaneously. In general, computing a *k*-way cluster directly is slower than grouping through repeated bisections. In terms of quality, for reasonably small values of *k* (generally less than 10–20), the direct approach leads to better conglomerates than those obtained by repeated bisections. However, as *k* increases, the repeated bisection approach tends to be better than direct grouping.

Our observation focuses exclusively on a reduced semantic universe composed by the verbs of action and movement, where the above-mentioned prepositions are frequently found. We have chosen a total of 92 verbs.

The list of the 92 Spanish verbs of movement and action used is the following: *acarrear*, *acelerar*, *acudir*, *agachar*, *agitar*, *alejar*, *andar*, *apartar*, *apoyar*, *aproximar*, *apuntar*, *arrastrar*, *arrimar*, *arrinconar*, *asentar*, *atraer*, *bajar*, *brincar*, *caer*, *caminar*, *circular*, *conducir*, *contener*, *continuar*, *correr*, *curvar*, *descender*, *deslizar*, *desmontar*, *desplazar*, *desviar*, *detener*, *dirigir*, *doblar*, *doblegar*, *empujar*, *encaminar*, *encauzar*, *enderezar*, *enfocar*, *entrar*, *enviar*, *escalar*, *escapar*, *escorar*, *escurrir*, *extender*, *fluir*, *galopar*, *girar*, *golpear*, *guiar*, *huir*, *hundir*, *inclinar*, *ir*, *juntar*, *lanzar*, *llevar*, *marchar*, *montar*, *mover*, *orientar*, *oscilar*, *partir*, *pasar*, *pasear*, *permanecer*, *portar*, *rastrear*, *recorrer*, *remolcar*, *remover*, *retirar*, *sacudir*, *salir*, *saltar*, *seguir*, *sobrevolar*, *subir*, *tirar*, *traer*, *transportar*, *trasladar*, *traspasar*, *vagabundear*, *venir*, *viajar*, *volar*, *voltear*, *volver* and *zarpar*.

The selection comes from the dictionary meanings of the DRAE (2014). With regard to the notions of movement and action, we assume the concepts 'trajectory' and 'path' proposed by Jackendoff (1996).

From this triple choice – the verbs of movement, the prepositions *a*, *hacia*, and *hasta*, and three groups of prepositional values (indicated in previous Tables I, II and III) –, we proceed to the extraction of the triplets with the following syntactic structures: V + P + N (verb, preposition, and noun), V + P + Vinfinitive (verb, preposition, and verb in the infinitive form), and V + P + A (verb, preposition, and adjective).

Our analysis uses words (inflected forms); the check with lemmas has not improved or added relevant information in the confirmation and validation of our hypothesis. We have also developed the experiment with syntactic functions, but we have not obtained relevant results either.

Then, we create a hand-sorted evaluations file with manual annotation. The annotator classifies a total of 1,731 triples of preposition *a*, 654 of preposition *hacia*, and 779 of preposition *hasta*. To improve our classification, we made a second manual annotation from a sample of 150 triplets. The agreement between two annotators is 79.2%.

Once the files are created and available, the CLUTO vector is generated. As the last step, there is a comparison between the manually annotated comparison files and the results of the CLUTO clusters.

The comparison program offers us a series of tables that relate the values (F1, SF1, SL1, SL2, L1, and L2) and the number of clusters, together with a percentage index of purity and inverse purity. By purity, we understand the concentration of each column in the table. It tells you how the elements that have been left in the same cluster are distributed among different categories. If all the elements of the cluster are in the same category, the purity of that column would be 100%. On the other hand, inverse purity means the same, but according to the concentration by rows. It measures how the elements of the same category are distributed in different clusters. If all the elements of the category are in the same cluster, the purity of the row would be 100%.

We have made the comparison for the case of *a* with 6, 5, 4, and 3 clusters respectively; and with 3 for *hacia* and *hasta*.

We shall now examine the results in relation to our hypothesis. Our assessment consists of an observation of the percentages of predominance and the analysis of the content of the subgroups proposed by the clustering mechanism.

It is important to remark that we are not building a machine-learning classifier for prepositions, but just using clustering to validate our hypothesis. Since the features used by the clustering algorithm are very simple (just the word embedding of the noun inside the prepositional phrase), we do not expect to predict the class of the prepositions fully. However, a high correlation between the obtained and expected clusters will prove that context semantics has an important weight in the behaviour of the preposition, thus confirming our hypothesis.

4. Empirical test

The empirical testing of our hypothesis is strongly supported because the groupings proposed by the CLUTO algorithm – the automatic learning tool used – correspond to hand-sorted examples. We reiterate that in our results, we can see how the prepositional meaning depends on the semantics of its context.

The figures range between 51% and 64%, depending on the number of clusters and the preposition. It exceeds 70% in some cases if we analyse adjacent cells of the clusters of each preposition and their respective semantic values.

Figure I shows the average percentages of coincidence between the automatic clusters that the CLUTO algorithm does, and the classification of the human annotator for the three prepositions – *a, hacia,* and *hasta*.

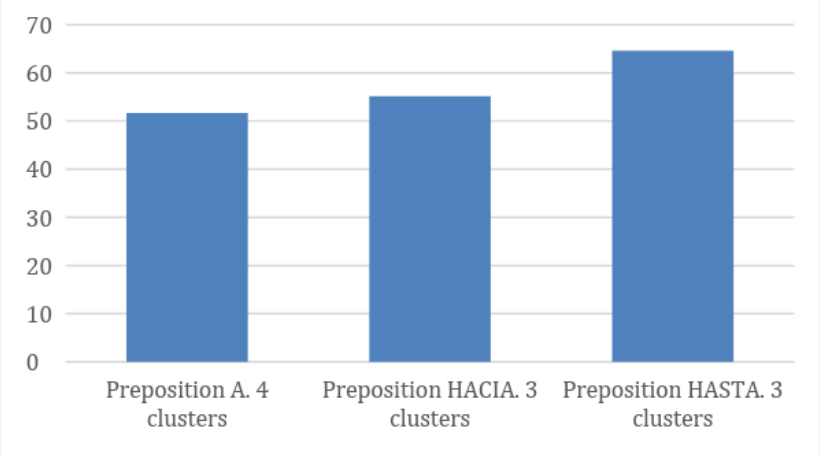


Figure 1. Comparison of the percentage of purity of the three prepositions

Below, we detail the results for each of the prepositions, as well as specific examples that show the validity of our predictions. The tables include the name of the preposition, the clustering parameter, the number of clusters, the number of triples evaluated in the comparison, and the columns with the numbering assigned by the CLUTO programme.

4.1 The preposition *a*

We have started the process of clustering with six groups. The purpose of these distributions is to achieve the most relevant semantic grouping. However, and to obtain better results, we have gone from the six levels or classes that express all the meanings of movement verbs to a progressive synthesis (from 6 to 5, from 5 to 4 and from 4 to 3). First, these reductions have simplified the functional spectrum (F1 + SF1), after the semi-lexical spectrum (SL1 + SL2) and, finally, the lexical spectrum (L1 + L2).

As it is shown in Figure II, in absolute terms, the best results obtained correspond to the grouping of 5 clusters. However, in a detailed analysis, a more accurate and adequate scenario is recognized in the grouping of 4 clusters, which is, in fact, what we will comment and exemplify in Table IV.

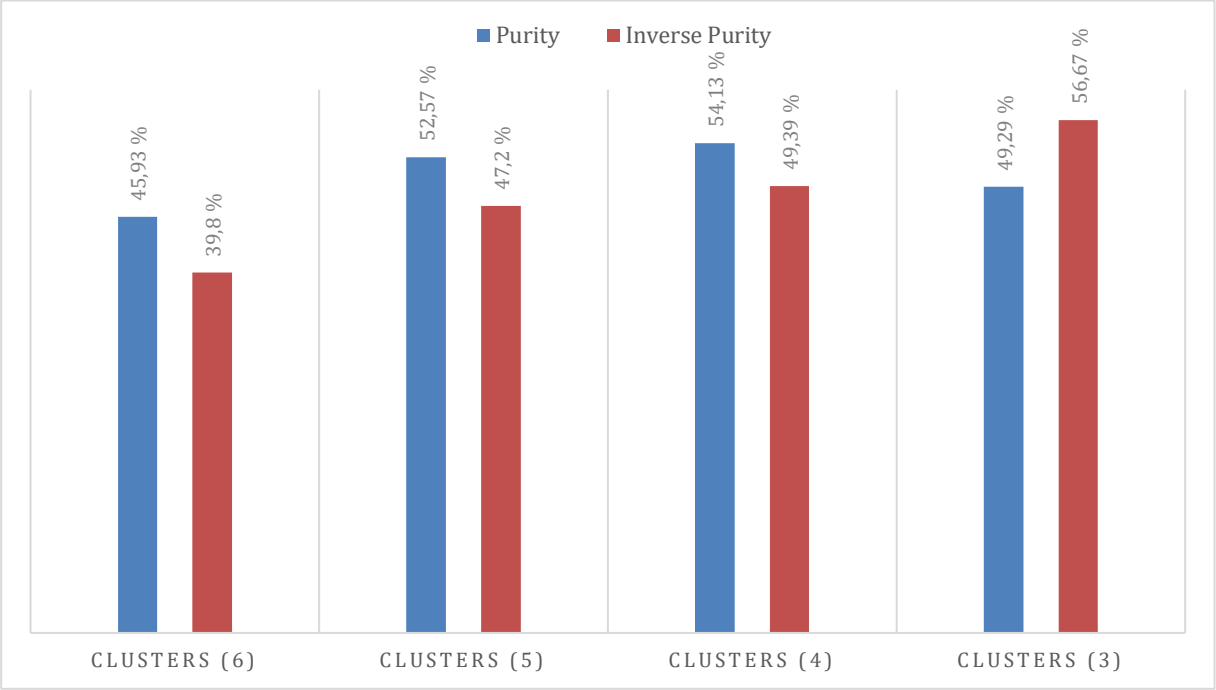


Figure 2. Percentages of purity (blue), and inverse purity (brown) by number of values and clusters – preposition *a*.

In turn, in Table IV, we have gathered the results of the grouping of 4 clusters and four prepositional values (functional, semilexical, lexical 1 and lexical 2 ones).

Table 4
Grouping into 4 clusters and 4 values of the preposition *a*

***A*/DIRECT/41731**	**0**	**3**	**1**	**2**
F1	279	73	118	141
SL1	100	281	75	52
L1	75	162	236	45
L2	54	18	33	59

In Table IV we can observe that the grey shaded diagonal contains a total of 855 cases (49.34%). In three of the four results, most cases are concentrated, both in columns and rows. If, in addition, we add the two next diagonals (upper and lower), where semantically close examples are found, we reach 1,343. We are talking about 77.58%.

A hypothesis test yields that we can reject that both classifications are independent with a 99%-confidence degree. Cohen's Kappa value is 0.28 (random agreement would be 0.0) which is usually considered as fair agreement.

The prepositional functionality is collected in the (F1 0) and (F1 2) clusters, adding a total of 420 cases, that is, 68.62%. In the first cluster (F1 0), the phrases are joined with infinitives, participles, and some pronominal phrases, with examples such as *acudir a presentar*, *aparcar a ver*, *continuar a denunciar*, *desplazar a otro*, *detener a atravesar*, *enfocar a conseguir*, *entrar a actuar*, *fluir a agitados*, *ir a cantar*, *llevar a actuar*, *mover a otro*, *pasar a denominar*, *salir a buscar* and *transporta a afectados*.

The F1 class is distributed in 1 cluster of verb phrases (F1 0) and two human prepositional phrases (F1 1) and (F1 2). There are not too many differences between clusters 1 and 2. The first seems more functional (professions), and the second more relational (family relationships and others). A curious note is that in (F1 3), we find many animals, humans, and artefacts. We also see that in (F1 2), there are phrases with a lexical head of the ontological type 'human' that correspond to direct objects, with examples such as *acarrear a director*, *apuntar a madre*, *atraer a anfitriones*, *bajar a futbolistas*, *caer a amigos*, *conducir a grupo*, *conducir a sospechoso*, *contener a enemigo*, *detener a humanos*, *enfocar a ancianos*, *entrar a abogado*, *extender a cliente*, *girar a hijo*, *golpear a villano*, *llevar a pareja*, *orientar a chicos*, *pasear a novia* and *sacudir a sujeto*.

The field of semi-lexical, halfway between the functional mark and the clear lexical expression, is found in the (SL1 3) cell, with 281 cases. The kind of examples of this type of phrase collects the manners that indicate the way or qualify, for example *acelerar a pie*, *andar a aire*, *aproximar a mano*, *arrastrar a cuerda*, *bajar a pie*, *circular a profundidad*, *descender a exceso*, *descender a paso*, *entrar a ojo*, *extender a complejos*, *galopar a velocidad*, *fluir a pie*, or *montar a caballo*.

Other semi-lexical examples would be those that indicate a non-physical space or place, figurative or orientation-like: *arrastrar a afuera*, *circular a anillo*, *conducir a cerebro*, *circular a profundidad*, y *desplazar a alrededor*.

The second cluster that collects semi-lexical examples is (SL1 0), with 100 cases. Again, we find certain cases of modal nuance, such as *desviar a manos*, *conducir a paranoia*, *correr a antojo*, *atraer a desaliento*, *conducir a indiferencia*, *conducir a colapso*, *llevar a caos*, *inclinar a modestia*, *permanecer a manos*, and *volar a acecho*.

In the (SL1 1) cluster, there are events and ways such as *apoyar a fuerte*, *correr a mandato*, *empujar a exilio*, *entrar a pacto*, *extender a matanzas*, *girar a defensa*, *guiar a racha o volver a clandestinidad*; while in (SL1 2), there are events such as *acudir a baile*, *acudir a funeral*, *acudir a torneo*, *dirigir a aventura*, *enviar a audiencia*, or *permanecer a regreso*. It would seem that in some cases, the algorithm differs very clearly between modes and events. The heterogeneous or transit character of this semantic value explains this aggrupation of such varied examples.

Regarding the lexical environment, the highest concentration is found in the (L1 1) cluster, with 236 cases, and in (L1 3), with 162 cases. The examples are always spatial, from specific places (buildings, facilities, or geographical locations – such as *acelerar a ciudad*, *arrastrar a centro*, *andar a iglesia*, *aproximar a ciudad*, *caer a afueras*, *descender a ultramar*, *desplazar a aeropuerto*, *entrar a casa*, *mover a campamento*, *hasta topónimos como acudir a Barcelona*, *atraer a Vizcaya*, *continuar a Europa*, *dirigir a Nicaragua*, or indications of limit or directions such as *asentar a oeste*, *caer a norte*, *fluir a noreste*, *partir a sur*, or *remover a centro*.

The semantics of time is collected in a very scattered way, without any specific predominance. Any of the four clusters presents phrases with a temporal indication of many types: month names, moments and periods of time, or moments of the day. This concentration

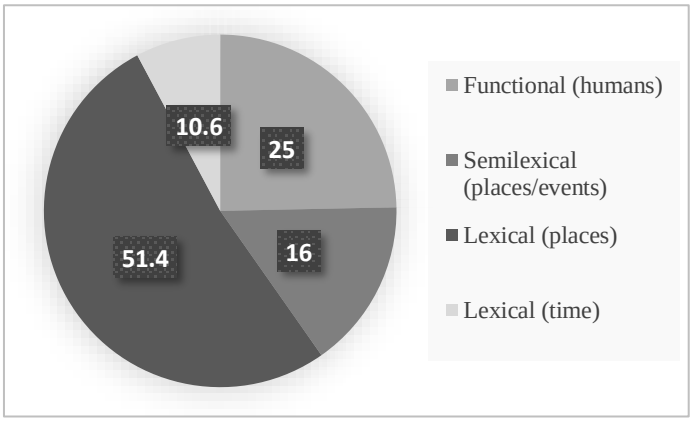
is found in the (L2 0) and (L2 2) clusters with 54 and 59 cases, respectively, representing 68.90% of the total.

Some examples of this kind of lexical value that express periods, moments, phases, or names of months are *acudir a ayer*, *caer a final*, *bajar a meses*, *correr a temporada*, *extender a noche*, *hundir a atardecer*, *inclinar a comienzo*, *permanecer a etapa*, *permanecer a semana*, *retirar a minutos*, *salir a mes*, or *trasladar a mayo*. We can see, for instance, some of the repetitions according to the (L2 0) cluster, 54 cases, distributed in the following repetitions: 8 *fin*, 7 *momento*, 5 *principio*, and 23 *tiempo*. (L2 1), 33 temporal cases and events: 13 *edad*, 5 *inicio* and 4 *llegada*. (L2 2), 59 examples, 5 *amanecer*, 4 *atardecer*, 6 *comienzo*, 2 *diario*, 20 *final*, 6 *mes*, 3 *meses*, 2 *semana* and 2 *temporada*. Y (L2 3), 18 examples, 4 *horas*, 3 *intervalos* and 2 *minutos*. In table V we collect the examples ordered.

Table 5
Temporary lexical values. Examples.

Examples of temporary lexical values. Preposition *a*	**Number and concrete examples**
(L2 0) 54 cases	8 fin, 7 momento, 5 principio, and 23 tiempo.
(L2 1) 33 cases	13 edad, 5 inicio and 4 llegada.
(L2 2) 59 cases	5 amanecer, 4 atardecer, 6 comienzo, 2 diario, 20 final, 6 mes, 3 meses, 2 semana and 2 temporada.
(L2 3) 18 cases	4 horas, 3 intervalos and 2 minutos.

Finally, in Figure III, the semantic values of each one of the four clusters are presented, indicating the percentage distribution of the examples in Table IV. We can also find some descriptive indications such as VP (predominance of verbal phrases), human (presence of human entities or events). The verification of the hypothesis is reinforced insofar as there is a correlation between the predicted values (functional, semilexical, and lexical 1 and 2 ones), and the classes of semantic entities that are grouped.



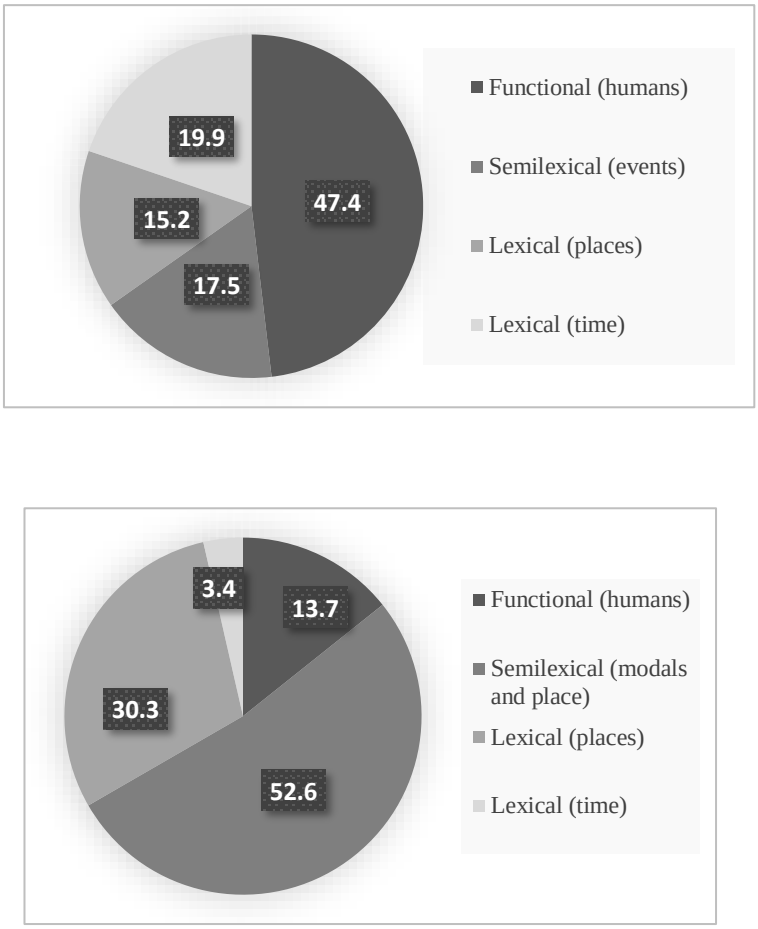


Figure 3. Clusters 0, 1, 2, and 3 – from top to bottom – found in the research of preposition *a*. Semantic values and predominance.

4.2 The preposition *hacia*

The percentage of agreement between the automatic clusters and the human classification increases with respect to the preposition up to 55.17% of average of purity. In Table VI, we find the distribution of cases in values and in clusters.

Table 6
Grouping into 3 clusters and 3 values of the preposition *hacia*

***Hacia* /DIRECT/3 654 Prepositional Phrases**	**2**	**0**	**1**
SL1	117	20	72
L1	169	171	51
L2	33	1	20

The first prepositional value, the semilexical (SL1) concentrates the phrases that denote beneficiaries and terms of verbal action, animated or abstract ones. We find examples such as *acelerar hacia blanco, aproximar hacia objetivo, apuntar hacia observador, apuntar hacia tipo, desviar hacia tema, desplazar hacia familia, detener hacia Dios, encaminar hacia*

dictadura, dirigir hacia dictadura, girar hacia amo, ir hacia socialismo, pasar hacia actores or *seguir hacia conocimiento*. We are in front of 117 cases of the (SL1 2) cluster, that is, 55.98% of the total of the row.

A hypothesis test yields that we can reject that both classifications are independent with a 99-% confidence degree. Cohen's Kappa value is 0.18 (random agreement would be 0.0), which is usually considered as slight agreement.

The locative expression, in turn, is collected in the (L1 0) and (L1 2) clusters. They represent 61.22% of the total. No substantial semantic differences are appreciated. The examples that appear refer to many forms of localization: defined and concrete places, open and closed spaces, and place nouns. Thus, we find triplets such as *acelerar hacia ciudades, acelerar hacia nordeste, apuntar hacia sierra, atraer hacia Antillas, bajar hacia sur, correr hacia lago, correr hacia tierras, fluir hacia capital, ir hacia avenida, oscilar hacia centro*, or *zarpar hacia Asia*.

On the other hand, the examples of (L1 2) are mostly concentrated in direction or orientation terms. In this way, we have *acelerar hacia abajo, agitar hacia adelante, agitar hacia lados, alejar hacia mitad, apuntar hacia afuera, apuntar hacia extremo, arrastrar hacia afuera, bajar hacia interior, deslizar hacia izquierda, enderezar hacia costado, orientar hacia camino*, and *sobrevolar hacia este*.

The presence of 169 cases in the (L1 2) cluster is an exception to our expectation. If we look at the (L1 2) class of examples, we can observe that they are locatives or directional terms, or figurative or abstract locatives, such as *parte, abismo, posiciones, proximidades*. Probably, the algorithm considers that this abstract sense is more similar to the semilexical value of (SL1 2), where we also find abstract entities. Some examples of that are *aproximar hacia obsolescencia, apuntar hacia existencia, apuntar hacia objetivo, caminar hacia libertad, conducir hacia socialismo, encauzar hacia posturas, guiar hacia experiencia, orientar hacia independentismo, seguir hacia conocimiento*, or *llevar hacia consenso*.

With regard to the semantic values of time, a phenomenon similar to the previous one occurs. In most cases – 33 out of 54 –, they are in the (L2 2) cluster, although our comparative prognosis was that they appeared in (L2 1). These 33 cases concentrate examples of temporal values with an abstract sense, such as *acelerar hacia final, andar hacia fase, aproximar hacia mientras, apuntar hacia eternidad, apuntar hacia futuro, caminar hacia vejez, conducir hacia fin, dirigir hacia proceso, llevar hacia expectativas*, or *venir hacia presente*.

The CLUTO algorithm prefers to leave it in the same column – cluster – because of its similarity, again due to the abstract character, with the upper (L1 2) cluster. On the other hand, in the (L2 1) cluster, there is a concentration of word *finales*, in 17 cases, and only three new words: *retiro, septiembre*, and *conquista*.

.

4.3 The preposition *hasta*

Hasta is the preposition that obtains the best results in the validation of our hypothesis. We are facing a 64.6-% purity of average.

The goodness of these results is probably related to the proper meaning of preposition. *Hasta* expresses a limit in any of its senses: an action, a beneficiary (semilexical values) and also a movement in space (locative values) together with a time limit.

Table 7
Grouping in 3 clusters and 3 values of preposition *hasta*

***Hasta* /DIRECT/3 779 Prepositional Phrases**	**1**	**2**	**0**
SL1	144	22	29
L1	28	240	41
L2	129	36	109

A hypothesis test yields that we can reject that both classifications are independent with a 99-% confidence degree. Cohen's Kappa value is 0.33 (random agreement would be 0.0), which is usually considered as fair agreement.

Semilexical values are gathered in the (SL 1) cluster, with 144 cases representing 70.24% of average purity. The examples belong to two different types – those that refer to the limit of an action, such as *como acelerar hasta alcanzar, alejar hasta terminar, asentar hasta ofrecer, caminar hasta agoten, continuar hasta situar, enfocar hasta trabajar, girar hasta conseguir,* and, in addition, those that indicate the receiver or the beneficiary of the action. Here, we have cases like *contener hasta animales, correr hasta personajes, entrar hasta dioses, enviar hasta condenados, juntar hasta personas* and *viajar hasta receptor*.

In the case of triples that indicate space and location, the percentage is very high. We are talking about 77%, or 240 cases (L1 2). The examples show the polysemic diversity of this expression: indoor or outdoor premises, buildings, facilities, or place names. It would be the case of the following series: *acelerar hasta fondo, acudir hasta edificios, apoyar hasta playa, asentar hasta terranova, bajar hasta llanura, caminar hasta Acapulco, circular hasta campos, continuar hasta tienda, desplazar hasta bolsa, dirigir hasta plaza, escalar hasta casilla,* and *extender hasta Balcanes.*

In the next cluster (L1 0), an almost complete concentration of place names can be found – 37 out of 40 –; some examples are *acudir hasta China, continuar hasta Marsella, extender hasta Pensilvania, ir hasta Jamaica, pasar hasta Vietnam,* or *viajar hasta Ucrania*.

Finally, temporality is organized in the (L2 1) cluster, being the most noumerous one with 129 cases. It concentrates triplets that indicate the phases of the time sequence, such as *acelerar hasta llegado, acudir hasta momento, agitar hasta veces, caer hasta final, continuar hasta comienzo, girar hasta ahora, permanecer hasta episodio, salir hasta pronto*. The second cluster of temporary triples is (L2 0), which mostly includes months, seasons, and some generic temporary expressions. Examples would be *apoyar hasta enero, conducir hasta setiembre, continuar hasta periodo, correr hasta temporada, dirigir hasta principios, doblar hasta fecha, llevar hasta octubre,* and *volver hasta semana*.

The fact that the largest one appears in the first column is related to the abstract character of the expression of time, universal and very common in human languages. This feature would possibly explain why it is next to the general locatives, which appear in the upper cluster. We have found examples such as *alejar hasta entonces, apoyar hasta niveles, asentar hasta descenso, desviar hasta nivel, continuar hasta salida, guiar hasta destino,* and *galopar hasta meta*.

4.4 Some complementary explanations

In the analysis by columns, in relation to our prediction, certain anomalous events are produced within certain clusters, which should be explained and which are compatible with our hypothesis. Thus, and following the data of Table IV, VI, and VII, we have analysed the following clusters: from the preposition *a*, (F1 0) versus (SL1 0), (SL1 3) versus (L1 3), (L1 1) versus (F1 1), and (L2 2) versus (F1 2); from the preposition *hacia*, (SL1 2) versus (L1 2), and (L2 1) versus (SL1 1); and from the preposition *hasta*, (SL1 1) versus (L2 2), and (L2 0) versus (L1 4).

To begin with, we find some errors of a syntactic nature. Once the examples are observed, they are found to be unusual or inappropriate phrases in Spanish grammar. They would be cases like *alejar a aviones*, *apartar a pecados*, *apartar a menudo*, *detener a dolor*, *desviar a manos*, or *extender a reales*. By addressing the context of these triplets, we see that the choice made by the analyser is incorrect from the point of view of syntactic congruence. On the basis of a sampling of examples, we estimate that the percentage of this kind of errors could reach 10%.

A second misleading in the semantic grouping arises from the repetitions of some nouns of the chosen triplets. This fact causes distorted results. CLUTO works by looking for similarities in the semantic context of the nouns of the V + P + N scheme. Repetitions of the same noun change the cluster sizes and distributions, altering the quality of the observation. This fact, however, does not detract from the similarities that they obtain in others. This happens because we did not include the verb word embedding in the vector, since we checked that it produced worse results. After an analysis of the data, we conclude that the system, being a small group (92 verbs), always grouped the occurrences of the same verbs, and this prevented the same verb from obtaining two different interpretations in the classification.

Another phenomenon that occurs is the presence of words that come from the same lexical family and are grouped in different boxes. It would be the case of couples like the following: *abandonar* / *abandonados*, *capturar* / *captura*, *concluir* / *conclusiones*, *completar* / *completo*, *golpeado* / *golpe*, *jueces* / *juicio*, *necesidad* / *necesidades*, *trabajo* / *trabaja*, *vida* / *vivir*, or *tomar* / *toma*. The formal difference, of a lexical-morphological nature, leads the algorithm to a certain confusion. Despite the fact that they are semantically similar, it groups them into different clusters.

In addition, there are some cases of partial homonymy in examples such as "weapons" and "to arm" (*armas* the plural common noun and second person of the indicative of the verb *armar*, *armas* also).

5. Conclusions and discussion

The analytical power of the so-called "Word embedding" – vectors of values obtained by similarity appearance – makes it possible to direct attention to prepositional semantics considering very large data sets of a real corpus. In our hypothesis – HGSS (*Hypothesis of Gradualness and Semantic Similarity*) –, and for our experiment, the focused observation of the verbs of movement and action in Spanish suggests a set of quantitative and qualitative evidences that reinforce it.

This methodological hypothesis helps to reduce certain theoretical difficulties that are generally attributed to the categorial nature of the preposition, in particular, how to determine its semantic contribution.

The quantitative testing of our hypothesis has made it possible to guide and refine the qualitative analysis of the clusters. In this way, it is confirmed that prepositions *a*, *hacia* and *hasta* organise their semantic values in degrees: from functionality to lexicity.

On the other hand, when we have experimented with enriched triplets with syntactic functions as a source for the corpus, analyzed by FreeLing, results have not improved the previous outcomes. For the operation of Freeling you can consult Padró & Stanilovsky (2012). The data obtained after the syntactic enrichment is below almost 10 percentage We have made the comparison with 6, 5, 4, and 3 values for A with the following percentages: 6 (45.38% Purity), 5 (49.53% Purity), 4 (42.37%), and 3 (36.92%).

The most important consequence of the results of our research is that a prominent part of the problems attributable to prepositional syntactic configurations is solved from a perspective of grouping by semantic similarity, that is, the meaning emerges in the proximity of the context. Consequently, the distributive meaning of the preposition is more relevant to the structural interpretation of prepositions than its role in case assignment or other idiosyncratic characteristics of the preposition. Obtaining recognizable semantic patterns reveals characteristics of this class of words that had not been indicated in the theoretical investigations, thanks to automatic learning by distribution and similarity.

In new perspectives, we will try to follow the trends pointed out in the experiment, improving the quality of the data (the selection of the prepositional phrases) and the scrutiny of the semantic similarities in the environment of other prepositions and other syntactic constructions.

Finally, we plan to explore the interlinguistic comparison with the intention of validating the HGSS of automatic grouping independently of the grammatical variation between languages (with prepositions, with morphological cases, with adpositions, or with some combination of the previous ones).

REFERENCES

Demonte, Violeta. (2011). Los eventos de movimiento en español. Construcción léxico-sintáctica y microparámetros preposicionales. In: Juan Cuartero Otal, Luis Garcia Fernandez and Carsten Sinner (eds.): *Estudios sobre perífrasis y aspecto*, München: Peniope, 16–42.

Ashbury, Ana; Dotlačil, Jakub; Gehrke, Berit and Rick Nouwen. (eds.) (2008). *Syntax and Semantics of Spatial P*, Utrecht: John Benjamins, Utrecht Institute of Linguistics.

Baldwin, Timothy; Kordoni, Valia and Villavicencio, Aline. (2009). Prepositions in Applications: A Survey and Introduction to the Special Issue. *Computational Linguistics, 35.2, 119–149.*

Baker, Mark. (2005). *Lexical categories.* Cambridge: Cambridge University Press.

Boleda, Gemma and Herbelot, Aurélie. (2016) (eds.) Special Issue on Formal Distributional Semantics. *Computational Linguistics* 42:4. MIT Press.

Baroni, Marco; Dinu, Georgiana and Kruszewski, Germán (2014*). Don't count, predict! A systematic comparison of context-counting vs. context predicting semantic vectors.* In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, June 23–25, 238–247.

Bosque, Ignacio and Demonte, Violeta. (eds.) (1999). *Gramática descriptiva de la lengua Española.* Madrid: Espasa & Calpe.

Choi, Soonja. (2006). Influence of language-specific input on spatial cognition: categories of containment, *First Language*, 26.2, 207–232.

Cuyckens, Hubert and Radden, Günter. (2002). *Perspectives on Prepositions.* Tübingen: Niemeyer Verlag.

DRAE. (2014). *Diccionario de la Real Academia Española de la lengua*, Barcelona: Espasa & Calpe.

Evans, Vyvyan and Tyler, Andrea. (2003). *The Semantics of English Prepositions*. Cambridge: Cambridge University Press.

Fernández Lópe and, María del Carmen. (1999). *Las preposiciones en español. Valores y usos. Construcciones preposicionales.* Madrid: Ediciones Colegio de España.

Hale, Ken; Keyser, Samuel Jay. (2002). *Prolegomenon to a Theory of Argument Structure.* Cambridge: MIT Press.

Herkovits, Annette. (1986). *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English*. Cambridge: Cambridge University Press.

Ibarretxe-Antuñano, Iraide and Hijazo-Gascón, Alberto. (eds.) (2015). *New Horizons in the Study of Motion: Bringing together applied and theoretical perspectives*. Newcastle upon Tyne: Cambridge Scholars.

Jackendoff, Ray. (1996). The Architecture of the Linguistic-Spatial Interface. In: P. Bloom, M. Peterson, L. Nadel, and M. Garrett (eds.). *Language and Space*. Cambridge: MIT Press, 1–30.

Karypis, George. (2003). http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview, University of Minnesota, Minnesota.

Kelleher, J.; Sloan, C. and Mac Namee, B. (2009). An investigation into the semantics of English topological prepositions, *Cognitive Processing*, 10.2, 233–236.

Levinson, Stephen. (ed.) (2006). *Grammars of Space: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

Mikolov, Tomas; Chen, Kai; Corrado, Greg and Dean, Jeffrey (2013a). *Efficient estimation of word representation in vector space*, arXiv:1301.3781.

Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg and Dean, Jeffrey (2013b). *Distributed Representations of Words and Phrases and their compositionality*, arXiv: 1310.4545v1.

Mikolov, Tomas; Le, Quoc. (2014). Distributed Representations of Sentences and Documents. In: *Beijing: Proceedings of the 31st International Conference on Machine Learning*, 1188–1196.

Moliner, María. (2012). *Uso de las preposiciones*, Madrid: Gredos.

Padró, Lluís and Stanilovsky, Evgeny. (2012). Freeling 3.0: Towards Wider Mutilinguality. *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul: ELRA.

Real Academia Española. (2009). *Nueva gramática de la lengua española*, Madrid: Espasa Libros.

Reese, Samuel; Boleda, Gemma; Pictures, Montse; Padró, Lluís and Rigau, German. (2010). Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In: *Proceedings of 7th Language Resources and Evaluation Conference LREC'10*, Malta: La Valleta.

Romo, Francisco. (2016). *Un estudio cognitivista de las preposiciones espaciales del español y su aplicación a la enseñanza E/LE*. Philosophical dissertation. Barcelona: UAB.

Saint-Dizier, Patrick. (2006). *Syntax and Semantics of Prepositions.* Dordrecht: Springer.

Slager, Emile. (1997). *Pequeño diccionario de construcciones preposicionales*. Madrid: Editorial Visor.

Svorou, Soteria. (1994). *The Grammar of Space*. Amsterdam: John Benjamins.

Taulé, Mariona, M.; Antònia Martí; Marta Recasens (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. In: *Proceedings of 6th International Conference on Language Resources and Evaluation*, Marrakesh, 96–101.

Glottometrics 44, 2019, 16-33

A Quantitative Analysis of English Variants Based on Dependency Treebanks

Guoqiang Zhang[1], *Haitao Liu*[2]

Abstract. Different English variants have been found to have different characteristics with respect to the aspects of pronunciation, vocabulary, spelling, and grammar, but there is a dearth of research from the perspective of sentence length and syntactic dependency. Therefore, the present research studies these two aspects based on 12 self-building academic dependency treebanks and finds that: (1) in terms of sentence length and syntactic dependency, different variants of English manifest great similarities; (2) regarding the distribution of dependency distance of English variants, the parameters of right-truncated modified Zipf-Alekseev model can provide evidence for the different stages of English globalization; (3) in addition, their syntactic dependencies are little affected by their respective native languages, suggesting the important roles of grammar and cognition in the syntactic constraint. The present research provides a new perspective to study English variants, and its conclusion is hopefully expected to enrich the current research theory.

Keywords: *English variants, sentence length, dependency distance, dependency direction*

1.Introduction

Along with the advancement of the world, all existing languages always change, evolve, and adapt to the needs of their users. English is such a case, and it has become an indisputable fact. After World War II, the rising of Britain and the US, and the spreading of advanced science and technology worldwide caused an unprecedented flourishing in the learning of English all over the world. Until today, English has been considered as a universal language and is being used in almost all aspects of international communication by both native speakers, and second language learners (Simons & Charles, 2018). However, its globalization is inevitably accompanied by localized assimilation, also called localization. In other words, in non-English speaking countries, the western British or American language culture English carries cannot fundamentally change the local language culture and thinking pattern, which will, in nature, resist the outbound English language culture. Then, during a long-time mutual interaction and penetration, they finally melt and reach a relatively stable fusion. In this process, new variants of English, which signify specific national and regional characteristics, came into being, and

[1] Jiyang College of Zhejiang A&F University. Email Address: 20110025@zafu.edu.cn
[2] Ningbo Institute of Technology, Zhejiang University, China. Correspondence to: Haitao Liu. Email Address: htliu@163.com, ORCID No.: https://orcid.org/0000-0003-1724-4418

meanwhile gradually caught the attention of scholars. After the term “Englishes” or “English variants”[3] was formally put forward in the academic field in 1978, various theories and research methods emerged one after another during the subsequent 40 years, among which Kachru's (1985) theoretical framework is more consistent with the history of English development. He proposed the famous three-concentric-circle theory, dividing the development and evolvement of the English into three levels: the “inner circle” countries are countries with large communities of native speakers of English, e.g., Britain, America, Australia, and considered to provide the norm; the "outer circle" countries have small communities of native speakers, but widely use English as a second language in education or broadcasting or for local official purposes, such as India, Nigeria, Philippines - these will develop the norm; while the “expansion circle” just relying on the norms provided by the inner circle, refers to countries which consider learning and using English as a second language, e.g., China, Germany, Japan. Referring to this kind of classification, it is natural to categorize English according to the geographical location or cultural background in the spreading place. Therefore, we may easily see the point that on the national level, different variants of English are closely related to their respective countries, such as Italian English, French English, German English, and so on.

The study of English variants has gradually aroused the attention of academics, and produced a series of research results. Reviewing the literature, we found that previous researches mainly focus on the changes of different English variants at the phonemic, lexical, and discourse levels. For example, Jiang Yajun (2003) conducted a qualitative analysis of three genres (letter, scientific article, and news) from the aspects of phonemics, discourse and vocabulary, and demonstrated the peculiarity of Chinese English (Jiang & Du, 2003), which was further supplemented by Mesthrie & Bhatt (2008) at the syntactic and pragmatic levels. In the book *World Englishes - A Study of New English Variants*, they, through a detailed case study, analyzed the similarities and differences of English variants in British colonies. Even between the two mainstream variants of British English and American English, there are some variations in linguistic features (Mesthrie & Bhatt, 2008). By referring to previous research perspectives, Rohdenburg and Schlüter (2009) made a contrastive analysis between British English and American English, and systematically cultivated their common and specific features in terms of pronunciation, grammar, lexicon, and pragmatic function (Rohdenburg & Schlüter, 2009).

Therefore, previous researches concern more about the linguistic commonality and peculiarity of different variants of English. However, these perspectives only demonstrate the variation of linguistic use in different cultural settings and ignore the study of discoursal stylistic and cognitive features, such as sentence length and language comprehension complexity. Sentence length is closely related to discoursal genres and authors’ writing styles, and has been considered as an evaluating factor in text analysis. Language comprehension complexity reflects the cognitive mechanism of information processing, and from the standpoint of dependency grammar, it manifests the linear distance between governing words and dependent words within a language (Grodner & Gibson, 2005; Temperley, 2007; Levy et al., 2013), also termed as dependency distance. Then, what are the differences and similarities of different English variants in terms of sentence length and syntactic dependency? - In order to find a tentative answer to this question, we conduct an empirical study of international journal

[3] “Englishes” and “English variants” are always used interchangeably.

articles composed by writers from different countries, and intend to analyze English variants from these two aspects. The following three specific questions will be discussed to address the aforementioned issues:

(1) What is the general feature and distribution of the sentence length in academic English variants?
(2) From the perspective of syntactic dependency, what are the similarities and differences among English variants?
(3) Is the syntactic dependency of English variants in synergy with their respective native languages?

This section is followed by the theoretical interpretation of sentence length and syntactic dependency. Section 3 presents the details of material and method used in this study, and section 4 focuses on the result and discussion. Finally, section 5 gives the concluding remark.

2. Theoretical Basis

A sentence is a basic linguistic unit consisting of one or more words that are grammatically linked. The investigation related to it include the study of syntactic formation, sentence meaning, and sentence pragmatics. Apart from these, sentence length, as an overt feature of the text, is an important dimension in exploring the linguistic features of the text. In English, sentence length is usually measured by the total number of words (Grabska-Grudzińska et al. 2012) or clauses in the sentence (cf. Köhler, Altmann & Piotrowski 2005) and can function as an indicator of text differentiation. Previous researchers (Kelih. Grzybek, Antić, & Stadlober 2005) found that sentence length is closely related to styles of text, and can be used for text typology. Kelih et al. (2005) conducted an empirical analysis of sentence length in 333 Slovenian texts, and found that the average sentence length in scientific articles and open letters is significantly longer than that of private letters, recipes, novels and dramatic texts, suggesting that formal genres prefer to use longer sentences than their informal counterparts. Furthermore, sentence length can embody the feature of an individual's writing style. In other words, different writers will demonstrate their specific patterns of sentence length choices (Mannion & Dixon, 2004; Pande & Dhami, 2015). Therefore, the research of sentence length is often used to solve copyright disputes or in stylistic typology. Since it can reflect certain discourse features, we will use it to conduct a comparative analysis of English variants in an attempt to explore certain patterns or variations within a specific genre.

Apart from representing an explicit feature of a certain text, sentence length has been proved to serve as a criterion to measure the difficulty of a sentence (Perera, 1980), indicating that sentence length is proportional to the difficulty of the sentence: the longer the sentence, the more difficult it is (Troia, 2011). This finding is consistent with the syntactic analysis from the perspective of dependency distance, which is employed in the present research to study English variants from the standpoint of dependency grammar. We take it for granted that the syntactic structure of a sentence consists of nothing but dependencies between individual units. Therefore, the analysis of these dependencies is mainly intended to determine the relationship between words, termed as dependency relation, of which the followings are generally considered to be its main features (Hudson, 2007; Liu, 2009).

1. *It is a binary relation between two linguistic units.*
2. *It is usually asymmetrical and directed, with one of the two units acting as the governor and the other as the dependent.*
3. *It is labelled, and the type of the dependency relation is usually indicated using a label on top of the arc linking the two units.*

One of the important properties of dependency relation is dependency distance, which indicates the linear distance between the governor and the dependent, measured by the positional difference (Hudson, 1995; Liu, 2009), and has been proved to be related to human working memory (Hudson, 2003; Liu, 2008). Its value reflects the intensity of human working memory load, and is usually used to reflect the cognitive difficulty of a sentence: the greater the dependency distance, the more difficult the sentence-processing analysis is (Liu, Zhao, & Li, 2009; Jiang & Liu, 2015). One of special dependency distances with the absolute value of 1 is between two adjacent words, also called adjacent dependency, which accounts for a large proportion in the overall distribution of the dependency distances. Collins (1996) found that 74.2% of dependency relations for English are between adjacent words. The proportion is 78% in Eppler's (2005) study and 61.7% in Jiang and Liu (2015). Furthermore, dependency distance can reflect word order or dependency direction. Its positive or negative value can help indicate the positional relationship between the governor and the dependent. If the value is positive, the governor is in front of the dependent (Head-initial); on the contrary, when the value is negative, the governor is located after the dependent (Head-final)[4]. This kind of direction between two dependent words can be an effective way to distinguish and categorize different languages (Liu, 2010). Moreover, the study of dependency distance does not only help to understand the cognitive processes of human beings, but also reflects the universality and specificity of natural languages (Jiang & Liu, 2015).

3. Research Materials and Methods

3.1 Corpus

The empirical studies based on large multilingual corpuses found that the dependency distance of human languages will be affected to some extent by annotating systems and research genres (Liu, 2008; Wang & Liu, 2017). Therefore, this study chooses academic genre as the research object, and adopts same annotating system, just to avoid the interference of these two aspects.

As no language study can exhaust all language materials, this study is carried out based on self-built corpuses. We extract texts randomly from international journals included in the ScienceDirect database[5] to ensure the representation of each English variant. The articles were all published between 2007 and 2017 to enhance the comparability. According to the affiliation/categorization of the database, 12 countries were selected: China, Britain, India, Iran, Israel, Turkey, Italy, Spain, Japan, the Czech Republic, France and Germany; regarding the

[4] In dependency relation, the terms governor or head are used interchangeably, with on distinction, as can be seen in (Liu, 2008, 2010; Liu, Zhao, et al., 2009).
[5] https://www.sciencedirect.com/

subject attribution of the thesis, four disciplines are selected: social science, physics, art and humanities, and genetics. Therefore, for each country, 120 papers were extracted (social sciences: 30; physics: 30; art and humanities: 30; genetics: 30). To ensure the efficiency and credibility of the study, we only extract the introduction part of the research papers to construct a large corpus of English variants. After that, sentences are delimited by graphological features such as upper-case letters and markers such as periods, question marks, and exclamation marks. Since the present research concerns the study of English variants from the perspective of syntactic dependency, sentence length is calculated by the total numbers of words in a sentence, and the average sentence length is the ratio of the number of words to the number of sentences. The statistics of each corpus is shown in Table 1.

Table 1

The statistics of English variant treebanks from 12 countries

English	Tokens	No. sentences	Average sentence length
Chinese English	74621	2803	26.62
British English	83191	2917	28.52
Indian English	90718	3703	24.5
Italian English	77944	2672	29.17
Israeli English	97498	3658	26.65
Iranian English	79818	3182	25.08
Spanish English	89613	3297	27.18
Turkish English	80730	3223	25.05
Japanese English	82906	3210	25.83
Czech English	85820	3297	26.03
French English	91206	3494	26.1
German English	80492	3058	26.32

In order to minimize the error and ensure the consistency of the annotation, Stanford Parser (3.4)[6] was used to analyze and annotate the dependencies of the corpus. Although it is still inevitable to produce parsing errors, Stanford Parser, as a probability-based parsing tool that uses authoritative Penn Treebank training data, is considered to be ideal in syntactic annotation. Moreover, after many updates, its accuracy is greatly improved. Its output (e.g. *my dog also likes eating sausage*) is shown as follows:

nmod:poss(dog-2, My-1)
nsubj(likes-4, dog-2)
advmod(likes-4, also-3)
root(ROOT-0, likes-4)
xcomp(likes-4, eating-5)
dobj(eating-5, sausage-6)

In this example, the *nmod: poss, nsubj, advmod, xcomp, dobj* outside the brackets indicate

[6] https://nlp.stanford.edu/software/lex-parser.shtml

the dependency relation between two words; in the brackets, the left side of the comma is the governor (*dog, likes, likes, likes, eating*), followed by a number suggesting the position of the governor in the sentence. Similarly, the right side of the comma is the dependent and its position number; *Root* indicates that the word *likes* is the root node of the sentence, which has no dependent, and its dependency distance has been defined to equal 0. Then, the outputs of syntactic analysis are introduced into Excel tables according to the columns (dependency relation, governor, governor positional number, dependent, and dependent number) to facilitate the statistical analysis.

3.2 Dependency Distance Calculation

Dependency distance is an important concept in dependency grammar, which is mainly used to measure the linear distance and describe the positional relationship between the governor and the dependent. Liu (2009) proposed a method to calculate the dependency distance. Assuming the existence of $W_1 ... W_i ... W_n$ word sequence, for any dependency relation between W_x and W_y ($0 < x$) - if W_x is the head word and W_y is the dependent word -, then the dependency distance between these two words is equal to *x minus y*. If x is greater than y, the difference of *x minus y* is greater than 0, suggesting dependency direction as the head-initial (HI), while conversely, if *x* is smaller than *y*, the difference of *x minus y* is a negative number, indicating the head-final (HF) direction. In the calculation of dependency distance, the absolute value is usually calculated, and adjacent dependency (dependency distance equals 1) as a special dependency refers to dependency relation between two adjacent words and occupies a higher proportion in discourse.

According to the calculating method, the average dependency distance (MDD) of the whole sentence is as follows:

$$\mathrm{MDD(Sentence)} = \frac{1}{\mathrm{n-1}} \sum_{\mathrm{i=1}}^{\mathrm{n-1}} \left| DD_i \right| \qquad (1)$$

In the formula, *n* is the total number of words in the sentence, and *DDi* is the dependency distance of the i^{th} dependency link. The root node of each sentence has no governing word, so its dependency distance is 0. The average dependency distance of the whole sentence is equal to the sum of all dependency distances divided by the difference between *n* and *1(n-1)*. Regarding the sample sentence mentioned above (*my dog also likes eating sausage*), the DD of *nmod:poss(dog-2, My-1)* is 2-1 = 1; the DD of *nsubj(likes-4, dog-2)* is 4-2 = 2; the DD of *advmod(likes-4, also-3)* is 4-3 = 1; the DD of *xcomp(likes-4, eating-5)* is 4-5 = -1; the DD of *dobj(eating-5, sausage-6)* is 5-6 = -1. According to formula (1), the MDD of this sentence is (1+2+1+1+1)/(6-1) = 1.2.

This calculating method can be applied beyond the sentence level. Then, after a modification, formula (1) can also be used to calculate the average dependency distance of a dependency treebank:

$$\mathrm{MDD(Treeba\ nk)} = \frac{1}{\mathrm{n-s}} \sum_{\mathrm{i=1}}^{\mathrm{n-s}} \left| DD_i \right| \qquad (2)$$

As in formula (2), quite differently, *n* refers to the total number of words in a treebank, and

s represents the total sentence number within the treebank. Similarly, the dependency distance of the root node is defined to be 0, for it has no governing word. Therefore, the average dependency distance of a treebank is the sum of all the dependency distances divided by the difference between *n* and *s(n-s)*.

4. Results and Discussion

4.1 The Distribution of Sentence Length

As an explicit feature of the text, sentence length presents different distribution tendencies in different styles or works of different writers, and is often used to study the stylistic feature of a different discourse (Kelih et al., 2005; Mannion & Dixon, 2004; Pande & Dhami, 2015). Therefore, 'sentence-lengths are not randomly distributed throughout a given text written by a certain author' (Sichel, 1974), but show a certain regularity. If this is the case, can the conclusion be extended to a group of writers, e.g., writers from different countries and with different native languages? – In the present study, we control the genre interference by choosing academic articles, and intend to find a preferred distribution mode in terms of sentence length.

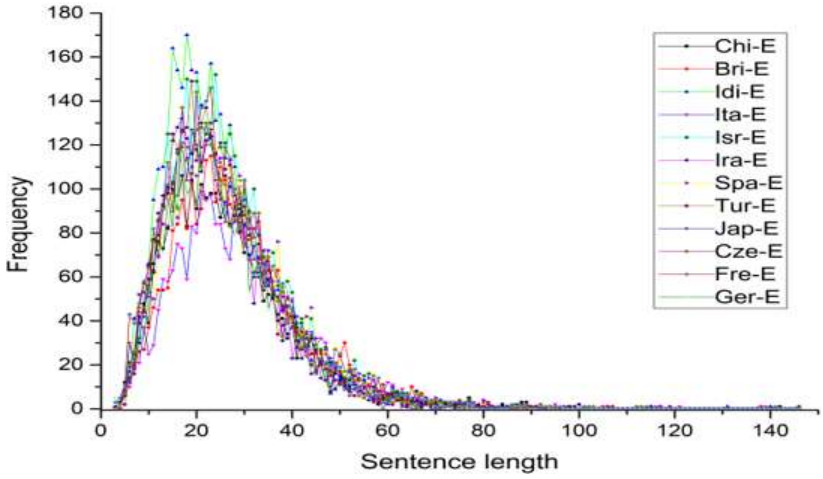


Figure 1. Statistics of sentence length of 12 English Variants[7]

After a statistical computation, we obtain the data of sentence length in different variants, as shown in Figure 1. Overall, sentence lengths of different variants manifest an interesting consistency, which is roughly in line with the unimodal distribution. In order to obtain a more scientific conclusion, we use Altmann-Fitter (3.1) to fit the data, and find that sentence lengths of all variants fit well the mixed negative binomial (k, p1, p2, α) distribution, and the fitting effect is very good ($P(X^2)$[8] > 0.05, $R^2 > 0.97$) [see Appendix A]. Meanwhile, using the data, we

[7] In this and following figures and tables, English variants are hereafter represented by the first three letters of each country plus "-E". Chinese English is referred to as Chi-E; British English is referred to as Bri-E; Indian English as Idi-E; Italian English as Ita-E; Israeli English as Isr-E; Iranian English as Ira-E; Spanish English as Spa-E; Turkish English as Tur-E; Japanese English as Jap-E; Czech English as Cze-E; and French English as Fre-E; and German English as Ger-E.

[8] When P (X^2) is more than 0.05, the fitting effect is very good (Liu, 2009); When P (X^2) is more than 0.01, the fitting effect is good.

also verify the hyperpascal distribution (Ishida & Ishida, 2007) and the extended positive negative binomial distribution (Pande & Dhami, 2015), which are considered to be appropriate models for the frequency of sentence lengths. Ishida & Ishida (2007) found that the hyperpascal distribution could be effectively applied to Japanese texts, while Pande & Dhami (2015) demonstrated that the extended positive negative binomial distribution is the right model for describing the sentence length distribution in Hindi language texts. However, in our study, none of the variants could be fitted by the hyperpascal distribution, and in the case of the extended positive negative binomial distribution, for 3 variants (British English, Italian English, and Japanese English) the condition of good fit, and for 2 variants (Spanish English and Turkish English) the condition of acceptable fit are satisfied (see Appendix B). Furthermore, we find that the negative binomial distribution is satisfied in 7 out of 12 variants (good fit: British English, Italian English, Spanish English, and Japanese English; acceptable fit: Israel English, Iranian English, and Turkish English) [see Appendix C], which contradicts Yule's conclusion that this distribution is inadequate for a representation of sentence-lengths in prose (Yule, 1944, cited in Sichel, 1974). All above findings indicate that the distribution of sentence length may be affected by research genres and writing systems in different languages[9]. However, based on the present research, we cannot tell to what extent a genre or a writing system affect the distribution of sentence length. Either one of the variables must be controlled in order to study the effect of the other variable on a sentence length distribution.

The average sentence lengths of English variants vary within a small range from 24.5 to 29.2 (see Table 1), with the average value of 26.4, which can be considered as the representative sentence length of academic articles. Comparing it with the average sentence length of English news (msl=20.9, see Table 1 in [Liu, 2008]), we found that the academic genre tends to use longer sentences, suggesting that the more formal the genre, the longer the sentence; and this may possibly be determined by the rigorous and interpretative nature of academic writing.

In addition to displaying a consistent distribution in terms of the frequency of sentence lengths, academic English variants from different countries manifest another common feature. While publishing articles in international journals, writers scarcely tend to employ not only very short sentences (less than 10 words), but also extremely long sentences (more than 70 words) [see Figure 1]. Short sentences, frequently found in the oral context, advertising texts and fiction stories contains a limited amount of information and are not appropriate to be used to interpret the complicated concepts and rigorous logic of academic texts. Conversely, extremely long sentences, conveying too much information, are often confusing to readers; for native speakers, this is generally regarded as a way of writing 200 years ago, and it has been proved to be inappropriate in the current academic English writing either (Wallwork, 2013).

Furthermore, with the increase of frequency, the sentence length distribution of the variants also shows a certain discreteness. The fitting model confirms this point, too. Although fitting well the mixed negative binomial (k, p1, p2, α) distribution, the variance of fitting parameters of different variants also demonstrated that while writing academic articles, writers from different cultural backgrounds also manifest their different group peculiarities. This can also be seen in Table 1, in which the average sentence length of English variants ranges from 24

[9] Different languages have different writing systems, e.g., Japanese has three different writing systems, hiragana, katakana, and kanji; Chinese has characters and words. Therefore, regarding the sentence length measurement, there are multiple criteria.

to 30 words with the Indian variant being the shortest (24.5) and the Italian variant the longest (29.1). However, the differences are not significant ($p < 0.05$), and this fact can provide evidence for the consistency of sentence length in the same genre (Kelih et al., 2005).

Sentence length has been proved to be related to syntactic difficulty or complexity (Perera, 1980). In general, as the average sentence length increases, the complexity of the sentences also increases (Troia, 2011). Likewise, from an empirical perspective, Liu (2008) provided evidence and proposed a calculating method in terms of measuring syntactic dependency. He found that the mean sentence length of a text is positively related to dependency distance (the longer the sentence, the greater the dependency distance). In this sense, we can hypothesize that since there is no significant difference in the aspect of sentence length, there will be no significant difference in average dependency distance. In next section, we will test this hypothesis through a statistical analysis from the standpoint of syntactic dependency.

4.2 Syntactic Dependency of English Variants

4.2.1 The Probability Distribution of Dependency Distance

Dependency distance, as a measure of the difficulty of sentence, reflects the extent to which human working memory restricts language comprehension and production. In other words, a word can be removed from short-term memory only after it is connected with other words and forms a dependency relationship (Liu, 2008). Research showed that human information-processing capacity is limited to 7±2 (Miller, 1956), and it is this similar working memory that makes the distribution of dependency distance show a consistent tendency among different languages (Liu, 2007; Jiang & Liu, 2015), and different genres within the same language (Wang & Liu, 2017). By referring to previous research findings, this part further explores the distribution characteristics of dependency distance among different English variants within the academic genre.

We use Altmann-Fitter (3.1) to fit the data. As the sample is relatively large, it is necessary to select the best distribution model by referring to the determination coefficient (R^2) and discrepancy coefficient (C) of the fitting at the same time[10]. Within the 12 academic English variants, dependency distance well fits the following distributions: the right-truncated Warring (b, n); the mixed negative binomial (k, p1, p2, α); the mixed geometric (q1, q2, α); the Shenton-Skees-logarithmic (a,b,θ); and the right-truncated modified Zipf-Alekseev (a, b; n = x-max, α fixed), with the determination coefficients (R^2) all above 0.99 (see Table 2). In addition, the right-truncated Zeta model was verified at the same time, and the results are similar to Liu (2007). Although the determination coefficient is lower than the one of other models, it is still acceptable. Then, the data of dependency distance also satisfies the exponential model ($R^2 > 0.99$), and this further confirms Jiang & Liu (2015) and Ferrer i Cancho's (2004) conclusion – even though they only focused on the fixed sentence lengths. Furthermore, other probability models found in this research are consistent with Ouyang & Jiang (2017) and Jiang & Liu (2015). All these shared probability distribution models indicate the universality of

[10] When the scale of a data set is too large and the Chi-square test fails, the discrepancy coefficient C is usually used to judge the fitting effect (when C is less than 0.02, the result is good; when the result is less than 0.01, the result is very good) [Liu, 2009].

human natural languages, through which we can infer that the syntactic acquisition of human language is always constrained by the tendency of dependency distance minimization, and this can be ascribed to the constraint of common working memory capacity and 'the principle of least effort' on syntactic structure.

Table 2

The fitting of dependency distance among 12 academic English variants

English variants	Right truncated modified Zipf-Alekseev	Right truncated Warring	Mixed negative binomial	Mixed geo-metric	Shenton-Skees-logarithmic	Right truncated Zeta	Expo-nential
Chi-E R^2	0.995	0.994	0.998	0.999	0.995	0.939	0.998
Bri-E R^2	0.997	0.995	0.994	0.999	0.998	0.948	0.998
Idi-E R^2	0.996	0.996	0.999	0.999	0.998	0.951	0.980
Ita-E R^2	0.997	0.996	0.998	0.999	0.997	0.951	0.998
Isr-E R^2	0.995	0.995	0.999	0.999	0.999	0.949	0.998
Ira-E R^2	0.996	0.994	0.996	0.998	0.999	0.947	0.998
Spa-E R^2	0.997	0.994	0.999	0.999	0.998	0.949	0.998
Tur-E R^2	0.996	0.995	0.997	0.999	0.997	0.949	0.998
Jap-E R^2	0.996	0.995	0.997	0.999	0.999	0.946	0.998
Cze-E R^2	0.995	0.995	0.996	0.999	0.996	0.953	0.998
Fre-E R^2	0.996	0.995	0.999	0.999	0.996	0.952	0.998
Ger-E R^2	0.997	0.995	0.988	0.999	0.999	0.947	0.998

However, the fitting results of these distributions to different English variants are not the same. The parameters of same model fitting to different variants show a certain discrete trend. And Ouyang and Jiang (2017) found that the variation of the parameters of probability distribution of dependency distance could measure well the language proficiency of second language learners. Similarly, in English variants, do the parameters of different distributions reflect certain tendencies? In order to find a tentative answer, we carried out statistical analysis of the parameters of the fitted models, and obtained interesting findings about the right truncated modified Zipf-Alekseev distribution. As can be seen in Table 3, the values of its four parameters (*a*, *b*, *n* and α) are presented. In order to explore the possible pattern or tendency of the parameters, we sort the data from small to large by the parameter *a*, and the clearer illustration of the variations is shown in Figure 2.

Table 3
Fitting the right truncated modified Zipf-Alekseev model to the dependency distance of different variants

Variants	a	b	n	α	X^2	$P(X^2)$	DF	C	R^2
Chi-E	0.9207	0.3637	131	0.4549	1562.849	0	78	0.0198	0.9945
Jap-E	1.1247	0.3219	160	0.4679	1134.335	0	83	0.0142	0.9966
Bri-E	1.0468	0.3367	140	0.4689	1356.489	0	82	0.0169	0.996
Idi-E	1.0758	0.3433	124	0.4773	1229.138	0	78	0.0141	0.9967
Ita-E	0.9312	0.3658	134	0.4712	1493.091	0	77	0.0198	0.9953
Isr-E	1.599	0.3301	90	0.467	1571.158	0	77	0.0167	0.9959
Ira-E	1.2218	0.2964	107	0.4689	869.68	0	80	0.0113	0.9972
Spa-E	1.0978	0.323	99	0.4699	1406.01	0	79	0.0163	0.9959
Tur-E	0.9984	0.3602	111	0.4702	1256.438	0	75	0.0162	0.996
Cze-E	1.0393	0.3414	108	0.4796	1728.724	0	78	0.0189	0.9952
Fre-E	0.9538	0.3698	101	0.4757	1645.41	0	74	0.0188	0.9955
Ger-E	1.0818	0.3307	103	0.4651	1120.532	0	77	0.0145	0.9965

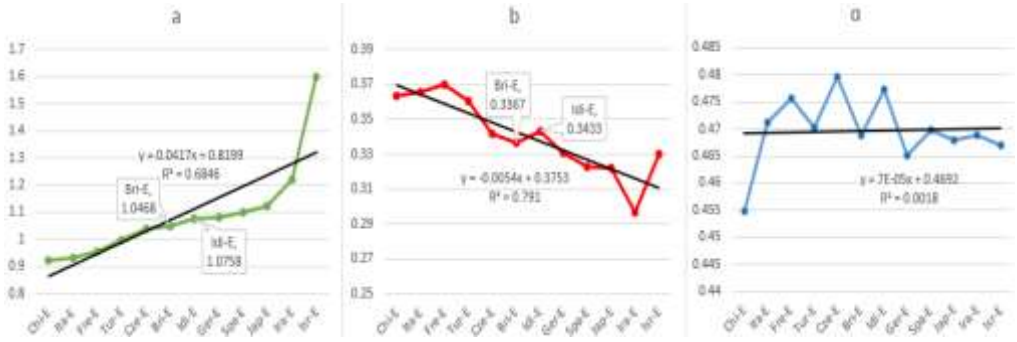


Figure 2. The variations of parameters (a, b, α) of the right truncated modified Zipf-Alekseev fitting the dependency distance of different variants

Through linear fitting, we find that the distribution of parameter *a* shows an upward trend from the Chinese variant to the Israeli variant (from small to large), while that of parameter *b* displays a corresponding falling tendency, and parameter *α* shows no significant variation – its fitting trend is basically parallel to the x-axis. Most interestingly, the British English variant always serves as the midpoint, whether it is in the parameter *a* distribution or the parameter *b* distribution, followed by the Indian English variant. In other words, the data of parameter *a* and parameter *b* of different English variants fluctuate around those of the British variant and the Indian variant. This finding provides evidence – from the empirical view – for Kachru's (1985) three-concentric-circle theory, suggesting that, in the globalization of English, British English as an English native language serves as the universal norm; in this regard, Britain's colonial countries, such as India, considering English as a second or official language, will develop the norm, while the other non-English speaking countries – such as China, Japan, Germany, Italy, etc. – will treat English as a foreign language, and learn and use English by following the norm

of British English. Therefore, their English variants manifest a trend of fluctuating around British English variant.

4.2.2 Statistical Analysis of Dependency Distance and Dependency Direction

With regard to the findings of the previous studies (Liu, 2007; Jiang & Liu, 2015; Wang & Liu, 2017), we can see that dependency distances follow particular models in different languages, genres, and variants. In other words, these models all share the features of a long-tail distribution, which suggests a distributional tendency that the smaller the dependency distance, the higher its frequency, and vice versa. Therefore, the dependency relation between two adjacent words with the dependency distance of 1 has been found to account for a large proportion in the total dependencies (Collins (1996): 74.2%; Eppler (2005): 78%; Jiang and Liu (2015): 61.7%), and considered one of the major reasons for the dependency distance minimization of human language (Jiang & Liu, 2015). In the present research, we find that the percentages of adjacency dependency in the English academic variants are varying within a small range (no more than 3%), being generally close to 50%, as shown in Table 4. Thus, we make a comparison with Liu's (2008) finding in English news (51.2%), and find that the proportion of adjacent dependency in academic articles is slightly lower, which may be attributed to genre differences.

Table 4

The statistics of dependency distance of English variants

Variants	MDD	HI(%)	DD(HI)	HF(%)	DD(HF)	1DD(%)
Chi-E	2.73	52.55%	-3.62	47.45%	2.47	45.49%
Bri-E	2.73	52.94%	-3.44	47.06%	2.41	46.89%
Idi-E	2.58	52.78%	-3.24	47.22%	2.39	47.73%
Ita-E	2.72	52.74%	-3.43	47.26%	2.44	47.12%
Isr-E	2.72	53.30%	-3.44	46.70%	2.44	46.70%
Ira-E	2.66	53.66%	-3.26	47.77%	2.51	46.89%
Spa-E	2.70	52.72%	-3.39	47.28%	2.44	46.99%
Tur-E	2.65	51.44%	-3.37	48.56%	2.43	47.02%
Jap-E	2.67	51.95%	-3.31	48.05%	2.47	48.04%
Cze-E	2.63	53.11%	-3.34	46.89%	2.42	47.96%
Fre-E	2.62	51.10%	-3.32	48.90%	2.44	47.56%
Ger-E	2.69	52.33%	-3.37	47.67%	2.47	46.51%

The dependency direction of academic English variants also shows a similar distribution in the fact that both the proportion of head-initial dependency (HI) and that of head-final dependency (HF) vary in a small range, with the overall proportion close to 50% (see Table 4). However, the average dependency distance of head-initial dependency is greater than that of head-final dependency, which may be determined by the inner syntactic nature of English.

There are great similarities shared by all academic English variants in terms of de-

pendency distance distribution, mean dependency distance (MDD), dependency direction (HI and HF), and the proportion of adjacent dependency. However, English is inevitably influenced by the native language, which definitely reflects the interplay of local society, economy, and culture in the spreading areas. Therefore, to what degree do the native languages influence the syntactic dependency of different English variants in the academic genre? In order to find an answer to this question, the present study gives reference to previous research results. By studying the syntactic dependencies in 20 languages, Liu (2008, 2010) found that different languages have different features in terms of dependency distance and dependency direction (see Table 5). Although both the annotating scheme (Liu, 2008), and the genre (Wang & Liu, 2017) have an effect on syntactic dependency, the effects are so trivial that they will not change the research conclusion. Therefore, we make a comparative analysis of the syntactic dependency among the different academic English variants and their respective local native languages, and find that: (1) comparing to Liu (2008, 2010), the percentage of adjacent dependency (1DD) in academic English variants is generally lower (around 47%) than the one of their respective native languages; (2) and likewise, the mean dependency distance of English variants varies within a smaller range (from 2.63 to 2.73), and fluctuates around 2.68 (median); (3) furthermore, dependency directions in English variants display a good consistence (HI% is slightly higher than the one of HF%), regardless of whether their native languages are inclined to be head-initial (Italian, Spanish, and German) or head-final (Japanese, Turkish, and Chinese) [see Table 5]. These findings indicate that native languages have little or no influence on their corresponding English variants, and also suggest that within the same annotating scheme and genre frame, it is the interaction of the same grammatical system and the similar cognitive mechanism that make different English variants display a very consistent tendency in syntactic dependency (Liu, 2008).

Table 5

A comparative study between English variants and their respective native languages

Languages	Dependency Distance（Liu 2008）		Dependency Direction（Liu 2010）		Academic English variants				
	1DD%	MDD	HI%	HF%	Variants	1DD%	MDD	HI%	HF%
Japanese	80.2	1.805	11	89	Jap-E	48	2.668	51.95	48.05
Italian	72.4	2.19	64.8	35.2	Ita-E	47.1	2.715	52.74	47.26
Turkish	64.2	2.322	5.9	94.1	Tur-E	47	2.654	51.44	48.56
Czech	53	2.441	54.5	45.5	Cze-E	47.9	2.634	53.11	46.89
English	51.3	2.543	48.8	51.2	Eng-E	46.9	2.726	52.94	47.06
Spanish	55.2	2.665	63.6	36.4	Spa-E	47	2.695	52.72	47.28
German	44.4	3.353	54.2	45.8	Ger-E	46.5	2.689	52.33	47.67
Chinese	47.9	3.662	31.5	68.5	Chi-E	45.6	2.734	52.55	47.45

What is more, concerning British English, the proportion of adjacent dependency (1DD%) of news is slightly higher than that of academic discourse (51.3 > 46.9), while it is opposite in

terms of dependency distance. From the perspective of genre, this can be accounted for by the fact that the more formal the genre is, the greater the average dependency distance tends to be, which is consistent with Hiranuma's (1999) and Liu. et al.'s (2009) conclusion.

5. Conclusion

In the process of globalization, English is inevitably influenced by the politics, economy and culture of local countries, and has gradually formed variants with specific native features. Regarding previous investigations which study English variants from the perspectives of pronunciation, vocabulary, grammar and even discourse, this paper serves as a piece of complementary research which focuses on English variants from the aspects of sentence length and syntactic dependency. Focusing on self-built treebanks which extract texts randomly from international journal articles written by different native speakers, we carry out a comprehensive study by means of quantitative methods and obtain very interesting findings: (1) the sentence length distribution of all English variants fits well the mixed negative binomial, while some fit the negative binomial and the extended positive negative binomial distributions, but none fit the hyperpascal distribution. The differences of fitting suggest that the distribution of sentence length will be affected not only by genres, but also by different writing systems of certain languages; (2) In addition, while writing articles for international journals, writers from all over the world tend to scarcely use ultra-short sentences (less than 10 words) and over-long sentences (more than 70 words), and the variation of sentence length distribution tends to be discrete with the increase of frequency, but within a restricted range, which could be seen through the variation of the parameters of the same model; (3) the distribution of dependency distance of different English variants fits well several models, of which the parameters (*a, b*) of the right-truncated modified Zipf-Alekseev distribution provide evidence for Kachru's (1985) three-concentric-circle theory of English variants; (4) furthermore, the mean dependency distance (MDD), adjacent dependency (1DD), and dependency direction of different variants all vary within a minimal range, indicating the great similarity among English variants of academic genre. We also find that native languages have little impact on the syntactic dependency of English variants, suggesting the great contribution of grammar and cognition in the minimization of dependency distance of human languages (Liu, 2008).

The findings of the present research have shown great similarity in terms of sentence length and syntactic dependency, through which we can infer that different English variants share certain inner homogeneity caused by the same grammar and similar cognitive mechanism; their peculiarities mainly lie in some superficial aspects, such as pronunciation, lexicon, and discourse. It is this internal universality among different variants that may make academic communication easier for scholars from all over the world. However, some linguistic features are not touched, thus they need to be further explored. Furthermore, by following the research method, the three-concentric-circle theory, representing the globalization of English, should be empirically tested and verified in a larger and more balanced variant treebanks in further research.

Acknowledgments

This work is partly supported by the Teachers' Development Foundation of Education Department of Zhejiang Province(Grant No. FX2017037)

REFERENCES

Collins, M. (1996). A New Statistical Parser Based on Bigram Lexical Dependencies. In: *Proceedings of the 34th Annual Meeting on Association for Compu- Tational Linguistics. Association for Computational Linguistics*, 184–191. https://doi.org/10.3115/981863.981888.

Eppler, E. D. (2005). *The syntax of German-English code-switching.* Doctoral thesis, University of London.

Ferrer i Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 70(5)*, 5. https://doi.org/10.1103/PhysRevE.70.056135

Grabska-Gradzińska, Iwona et al.(2012). Multifractal Analysis of Sentence Lengths in English Literary Texts." *ArXiv:1212.3171* (March 2014):5. Retrieved (http://arxiv.org/abs/1212.3171).

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2), 261–290. https://doi.org/10.1207/s15516709cog0000_7

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

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

Hudson, R. (2007). *Language Networks: The New Word Grammar. Studies in Second Language Acquisition*. Oxford: Oxford University Press.

Ishida, M., & Ishida, K. (2007). On distributions of sentence lengths in Japanese writing. *Glottometrics, 15*, 28–44.

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*(May), 93–104. https://doi.org/10.1016/j.langsci.2015.04.002.

Jiang, Y.,& Du, R. (2003). On the Question of "Chinese English"--Response to the "Question of 'Chinese English.'" *Foreign Language Education, 1*, 27–35.(In Chinese).

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In:R. Quirk & W. H.G. (eds.), *English in the world: Teaching and learning the language and literatures (pp. 11–30)*. Vambridge_ Cambridge University Press

Kelih, E., Grzybek, P., Antić, G., & Stadlober, E. (2005). Quantitative text typology: The impact of word length. In *Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 53–64). https://doi.org/10.1007/3-540-28084-7-5

Köhler, R., Altmann, G. & Piotrowski, R. (2005). *Quantitative Linguistik / Quantitative Linguistics. Ein internationales Handbuch / An International Handbook*. Berlin, Boston: De Gruyter Mouton.

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language, 69*(4), 461–495. https://doi.org/10.1016/j.jml.2012.10.005

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics, 15*, 1–12. Retrieved from http://www.lingviko.net/glotto15f.pdf

Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, *9*(2), 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Liu, H. (2009). *Dependency Grammar: From Theory to Practice.* Beijing: Science Press.

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, *120*(6), 1567–1578. https://doi.org/10.1016/j.lingua.2009.10.001

Liu, H., Hudson, R., & Feng, Z. (2009). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory*, *5*(2), 161–174. https://doi.org/10.1515/CLLT.2009.007

Liu, H., Zhao, Y., & Li, W. (2009). Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznan Studies in Contemporary Linguistics, 45*(4), 495–509. https://doi.org/10.2478/v10010-009-0025-3

Mannion, D., & Dixon, P. (2004). Sentence-length and Authorship Attribution : the Case of Oliver Goldsmith, *Journal of Memory and Language, 19*(4), 497–508.

Mesthrie, R., & Bhatt, R. M. (2008). *World Englishes: The Study of New Linguistic Varieties (pp. 1–276)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511791321

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus 2 Some Limits on Our Capacity for Processing Information. *Psychological Review,* 63(2), 81–97. https://doi.org/10.1037/h0043158.

Ouyang, J., & Jiang, J. (2017). Can the Probability Distribution of Dependency Distance Measure Language Proficiency of Second Language Learners? *Journal of Quantitative Linguistics*, *6174*, 1–19. https://doi.org/10.1080/09296174.2017.1373991.

Pande, H., & Dhami, H. S. (2015). Determination of the Distribution of Sentence Length Frequencies for Hindi Language Texts and Utilization of Sentence Length Frequency Profiles for Authorship Attribution. *Journal of Quantitative Linguistics*, *22*(4), 338–348. https://doi.org/10.1080/09296174.2015.1106269.

Perera, K. (1980). The Assessment of Linguistic Difficulty in Reading Material. *Educational Review*, *32*(2), 151–161. https://doi.org/10.1080/0013191800320204.

Rohdenburg, G., & Schlüter, J. (2009). *One language, two grammars? Differences between British and American English* (pp. 1–461). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511551970.

Sichel, H. S. (1974). On a Distribution Representing Sentence-length in written Prose. *Journal of the Royal Statistical Society: Series A (General)*, *137*(1), 25–34. https://doi.org/10.2307/2345142

Simons, Gary F. and Charles D. Fennig (eds.). 2018. Ethnologue: Languages of the World, Twenty-first edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

So Hiranuma. (1999). Syntactic difficulty in English and Japanese: A textual study. *UCL Working Papers in Linguistics*, *11*, 309–322.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*. *105*(2), 300–333. https://doi.org/10.1016/j.cognition.2006.09.011

Troia, G. (2011). *Instruction and Assessment for Struggling Writers: Evidence-based Practices*.

New York: Guilford Press.

Wallwork, A. (2013). *English for Research: Grammar, Usage and Style*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-1593-0.

Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*, *59*(866), 135–147. https://doi.org/10.1016/j.langsci.2016.09.006

Appendix A

Fitting the mixed negative binomial model to the sentence length of different variants

Variants	k	P1	P2	α	X^2	$P(X^2)$	DF	C	R^2
Chi-E	4.3766	0.1639	0.0788	0.9389	81.99	0.7859	93	0.0292	0.9875
Bri-E	4.4975	0.1522	0.0810	0.9709	83.35	0.5303	85	0.0286	0.9808
Idi-E	5.1058	0.2001	0.0919	0.9507	89.41	0.4976	90	0.0241	0.9889
Ita-E	4.1909	0.1392	0.0587	0.9917	91.04	0.2810	84	0.0340	0.9716
Isr-E	4.5519	0.1645	0.0819	0.9701	96.51	0.2276	87	0.0264	0.9801
Ira-E	4.7728	0.1844	0.1050	0.9423	65.09	0.8523	78	0.0204	0.9889
Spa-E	4.2304	0.1503	0.0761	0.9887	88.91	0.2321	80	0.0270	0.9835
Tur-E	5.4430	0.2075	0.1197	0.9325	53.37	0.9774	76	0.0165	0.9912
Jap-E	4.2958	0.1613	0.0773	0.9730	75.80	0.7521	85	0.0236	0.9887
Cze-E	4.2560	0.1618	0.0936	0.9320	102.77	0.1050	86	0.0312	0.9795
Fra-E	3.8124	0.1439	0.0686	0.9813	106.07	0.0606	85	0.0304	0.9767
Ger-E	4.3515	0.1587	0.0783	0.9828	97.86	0.0739	79	0.0320	0.9738

Appendix B

Fitting the extended positive negative binomial model to the sentence length of different variants

Variants	k	p	a	X^2	$P(X^2))$	DF	C	R^2
Chi-E	3.8802	0.1396	0.9996	118.01	0.0011	75	0.0421	0.9844
Bri-E	4.2903	0.1432	0.9993	93.70	0.1238	79	0.0321	0.9795
Idi-E	4.4239	0.1671	0.9997	133.96	0.0000	69	0.0362	0.9780
Ita-E	3.8706	0.1300	0.9989	102.45	0.0626	82	0.0383	0.9669
Isr-E	4.0419	0.1478	1.0000	107.60	0.0081	75	0.0294	0.9773
Ira-E	4.1814	0.1536	0.9997	109.85	0.0034	73	0.0345	0.9794
Spa-E	3.8594	0.1392	1.0000	102.01	0.0298	77	0.0309	0.9805
Tur-E	4.7803	0.1770	0.9994	96.67	0.0127	68	0.0300	0.9787
Jap-E	4.0287	0.1493	1.0000	74.65	0.4274	73	0.2320	0.9886
Cze-E	3.6933	0.1378	1.0000	118.89	0.0012	76	0.0361	0.9777
Fra-E	3.5519	0.1334	0.9991	122.50	0.0010	78	0.0351	0.9740
Ger-E	3.8042	0.1401	0.9984	108.25	0.0089	76	0.0354	0.9715

Appendix C

Fitting the negative binomial model to the sentence length of different variants

Variants	k	p	X^2	$P(X^2)$	DF	C	R^2
Chi-E	3.6683	0.1335	118.32	0.0017	77	0.0422	0.9846
Bri-E	4.1193	0.1392	93.76	0.1229	79	0.0321	0.9775
Idi-E	3.9135	0.1545	140.37	0.0000	73	0.0379	0.9778
Ita-E	3.9534	0.1311	99.48	0.0918	82	0.0372	0.9693
Isr-E	4.1693	0.1500	102.65	0.0270	77	0.0281	0.9788
Ira-E	4.0076	0.1546	101.08	0.0165	73	0.0318	0.9860
Spa-E	4.0004	0.1418	97.41	0.0783	79	0.0295	0.9827
Tur-E	4.5864	0.1699	95.61	0.0187	69	0.0296	0.9680
Jap-E	3.9573	0.1479	76.68	0.4245	75	0.0239	0.9883
Cze-E	3.7717	0.1410	121.34	0.0010	77	0.0368	0.9785
Fra-E	3.5488	0.1334	122.61	0.0012	79	0.0351	0.9739
Ger-E	3.8018	0.1401	114.38	0.0037	77	0.0374	0.9714

Glottometrics 44, 2019, 34-58

The Diachronic Relationship Between the Contemporary American English Present Perfect and Past Simple across Registers[5]

Xiaowen Zhang[6], Yunhua Qu[7], Zhiwei Feng[8]

Abstract. The relationship between the diachronic change of the present perfect (PP) and the past simple tense (SP) in English has always been an important, but still puzzling subject of studies on the English aspect-tense system, because the two constructions are both related to past-time reference, and the distinction between the two is not clear-cut in many English varieties. Even contexts labeled by temporal adverbials – like *yesterday*, which tends to be used with the SP, or *since*, which is usually combined with the PP – have become increasingly compatible with the other construction. Therefore, it is assumed that the diachronic change of the PP or the SP should not be studied individually, as is done in many previous studies, but observed from a broader perspective – the competition between the two.

This study, using the largest and most balanced American English corpus, COCA (The Corpus of Contemporary American English), aims to investigate the diachronic change patterns of the PP and the SP, as well as to detect their relationship, in five different registers (Spoken, Fiction, Magazines, Newspapers, and the Academic one). Findings show that the diachronic change of the PP or the SP is closely related, either negatively (i.e. competitively) or positively in the entire corpus, except for the Academic register. To be more specific, the development of the PP and the SP are in competition in COCA and the Spoken register, while positively related in Fiction, Magazine, and Newspaper. No statistically significant correlation is found in the Academic one.

This research adds the dimension of register into the study of the diachronic change of the English tense-aspect system, adopting relatively convincing statistical methods to reveal a panorama about the relationship between the development of the PP and the SP in contemporary American English.

Key Words: *Present perfect, simple past, diachronic change competition, register*

1. Introduction

Synergetics, a theoretical modelling, treats spontaneous changes of structures. Linguistic studies have shown that synergetics applies to functional analytic models and explanatory approaches of quantitative linguistics, which provides concepts applicable to the phenomena of self-regulation and self-organisation in quantitative linguistics. Similar to other self-organising systems, language is characterised by cooperative and competitive processes

[5] Supported by Philosophy and Social Sciences Planning Project of Zhejiang Province(17NDJC201YB).

[6] School of International Studies, Zhejiang University, China; Zhejiang Institute of Communications, China,
E-mail: happywendyzhang@126.com.

[7] Corresponding author School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China, Tel:+86 13735805856, Email: qu163hua@163.com, ORCID No.: https://orcid.org/0000-0003-1724-4418 .

[8] Hangzhou Normal University, E-mail: zwfengde2010@163.com.

which, combined with the external forces of biology, psychology, physics, the social system, and others, form the dynamics of the system (Köhler, 2005). It is the same with the present perfect (PP) and the simple past tense (SP), which show a certain development pattern.

The PP and the SP in English have always attracted significant attention from scholars interested in the English aspect-tense system. Theoretically, the two are different in terms of both aspect and tense. Many previous studies have been devoted to distinguishing them. According to Reichenbach (1971), for the SP, the point of the event and the reference are simultaneous, both before the point of speech, while for the PP, the point of the event is before the point of speech, but the point of reference is simultaneous with the point of speech. Similarly, Meyer-Viol and Jones (2011) point out that the PP is seen as having its reference time at speech time, thus showing a current relevance, while the SP is seen as having its reference time at event time, showing no current relevance at all. Moreover, the former describes a state that exists until present, which serves to connect a past event to a present state in some way, while the latter describes one that no longer exists at present (Biber et al., 1999, 3). The two should not be interchangeable.

It is worth noting that the differences between the two are often made explicit by time adverbials used with the main verb, which we call temporal adverbials. An adverbial used with the simple past tense usually indicates a definite past moment or period when the event or state occurred – like *then*, or *yesterday*, which mark a clear ending point before the present time (Biber et al., 1999). Definite temporal adverbials indicating past should be incompatible with the PP. As mentioned by Klein (1992), the PP does not co-occur with definite past-time adverbs. In many other studies, it has been repeated that the co-occurrences of PP with definite past-time adverbials, especially in "standard" British English, are not acceptable (Portner, 2003; Schaden, 2009). In contrast, the PP is often used with adverbials indicating a period with a beginning point or a duration time, but no definite ending time (Biber et al., 1999). Indefinite temporal adverbials have been shown to be one of the typical triggers of the PP in "standard" English. That is to say, indefinite temporal adverbials like *already*, or *never* tend to be used with the PP. The collocational compatibilities of the PP with indefinite temporal adverbials and SP with definite ones denoting past are also often exploited in EFL materials to help students distinguish PP and SP (Rastall, P. 1999). Some usage guides also attack the combination of the SP with indefinite time expressions (like *yet*) as ungrammatical (Burchfield, 1996).

However, since both the PP and the SP are used to refer to a past event or state, when it comes to a reference to past-time in English, two choices appear – the SP and the PP. A question of why the PP cannot be combined with definite past-time adverbials appeared is termed as the PP puzzle (Klein, 1992). Actually, the PP's co-occurrences with definite past-time adverbials have been observed, which can be more commonly identified / proved in spoken English (Hundt & Smith, 2009; Werner, 2013). On the other hand, the co-occurrences of the SP with indefinite temporal adverbials can be even more commonly observed in modern English, not only in spoken, but also in other registers.

Since the boundary between the PP and the SP is becoming increasingly vague, more and more studies have been devoted to studying the relationship between the two. It is assumed that the two constructions are in competition with each other regarding their development. Therefore, when the development of the PP is studied, usually that of the SP will also be studied at the same time (Elsness 2014; Yao, 2014; Schaden, 200. In comparison with many traditional studies, which focus only on the PP or the SP itself (Arts & Bowie, 2012; Aarts et al., 2013), such studies can reveal the correlation, such as competition, between the changes of the two constructions.

Some interpretations for the identified competition are attempted. The use of the SP will not call for any pragmatic inferences. As a result, the SP may increase at the expense of the PP,

as detected by some studies (Yao & Collins, 2012). The typological oddity of the English PP, the rating data from eWAVE, also implies the interchangeability of the PP and the SP (Szmrecsanyi & Kortmann, 2009). One of the main purposes of the present study is to detect whether in the competition, the PP is losing ground to the SP.

However, few studies concerned with the relationship between the development of the PP and the SP focus on American English (AmE), a variety in which the speed of the increasing ratio of the SP to the PP is considered higher (Yao, 2014), and the preference for the SP over the PP is stronger (Hundt & Smith 2009). A much more significant decrease in the frequency of the PP relative to that of the SP has been observed in AmE data than that in the BmE ones (Yao, 2014).

Moreover, whether the competition patterns are subject to the influence of register even within one variety (like AmE) has not been explored systematically. After all, Biber (2011) has shown that linguistic features can differ significantly across registers. Biber & Gray (2013) confirms that the lexico-grammatical patterns do not generally apply to English. Bao et al. (2017) reveals that different registers exhibit different change in the perfect construction.

To fill the gap in the studies of competition between the PP and the SP mentioned above, the present paper aims to conduct a comprehensive corpus study on the following issues:

1) How does the PP/SP change in the corpus and across registers?
2) How do the two change patterns interrelate with each other?
3) Do the PP and the SP compete, as claimed by Schedan?

2. Research Basis

Quite a few studies have detected that the frequency change of the PP is related to that of the SP, with the former decreasing and the latter increasing (Elsness, 2014; Yao & Collins, 2012). So it can be reasonably assumed that the two are in competition, with one increasing at the expense of the other.

Biber & Gray (2013) present specific case studies of twentieth-century historical language change to show that seemingly minor differences in register can correspond to meaningful and systematic differences in the patterns of linguistic change.

Therefore, hypotheses about the competition between the PP/SP proposed by Schaden (2009), and register analysis framework proposed by Biber and Conrad (2009) will be the main basis for this study, especially for the interpretation of the results.

This study also intends to explore whether the PP/SP competition description holds water and even applies in various registers.

a. PP/SP Competition

According to Schaden (2009), both the PP and the SP are "one-step past-referring tenses" which have to compete against each other. He argues that the competition between the two influences their distribution. Between the PP and the SP, the more restricted the PP is, the less restricted the SP will be, and vice-versa. The distributions of the two are connected as "the more one sees of one form, the less one sees of the other". More importantly, he claims that in the English language, in the context with definite past-time adverbials, the SP is the unmarked form, while the PP is the marked one. So if the addresser wants to achieve an effect of "current relevance", he may choose the marked PP rather than the default SP to trigger some extra inferences. As a result, in English, the PP is the loser of the competition, since there are fewer restrictions on the use of the "stronger" form, the SP.

b. Register Analysis Framework

As for why language used in different registers vary, some analyses have been made, one

of the most representative and applicable frameworks of which are Biber's.

Biber and Conrad (2009) put forward a more specific analytical framework, involving two factors generally, i.e., linguistic features and situational features. To analyse a register, first of all, the situational features, which determines the register as it is, need to be identified, including the participants, their relations, the communication channel, the setting, etc. Then, typical and dominant linguistic features are picked out using methods like comparative approach, quantitative analysis, and sample analysis. The last step is the key step, which explains the spotted facts by associating the linguistic features with situational features.

3. Research Methods

3.1 Research Materials

The corpus adopted in this article is the Corpus of Contemporary American English (COCA), which is the largest and the most balanced corpus of American English. This corpus consists of texts with 450 million words in total, which are equally divided into five registers: spoken, fiction, magazines, newspapers, and academic texts. 20 million words per each year from 1990 to 2012 were collected to compose it. Moreover, it is updated regularly, with the most recent updating time in summer, 2012. Although COCA only collects the data of the recent 20 years, our preliminary studies show significant changes of the PP and the SP in different registers can be detected. In addition, the data are most suitable to explore recent developments. Another corpus of American English, the Corpus of Historical American English (COHA), does not include an important register, the Spoken register, despite the fact that it is composed of data spanning almost 200 years.

Therefore, COCA is chosen as the research material in this study considering its size, balance, representativeness, and currency.

3.2 Research Procedures

3.2.1 Data Retrieval

First of all, the frequencies of the PP and SP from 1990 to 2012 in five registers need to be retrieved. Thanks to the ingenious design of the COCA online retrieval system, the frequency in each year in each register can be obtained at the same time in each retrieval.

According to the annotation syntax of the COCA, the query syntax for the PP, which are *[have]/[has] [v?n*]* and *[have]/[has]*[v?n*]* (in case of such structures as *have/has not/already done*), are used to retrieve the PP.

To retrieve the frequencies of the SP, the query syntax *[v?d*]* is first used. Then *had [v?n*]* and *had *[v?n*]* are used to exclude the past perfect.

One point to note here is that we use the frequency-per-million to make the results comparable considering different sizes in different years. All the following statistical results are based on the data retrieved (see Appendices).

3.2.2 Data Analysis

Data obtained will be calculated and, more importantly, standardized and then preliminarily analyzed with Excel, to get a direct impression of the distribution patterns and development trends of the PP and the SP in relation to each other in registers. In particular, some graphs based on the data are drawn.

In order to present the alternation patterns in a more statistical way and make the results more convincing, linear regression analysis and correlation analysis from SPSS (19.0) will be conducted. Regression analysis is a statistical technique that can relate the dependent variable to one or more independent variables. Linear regression analysis assumes the development? model of the dependent variable (the frequency of the PP or SP in our study, for example) is linear on the basis of / in relation to the independent variable (the year in our study). Since there is only one independent variable, the simple linear regression model is adopted:

$$y = b_0 + b_1 x + \epsilon;$$

y and *x* refer to the dependent and independent variables; β_0 refers to the intercept, and β_1 refers to the slope, i.e. the change speed of *y* in relation to *x*. Whether β_1 is positive or negative can show the PP or SP increases or decreases along the years. At the same time, we can get *p*-value of β_1 to see whether the change is significant or not. And if the value of *p* is less than 0.05, the change is significant. R^2 (from 0 to 1) can demonstrate how the regression models obtained conforms to the data retrieved (Coakes, 2013).

In our study, regression analysis can reveal the change trends of the PP and the SP across registers, and whether the changes are significant. Moreover, the trend charts based on the regression analysis can provide a visual insight into the relation between the changes of the PP and the SP in a specific register.

In addition, interpretation will be done from the perspectives of some previous studies.

4. Results and Discussion

4.1 General Distribution Patterns of the PP/SP in Different Registers

4.1.1 General Distribution Patterns of the PP in Different Registers

General distribution patterns of the PP across registers are demonstrated in the following Figure 4.1 (data from Appendix II).

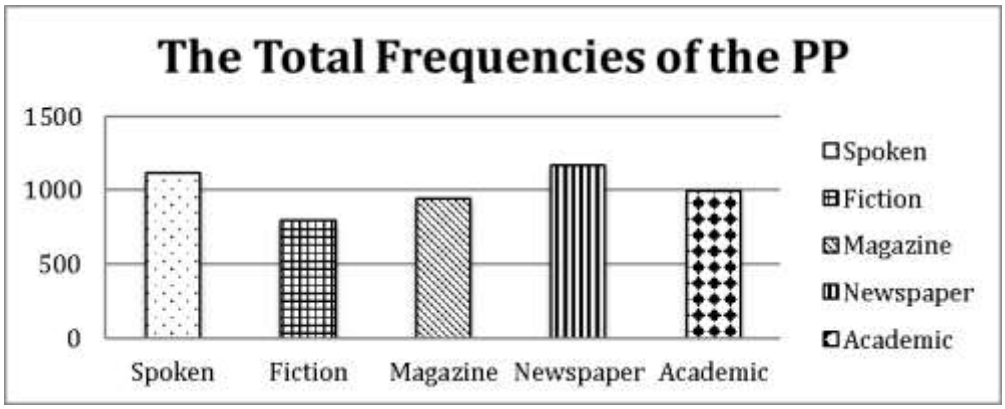


Figure 4.1 General distribution patterns of the PP across registers

The frequencies of the PP in the Spoken, Newspaper, and Academic register are higher than those in the other two registers. This is in line with Biber's (1999) findings that the large majority of the verb phrases with the perfect aspect in conversation, news reportage, and academic prose are in the present tense, while fiction shows the opposite preference for the past tense, because the PP is used to report events or states occurring at an earlier time, but still showing some current relevance. And in conversation, the speakers tend to focus on the

present time. Moreover, in conversations, the simpler form, like the SP, is usually preferred to communicate faster (Biber, 2009). Thomson and Martinet (1980: 157–8) state that the PP is often used in newspapers and broadcasts to introduce an action which takes place in the past, by which a pragmatic emphasis on the current relevance of the action can be achieved (Rastall, 1999). The frequent use of the PP in news reportage is often referred to as "hot news perfect" (Portner, 2003). Yao & Collins (2012) also point out that the relatively high frequency of the PP in Newspaper is most likely the reflection of the use of the hot news perfect to report events of the recent past to show newsworthiness. In academic prose, the PP is often used to show the continuing relevance of a past finding, because "the truth conditions are often not constrained to a particular time frame" (Yao & Collins, 2012).

4.1.2 General Distribution Patterns of the SP in Different Registers

General distribution patterns of the SP across five registers are shown in Figure 4.2 (data from Appendix II).

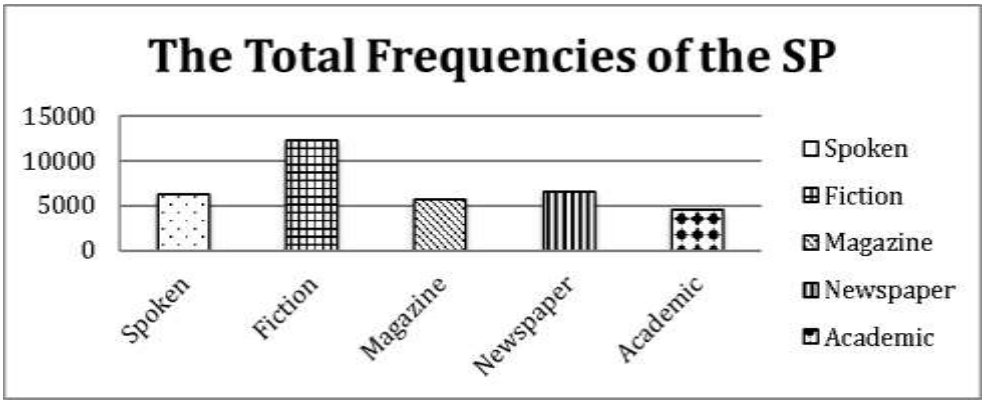


Figure 4.2 General distribution patterns of the SP across registers

It is self-evident that the frequency of the SP in the Fiction register is far higher than those in the other four registers, which corresponds to Biber's (1999) findings that fiction pieces show a strong preference for past tense verbs. Vabalienė et al. (2009) also find that the most frequent tense used in the texts of fiction is the simple past tense. Although the Fiction register has many situational characteristics in common with other written registers, including having enough time for planning and revision, being written for a large and general audience, involving no interaction between the author and the readers, etc., "these situational characteristics have almost no influence on the linguistic characteristics of a fictional text" (Biber & Conrad, 2009: 132). One important parameter of variation among fiction pieces is "whether the story is told as a narration of past events, or as a description of events as they occur at the time of telling" (Biber & Conrad, 2009: 138). The more common style is to tell story as events taking place in the past (Biber & Conrad, 2009: 138). This explains perfectly why the SP in the Fiction register is obviously used much often than in other written registers. As Biber and Conrad (2009:139) conclude, the linguistic features of the Fiction register are mostly influenced by style choices rather than the normal situational features, which determine register variations.

4.1.3 Relative Distribution Patterns of the PP and the SP in Different Registers

When retrieving data from the corpus of the frequencies of the PP and SP, we find that

despite the fact that there are large gaps between the frequencies of the PP or SP in some registers, one point in common is that the frequency of the SP is always much higher than that of the PP in any register, as well as in the whole corpus. This is actually within our expectation, as Schaden (2008; 2009) assumes that the SP is the default form, compared with the PP, which is the marked SP. The specific ratio of the SP to the PP (SP/PP) in each register as well as the whole corpus is shown in Figure 4.3 (data from Appendix II).

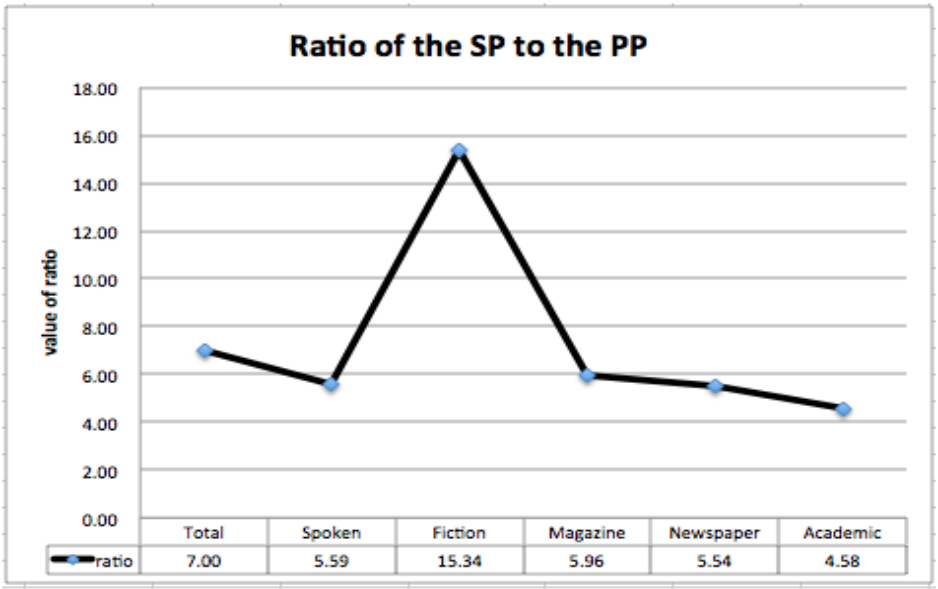


Figure 4.3 General distributions of the PP and the SP across registers

Most ratios are almost at the same level, except for one salient point, namely that of the Fiction register, which is the peak. This actually corresponds to the aforementioned finding that the frequency of the SP is particularly higher compared with other registers because of its common style / feature? of narrating stories that occur in the past (Biber & Conrad, 2009: 139).

4.2 Relationship between the Development of the PP and the SP across Registers

4.2.1 Relationship between the PP and the SP in the Corpus

The relative development trends of the PP and SP are shown in Figure 4.4 (data from Appendix II).

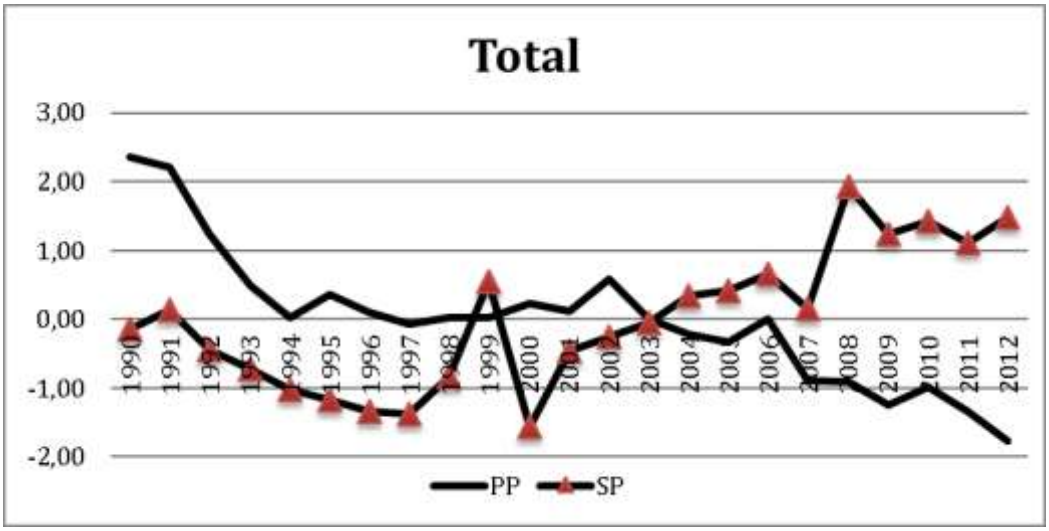


Figure 4.4 Relative development trends of the PP and the SP in COCA

Despite fluctuations at some point, it can be clearly observed that the PP has been decreasing generally over the two decades, while the SP has been increasing generally. Similar findings have been achieved by some researchers previously who claim that the SP has been increasing at the expense of the PP (See Eleness, 1997; Yao & Collins, 2012).

To see whether the decrease of the PP and the increase of the SP are significant along the years, linear regression analysis is conducted. The fitting model is presented in Figure 4.5 (data from Appendix II).

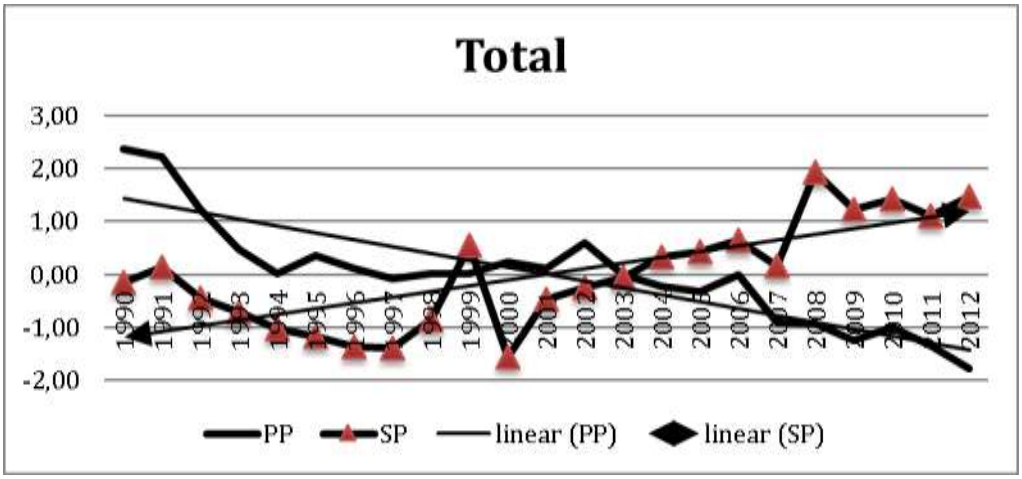


Figure 4.5 Total: linear regression model

Table 4.1
Results (Total) from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Total (PP)	5384.17	-32.36	< 0.05
Total (SP)	34017.45	89.93	< 0.05

According to Table 4.1, the result $p < 0.05$ shows that the frequency of the PP is significantly linearly related to the time, and so it is with the SP. Therefore it can be concluded that the PP has indeed been decreasing significantly ($\beta_1 = -32.96$, $p < 0.05$) over the

years, and the SP has been increasing significantly ($\beta_1 = 89.93$, $p < 0.05$).

However, it cannot be concluded safely that the PP has been losing ground to the SP just because the PP has been decreasing and the SP has been increasing. So, a necessary statistical method is applied to verify whether the above conclusion is true. Correlation analysis proves a significant relationship between the increase of the SP and the decrease of the PP ($R = -0.51$, $p = 0.014$). This result confirms Yao and Collins's (2012) claim that the PP has been constantly losing ground to the SP.

Our finding, which concerned the development trends of the PP and the SP in the whole corpus data, echoes the competition approach, and thus adds more evidence to the fact that the PP and the SP are in competition. Many studies have verified Schaden's competition approach to the distribution of the PP and the SP – like that of Yao and Collins (2012), which finds arrives at the same conclusion.

However, when studied in different registers, the competition phenomenon proposed by Schaden does not always apply, but only to the Spoken register. In other words, in the other four written registers, , the PP and the SP are not in competition, but show various relation patterns.

4.2.2 Relationship between the Development of the PP and the SP in Spoken Register

The relative development trends of the PP and the SP in the Spoken register do not seem as explicit as those in the whole corpus, as shown in Figure 4.6(data from Appendix II).

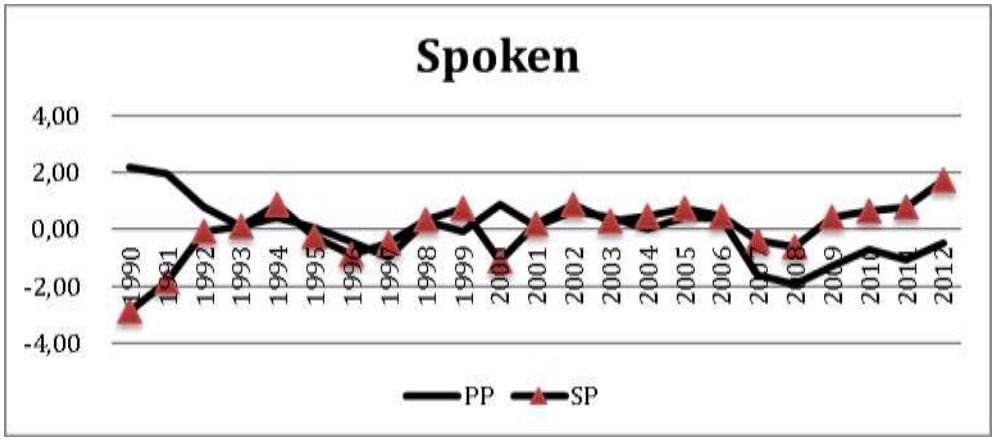


Figure 4.6 Relative development trends of the PP and SP in Spoken

There are many ups and downs in the middle, which may be a good reflection of the competition between the PP and SP. But generally, similar to the total, the PP has decreased and the SP has increased.

Linear regression analysis confirms the above conclusion, as shown in Figure 4.7(data from Appendix II).

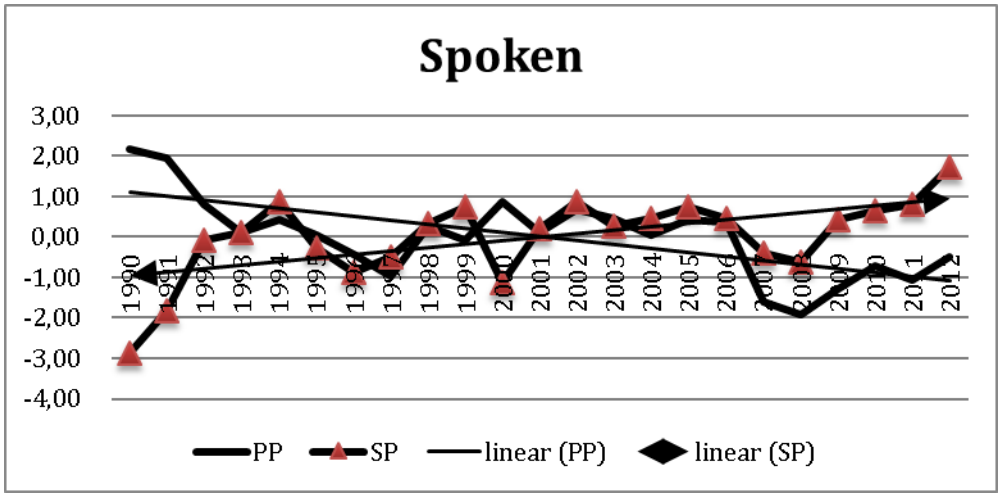


Figure 4.7 Spoken: linear regression model

Table 4. 2
Results (Spoken) from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Spoken (PP)	1245.14	-11.66	0.00041
Spoken (SP)	5831.22	30.82	0.00321

According to Table 4.2, the result p =.00041, which is less than the expected p value, shows that the frequency of the PP is significantly linearly related to the time, and so it is with the SP. So the decrease of the PP (β_1 = -11.66, $p < 0.05$), and the increase of the SP (β_1 = 30.82, $p < 0.05$) are both significant. As Yao and Collins (2012) point out, the diachronic decrease of the PP in the Spoken Register reflects well the phenomenon that the PP occurs more frequently in formal written registers than in informal spoken registers.

Correlation analysis also reveals a significant relationship between the decrease of the PP and the increase of the SP (R = -0.42, p = 0.047). That is to say, in Spoken register, the PP has been replaced in some contexts by the SP.

According to Biber & Conrad (2009), some of the main situational features of the spoken and written registers are summarized in Table 4.3; this may lead to their distinctive relation patterns of the PP and the SP.

Table 4.3

Different situational features of spoken & written registers

Situational Features	**Spoken**	**Written**
I. Participants	people with all kinds of backgrounds	usually people with professional backgrounds, like writers, reporters, etc.

II. Relations among participants	usually interact directly, and all participants have some shared knowledge	no interaction and the writer has more knowledge
III. Channel	speaking	writing
IV. Production circumstances	texts are usually produced in real time, and cannot be revised or edited	texts are carefully planned and edited
V. Setting	participants usually share the same physical space	unknown

From Table 4.3, it can be concluded that compared with written registers, the spoken register has some distinctive situational features that can well explain why it is the only register in which the PP and the SP are in competition. According to Schaden (2009), the competition of the PP and the SP is a matter of the choice between the marked form, the PP, and the default form, the SP, in a certain context. The use of the marked form usually trigger a pragmatic reasoning process, causing extra pragmatic inferences. Participants in a conversation are usually people from all kinds of backgrounds, who may not pay as much attention to the underlying pragmatic inferences as those from a professional background – like fiction writers. As a result, the default form, the SP, will be more often adopted. In addition, the participants usually share some knowledge and often the same physical space, which may also reduce the necessity of extra pragmatic inferences. Moreover, in the spoken register, the texts are produced in the real time and cannot be revised and edited. Therefore, participants usually have no time to consider whether extra pragmatic inferences are necessary, and choose the default form directly without much deliberation. For example, in some contexts with some indefinite temporal adverbials, where the PP are usually used, speakers choose the SP, as shown in the following sample sentences in Table 4.4 (from COCA).

Table 4.4
Samples of using of the SP over the PP

1 ***He stepped out. He never was pro-choice***. He was always anti-abortion.
2 I ***was upset because he was already*** getting a reputation as being a rebel.
3 I love ***them and how I never wanted this***. *I never* wanted this. I never wanted to leave.
4 [Voiceover] *And I* ***saw a young boy who had no one***, never had no one.
5 The ship approached Port-au-Prince, ***Constant already had 40 of his FRAPH.***
6 Reporters are now not allowed to circulate. *We* ***were kicked out very recently***.
7 *I mean,* ***I just turned around, I never looked at again***.
8 Mayor NAIRN: We – ***we saved it. We never was out of phone system.***
9 We got it down to $30 billion, *but* ***it wasn't good enough*** *yet because* there's still room.
10 BILL-LAGATTUTA: *Connie,* ***this crime was never prosecuted***, never ever.
11 That was killed in the war ***before I was born and he never got to see me.***
12 Mr-MORTON: ***There was one occasion relatively recently*** where no people agreed.

13	[Voiceover] ***He went back to Baltimore*** **recently** to host and play in parties.
14	Rep-ROSTENKOWSKI: ***It – it was – it was already used in*** – in a larger office.
15	Ms-BORGER: What about North Korea? ***You caused quite a stir recently*****.**

The marked form, the PP, is sort of "ignored" in speaking, and therefore loses its ground to the default form, the SP. Miller (2000) also points out that the PP is more vulnerable to the competition with the SP in the spoken register. In a word, because of the inner situational features of the spoken register, the SP has been intruding into the territory of the PP there.

.

4.2.3 Relationship between the Development of the PP and the SP in Fiction

In the Fiction register, the relative development trends of the PP and SP present a different pattern from those of the whole corpus and the Spoken register, as shown in Figure 4.8 (data from Appendix II).

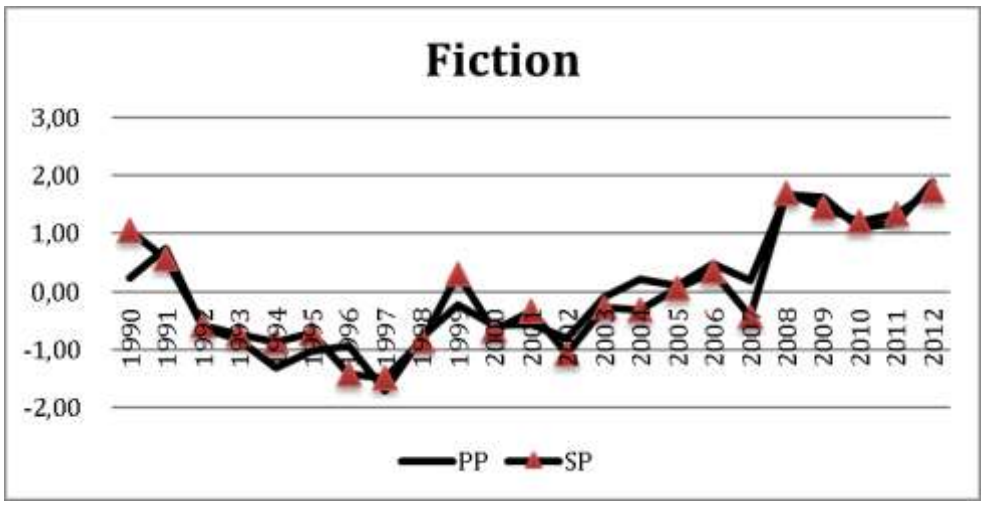


Figure 4.8 Relative development trends of the PP and SP in Fiction

It can be directly observed that the development trends of the PP and the SP are almost synchronous all the time, decreasing and increasing at the same time. Generally, the frequencies of the PP and the SP have increased in the past two decades.

Results of linear regression analysis are shown in Figure 4.9 (data from Appendix II).

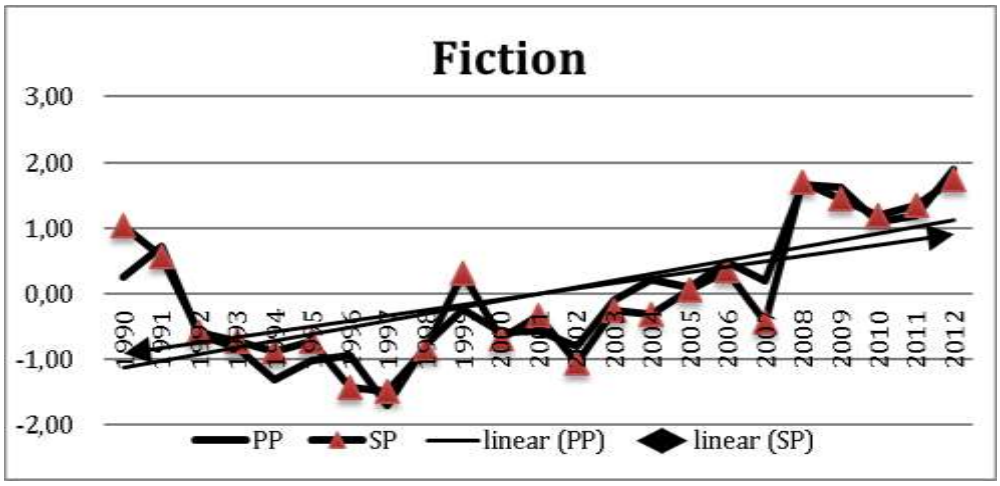


Figure 4.9 Fiction: linear regression model

Table 4.5

Results (Fiction) from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Fiction (PP)	710.36	7.36	0.00027
Fiction (SP)	11140.51	92.87	0.0054

According to Table 4.5, the result $p = 0.00027$, less than 0.05, shows that the frequency of the PP is significantly linearly related to the time, and it is the same with the SP. Therefore, regression analysis verifies the above assumption, proving that both the PP ($\beta_1 = 7.36$, $p < 0.05$) and the SP ($\beta_1 = 92.87$, $p < 0.05$) in Fiction have increased significantly.

Despite the synchronous increase, correlation analysis shows that the development trends of the PP and SP are significantly positively related ($R = 0.94$, $p < 0.05$), which implies that there must be some factors influencing their development at the same time. These factors will be elaborated in the following.

Fiction has many common situational characteristics with other written registers. It is produced by an author who has a sufficient time to plan, revise, and edit the texts. Normally no interaction is involved between authors and readers. Readers usually have no idea about when and where the texts are written. Also like newspaper reports, fiction is usually written for a general audience who shares little common knowledge with the author. However, why does the Fiction register show distinctive development patterns between the PP and the SP? According to Biber and Conrad (2009:132-138), fiction is one of the most complicated registers from a situational perspective because the above-mentioned external situational features exert almost no influence on the linguistic features of texts. It is almost of no significance whether the reader and the author interact, whether they know each other, etc., because what really matters is the situational contexts in the fictional world that is described. Therefore, the factor that determines linguistic features of fictional texts is how the fictional world is constructed, not what the "real-world" situational context is. In other words, it is mainly the style of a fictional text that determines its linguistic features.

The most common style of fiction is to narrate events. In narration, what is most important is to illustrate explicitly the temporal relations among different events, and tenses are critical to such illustrations. Therefore, fiction is the register that is richest in verbs and tenses, including the PP and the SP (Yang & Huang, 2013), as shown in Figure 4.10 (data from Appendix II).

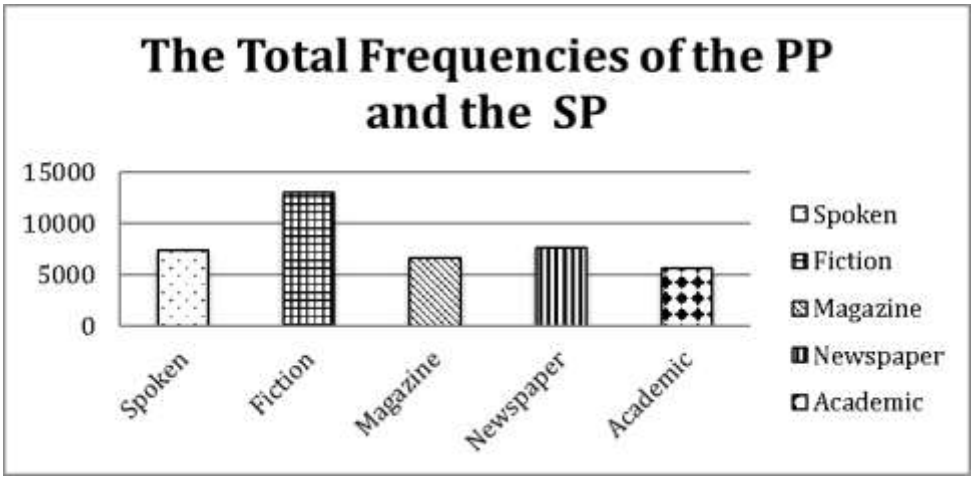


Figure 4.10 The total frequencies of the PP and the SP

According to Bybee (2003: 603–618), frequency may not be the result of language change, but the trigger of it. It is also claimed that certain constructions can be more and more strongly favoured or disfavoured by some genres (Biber & Finegan, 1989). Therefore, the high repetition of the PP and the SP can strengthen fiction's stylistic preference for the two constructions.

Also, according to Biber & Conrad (2009: 144–160), there are some stylistic drifts that may have contributed to the almost synchronous increase of the PP and the SP. Compared with previous novels, which show a strong preference for long sentences and complex noun phrases, modern novels become increasingly dependent on a "simpler style with more verbs, short clauses and adverbials" (Biber & Conrad, 2009: 155). The increase of the use of verbs is very likely a significant factor leading to the increase of the PP and the SP.

In short, the increase of both the PP and SP is very likely the result of the strengthening of fictional preference for narrative tenses and its style-determined increase in the use of verbs.

4.2.4 Relationship between the Development of the PP and the SP in the Magazine Register

Similar to the Fiction register, the PP and the SP in the Magazine register also show almost the same development trend, as can be observed in Figure 4.11 (data from Appendix II). But interestingly, in contrast to Fiction, they have decreased rather than increased.

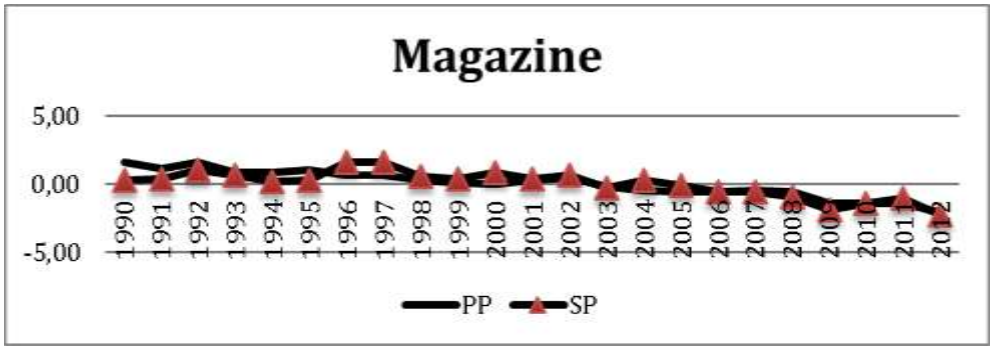


Figure 4.11 Relative development trends of the PP and SP in Magazine

The results of linear regression analysis are shown in Figure 4.12 (data from Appendix II).

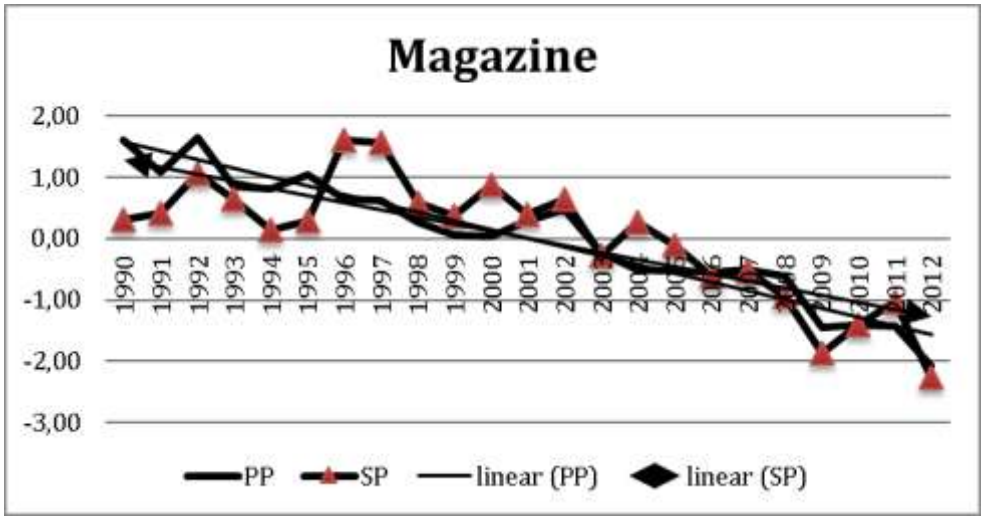


Figure 4.12 Magazine: linear regression model

Table 4.6
Results (Magazine) from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Magazine (PP)	1121.08	-15.43	< 0.05
Magazine (SP)	6081.52	-40.32	< 0.05

From Table 4.6, it can concluded that the result p < 1e-5 shows that the frequency of the PP is significantly linearly related to the time, and so it is with the SP. Results show that both the PP (β_1 = -15.43, p < 0.05) and the SP (β_1 = -40.32, p < 0.05) in Magazine have decreased significantly.

Correlation analysis also shows a significant positive relationship between the decrease of the PP and the SP (R = 0.83, p < 0.05)**.**

It can be reasonably assumed that there must be certain factors that are responsible for the almost synchronous decrease of the PP and the SP. Since Magazine is a register that emphasizes timeliness, which is usually embedded in the simple present tense, so it is hypothesized that the decrease of the PP and the SP is partly due to the increase of the simple present tense. To test this hypothesis, first we retrieve the frequencies of the simple present tense from 1990 to 2012 and conduct a statistical analysis on the data. The linear regression analysis shows that the frequency of the simple present in the Magazine register does increase significantly, in contrast to that of the PP and the SP, as shown in the following Table 4.7 (data from Appendix III).

Table 4.7
Results (Magazine) for the simple present from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Magazine	4360.44	85.442	0.05

The result p < 1e-5 shows that the frequency of the simple present is significantly linearly related to the time, which shows that the simple present tense in Magazine has increased significantly (β_1 = 85.44, p < 0.05).

Moreover, further correlation analysis proves a significantly negative relation between the increase of the simple present and the decrease of the PP (R = -0.89, $p < 0.05$), and the decrease of the SP (R = 0.60, p = 0.002). Therefore, it is confirmed that the increase of the simple present contributes a lot to the decrease of the PP and the SP. Since Magazine puts much emphasis on timeliness, in some contexts where the SP or the PP should be used, the simple present is selected instead, as shown in the samples in Table 4.8.

Table 4.8

Samples (Magazine) of using simple present over the PP and the SP(data from Appendix III)

1 ***Graf claims she is under too much pressure already.*** "Look, " she says.
2 **He says. "** ***I'm alive and already getting around*** pretty well. I'm happy the way I am.
3 C***ultural change is already evident*** in Montana, in towns like it.
4 Expenses such as editorial, **layout*, and design are already paid*.**
5 "The old blood ***pressure is a little high this morning***, my friend."
6 Winds are usually **calmest then*, but this morning the chill wind blasts*** through my body.
7 to a helper digging nearby, "***What are you up to this morning*?**" I ask. # "Rotating
8 We have been assembled ***for his visit this morning evince a cheer*.**
9 ***The three objects form a pretty group on this morning.***
10 ***I am fasting this morning while*** I read a prayer book that my late Uncle Hy gave.
11 she smiled at me gently ***and inquired, "How are you this morning?"***
12 ***Yet this morning the ballpark is*** relatively quite, like a slumbering.
13 ***But no warning is given this morning.***
14 ***"You seem a little down this evening.*** What's bothering you?"
15 ***I have this misfortune this evening of*** sitting with the city's pena.
16 ***So far this evening she is up*** fifty bucks.
17 Williard McCloud and Todd ***Campbell-added, "This evening there are*** 36 men.

4.2.5 Relationship Between the Development of the PP and the SP in Newspaper

In the Newspaper register, there are some fluctuations and intersections in the process of the development of the PP and the SP, but generally, they both seem to have decreased, as shown in Figure 4.13 (data from Appendix II).

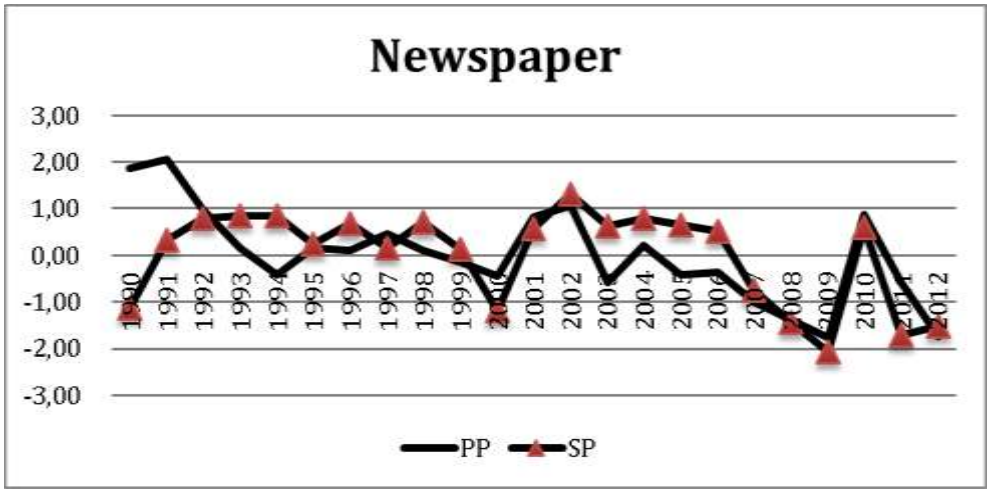


Figure 4.13 Relative development trends of the PP and SP in Newspaper

Linear regression analysis shows they have, in general, both decreased, as can be seen in Figure 4.14 (data from Appendix II).

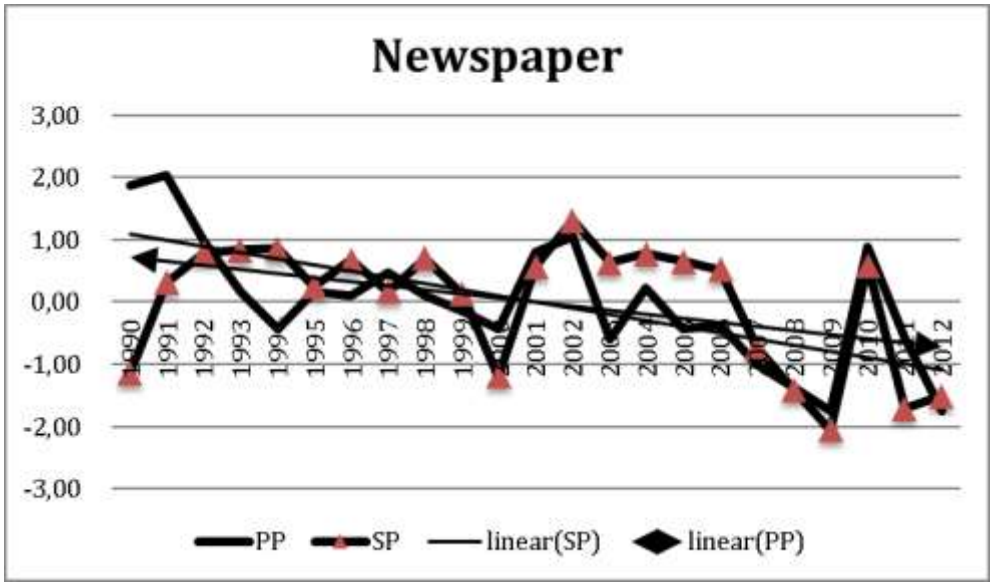


Figure 4.14 Newspaper: linear regression model

Table 4.9
Results (Newspaper) from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Newspaper (PP)	1223.68	-4.74	0.000393
Newspaper (SP)	6629.75	-12.95	0.0338

From Table 4.9, p = 0.0004, less than 0.05, shows that the frequency of the PP is significantly linearly related to the time, and it is also true of the SP. It can be concluded that both the PP (β_1 = -4.74, $p < 0.05$), and the SP (β_1 = -12.95, $p < 0.05$) in Newspaper have decreased significantly.

Correlation analysis also shows a significant positive relationship between the decreases

of the PP and the SP (R = 0.52, p < 0.05), implying that there must be some factors influencing their development at the same time, as explained in the following.

Similar to the Magazine register, the Newspaper register presents a positive relationship between the diachronic change of the PP and the SP. Since Newspaper places even more emphasis on newsworthiness, in news discourse, it is conventional to use the present tense to describe both present events and events which occur in the past. As Yao (2016) points out, some linguistic contexts in news that intend to communicate with the simple present also allow the present perfect and the past tense, but the validity of the information communicated would be much weaker with these verb forms. Consequently, it is reasonable to assume that it is also the increase of the simple present that contributes to a great degree to the decrease of the PP and the SP. After retrieving the frequency of the simple present, a linear regression analysis is conducted, which shows that the proportion of the simple present in the Newspaper register has indeed increased significantly along the decrease of the PP and the SP, as shown in Table 4.10 (data from Appendix III):

Table 4.10

Results (Newspaper) for the simple present from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Magazine	3024.06	20.89	0.001

The result p = 0.001, less than 0.05, shows that the frequency of the simple present is significantly linearly related to the time. It can be concluded that the simple present tense in Magazine (Newspaper?) has increased significantly over the years (β_1 = 20.89, p < 0.05).

What's more important, correlation analysis confirms the increase of the simple present is significantly negatively related to the decrease of the PP (R = -0.71, p < 0.05) and the SP (R = -0.52, p = 0.01) by conducting linear regression analysis respectively. Similar to the Magazine register, in some contexts where the PP and the SP should have been used, the simple present is adopted to emphasize the newsworthiness, as can be observed from the sentence samples in Table 4.11 (from COCA).

Table 4.11

Samples (Newspaper) of using the simple present over the PP and the SP

1	***The event already has its own*** T-shirt." Harding-Kerrigan,"
2	You don't have to push. ***You already have a seat.*** You already have your cake.
3	***Lists of potential candidates already exist*** for certain jobs,
4	"We're talking about ***taking plutonium that is already*** **a** weapons concern.
5	***What we see on our television screens this morning*** are tentative beginning of the day.
6	During the hour we kept ***checking and are still there*** this morning.
7	Our new ***video phones are a little slow this morning.***
8	The ***experts sit down this morning to a second*** day of meeting to review findings.
9	***What counts is standing with her at the bus stop this morning.***
10	***When I walk into the kitchen this morning at 7:45,*** there is no activity going on.
11	"Or is it ***Saturday's?" he's all confused this*** morning.
12	***This morning I walk through a new industrial*** park to
13	***Brenda is at the county jail this evening,*** visiting her son.
14	***It isn't too busy yet this evening*** for the little families and high school kids
15	***This evening my cooking begins*** lightly enough. I slice symmetric circles.

Therefore, it can be reasonably assumed that the synchronous decrease of the PP and the SP in the Newspaper register is to a great degree caused by the increase of the use of the simple present tense due to its need for newsworthiness.

3.2.6 Relationship between the Development of the PP and the SP in Academic

In the Academic register, the development trends seem different from all previous registers. The frequency of both the PP, and the SP keep relatively stable for most of the time; then, they suddenly increase, but decrease again, as shown in Figure 4.15 (data from Appendix II):

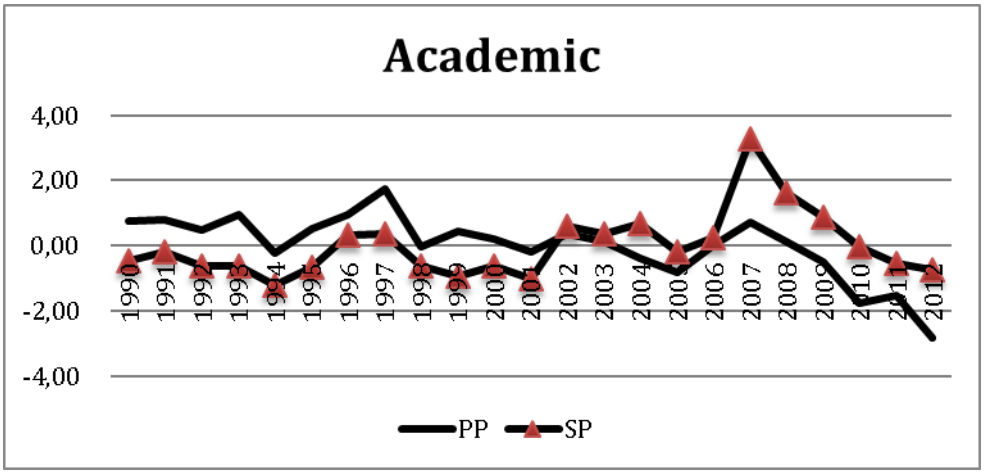


Figure 4.15 Relative development trends of the PP and SP in Academic

The results of regression analysis are presented in Figure 4.16 (data from Appendix II).

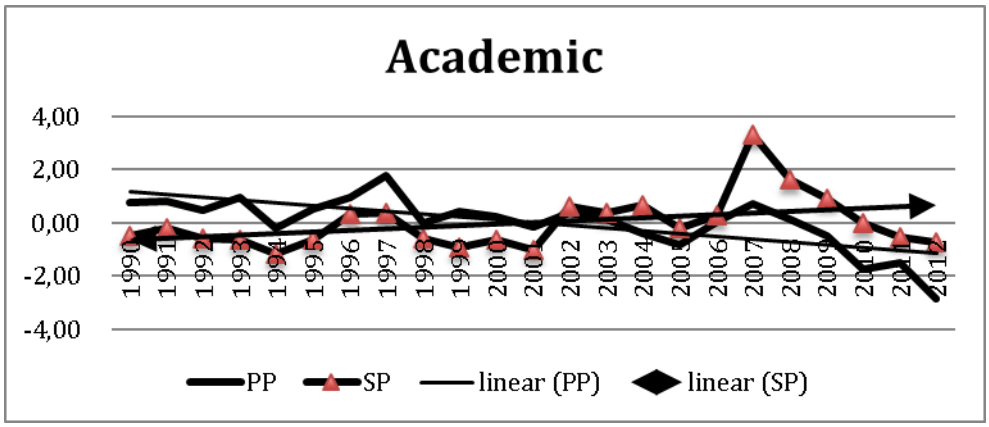


Figure 4.16 Academic: linear regression model

Table 4.12
Results (Academic) from linear regression analysis

	Intercept β_0	**β_1**	**P-value (β_1)**
Academic (PP)	1083.92	-7.89	0.000114
Academic (SP)	4334.30	19.46	0.0587

From Table 3.6, the result $p = 0.0001$ shows that the frequency of the PP is significantly linearly related to the time. However, for the SP, $p = 0.0587$, is greater than 0.05, it does not change significantly over the years. Results from linear regression analysis of the whole dataset in Academic show that generally the PP has decreased significantly ($\beta_1 = -7.89$, $p < 0.05$), but the SP does not show a significant change ($\beta_1 = 19.46$, $p > 0.05$).

Correlation analysis shows no significant relationship between the development of the PP and the SP ($R = 0.20$, $p = 0.36$).

The Academic register is the most distinctive one among the four written registers. The development patterns of the PP and the SP are not as significantly related as in the other registers. The PP has decreased significantly, while the SP has remained relatively stable. This is not surprising, as Academic is relatively the most stable register of all in language drifts (Biber & Gray, 2009: 126). The decrease of the PP may be attributed to the change of the emphasis on different sections of research articles. Usually, a research article is composed of the following sections / parts : Introduction, Methods, Results, Discussion, and Conclusion, with each section playing a different role, especially through the dominant verb tenses in them. These are summarized in Table 4.13, according to Biber & Gray (2009: 129) and Waard & Maat (2012).

Table 4.13
Main section with typical functions and verb tenses

Section	Function	Verb tenses
Introduction	describe research background and goal	past tense and present perfect
Methods	describe procedures of data collection, experiments, analysis	simple present
Results	present the findings	simple present
Discussion	interpret the findings by relating them to previous finding and theories	simple present and present perfect
Conclusion	summarize the results and present the implications	simple present

According to Li & Ge (2009), the present perfect in research articles is mainly for the purpose of current relevance and immediacy, especially in the sections of Introduction and Discussion. However, the use of the PP has undergone some changes. In presenting previous related researches, more researchers tend to use the simple past for the sake of humbleness since " science is a collection of hypotheses and it is not a field of certainty" (Li & Ge, 2009). When reporting and interpreting the results, the simple present is increasingly preferred by researchers to enhance the generality of their findings (See Table 4.14). Individually, the increase of the SP or the simple present is not significant enough to contribute to the decrease of the PP, but put together, they can cause the change.

Table 4.14
Samples of using the SP and the simple present over the PP

SP to present previous studies
1 Then in May 1993, ***Holliday presented a resolution that said, in part.***"
2 ***His attitude, which was shared by some other Vista teachers,*** was that you were wrong.
3 ***Prentice stated that creation and evolution*** are both beliefs and that both can be created.
4 ***He said the book was biased on the grounds*** it offended a religious point of view.
5 **She *argued that religion is*** an inappropriate criterion for the selection of science book.
6 The public school biology teacher ***sought to invalidate the state law that*** prohibited all.
7 Walker ***asserted that most species*** are superfluous, more like passengers than like drivers.
8 Tilan ***also saw evidence that there*** may be a threshold beyond which more
Simple present to present findings
1 ***But many desirable species remain elusive***.
2 *Sharks* ***do not have air bladders to give them buoyancy yet***.
3 The full microbial flora ***that humans carry, which outnumber even own cells*** **already.**
4 Therefore, ***house mice show very high-amplitude population*** fluctuations recently.
5 Avian ***predators move more often and over longer distances*** when vole population
6 Therefore, ***population models indicate that change*** in reproductive output **do not decrease** in recent years.
7 ***We identify unprotected areas of the world that*** have remarkably high population.
8 Podcasts, videos and ***other resource enhance the lifelong learning experience already.***

So the above change of researchers' preference for another tense to the PP may be an important contributor to the decrease of the PP in Academic.

5. Conclusion

The present study carries out a systematic investigation of the relationship patterns of the development of the PP and the SP in contemporary American English, as well as in five different registers (Spoken, Fiction, Magazines, Newspapers, and Academic) and attempts to explain various patterns according to the specific characteristics of each register.

Firstly, general distribution patterns of the PP and the SP across registers are detected. It is found that the frequencies of the PP in the Spoken, Newspaper, and Academic register are higher than those in the other two registers, due to their focus on the present time and current relevance (Biber, 1999; Rastall, 1999). As for the SP, Fiction possesses the highest frequency due to its distinct stylistic feature of narrating past stories. As an echo to the above findings, it is found that the ratio of the SP to the PP in Fiction is much higher than those in the other four registers.

Secondly, relation patterns of the development of the PP and the SP are explored using correlation analysis and linear regression analysis. Results show that the PP and the SP are in competition in the whole corpus data, developing in opposite directions, as some researchers claimed (Schaden, 2009; Yao & Collins, 2012). However, following Biber (2011), who claimed that language development could present different patterns in different registers, we conduct a specific analysis of the relationship patterns across registers. Results echo Biber's claim that general language findings do not always apply to specific registers. It is found that the general competition pattern between the PP and the SP does not apply to each register. In fact, the competition pattern can only be observed in the Spoken register. In the other four ones, which are all written ones, the relation patterns between the PP and the SP vary. In the

Fiction register, the PP and the SP are positively related, increasing almost synchronously. On the other hand, in the Magazine and the Newspaper registers, the PP and the SP are also positively related, but they both decrease significantly. In the Academic register, the PP and the SP are not significantly related. To be more specific, the PP decreases significantly, while the SP does not show a significant change.

Thirdly, interpretations of the relationship as well as individual development patterns in each register are attempted, on the basis of their own characteristics, whether situational or stylistic ones. The distinct competition pattern detected in the Spoken register may be the result of its fewer necessities for pragmatic references usually carried by the marked form, the PP (Biber & Gray, 2009; Schaden, 2009). As for the Fiction register, its stylistic feature of narrating determines that it calls for more verbs and verb inflexions than the other registers (Yang & Huang, 2013). This feature is enhanced in the course of time, causing the increase of both the PP and the SP (Bybee, 2003; Biber and Finegan, 1989). The increase of the PP and the SP in Magazine and Newspaper is very possibly the result of the increase of the simple present tense, as shown through correlation analysis; this is due to the fact that the two registers place a great emphasis on timeliness. Academic register shows a significant decrease of the PP only, but no significant change of the SP, which is likely due to researchers' preference for other tenses over the PP to introduce the background and discuss their own findings and implications (Li & Ge, 2009).

By solving the above problems, a comprehensive picture of the relationship between the development of the PP and the SP, two closely related constructions, is revealed. The present paper fills the gaps in the study of the relationship between the PP and the SP by adopting correlation analysis to test their development relationship statistically, and also introducing the dimension of register. It confirms the competition between the PP and the SP, but also shows that register is an indicator of variations of language development.

The present work helps to refine our understanding of recent and ongoing grammatical change in American English, fill some gaps, and correct some misperceptions in studies on English grammatical change. Moreover, focusing on different registers can direct us to some interesting and important language change patterns that have not even been detected before. The development of the PP and the SP is merely an "iceberg" of the evolution of the English language. However, as they are two extremely important constructions in the English tense and aspect system, the results of their development can be seen as an indicator of the development of the English language. Language changes, especially those of a narrow range, are not necessarily a result of grammaticalization, but they may perhaps be due to stylistic or social influences, which can be verified in the present study. The article thus introduces more possibilities into the studies on language development.

REFERENCES

Aarts, B. & Bowie, J. (2012). Change in the English Infinitival Perfect Construction. In: *The Oxford Handbook of the History of English*. London: Oxford University Press, 200–210.

Aarts, B., Close, J., Leech, G. et al. (2013). *The Verb Phrase in English: Investigating Recent Language Change with Corpora.* New York: Cambridge University Press.

Bao, C., Zhang, X., Qu, Y., & Feng, Z. (2018). American English perfect construction across registers. *Journal of Quantitative Linguistics*, 1, 1–28.

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

Biber, D. (2011). Register as a Predictor of Linguistic Variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Biber D. & Finegan E. (1989). Drift and the Evolution of English Style: A History of Three Genres. *Language*, 65(3): 487–517.

Biber, D & Gray, B. (2013). Being Specific About Historical Change. *Journal of English Linguistics*, 41(2), 104–134.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Grammar of Spoken and Written English. Harlow: Longman.

Burchfield, R., W. (1996). *The New Fowler's Modern English Usage*. London: Oxford University Press.

Bybee, & Joan. (2008). *Mechanisms of Change in Grammaticization: The Role of Frequency*. The Handbook of Historical Linguistics. New Jersey: Blackwell Publishing Ltd.

Coakes, S. J. (2013). *SPSS Version 20.0 for Windows*. New Jersey: Wiley & Sons.

Elsness, J. (1997). *The Perfect and the Preterite in Contemporary and Earlier English.* Berlin, New York: Mouton de Gruyter

Elsness, J. (2009). The Present Perfect and the Preterite. In: Günter Rohdenburg and Julia Schlüte (eds.) *One Language, Two Grammars? Differences between British and American English*. Cambridge: Cambridge University Press, 228–245.

Elsness, J. (2014). The Present Perfect and the Preterite in Late Modern and Contemporary English. In: *Corpus Interrogation and Grammatical Patterns*. Amsterdam: John Benjamins Publishing Company, 81–10011.

Hundt, M. & Smith, N. (2009). The Present Perfect in British and American English: Has There Been Any Change Recently. *ICAME journal*, 33, 45–63.

Klein, W. (1992). The present-perfect puzzle. *Language*, 68: 522–525.

Köhler, R. (2005). Synergetic Linguistics. *Quantitative Linguistics An International Handbook*. Berlin & New York: de Gruyter.

Li, L. J., & Ge, G. C. (2009). Genre Analysis: Structural and Linguistic Evolution of the English–medium Medical Research Article (1985–2004). *English for Specific Purposes, 28*(2), 93–104.

Meyer–Viol, W. P., & Jones, H. S. (2011). Reference Time and the English Past Tenses. *Linguistics and Philosophy*, 34(3), 223–256.

Portner, P. (2003). The (Temporal) Semantics and (Modal) Pragmatics of the Perfect. *Linguistics and Philosophy*, 26(4), 459–510.

Rastall, P. (1999). Observations on the Present Perfect in English. *World Englishes, 18*(1), 79–84.

Reichenbach, H. (1971). *The Direction of Time*. California: The University of California Press, 117-125.

Schaden, G. (2008). Say Hello to Markedness. In Fischer, S., van de Vijver, R., Vogel, R. (eds.) *Optimality Theory and Minimalism: Interface Theories*. Germany: University of Potsdam, 1–25.

Schaden, G. (2009). Present Perfects Compete. *Linguistics and Philosophy* 32: 115–41.

Schaden, G. (2012). Modelling the "Aoristic Drift of the Present Perfect" as Inflation. An Essay in Historical Pragmatics. *International Review of Pragmatics*, 4(2), 261–292.

Szmrecsanyi, B., & Kortmann, B. (2009). The Morphosyntax of Varieties of English Worldwide: a Quantitative Perspective. *Lingua,* 119(11), 1643-1663.

Thomson, A. J. & Martinet, A. V. (1980). *A Practical English Grammar*. Oxford: Oxford University Press.

Vabalienė, Dalia Judita, Strimaitienė, Marija. (2004). Some Aspects of the English Tense Form Use in the Texts of Humanities. *Man and the Word*, 3(6), 4–8.

Waard, A. D., & Maat, H. P. (2012). Verb form Indicates Discourse Segment Type in Biological Research Papers: Experimental Evidence. *Journal of English for Academic Purposes*, 11(4), 357–366.

Werner, V. (2013). Temporal adverbials and the present perfect/past tense alternation. *English World-Wide*, 34(2), 202–240.

Yao, X. & Collins, P. (2012). The present perfect in world Englishes. *World Englishes*, 31(3), 386–403.

Yao, X. (2014). Developments in the Use of the English Present Perfect 1750–Present. *Journal of English Linguistics*, 42(4), 307–329.

Yao, X. (2016). The evolution of the "hot news" perfect in English: a study of register-specific linguistic change. *Journal of Historical Pragmatics*, 17(1), 129–152.

Yang, S. Y., & Huang, Y. Y. (2013). A survey of the four major Chinese aspect markers in different modes of discourse. *Contemporary Linguistics*, 3(15), 268–283.

Appendix I

Frequencies of the PP and the SP across Registers

	Total		Spoken		Fiction		Magazine		Newspaper		Academic	
	PP	SP	PP	SP	PP	SP	PP	SP	PP	SP	1044.91	4410.81
1990	5,586.09	34969.16	1358.85	5175.69	816.48	13429.78	1,110.30	5703.4	1255.57	6249.47	1049.35	4502.7
1991	5,548.86	35213.46	1331.68	5547.61	851.72	12890.61	1,052.16	5736.29	1263.95	6536.24	1025.8	4367.48
1992	5,305.05	34733.06	1198.64	6169.9	753.95	11601.75	1,112.92	5963.46	1213.75	6630.49	1059.3	4357.99
1993	5,117.98	34498.38	1116.06	6243.02	739.7	11438.45	1,028.54	5818.94	1174.4	6639.99	973.2	4175.24
1994	5,003.42	34241.91	1155.75	6503.63	703.82	11272.24	1,024.08	5647.23	1146.57	6643.58	1027.61	4349.78
1995	5,086.58	34116.72	1110.33	6106.13	725.53	11441.93	1,048.75	5694.81	1174.35	6524.06	1059.99	4677.36
1996	5,022.28	33974.38	1053.3	5885.68	731.54	10649.96	1,005.75	6154.39	1171.71	6606.97	1119.69	4689.14
1997	4,978.41	33952.6	992.35	6026.31	675.45	10592.5	1,001.78	6136.49	1189.13	6508.15	986.75	4365.85
1998	5,001.99	34397.15	1135.98	6320.97	742.75	11303.19	964.73	5793.81	1171.78	6613.33	1021.19	4260.75
1999	4,999.37	35552.74	1094.87	6464.84	781.51	12600.4	941.63	5727.6	1160.17	6499.15	1006.22	4363.94
2000	5,054.12	33797.35	1206.69	5801.42	756.19	11495.04	939.00	5898.09	1146.01	6238.86	976.56	4235.13
2001	5,022.75	34713.16	1115.89	6275.68	758.38	11879.79	966.77	5735.09	1205.15	6587.49	1016.63	4767.47
2002	5,143.79	34888.6	1182.9	6507.67	741.31	11062.68	986.13	5817.89	1216.83	6732.88	1000.25	4693.99
2003	4,996.50	35048.4	1156.5	6295.08	791.28	11959.51	909.69	5499.76	1138.78	6600.04	960.07	4788.19
2004	4,941.72	35373.69	1108.92	6365.33	814.35	11901.42	880.75	5690.53	1177.63	6628.2	928.9	4506.58
2005	4,914.73	35447.47	1155.09	6467.23	806.04	12312.35	878.23	5558.83	1146.48	6602.46	991.47	4651.28
2006	4,997.06	35635.61	1146.77	6360.83	833.91	12643.4	875.69	5381.1	1149.23	6577.1	1042.15	5656
2007	4,770.43	35224.26	917.42	6058.8	812.3	11769.27	879.56	5411.65	1118.99	6328.57	997.89	5109.26
2008	4,766.61	36701.02	879.06	5974.01	919.2	14158.46	869.13	5265.91	1101.32	6193.39	952.44	4860.63
2009	4,685.53	36123.79	955.03	6354.86	916.65	13888.37	778.22	4952.01	1083.19	6067.91	859.13	4556.62
2010	4,749.02	36283.54	1021.24	6425.91	877.82	13601.34	782.09	5108.6	1208.72	6591.08	875.72	4391.28
2011	4,660.02	36011.84	979.34	6481.24	884.34	13766.19	781.06	5237.29	1139.56	6135.83	778.29	4323.07
2012	4,553.37	36325.14	1046.25	6812.72	935.63	14206.33	709.38	4812.89	1083.83	6175.86	1044.91	4410.81

Appendix II

Standardized Amounts of the PP and the SP

	Total		Spoken		Fiction		Magazine		Newspaper		Academic	
	PP	SP	PP	SP	PP	SP	PP	SP	PP	SP	PP	SP
1990	2.36	−0.15	2.16	−2.88	0.25	1.04	1.60	0.30	1.87	−1.14	0.75	−0.48
1991	2.21	0.14	1.93	−1.84	0.73	0.57	1.07	0.40	2.05	0.31	0.81	−0.20
1992	1.23	−0.44	0.80	−0.09	−0.62	−0.58	1.63	1.05	0.99	0.79	0.49	−0.61
1993	0.49	−0.72	0.09	0.12	−0.81	−0.73	0.85	0.64	0.16	0.84	0.94	−0.64
1994	0.03	−1.03	0.43	0.85	−1.31	−0.87	0.81	0.14	−0.43	0.86	−0.22	−1.19
1995	0.36	−1.18	0.04	−0.27	−1.01	−0.72	1.04	0.28	0.16	0.25	0.51	−0.66
1996	0.11	−1.35	−0.44	−0.89	−0.93	−1.43	0.64	1.60	0.10	0.67	0.95	0.33
1997	−0.07	−1.37	−0.96	−0.49	−1.70	−1.48	0.61	1.55	0.47	0.17	1.75	0.37
1998	0.02	−0.84	0.26	0.34	−0.77	−0.85	0.26	0.57	0.10	0.70	−0.03	−0.61
1999	0.01	0.55	−0.09	0.74	−0.24	0.31	0.05	0.37	−0.14	0.13	0.43	−0.93
2000	0.23	−1.56	0.87	−1.12	−0.59	−0.68	0.03	0.87	−0.44	−1.19	0.23	−0.62
2001	0.11	−0.46	0.09	0.21	−0.56	−0.33	0.28	0.40	0.81	0.57	−0.17	−1.01
2002	0.59	−0.25	0.66	0.86	−0.79	−1.06	0.46	0.63	1.05	1.31	0.37	0.60
2003	0.00	−0.06	0.44	0.26	−0.10	−0.26	−0.24	−0.28	−0.59	0.64	0.15	0.38

2004	-0.22	0.33	0.03	0.46	0.22	-0.31	-0.51	0.27	0.23	0.78	-0.39	0.67
2005	-0.32	0.42	0.43	0.75	0.10	0.05	-0.53	-0.11	-0.43	0.65	-0.81	-0.19
2006	0.00	0.65	0.35	0.45	0.49	0.35	-0.55	-0.62	-0.37	0.52	0.03	0.25
2007	-0.90	0.15	-1.60	-0.40	0.19	-0.43	-0.52	-0.54	-1.01	-0.74	0.71	3.30
2008	-0.92	1.93	-1.93	-0.64	1.66	1.69	-0.61	-0.96	-1.38	-1.42	0.12	1.64
2009	-1.24	1.23	-1.28	0.43	1.63	1.45	-1.45	-1.86	-1.76	-2.06	-0.49	0.89
2010	-0.99	1.43	-0.72	0.63	1.09	1.20	-1.41	-1.41	0.88	0.59	-1.75	-0.03
2011	-1.34	1.10	-1.07	0.79	1.18	1.34	-1.42	-1.04	-0.57	-1.71	-1.52	-0.53
2012	-1.77	1.48	-0.50	1.72	1.89	1.74	-2.08	-2.26	-1.75	-1.51	-2.83	-0.74

Appendix III

Frequencies of the Simple Present Tense in Magazine and Newspaper

	Magazine	Newspaper
1990	3,908.17	3,908.17
1991	4,147.96	4,147.96
1992	4,420.16	4,420.16
1993	4,524.24	4,524.24
1994	4,655.14	4,655.14
1995	4,873.94	4,873.94
1996	4,997.62	4,997.62
1997	5,211.28	5,211.28
1998	5,488.65	5,488.65
1999	5,440.42	5,440.42
2000	5,506.20	5,506.20
2001	5,544.49	5,544.49
2002	5,355.28	5,355.28
2003	5,476.44	5,476.44
2004	5,513.22	5,513.22
2005	5,703.11	5,703.11
2006	5,625.14	5,625.14
2007	5,735.39	5,735.39
2008	5,924.19	5,924.19
2009	5,687.83	5,687.83
2010	5,719.64	5,719.64
2011	6,151.39	6,151.39
2012	6,297.03	6,297.03

Glottometrics 44, 2018, 59-75

Typological Features of Zhuang from the Perspective of Word Frequency Distribution

Aiyun Wei[1,2], *Haitao Liu*[1,3*]

Abstract. Investigating lexical features with statistical methods has always been a key object of quantitative linguistic research. However, though Zhuang is the mother tongue of the minority with the largest population in China, its lexical features have attracted little attention from the researchers employing quantitative means. Based on a corpus (CZL) of over 500,000 tokens of the Zhuang language, this study addresses the features of word frequency distribution of Zhuang. The results show that Zhuang shares the universal feature of other tested languages in that its word frequency distribution abides by the Zipf's Law and the "Least Effort Principle". The study also tests the word frequency distribution of Zhuang texts of different genres, which shows that for different genres, the values of some parameters, such as *b*, are different. Moreover, in order to test whether Zhuang language has any distinctive or typological features in word frequency distribution, the values of the *h*-point and *a*-index of the texts in CZL are computed as well. It is found that the two indexes are effective in distinguishing Zhuang from other languages, and the position of Zhuang on the analytism-synthetism continuum proposed by Popescu is close to those of the Polynesian language family, which may be helpful for intersubjective placement of Zhuang into a language group. This study would open a new perspective in the statistical lexical research of Zhuang language and present a "new" corroborated language with respect to the laws in quantitative linguistics.

Keywords: *Zhuang lexicon*; *word frequency distribution*; *Zipf's Law*; *h-point & a-index*; *typological features*

1. Introduction

Zhuang is a Chinese minority language that has the most speakers among the minority languages in China. With the increasing government's concern about protecting ethnic cultures and languages, the studies of Zhuang have attracted more and more attention in China. According to the results of the bibliometric statistics on the studies of Zhuang, since the end of the 1970s, there has been an increase of interest in Zhuang studies, esp. since the year 2010, as the publications related to Zhuang studies within the four years (2010–2013) outnumbered the total number of the studies in the previous 20 years. Moreover, the existing studies have covered different aspects of the language, such as the lexicon, grammar, phonetics and phonology, translation, and so forth (Wei 2015). Some of these studies focused on the features of

* 1 Department of Linguistics, Zhejiang University, Hangzhou, China; 2 College of Foreign Studies, Guangxi Normal University, Guilin, China; 3 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

the loan words, dialect words, modal particles, or adjectives (Lan 2007, 2011; He 2007; Huang 2010; Wei 2012); some investigated the typological features of Zhuang lexical classifiers (Qin 2015); some probed into the cross-border Zhuang lexical system (Wu 2005); and some tried to describe the features of the word classes of Zhuang (Qin, 2004).

In addition, Zhuang has also received some attention from related studies outside China, though not much. However, few of these studies focused on the Zhuang lexical system directly, as most of them just mentioned some general features of Zhuang lexicon while addressing other aspects, such as the writing system, the change & variation, the documentation & fieldwork, or the revitalization & documentation of the language, a specific tone in a Zhuang dialect, the Zhuang culture & people, or the minority language planning of China (Hu 1982; Edmondson 1992; Snyder 1997; Bauer 2000; Huang 2002; Huang 2003; Bodomo 2007, 2010; Adamson 2009).

In a word, most of these existing studies considered the features of some types or aspects of Zhuang lexicon, such as the comparison between Zhuang words and those of other languages (esp. Chinese), the comparison between the usage of the words in different Zhuang dialects, and so on. Moreover, few of them, either in or outside China, aimed at constructing a dynamic Zhuang corpus, objectively analyzing the quantitative lexical features of the whole Zhuang lexical system, or trying to define to which language family Zhuang belongs from the perspective of its quantitative lexical features.

Words play an indispensable part in human languages, and serve as the tool to convey information to achieve the communicative goal of any language (Zipf 1949). The quantitative research of language started with lexical statistics, mainly including the word/character frequencies. Popescu (2009) pointed out that, as the most traditional perspective, word frequency had always been the hottest topic since people had first conducted quantitative research of language. When conducting a critical review of Zipf's word frequency law in natural language and its future directions, Piantadosi (2014) stated that the frequency distribution of words has been the key study of statistical linguistics of the past 70 years, and the focus of the research lies in fitting Zipf's Law to the word frequency distribution of specific languages. He also proposed that to make progress in understanding the reason why the word frequency distribution of a specific language follows Zipf's Law, researchers should seek evidence beyond the law itself, test assumptions and evaluate novel predictions with new and independent linguistic data. There have been numerous studies on word frequency, covering those "big" languages such as English, German, French, Chinese, and so on, and some minority group languages such as the African aboriginal language Meroitic (Smith 2007). As almost all the languages researched in these studies are Indo-European languages, these investigations are helpful for understanding universal linguistic features of the languages in this family. However, the final aim of quantitative linguistics is to seek out the universal law that can explain any human language (Narisong et al. 2014), just as it was put by Popescu (2013, p. 224) that "Every 'new' language can falsify a beloved theory or force us to modify it". This is the most powerful impetus for quantitative linguists to keep exploring the properties of different languages, esp. those rarely-studied or non-studied ones.

Quantitative linguistics is a linguistic branch based on real-life linguistic materials, employing accurate methods to investigate linguistic structures, and developing regulations. Its objective is to conduct a quantitative analysis and dynamic description of various language phenomena, language structures, structural properties, and their interrelations with various

quantitative methods such as probability, mathematical statistics, information theory, function theory, and so on (Liu 2017).

Considering the inadequate achievements in the quantitative aspects of Zhuang lexical research, and the abundant outcomes that have been achieved in the word frequency studies of other languages – especially those of the Indo-European family –, and the features and objectives of quantitative linguistics, the present study probes into the word frequency distribution of Zhuang. Based on the paralleled dynamic Zhuang corpus (CZL), the study explores the possible universal features that Zhuang shares with other languages and the potential distinctive or typological traits in its word frequency distribution; moreover, it investigates the word frequency distribution features of different genres to test whether genre may have any impact on word frequency distribution. In addition, to clarify the comprehensive features of Zhuang from the perspective of word frequency distribution, we resort to calculating the values of the *h*-point and *a*-index of CZL. These two quantitative indexes have been tested by numerous natural languages and could be used to reveal the analytism-synthetism features or the lexical richness of languages; researchers have used them to classify different languages or different language varieties (Popescu 2009; Liu et al. 2011; Deng et al. 2012; Huang 2013). Through analyzing the values of the *h*-point and *a*-index of the Zhuang language data, we would be more objectively able to explore the lexical features of Zhuang from its position on the analytism-synthetism continuum, which could promote the sophisticated and scientific research of the lexicon or other perspectives of Zhuang.

Taking into account the related literature review of Zhuang lexical studies, the existing word frequency distribution studies of other languages, and the issues we are going to deal with in this study, we propose the following two research questions and hypothesis.

Question 1: Does Zhuang share the universal feature with other natural languages in that the word frequency distribution abides by the Zipf's Law? Does genre pose any influence on the fitting results?

As the word frequency-rank relationship of many languages – such as English, French, Spanish, Chinese, Korean, Indian and so on – have been tested to abide by the Zipf's Law (Zipf 1949; Kučera et al.1967; [Wang et al 2005]; Sun 1986; Choi 2000; Jayaram et al. 2008), to seek out whether Zhuang also shares this universal feature with other languages, we fit the Zipf's Law to the data of Zhuang word frequency in the present study. We first hypothesize that Zhuang shares the common feature with other tested languages in word frequency distribution, and the data in CZL would well fit the power law model expressed in Formula (1).

$$y = ax^{-b} \qquad (1)$$

Meanwhile, as CZL comprises texts of different genres, we also hypothesize that their fitting results may be different, so we would fit the model to the data of different Zhuang genres as well.

Question 2: The values of the h-point and a-index of language data have been tested as to its use in classification of different languages or language varieties. In this case, could the values of the h-point and a-index values derived from the Zhuang texts in CZL point out the typological lexical features of Zhuang language, and thus indicate to which language family Zhuang belongs?

As for the study of language typology, most researches have focused on the order of grammatical units in sentences (Greenberg 1963; Dryer 1992,1997; Song 2001; Bickel 2007; Liu 2010), but some may try to investigate this issue from other perspectives, such as word frequency distribution (Popescu 2009). As Frans Plank proposed in the first issue of *Linguistic Typology* (1997(1):1; [Koptjevskaja 2018, p.1]), the essence of typology lies in structural traits – ranging from sound and grammar to lexicon and discourse – that could vary independently from language to language, but actually do vary together, setting limits to cross-linguistic variation and defining the groundplans on which languages are constructed. Following this idea, any language structural trait or property that can distinguish languages from one another can be used as a classifying criterion.

h-point and *a*-index are two parameters derived from the word frequency distribution of linguistic texts, and are also used to investigate the vocabulary richness of texts or classify different languages. As Popescu (2009, p. 24) put it, without taking recourse to the morphology of languages, the *h*-point and the derived *a* measure can help a linguist find the position of a language on the analytism-synthetism scale, and thus make cross-linguistic comparison possible. According to incomplete statistics, the two indexes have been used to classify 20 languages (Popescu, 2009) and different language varieties (Deng et al. 2012; Huang 2013). Therefore, we hypothesize that the values of the *h*-point and *a*-index obtained from CZL could also present the general Zhuang lexical features, so that the value of the *a*-index could determine the place of Zhuang on the analytism-synthetism continuum and then reveal its typological feature in this aspect.

The present study consists of four parts. Part 1 is the literature review of previous Zhuang lexical studies, especially those related to word frequency distribution analysis; Part 2 is the source of research data and research method; Part 3 presents the analysis of the experiment results and the discussions correspondent to the research questions; and the last part draws a conclusion for the study.

2. Research Materials and Methods

2.1 Research Material

The data used in this study is the Corpus of Zhuang Language (CZL), a paralleled general corpus built by the authors of the paper. When constructing the corpus, we tried hard to follow the establishing principle of the widely-used English corpus (FLOB: Freiburg-LOB Corpus of British English) and the Chinese corpus (LCMC: The Lancaster Corpus of Mandarin Chinese)[1]. LCMC, constructed according to the model of FLOB, includes 15 genres of dynamic Chinese texts, which were selected from formal Chinese publications of the 5-year-span period (1989–1993) (Xiao 2004). However, due to the lack of appropriate Zhuang texts ready for use and the special situation we met when collecting the language materials, we had to proceed as follows: 1) the Zhuang corpus we built (CZL) contains about 500,000 Zhuang word tokens basically according to the structure of FLOB and LCMC, as sufficient suitable texts are unavailable. However, according to Strauss et al. (2006, p. 291), on the one hand, it turns out that a longer text does not necessarily yield better results; on the other hand, increasing

[1] LCMC available at http://ling.cass.cn/dangdai/LCMC/LCMC.htm.

text length need not necessarily yield worse results, so the corpus built is considered to be enough for use as far as this study is concerned. 2) In FLOB and LCMC, novels are classified into five categories, but as few Zhuang texts related to Kungfu stories, detective novels, or science fiction exist, we had to mark all novels as one type and make the total proportion of novels equivalent to that in FLOB and LCMC. 3) As there are no religious Zhuang texts, we replaced this genre with Zhuang folk songs, which are closely related to Zhuang people's daily life and full of emotional expressions, thus representing the basic beliefs of the Zhuang people. 4) Since Zhuang people do not communicate in Zhuang in the fields related to science and technology, commerce or academy, no independent Zhuang texts in these fields are available. 5) All texts of FLOB and LCMC were written from the years 1989 to 1993, while no electronic copies of Zhuang texts during that period are available, so we turned to the Zhuang texts from 2010–2015 instead. 6) With very limited number of audience, Zhuang texts on news tend to be simple and limited in variety, so we combined Zhuang news broadcast, editorials, and news reviews into the general category of news.

Though we could not make the proportions of the texts in exact accordance with those in FLOB and LCMC, all the selected texts are from the two most influential and authoritative Zhuang publications: one is *Sam Nyied Sam* – a magazine containing the most influential Zhuang texts of different types, and the other is Guangxi Nationality Newspaper (Zhuang version) – a newspaper including Zhuang texts about most of the things happening in the districts inhabited by Zhuang people. Both publications are written in standard Zhuang, the authoritative standardized form of the Zhuang language. In addition, most of the texts (except the news) are from *Sam Nyied Sam*, an inclusive magazine consisting of novels, prose, travel notes, folktales, folksongs, bibliographies, government work reports, Zhuang learning notes, and the related contents in the Zhuang language, which includes almost all aspects of Zhuang people's life. And we selected the news from Guangxi Nationality Newspaper instead of *Sam Nyied Sam* only because the former is regularly published within a shorter time span – it is published once a week, while the latter once in two months. Table 1 and Table 2 list the basic information and the component proportions of CZL.

Table 1

The Basic Statistical Information of CZL

	Type	Token	Type/token ratio
Word total	24,286	585,455	4.15%

Table 2

The Components of the Text Genres in CZL

Code	Genre	Code	Genre
A	Novels	E	Work reports
B	Prose	F	News
C	Folk tales	G	Bibliographies & essays
D	Folk songs		

2.2 Research Method

It is obviously shown in the introduction section that the emphasis of this study lies in searching for possible universal or distinguishing features of Zhuang in word frequency distribution.

The study is conducted following the two steps below.

According to Piantadosi (2014), as one of the key objects of statistical linguistic research in the past 70 years, word frequency distribution has been tested to approximately follow a simple mathematical form known as Zipf's Law. Moreover, related linguistic statistic experiments have shown that the Zipf's Law is also applicable in the research of other linguistic phenomena, such as Chinese characters, letters and so on (Sun 1986). Therefore, we use the software NLREG to fit Formula (1) to the word frequency distribution data derived from CZL, and conduct a detailed analysis of the fitting results with relevant quantitative linguistic hypotheses.

In addition, since the *h*-point and *a*-index are two important quantitative indexes that indicate the typological features of word frequency distribution, to find out the distinctive features that distinguish Zhuang from other languages, we use QUITA to calculate their values on the basis of the word frequency distribution data of CZL, thus enabling us to objectively analyze the results with relevant quantitative linguistic theories.

The *h*-point and *a*-index are two important textual indexes that could indicate the stylistic features of texts (Popescu et al. 2012), and the concepts of these two terms originated from Hirsch (2005), who used the terms to evaluate the output and influence of scientific research staffs. He pointed out that if we regard the number of a scholar's published papers as *N* and the number of citation of each paper as *P*, and then arrange the *P*'s of the *N*'s in a decreasing order, we can find an intersection on the decreasing sequence, at which the rank order (*N*) is equal to the value of *P*, and this intersection is called "*h*-point". And there is a proportion relation (*a*) between the value of the "*h*-point" and the number of the published papers (*N*), shown as in Formula (2):

$$a = \frac{N}{\mathrm{h}^2} \qquad (2)$$

The *a* here is called "*a*-index". Popescu (2006, 2007) introduced these two concepts into linguistic research and proposed that, on the word frequency distribution curve, we can always find a point at which the value of *r* (rank) is close/equal to that of the *f* (frequency); and it is this point that divides any word frequency distribution curve into two regions, with the one above the point covered with functional/synsemantic and high-frequency words and the other one under the point covered with notional/autosemantic and low-frequency words (Popescu & Altmann 2006; Popescu 2007, 2009, Popescu et al. 2009; Martináková Z et al. 2008; Popescu et al. 2012). This is clearly shown in Figure 1 below.

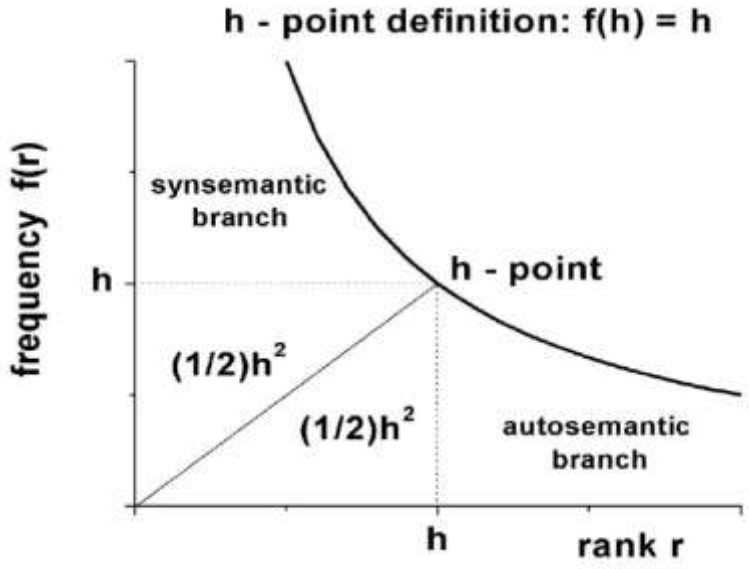


Figure 1. Graphical representation of *h*-point（Adapted from Popescu et al 2006:25）

However, in many rank-frequency distributions, there is no such a word that satisfies the condition of $r = f(r)$, in which case the *h*-point lies between two adjacent words. If we respectively set the adjacent two ranks as r_1 and r_2（$r_2 > r_1$）， and if the condition of "$r_1 < f(r_1)$ and $r_2 > f(r_2)$" could be satisfied, the *h*-point will be the intersection of the line passing (r_1, $f(r_1)$) and (r_2, $f(r_2)$) and the straight line $y = x$, which can be illustrated by Formula (3). The value of the *h*-point can also be automatically calculated by the software QUITA, so the introduction here is just for the sake of clarifying the process.

$$h = \frac{f(r_1)r_2 - f(r_2)r_1}{r_2 - r_1 + f(r_1) - f(r_2)} \qquad (3)$$

3. Results and Discussions

3.1 The Analysis of Zhuang Word Frequency

As mentioned above, the Zipf's Law has been tested to exist in the word frequency distribution of many languages (Yu. et al. 2018). Therefore, to test whether Zhuang shares this universal feature, we fitted the model of Formula (1) to the word frequency data of Zhuang. If the Zhuang word frequency distribution abides by the law, and the data fits the model well, we would fit the model to the data of different genres of Zhuang as well, as CZL contains different genres of texts. We use the software NLREG to fit the aforesaid model to the Zhuang word frequency distribution data and employ relevant quantitative linguistic theories to analyze the fitting results in detail.

Meanwhile, to conduct a comparative analysis of Zhuang word frequency distribution, we randomly extract word frequency distribution data of the nearly equivalent number of word tokens (about 580,000 words) from the Chinese corpus (LCMC) and the English corpus (FLOB) respectively, and then conduct a comparison among the word frequency distribution of CZL, LCMC and FLOB. The results are presented as in Table 3, Table 4, and Table 5. The

first column of these three tables refers to the ranking orders; the second refers to the specific words of correspondent ranking orders; the third column is the frequency of the word tokens, i.e. the occurrence frequencies of the word tokens appearing in the texts investigated; and the fourth column presents the percentage that specific words account for in the whole corpus. The three tables show the following results: first, all the Zhuang words (68 words) with the highest frequency are monosyllables, and their accumulative frequencies account for 36.2% of the Zhuang word total in CZL, abiding by the Zipf's "Least Effort Principle" (Zipf, 1949) ; second, "mbouj (no)" is the word with the highest frequency in CZL, and the proportion *Pr.* = 0.0203 is less than 0.1, which is in accordance with the related content verified by Zipf (1949); "的"(of) is the Chinese word with the highest frequency in LCMC, and the proportion *Pr.* = 0.0633 is less than 0.1; and "the" is the English word with the highest frequency in FLOB, with the proportion *Pr.* = 0.0627, also less than 0.1. Therefore, from the values of *Pr.* and the word proportion of the three corpora, it is obvious that even though there is a slight difference in between, there is no significant difference between Zhuang and Chinese or English as far as word frequency distribution is concerned.

Table 3

The Word Frequency of CZL (585,455 tokens)

Rank	Words	Frequency	Proportion（%）
1	mbouj (not/no)	11,905	0.0203
2	de (he/she/it)	9,503	0.0162
…	…	…	…
68	cienz (money)	1,190	0.0020
69	seizneix (now)	1,171	0.0020
…	…	…	…
24,286	zhengfuj(government)	1	0.000002

Table 4

The Word Frequency of LCMC (594,010 tokens)

Rank	Words	Frequency	Proportion（%）
1	的 (of)	37,625	0.0633
2	是 (be/yes)	8,745	0.0147
…	…	…	…
22	说 (say)	1,967	0.0033
23	一个 (a/one)	1,746	0.0029
…	…	…	…
34,816	四百万 (4 million)	1	0.000002

Table 5
The Word Frequency of FLOB (589,189 tokens)

Rank	Words	Frequency	Proportion（%）
1	the	36,962	0.06273
2	of	18,424	0.0313
…	…	…	…
59	who	1,237	0.0021
60	about	1,211	0.00205
…	…	…	…
32,486	zz	1	0.000002

After a brief comparison between the word proportions of the three corpora, to conduct a more specific analysis of the word frequency distribution of the three languages, and to achieve a more direct and visualized effect, we also fit the formula of Zipf's Law to the word frequency data of Zhuang, Chinese, and English, and the results are shown in Figure 2 and Table 6 respectively.

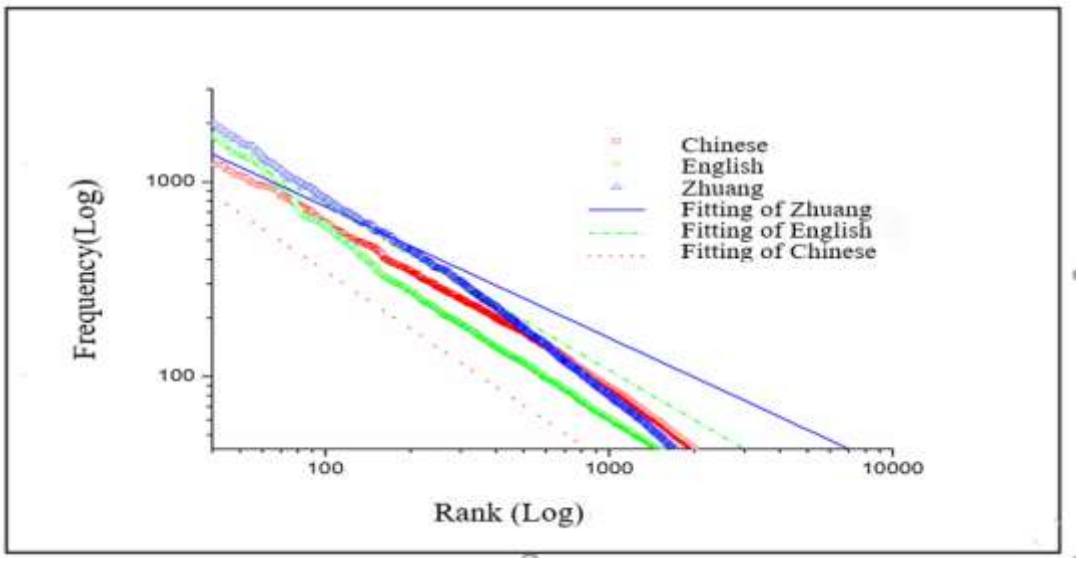


Figure 2. The Fitting Results of the Zipf's Law to the Word Frequency & Rank of CZL, LCMC and FLOB

In Figure 2, the *X* axis is the rank; *Y* is the frequency; the scatter plot is the actual statistical data points of the word frequency distribution; and the slash is the fitting results, in which the blue, red, and green line stands for Zhuang, Chinese, and English respectively. From the figure, we can also see that the fitting slashes are quite similar to each other as they almost overlap, which means that the word frequency distributions of Zhuang, Chinese, and English have much in common in this perspective. Table 6 presents the detailed values of related parameters in the fitting experiments of the three languages.

Table 6
Fitting the Zipf's Law to the Frequency & Rank of CZL, LCMC and FLOB

Corpus	The fitting results	*b*	R^2
CZL	$y = 16756*x^{-0.672}$	0.675	0.880
LCMC	$y = 33039*x^{-0.984}$	0.988	0.936
FLOB	$y = 38179*x^{-0.850}$	0.850	0.978

As could be seen from Table 6, the fitting of CZL presents good results, and those of LCMC and FLOB are very good (criterion: $R^2 > 0.9$ means "very good", $R^2 > 0.80$ "good", $R^2 > 0.75$ "acceptable", $R^2 < 0.75$ "unacceptable"). The results are not quite differentiating from one another, but we have observed that the values of the power exponent *b* present different results, and the *b* derived from CZL demonstrates the most striking difference from that of Chinese and English. Therefore, we consider that the differences among the values of the power exponent *b* may be able to serve as an index to categorize different languages, which coincides with the results of previous studies. Bujdosó (2006) investigated the word frequency distribution of 21 official languages in the *Charter of European Union*, and classified the 21 languages according to the different values of the power exponent *b* in the fitting model; the results show that some parameters in Zipf's Law can be used as indexes for language classification.

According to previous related studies, the fitting of the power law function to the word frequency data derived from a rather large corpus may yield very good results, i.e. $R^2 > 0.9$ (such as those of the FLOB and LCMC), while the value of R^2 in the fitting results of CZL is only 0.88. We conjecture that it may be due to the fact that letters indicating six different tones (except the first one) in Zhuang are directly placed at the end of correspondent words, in which case the mean word length of Zhuang tends to be longer than that of the other languages (when the word length is measured by the number of letters). This issue is worthy of thorough consideration in our further research.

In addition, as the CZL consists of different genres of texts, to find out whether genre may influence the fitting results, we fit the model to the data of different genres of the texts in CZL. Considering the text length may pose some impact, when comparing the results of different genres, we limit the size of each genre to about 50,000 word tokens. The results are shown in Table 7.

Table 7
The Fitting Results of Different Genres to Zipf's Law

Genre	Corpus length	R^2	*b*	The fitting results
News	48,998	0.843	0.613	$y = 1109*x^{-0.613}$
Bibliographies &essays	55,882	0.893	0.651	$y = 1534*x^{-0.651}$
Folk songs	53,920	0.926	0.663	$y = 1552*x^{-0.663}$
Folk tales	45,447	0.906	0.672	$y = 1621*x^{-0.672}$
prose	55,812	0.910	0.678	$y = 1922*x^{-0.678}$
Novels	55,530	0.932	0.693	$y = 2119*x^{-0.693}$
Work report	52,410	0.968	0.732	$y = 2110*x^{-0.723}$
Mixed genres in CZL	585,455	0.880	0.675	$y = 16756*x^{-0.675}$

As clearly shown in Table 7, the fitting results of all the seven genres of texts collected in CZL appear acceptable, even though there are some difference between them. However, either considering the values of R^2 or those of the parameter *b*, the fitting results of the bibliographies & essays, folk songs, folk tales, prose, and novels are similar to one another, but there is still some difference between these values and those of CZL. In addition, the values of R^2 and the power exponent *b* of the news and work reports show striking differences, as the value of R^2 of the news is only 0.842, lower than 0.85, thus being considered as acceptable only, and the value of *b* of this genre is also the smallest one among all the seven genres collected; while the value of R^2 of the work reports is 0.968, higher than 0.95, considered as a very good fitting, and the value of *b* here is also the highest among the seven genres. This phenomenon may be due to the diction of the texts, for the news in Zhuang are various in themes and contents, including almost all the events or related contents happening in Zhuang residential districts; the news texts are written with affluent words or phrases, while the work reports are quite limited in the scope of collecting materials, and the words are used in somewhat fixed patterns. This phenomenon may also be explained by the TTR of the seven genres of texts.

According to the statistical results of QUITA (the Quantitative Index Text Analyzer), the TTRs (Type-Token Ratio) of these five genres of texts are also similar to one another, between 9% and 10%. The TTR of the news texts is 11.5%, which is the highest in the seven genres of texts. However, for the government work reports, the words with high frequencies account for a larger proportion part in the corpus. To be more specific, 4,181 word types constitute 53,162 tokens, thus yielding a TTR of 7.9%, the smallest one within the seven genres.

Therefore, the values of the parameter *b* in the power law function $y=ax^{-b}$ is closely related to the values of R^2, and apart from being an index in language classification (Bujdosó 2006), it could, to some extent, reveal the classifying feature of different genres within one language, and the degree of lexical richness of the observed texts as well.

3.2 The Analysis of the *h*-point and *a*-index of CZL

In 3.1, we apply the Zipf's Law to fit the word frequency distribution data of CZL, and it is found that Zhuang shares the universal feature with other languages in word frequency distribution, as it abides by the power law distribution model and the "Least Effort Principle". To further investigate the possible distinctive features that distinguish Zhuang from other languages, we use QUITA to analyze the texts in CZL, and calculate the values of *h*-point and *a*-index.

Table 8

The Word Frequency Distribution of Zhuang in CZL

rank	words	frequency	%	
1	mbouj(no/not)	11,905	2.0335%	
…	…	…	…	
305	seng(bear/produce)	307	0.0524%	
306	sonhag(teaching)	307	0.0524%	
307	rog(outside)	305	0.0520%	

According to the methods of calculating the value of *h*-point and the word frequency distribution data of Zhuang in Table 8, there is no such a word that meets $r = f(r)$, so we use Formula (2) to calculate the value of the *h*-point. The results are shown as below, which is

almost equivalent to the value (306.33) calculated by QUITA:

$$h = \frac{307 \times 307 - 305 \times 305}{307 - 305 + 307 - 305} = 306.$$

To get the value of *a*-index, we bring the value of *h*-point into Formula (3), with the result shown as follows:

$$a = \frac{585455}{306^2} = 6.25.$$

As it is linked to the vocabulary size of a text, the value of *h*-point and the related *a*-index can be used to measure the vocabulary richness. The higher the value of the *a*-index, the richer the vocabulary in the text would be. Popescu (2009) stated that with the higher value of the *h*-point (the lower value of *a*-index), the language will be more likely an analytic one. In other words, the value of the *h*-point is an index to indicate the analytism-synthetism degree of a language. Generally speaking, the higher the value of *h*-point (the lower the one of *a*-index), the smaller the number of word types, which shows that the synthetic feature of the language has been replaced by inflectional parts or auxiliary words. Therefore, the values of *h*-point and *a*-index may reflect the textual features within one language, while they could indicate the analytism-synthetism feature of different languages. To test whether *a*-index could be used to classify different languages, Popescu (2009) compared the *a*-indexes of English, German, Russian, Latin, and some Polynesian languages, such as Samoan, Hawaiian, and Maori. The results show that the values of the *a*-indexes of the Polynesian languages are rather small, and these are the human languages considered to bear the most distinctive analytic features; while the *a*-indexes of Hungarian and Latin are fairly big, indicating the high synthetic degree of the two languages. In this case, the value of *a*-index derived from that of the correspondent *h*-point can be inductive in finding out the position of a particular language on the analytism-synthetism continuum.

Based on Popescu's study, the present study adds the *a*-index values of Zhuang (6.25) and Chinese (8.09) to the continuum, with the results shown in Figure 3 below.

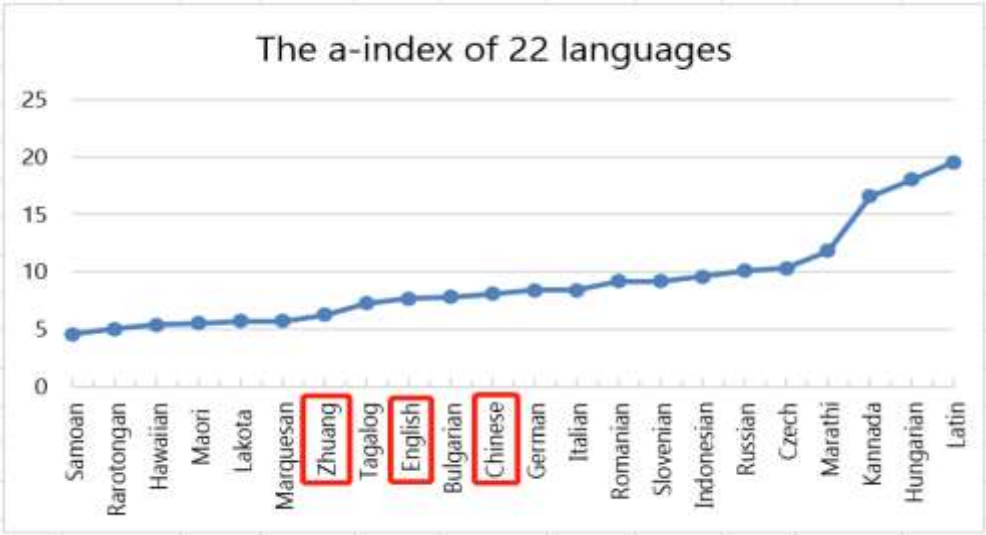


Figure 3. The *a*-index of 22 languages

According to the study of Popescu (2009) and the results in Figure 3, we could come to the following findings: firstly, the positions of English and Chinese fall on the somewhat middle part of this analytism-synthetism continuum, which is in accordance with both the results of the study of Liu et. al (2010) on language classification by means of the method of language clusters based on linguistic complex networks, and those of Liu's language typological study with the method of sentence dependency direction (Liu 2010); secondly, on this analytism- synthetism continuum, the position of Zhuang is close to those of the Polynesian languages, which is quite similar to the results in Liu's classification of Esperanto (a = 6.5) (Liu 2011), as both the positions of Zhuang and Esperanto lies quite near that of English – with only Tagalog in between. Therefore, even though this is only a continuum, which means the analytic or synthetic feature is only a comparative concept, we can still conclude that, similar to the languages in the Polynesian language family, Zhuang is a typical analytical language.

For many years, there has been a controversy among Zhuang researchers as to whether the Kam-tai language (Zhuang included) belong to the Sino-Tibetan language family, or whether there is a genetic relation between the Kam-tai language family and the Polynesian language. As early as over 70 years ago, Benedict(1942) proposed that the Thai (Kam-tai language – Zhuang included), Kadai and Indonesian had formed a new alignment of language in Southeastern Asia, while influential Zhuang researchers such as Li Fanggui and W. Gedney held that the Kam-tai language belongs to the Sino-Tibetan family and it has no close relation with the Polynesian language family([Ni1988]), so did Wei & Qin (2006). However, Ni (1988) compared the 2,400 words selected from Zhuang-dai, Kam-Sui, and Li language with those selected from Indonesian (included in the Polynesian language family) and found that these words of the Kam-tai language are closely related to the words selected from the Polynesian language family. Moreover, the results show that from the perspective of word order and combinative relation, the Kam-tai language is closer to Indonesian than Chinese. And in his follow-up researches (Ni 1994), he pointed out that the Kam-Tai language and the Polynesian language family came from the same source according to the comparative analysis of the archeological materials, the cultural history, and language materials. The study of Deng & Deng (2011) indicated that there is a genetic relation between the Polynesian language and the Kam-tai language family with the evidence derived from the similarity between the two language families. They also verified that different languages in these two language families have some degree of genetic relation between one another from the perspectives of anthropology, etymology statistics, archeology, genetics, and so on.

According to the aforementioned, though there is no unanimous view as to the issue whether the Kam-tai language family should belong to the Polynesian or the Sino-Tibetan language family, or whether it forms a language family of its own, it is an indisputable fact that there is a genetic relationship between the Kam-tai language and the Polynesian language family. The results are in accordance with our findings that reveal the typological feature of Zhuang based on the value of *a*-index derived from the dynamic real-life Zhuang language data. As Zhuang is close to the Polynesian languages on the analytism-synthetism scale continuum, it is quite likely that there is a genetic relation between the two language families.

4. Conclusion

Based on the analyses in the previous parts, we can make a conclusion corresponding to the two research questions in the introduction section as follows:

1) Zhuang shares the universal feature with most languages, such as English and Chinese, in word frequency distribution, i.e., the word frequency distribution of Zhuang abides by both Zipf's Law ($R^2 = 0.88$) and the "Least Effort Principle". The fitting results of the whole corpus is not so ideal as those of English or Chinese. This might be due to the fact that Zhuang is a pure phonographically written language, and the particular letters that indicate different tones (except the first one) are directly written at the end of specific words. This is an issue worth further study. Through investigating the fitting results of Zhuang texts of different genres in CZL, we found that the power exponents of different text genres are variegated. The fitting of the news texts, with highest lexical richness, presents the poorest fitting result ($R^2 = 0.842$), while that of the government work reports, with the lowest lexical richness, shows the best fitting result ($R^2 = 0.968$). Therefore, by referring to Bujdosó's (2006) research, we consider that the values of the power exponent *b* in Zipf's Law could not only be used as an index to classify different languages, but may also be used to classify different genres within the same language.

2) From the values of *h*-point and *a*-index derived from the texts in CZL, we found that, in word frequency distribution, Zhuang is closely related to the languages of the Polynesian family, as their positions on the analytism-synthetism continuum are quite close to one another. And the positions of English and Chinese fall on the middle part of the continuum, bearing both analytic and synthetic features. Therefore, the positions of the two languages are similar to those in the previous studies of language classification (Liu et al. 2010, Liu 2011). However, due to the limited length and inadequate diachronic language materials, with the present results, it is sufficient to present that Zhuang is quite similar to the Polynesian language family in word frequency distribution. Following this idea, the present study may provide some scientific support to the issue of which language family Zhuang belongs to.

In a word, the present study deals with the possible universal traits that Zhuang shares with other languages and the distinctive features that distinguish Zhuang from them. This coincides with the two most influential areas of linguistic typology, i.e. diversity and unity/universal, as the area of linguistic typology dealing with diversity investigates the structural variation in the world languages, while the area concerned with unity focuses on the discovery of language universals (Song 2018). Therefore, on the one hand, this study broadens the investigating scope of Zhuang since it may provide a solution to the language family affiliation of this language; on the other hand, the study contributes considerably to the development of linguistic typology – as J. J. Song (2018, p. 22) put it, typological classification naturally calls for data from a wide range of languages, which enables us to minimize the risk of elevating some of the least common structural properties to the status of language universals.

Acknowledgments

This work is partly supported by the National Social Science Foundation of China (Grant No. 17BYY182, 17BYY120), and the Social Science Foundation of Education Department of Guangxi Zhuang Autonomous Region (Grant No. KY2016YB056).

REFERENCES

Adamson, B., Feng, A. (2009). A comparison of trilingual education policies for ethnic minorities in China. *Compare,* 39(3), 321–333.

Bauer, R. S. (2000). The Chinese-based writing system of the Zhuang language. *Cahiers de Linguistique Asie Orientale*, 29(2), 223–253.

Benedict, P. K. (1942). Thai, Kadai, and Indonesian: a new alignment in Southeastern Asia. *American Anthropologist*, 44(4), 576-601.

Bickel, B. (2007). Typology in the 21st century: major current developments. *Linguistic Typology*, 11(1), 239–251.

Bodomo, A., Pan, Y. Q. (2007). *A proficiency course in Zhuang: Fieldwork documentation and revitalization of a language and culture of southwestern China*. Hong Kong: The Linguistic Society of Hong Kong.

Bodomo, A. (2010). Documentation and Revitalization of the Zhuang Language and Culture of Southwestern China Through Linguistic Fieldwork. *Diaspora, Indigenous, and Minority Education,* 4(3), 179–191.

Bujdosó, I. R. (2006). Vortstatistika ekzamenado de la plurlingva teksto de la konstitucipropono de Eŭropa Unio. *Proceedings Intemacia Kongresa Universitato Florenco, 134–143.*

Choi, S. W. (2000). Some statistical properties and Zipf's law in Korean text corpus. *Journal of Quantitative linguistics*, 7(1), 19–30.

Deng, X., Deng. X. (2011). Discussion of the genetic relationship between Kra-Dai and Austronesian. *Language Research, 34–41*.

Deng, Y., Feng, Z. (2012). A Quantitative Linguistic Study on the Relationship between Word Length and Word Frequency. *Journal of Foreign Languages* (3), 29–39.

Dryer, M. (1992). The Greenbergian word order correlations. *Language* (68), 81–138.

Dryer, M. (1997). On the 6-way word order typology. *Studies in Language* (21), 69–103.

Edmondson, J. (1992). Change and variation in Zhuang. In: *Papers from the Second Annual Meeting of the Southeast Asian Linguistic Society*. Tempe: State University of Arizona Program for Southeast Asian Studies, 147–185.

He, S. (2007). *A study on the modal particle in Xincheng Zhuang*. Ph.D. dissertation. Minzu University of China.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg, J. (Ed.), *Universals of Language*. Cambridge, MA: MIT Press, 58–90.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output, *Proceedings of the National academy of Sciences of the United States of America,* 2005. 102(46), 16569–16572.

Hu, Z. (1982). Some Linguistic Differences in the Written English of Chinese and Australian Students. *Language learning and Communication,* 1(1), 39–49.

Huang, M. (2010). *Research of the Zhuang adjectives in Daxin*. Ph.D. dissertation. Minzu University of China.

Huang, P. (2002). Sinification of the Zhuang people, culture, and their language. *SEALS XII, 89–100.*

Huang, W. (2013). *On Word Frequencies Distribution and Vocabulary Richness in Genres of Modern Chinese.* Ph.D. dissertation. Communication University of China.

Huang, X. (2003). Minority language planning of China in relation to use and development, November, Bangkok: Paper presented at *the Conference on Language Development Language Revitalization and Multilingual Education in Minority Communities in Asia.*

Jayaram, B. D., Vidya, M. N. (2008). Zipf's law for Indian languages. *Journal of Quantitative Linguistics*, 15(4), 293–317.

Kučera, H., Francis, W. N. (1967). Computational analysis of present-day American English. Dartmouth Publishing Group.

Koptjevskaja-Tamm, M. (2018). Introduction from the new Editor: Linguistic Typology today and tomorrow. *Linguistic Typology,* 22(1), 1–12.

Lan, Q. (2007). Research on the sources of color terms in Zhuang dialects. *Minority Languages of China* (5), 34–43.

Lan, Q. (2011). The loanwords from ancient Chinese and the relationship between Sino-Vietnamese and ping speech. *Minority Languages of China* (3), 48–61.

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. (2011). Quantitative analysis of Zamenhof's Esenco kaj estonteco, *Language Problems & Language Planning,* 35(1): 57–81.

Liu, H. (2017). *Introduction to Quantitative Linguistics*. Beijing: The Commercial Press.

Liu H., Li W. (2010). Language clusters based on linguistic complex networks. *Chinese Sci Bull 55, doi.10.1007/s11434-010-4114-3*

Martináková, Z., Popescu, I. I., Mačutek, J., & Altmann, G. (2008). Some problems of musical texts. *Glottometrics,* 16, 80–110.

Narisong, Jiang, J., Liu, H. (2014). Word Length Distribution in Mongolian. *Journal of Quantitative Linguistics* 21(2), 123–152.

Ni, D. (1988). The Zhuang-Dong languages in China and Austronesian. *Journal of Minzu College of China* (3), 54–64.

Ni, D. (1994). The relationship between austronesian and baiyue languages. *Minority Languages of China* (3), 21-35.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language. A critical review and future directions. *Psychonomic Bulletin & Review* 21(5), 1112–1130.

Popescu, I. I., & Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics,* 13, 23–46.

Popescu, I. I.. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text, 555–565*. Berlin/New York: Mouton de Gruyter.

Popescu, I. I. (ed.) (2009). *Word frequency studies* (Vol. 64). Berlin: Walter de Gruyter.

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

Popescu, I. I., Čech, R., & Altmann, G. (2012). Some geometric properties of Slovak poetry. *Journal of Quantitative Linguistics,* 19(2), 121–131.

Popescu, I. I., Naumann, S., Kelih, E., Rovenchak, A., Overbeck, A., Sanada, H., Smith, R., Čech, R., Mohanty P., Wilson, A., & Altmann, G. (2013). Word length: aspects and languages. *Issues in quantitative linguistics, 3, 224–281.* Dedicated to Karl-Heinz

Best on the occasion of his 70th birthday. Lüdenscheid: RAM.

Qin, F. (2015). Revisiting the classifiers in the Zhuang language from a typological perspective. *Studies of the Chinese Language* (5), 513–522.

Qin, X. H. (2004). *Zhuang lexicology.* Beijing: The Ethnic Publishing House.

Smith, R. D. (2007). Investigation of the Zipf-plot of the extinct Meroitic language. *Glottometrics,* (15), 53–61.

Snyder, W. C., & Lu, T. (1997). Wuming Zhuang tone sandhi: a phonological, syntactic, and lexical investigation. *Summer Institute of Lingusitics and the University of Texas at Arlington Publications in Linguistics* 124, 107–140.

Song, J. (2001). *Linguistic Typology. Morphology and Syntax.* London: Pearson Education.

Song, J. (2018*). Linguistic Typology*. London: Oxford University Press.

Strauss, U., Grzybek, P., Altmann, G. (2006). *Word length and word frequency*. Berlin: Springer.

Sun, J. (1986). The study of word frequency distribution law from the perspective of its word frequency statistics. *Studies of the Chinese Language* (3): 34–35.

Wei, A. (2015). A review on Zhuang Language research based on quantitative analysis of literature. *Journal of Guangxi Normal University (Philosophy and Social Sciences Edition)* (1): 153–159.

Wang, D., Li, M., Di, Z. (2005). True reason for Zipf's law in language. *Physica A: Statistical Mechanics and its Applications*, 358(2–4), 545–550.

Wei, M. (2012). *The Grammar of Zhuang in Xia'ao*. Ph.D. dissertation. Shanghai Normal University.

Wei, J., Qin, X. (2006). *General Theory of Zhuang Language*. Beijing: Press of Minzu University of China.

Wu, X. (2005). *On Zhuang Languages Across Border.* Ph.D. dissertation. Huazhong University of Science & Technology.

Xiao R,, McEnery T. (2004). *Aspect in Mandarin Chinese. A corpus-based study* (Vol.73). Amsterdam: John Benjamins Publishing.

Yu, S., Xu, C., Liu, H. (2018). Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. http://arxiv.org/abs/1807.01855.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, Mass.

Glottometrics 44, 2019, 76-86

Probability Distribution of Causal Linguistic Features

Hong Ma[1], *Haitao Liu*[1 2]

Abstract. This study investigated the probability distribution of cause-effect language produced by four different groups of speakers, ESL and non-ESL high school and primary students. Results showed that, regardless of speakers' English proficiency, the probability distribution of causal linguistic features produced by native speakers fitted the right-truncated modified Zipf-Alekseev model, while this model did not capture the distribution of causal linguistic features identified in nonnative speakers' speech. The results of current research suggested that the right-truncated modified Zipf-Alekseev model has the potential of differentiating between native and nonnative causal discourses. In addition, the fitting results shed light on relationship between parameters (*a* and *b*) and students' proficiency levels as well as tentatively offer an alternative to systemic functional linguistic (SFL) method of evaluating students' causal language use.

Key words: *Diversification, Causal linguistic features, Zipf-Alekseev distribution*

1. Introduction

In linguistics, Zipf (1949) well identified the least effort principle, which underlies unification or diversification of linguistic units in naturally occurring languages to alleviate speakers' physical or mental efforts. Diversification, an umbrella term, encompasses generation of variants (all free or conditional "non-standard" forms of an entity, e.g., allophones, allomorphs, dialectal or sociolectal expressions of a concept, etc.) and secondary forms (in some way derived from the primary form, e.g. polysemy, cases, times, moods, aspects, etc.), and acquisition of membership in different classes (built by a class-building criterion, e.g., derivates, compounds, declination classes, word classes, even semantic classes, etc.) [Strauss & Altmann, 2006]. It has been widely recognized that diversification, serving as a process operating in the self-regulation of language, follows probability distribution, and is appropriate for mathematical modelling (Rothe, 1991; Altmann, 1996; Strauss & Altmann, 2006). Strauss and Altmann (2006) speculated that the ranked frequencies of individual entities obey a rank-frequency distribution, a function ranking the frequencies of linguistic entities in a descending order.

Pioneered by Köhler (1986), researchers have exerted great efforts in identifying a unified distribution to model diversification processes pertaining to various linguistic entities (Altmann, 1991; Hřebíček, 1996). More probability distributions modelling diversification of various

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

[2] Ningbo Institute of Technology, Zhejiang University, China. Correspondence to: Haitao Liu. Email address: htliu@163.com, ORCID-No.: https://orcid.org/0000-0003-1724-4418

linguistic entities can be found in Rothe (1991). Recently, Zipf-Alekseev distribution, a well-known Zipf's Law-related distribution, has been found powerful in modeling diversification processes of various linguistic entities, including adverbials (Čech & Uhlířová, 2014), synonyms (Zhu & Liu, 2018), word length (Chen & Liu, 2014; Mohanty & Popescu, 2014), dependencies (Liu, 2009), semantic roles (Liu, 2012), and discourse relations with respect to the categories of the rhetorical structure theory (Yue & Liu, 2011; Zhang & Liu, 2015). Ouyang and Jiang (2018) ranks among the very few that examined the model fit of second language learners' language production. They found that the probability distribution of dependency distances of nine consecutive grades of English language learners fits the Zipf-Alekseev distribution, and that parameters *a* and *b* in this model reflect learners' English proficiency level in the way that with the increase of grades (learners' English proficiency), parameter *b* decreases significantly, while parameter *a* increases.

Viewing variations of cause-effect language as a diversification phenomenon, this study intends to investigate the probability distribution of causal linguistic features, given that little effort has been exerted to examine the quantitative aspects of this phenomenon. Variations of causal linguistic features examined in this current study were originally identified in a piece of research that adopted the SFL perspective. The SFL research suggested that causal linguistic features progress semantically from the temporal to the cause (external cause) and the proof (internal cause) ones, and lexicogrammatically from less metaphoric structures (i.e., relators and circumstances) to more metaphoric structures (i.e., qualities and entities). By combining the sematic and lexicogrammatical dimensions, Slater (2004) presented a linear path mapping the progression from less sophisticated causal linguistic features to more sophisticated features, when she examined oral causal discourse produced by different groups and calculated the frequency of causal linguistic features. The comparison between causal languages of native primary students and of native high school students has revealed that the sophistication of causal language progresses following the developmental path. A similar pattern of development was observed in the causal discourses produced by ESL speakers and native English speakers at the high school level. The native English speakers at this level produced more general metaphoric entities, while the nonnative high school students used more temporal circumstances to express causality. Given that the developmental path of cause can support the validity judgments that rate one performance of causal discourse over another (Slater & Mohan, 2010), Ma and Slater (2015, 2016) utilized the path to validate scores generated by an automated writing evaluation (AWE) system. Their findings suggested that the progression of causal language coincide with teachers' intuitive judgments and supports AWE scores. Studies from the SFL perspective were concluded with the pedagogical implication that language teachers should scaffold students' language development by encouraging them to use causal language located at the high end of the developmental path of cause (processes and entities). In spite of sufficient attention from the SFL and language teaching research, causal language has yet to be analyzed quantitatively.

This study, therefore, intends to improve our understanding of distribution of causal language by fitting the right-truncated modified Zipf-Alekseev model to cause- effect language in a naturally occurring discourse. The majority of previous research on probability distribution of various linguistic entities investigated spoken and written discourse produced by native speakers of the target language, leaving the current researchers wondering whether

causal linguistic features produced by language learners also abide by the right-truncated modified Zipf-Alekseev model. With reference to Ouyang and Jiang (2018), we hypothesize that the probability distribution of causal linguistic features, regardless of speakers' nativeness/ non-nativeness or proficiency levels, follows the right-truncated modified Zipf-Alekseev model. We also expect to see a relationship between language proficiency levels and parameters *a* and *b*, as observed in Ouyang and Jiang (2018). To test our hypotheses, the following research questions are answered:

Question 1: Does the probability distribution of causal language produced by each of the four groups follow the right-truncated modified Zipf-Alekseev distribution?

Question 2: Does the probability distribution of causal language produced by the higher-level groups demonstrate higher parameter *a* and lower parameter *b* values?

2. Materials and Method

Native English-speaking students from primary grades (ages six/seven) and high school (ages from fourteen/fifteen), as well as non-native English-speaking students at the same age levels were asked to explain their knowledge of what they had been studying in their science classes. Ten hours of interviews were recorded, transcribed, and analyzed using the linear progression of causal language as the framework (see Table 1). As presented in Table 1, the sophistication of causal linguistic features increases from external temporal conjunctions (located at the top of Table 1) to general metaphoric entities (located at the bottom of Table 1). Table 1 provides explanations for the related SFL terminology and examples for each variation of causal language.

Table 1

Linear display of the developmental path of cause (adapted from Slater, 2004)

Features	Meaning	Examples
External temporal conjunctions	Conjunctions indicating time sequence	When, then…
External causal conjunctions	Conjunctions indicating causality	If, because, therefore…
Internal conjunctions	Logical conjunctions organizing text	Firstly, additionally, furthermore…
Temporal circumstances	Adverbials indicating time sequence	After…
Causal circumstances	Adverbials indicating causality	As a consequence, due to, through…
Temporal processes	Verbs indicating time	Follow, proceed…
Causal processes	Verbs indicating causality	Causes, contributes to…
Proof processes	Verbs indicating proof	Prove…
Temporal entities	Nouns indicating time	The beginning, the following
Causal entities	Nouns indicating causality	Cause, effect, consequence…
General metaphoric entities	Nominalization (noun transformed from a verb)	Reactant, product, circulation…

Table 2 presents the frequency lists of causal linguistic features produced by ESL and non-ESL high school and primary students. These data were originally published in Slater (2004).

Table 2

Frequency lists of causal language produced by different speakers
(adapted from Slater, 2004)

Cause-effect language	Primacy (non-ESL)	High school (non-ESL)	Primacy (ESL)	High school (ESL)
External temporal conjunctions	25.35	51.11	33.59	29.68
External causal conjunctions	29.11	12.81	26.46	30.53
Internal conjunctions	0	.28	0	0
Temporal circumstances	15.96	22.56	18.3	30.31
Causal circumstances	3.76	.56	0.5	1.47
Temporal processes	0	1.39	0	0
Causal processes	1.88	6.41	2.4	4.21
Proof processes	.94	.7	.51	0
Temporal entities	0	2.51	0	0
Causal entities	.94	4.46	0	0
General metaphoric entities	0	16.99	2.04	11.37

Note. Numbers have been normalized to occurrences in 1,000 words.

Since variations of causal language qualify as a diversification process, we hypothesize that the Zipf-Alekseev model captures their distribution in naturally occurring language (Hřebíček 1996, cited from Strauss & Altmann, 2006). Hřebíček used two assumptions:

The logarithm of the ratio of the probabilities P_1 and P_x is proportional to the logarithm of the class size, i.e. –

$$\ln\frac{P_1}{P_x} \propto \ln x\,.$$

(i) The proportionality function is given by the logarithm of Menzerath's law (hierarchy), i.e. –

$$\ln\frac{P_1}{P_x} = \ln(cx^b)\ln x\,,$$

yielding the solution –

$$P_x = P_1 x^{-(\ln(c+b)\ln x)}, x = 1, 2, 3, \ldots \quad . \tag{1}$$

As $\ln c$ is a constant, one can write –

$$P_x = P_1 x^{-(a+b\ln x)}, x = 1, 2, 3, \ldots \quad .$$

If (1) is considered a probability distribution, the P_1 is the normalizing constant; otherwise, it is estimated as the size of the first class, x = 1. Very often, diversification distribution displays

a diverging frequently in the first class, while the rest of the distribution behaves regularly. In these cases, one usually ascribes the first class a special value α, modifying (1) as –

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

where

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

Distributions (1) or (2) are called Zipf-Alekseev distributions. If *n* is finite, (2) is called a modified right-truncated Zipf-Alekseev distribution. In this study, we use the Altmann-Fitter software to fit the model to the four different groups of data – causal linguistic features produced by native and nonnative high school and primary students.

3. Results and Discussion

3.1 The Fitting Results by Zipf-Alekseev Distribution

The fitting results of causal linguistic features produced by the four groups of speakers are presented in Table 3 and Table 4. The observed and expected frequencies of cause-effect language use by high school and primary students are organized in a descending order. Figure 1 to Figure 4 display the graphic representations of fitting the causal language produced by four different groups of speakers to the right-truncated modified Zipf-Alekseev distribution.

As shown in Table 3, ESL high school students demonstrated less variety in terms of cause-effect language use than their native counterparts. In addition, ESL high school students used significantly less external causal conjunctions, and more external temporal conjunctions and temporal circumstances. For both ESL and non-ESL students at primary level, external causal conjunctions, external temporal conjunctions, and temporal circumstances rank as the most frequently used causal linguistic features.

Table 3

Fitting the right-truncated modified Zipf-Alekseev distribution to causal language produced by high school student

	high school (non-ESL)			high school (ESL)		
X[i]	Causal linguistic features	f[i]	NP[i]	Causal linguistic features	f[i]	NP[i]
1	External temporal conjunctions	51.11	51.45	External causal conjunctions	30.53	30.69
2	Temporal circumstances	22.56	25.86	Temporal circumstances	30.31	38.82
3	General metaphoric entities	16.99	15.52	External temporal conjunctions	29.68	18.38
4	External causal conjunctions	12.81	9.48	General metaphoric entities	11.37	9.01

5	Causal processes	6.41	6.01	Causal processes	4.21	4.68
6	Causal entities	4.46	3.94	Causal circumstances	1.47	2.56
7	Temporal entities	2.51	2.67	Temporal entities	0	1.46
8	Temporal processes	1.39	1.86	Temporal processes	0	.87
9	Proof processes	.7	1.32	Proof processes	0	.54
10	Causal circumstances	.56	0.96	Causal entities	0	.34
11	Internal conjunctions	.28	0.71	Internal conjunctions	0	.22
	a = 0.0911, b = 0.6523, n = 11, α = 0.4295, X^2 = 2.6622, P(X^2) = 0.7519, C = 0.022			a = 0.21, b = 0.912, n = 11, α = 0.2853, X^2 = 13.3735, P(X^2) = 0.0039, C = 0.1243		

In this and following similar tables: X[i] – the observed classes; F[i] – observed frequency; NP[i] – calculated frequency according to the modified right-truncated Zips-Alekseev distribution; *a*, *b*, *n*, and α – the parameters of the modified right-truncated Zipf-Alekseev distribution; X^2 – Chi-square; DF – degrees of freedom; P – probability of chi-square.

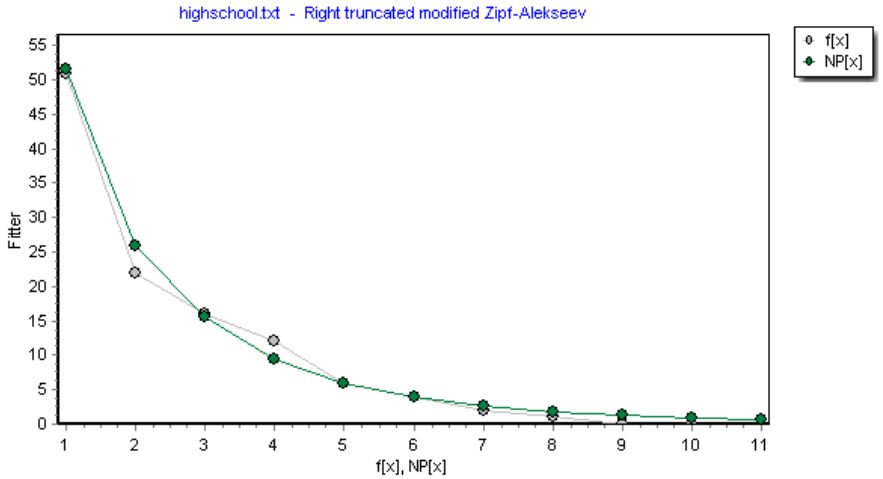


Figure 1. Fitting the modified right-truncated Zipf-Alekseev distribution to high school students' causal language

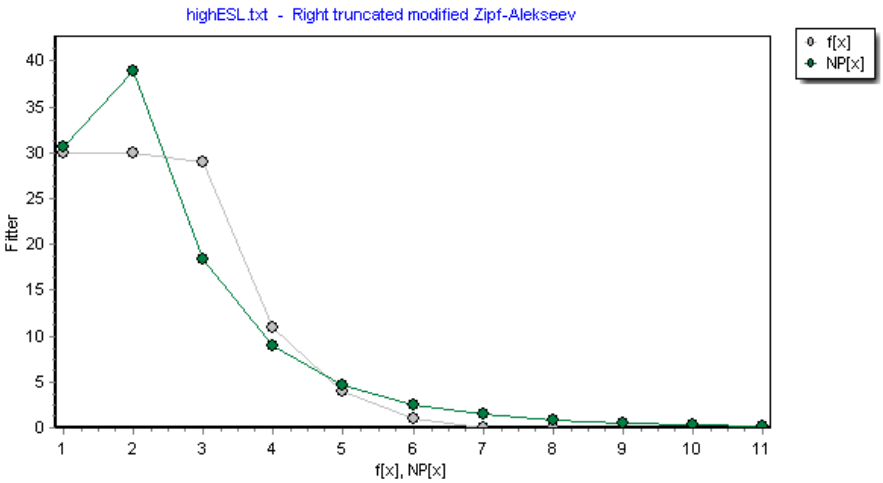


Figure 2. Fitting the modified right-truncated Zipf-Alekseev distribution to ESL high school students' causal language

Table 4
Fitting the modified right-truncated Zipf-Alekseev distribution to causal language produced by primary students.

	Primacy (non-ESL)			Primacy (ESL)		
X[i]	Causal linguistic features	f[i]	NP[i]	Causal linguistic features	f[i]	NP[i]
1	External causal conjunctions	29.11	29.47	External temporal conjunctions	33.59	33.91
2	External temporal conjunctions	25.35	27.82	External causal conjunctions	26.46	29.59
3	Temporal circumstances	15.96	11.34	Temporal circumstances	18.3	11.73
4	Causal circumstances	3.76	4.84	Causal processes	2.4	4.73
5	Causal processes	1.88	2.21	General metaphoric entities	2.04	2.02
6	Proof processes	.94	1.08	Proof processes	.51	.92
7	Causal entities	.94	.56	Causal circumstances	.5	.44
8	Internal conjunctions	0	.3	Internal conjunctions	0	.22
9	Temporal processes	0	.17	Temporal processes	0	.12
10	Temporal entities	0	.1	Temporal entities	0	.06
11	General metaphoric entities	0	.06	Causal entities	0	.04
	a = 0.2834, b = 1.0765, n = 11, α = 0.3781, X^2 = 2.4613, P(X^2) = 0.2921, C = 0.0316			a = 0.0131, b = 1.2658, n = 11, α = 0.4047, X^2 = 5.5074, P(X^2) = 0.0189, C = 0.0657		

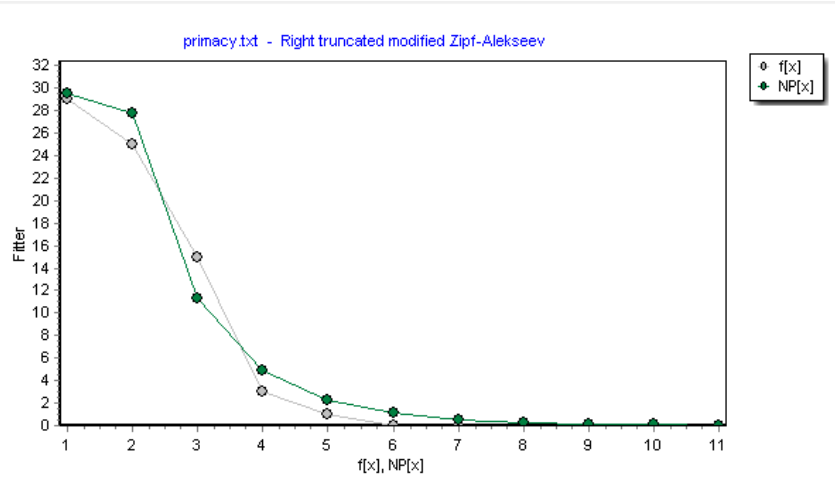


Figure 3. Fitting the modified right-truncated Zipf-Alekseev distribution to primary students' causal language

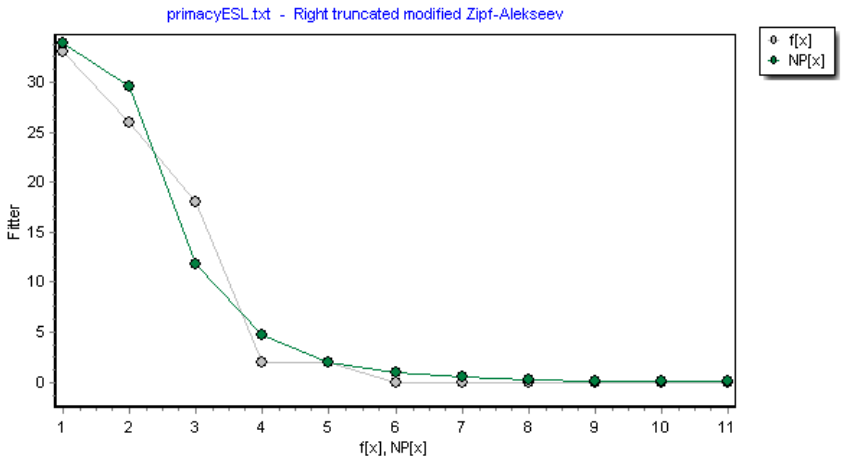


Figure 4. Fitting the modified right-truncated Zipf-Alekseev distribution to ESL primary students' causal language

The distribution of causal language in non-ESL high school students' speech abides by the right-truncated modified Zipf-Alekseev model (see Table 3). Native primary students' causal language use, though not fully-developed, also follows the right-truncated modified Zipf-Alekseev model (see Table 4). Nevertheless, causal language produced by ESL high school students', even though more developed than non-ESL primary school students' for using more metaphoric causal language (i.e., general metaphoric entities), does not follow the right-truncated modified Zipf-Alekseev model (see Table 3). As can be visualized in Figure 3, ESL high school students' frequency of using temporal circumstances is noticeably lower than the expected value, and their frequency of using external temporal conjunctions is higher than the expected value. These deviations from the expected values may have caused the falsification of this model. The causal language use by ESL primary students, with the lowest level of mastery of the English language among the four groups, does not follow the right-truncated modified Zipf-Alekseev model either (see Table 4), with the frequencies of external causal conjunctions, temporal circumstances, and causal processes obviously deviating from the expected values (see Figure 4). Different from Ouyang and Jiang's (2018) finding that dependency distances of language produced by native speakers and language learners are distributed following the right-truncated modified Zipf-Alekseev model, the results of current research suggested that the right-truncated modified Zipf-Alekseev model, to some extent, has the potential of differentiating between native and nonnative causal discourse. Therefore, instruction on causal language use should pay attention not only to using an increasing number of metaphoric features, but also to observing the probability distribution of causal linguistic features.

The unfitness in our finding could be attributed to language education that learners received. Language education has been widely known for focusing on how learners acquire grammar and grammatical sub-systems (Bestgen & Granger, 2014). This may explain why students in Ouyang and Jiang's (2018) research, from two high schools and one university in China, produced sentences following the right-truncated modified Zipf-Alekseev model syntactically. Receiving little attention from language education, causal linguistic features produced by nonnative speakers are less likely to abide by the right-truncated modified Zipf-Alekseev model as syntactic features do. Another reason that could explain the unfitness

of ESL primary students' language production is that proficiency level of these students is too low to produce fine-tuned causal linguistic features, the probability distribution of which follows the aforementioned model.

3.2 Parameters and English Proficiency

For causal linguistic features produced by native primary and high school students that abide by the right-truncated modified Zipf-Alekseev model, we observed a pattern of parameter *a* and *b* that partially supports Ouyang and Jiang's (2018) observation. Ouyang and Jiang (2018) found that with the increase of English proficiency, parameter *b* decreases, while parameter *a* increases. Similarly, in the current study, we observed that parameter *b* (*b* = 0.6523) of high school students' language production was lower than that of primary students' language production (*b* = 1.0765). However, different from Ouyang and Jiang's (2018) prediction, we observed that parameter *a* of primary students' language (*a* = 0.2834) was higher than that of high school students (*a* = 0.0911), indicating a decrease in parameter *a* with the increase of proficiency level.

Despite the conflicting result, the practice of differentiating and ranking students' proficiency levels based on parameters *a* and *b* provided an alternative method of evaluating students' causal language development. Slater (2004) generalized that students with higher proficiency level produce more metaphoric causal linguistic features. Reflecting the general tendency of causal language development, Slater's (2004) observation may not apply to individual causal linguistic features. For example, external temporal conjunctions and temporal circumstances (less metaphoric features) produced by non-ESL high school students outnumber those produced by non-ESL primary students, and ESL high school students produced noticeably more temporal circumstances (less metaphoric features) than ESL primary students. An examination of parameters *a* and *b* inarguably provides a more direct and precise method of evaluating students' causal language production.

4. Conclusion

This current study investigated the probability distribution of cause-effect linguistic features produced by four different populations, the non-ESL and ESL high school and primary students. Results show that, regardless of students' proficiency levels, native speakers' cause-effect language use abides by the right-truncated modified Zipf-Alekseev distribution, while the ESL students' does not follow the model. The results indicate that the right-truncated modified Zipf-Alekseev distribution can be employed to distinguish native and nonnative language productions. However, this observation was not consistent with the previous research finding out that dependency distances in native and nonnative speakers' language are distributed following the right-truncated modified Zipf- Alekseev distribution.

The current observation about parameter *b* lends support to an inverse relationship between parameter *b* and students' proficiency levels. Our result on parameter *a*, nevertheless, did not agree with the observation reported in Ouyang and Jiang (2018) in that parameter *a* decreases, rather than increases, with the increase of proficiency level. In view of the con-

flicting findings, future research could investigate model fit and trends in parameter changes by incorporating other linguistic entities, or recruiting populations with various language proficiency levels.

REFERENCES

Altmann, G. (1991). Modelling diversification phenomena in language. In: U. Rothe (ed.), *Diversification processes in language: Grammar* (pp. 33-46). Hagen: Rottmann.

Altmann, G. (1996). Diversification processes of the word. *Glottometrika*, 13, 105–120.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28–41.

Čech, R., Uhlířová, L. (2014). Adverbials in Czech: Models for their frequency distribution. In: G. Altmann, R. Čech, J. Mačutek, & L. Uhlířová (eds.), *Empirical approaches to text and language analysis* (pp. 49–49). Lüdenscheid: RAM-Verlag.

Chen, H., & Liu, H. (2014). A diachronic study of Chinese word length distribution. *Glottometrics*, 29, 81–94.

Hřebíček, L. (1996). Word associations and text. *Glottometrika*, 15, 12–17.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

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

Liu, H. (2012). Probability distribution of semantic roles in a Chinese treebank annotated with semantic roles. In: G. Altmann, P. Grzybek, S. Naumann, R. Vulanovic (eds.), *Synergetic linguistics: Text and language as dynamic systems* (pp. 101–108). Wien: Praesens Verlag.

Ma, H., & Slater, T. (2015). Using the developmental path of cause to bridge the gap between AWE scores and writing teachers' evaluations. *Writing & Pedagogy*, 7(2).

Ma, H., & Slater, T. (2016). Connecting *Criterion* scores and classroom grading contexts: A systemic functional linguistic model for teaching and assessing causal language. *CALICO Journal*, 33(1), 1–18.

Mohanty, P., Popescu, I.-I. (2014). Word length in Indian languages 1. *Glottometrics*, 29, 95–109.

Ouyang, J., & Jiang, J. (2018). Can the probability distribution of dependency distance measure language proficiency of second language learners?. *Journal of Quantitative Linguistics*, 25(4), 294–313.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

Slater, T.J.A. (2004). The discourse of causal explanations in school science. PhD Thesis, University of British Columbia.

Slater, T., Mohan, B. (2010). Toward systematic and sustained formative assessment of causal explanations in oral interactions. In A. Paran and L. Sercu (eds.), *Testing the untestable in foreign language education. Bristol* (pp. 259–272). Buffalo & Toronto: Multilingual Matters.

Strauss, U., Altmann, G. (2006). Diversification laws in quantitative linguistics. Retrieved from http://www.uni-trier.de/uni/fb2/ldv/lql_wiki/index.php/Diversificatio.

Yue, M., Liu H. (2011). Probability distribution of discourse relations based on a Chinese RST-annotated corpus. *Journal of Quantitative Linguistics*, 18(2), 107–121.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge: Addison Wesley.

Zhang, H., Liu, H. (2015). Quantitative aspects of RST rhetorical relations across individual Levels. *Glottometrics*, 33, 8–24.

Zhu, J., & Liu, H. (2018). The distribution of synonymous variants in Wenzhounese, *Glottometrics*, 41, 24–39.

Glottometrics 44, 2019, 87-93

Measuring Lexical Richness of the USA Presidents´ Inauguration Speeches

Hanna Gnatchuk[1]

Abstract: The present study deals with a quantitative and corpus-based study of inaugural speeches of the USA presidents in terms of lexical richness. The fifteen speeches have been analysed and the lexical richness of presidents has been compared by means of the one-way ANOVA. The results of the findings have shown us that the indices of the speeches by the USA presidents differ significantly. All the tests have been computed in statistical program R-Studio and in Python 3.

Key words: *English, stylometrics, diversity index, average words' repeats, ANOVA, Tukey (HSD) Test*

1. Introduction:

Some notes on stylometrics and quantitative features of a text

Stylometrics is considered to be the branch of applied linguistics. Its main task is to make a quantitative analysis of linguistic units of a certain register or functional style. Stylometrics as a term was firstly introduced by W. Ditenberg at the end of the XIX century, who was engaged with the detection of authors for Plato's dialogues. It is worth mentioning that the solution to the problem of author's attribution to a text is quite popular in linguistics, literary studies, sociological psychology, medical diagnostics or in criminal affairs. As the example, let us consider the importance of stylometrics in the juridical process. Here one deals with the detection of false evidence, the authorship of anonymous letters, contracts and so on, which are available in the juridical process.

Special attention should be paid here to the notion of *attribution of the text*, namely textual belonging to a certain group of texts. Such attributes as author, time and style are the types of the text attribution. In particular, the attribution of authors concentrates on the detection of the author of a text; time attribution deals with the time of writing a text and style – with the detection of the genre or style of a text.

In this case, a variety of statistical methods can be used. The attribution of the text was at first established by means of the frequency of words. Then the author's style was determined by means of the average number of words in a sentence, the number of words in a sentence, the number of sentences in the paragraph. Moreover, the word-stock of the texts of the writer is also taken into account. Here one pays attention to the repetitions, the meta-phorical meanings of words, mistakes and other features. It is worth mentioning that the style of each writer can be characterized by the relationship of more or less frequent words. This feature is taken by a reader as a rich or poor word-stock of texts.

[1] Universität Klagenfurt; agnatchuk@gmail.com

In general, there is a vast majority of quantitative parameters of texts which can be used in the analysis of a text: the length of a text, the number of word forms in a text, the diversity index, the average repeat of a word in a text, hapax legomena, the exception index, concentration index. In this examination, we shall deal with some of these quantitative parameters in terms of the inaugural speeches of the USA presidents.

1. **Diversity indices in the inaugural speeches of the USA presidents**.

In the present research we shall deal with two quantitative features of the text: lexical richness (diversity index or type-token ratio) and the average repeat of words in a text. Under lexical richness (diversity index) one understands a measurement procedure which determines the relationship between the number of different words in a text (or the number of types) to the number of the total length of a text (the number of tokens). Formula 1.1 illustrates this relationship:

$$R = \frac{V}{N} \qquad (1.1)$$

R is the diversity index;
V stands for the number of different words;
N is the total number of words in a text;

The higher R is, the richer is the vocabulary of a text and vice versa: the lower R, the poorer the vocabulary of a text. It is to be remarked that a comparison of richness computed in this way can be used only for texts of the same language, otherwise the extent of synthetism yields inappropriate results.

Another quantitative feature of a text is the average repeat of words. It shows the relationship between the length of a text (the total number of words) to the number of different words. This can be found with the help of Formula 1.2:

$$A = \frac{N}{V} \qquad (1.2)$$

A is the average number of words' repeats;
V stands for the number of different words;
N is the total number of words in a text;

The lower the average repeat is, the richer the vocabulary of a text is.

The present study is intended to reveal at first diversity indices and average repeat of words in the inauguration speeches of 15 USA presidents. In such a way, the corpus of our studies consists of 15 inauguration speeches, which can be found at https://en.wikisource.org/wiki/Portal:Inaugural_Speeches_by_United_States_Presidents. For the analysis of the tokens, types, diversity indices and average word repeat, we have used natural language tool kits in Program Python 3 (environment Anaconda). In such a way, Table 1 illustrates the findings for each president in terms of the above-mentioned quantitative parameters:

Table 1:
Lexical richness features for USA Presidents' inauguration speeches

Presidents	Tokens	Types	Diversity index	Average repeat
Obama (2009)	2726	938	0.34	2.9
Bush (2005)	2376	783	0.32	3.0
Clinton (1997)	2462	774	0.31	3.2
Bush(1989)	2713	793	0.29	3.4
Reagan (1985)	2946	924	0.31	3.2
Carter (1977)	1380	529	0.38	2.6
Nixon (1973)	2028	545	0.26	3.7
Johnson (1965)	1715	571	0.33	3.0
Kennedy (1961)	1546	570	0.36	2.7
Eisenhower (1957)	1917	622	0.32	3.1
Truman(1949)	2528	781	0.30	3.2
Roosevelt (1945)	637	280	0.43	2.3
Hoover (1929)	3890	1087	0.27	3.6
Coolidge (1925)	4442	1221	0.27	3.6
Harding (1921)	3756	1170	0.31	3.1

One can see that our table contains the number of tokens (text length), types (the number of different words), diversity index and average repeat. It is quite obvious that the higher the diversity index of a text, the lower the average repeat of words in a text is. For example, the average repeat of words in Roosevelt's inauguration speech is 2.3. This means that each word in his speech occurs on average 2 times. It is worth mentioning that his speech dates back to 1945, known as the end of the Second World War. The proportion of different words in his speech (diversity index) covers 0.43 or 43 % of the whole text.

Table 1 shows that the diversity index of Roosevelt speech is the highest whereas the average repeat of words is the lowest. In contrast, Nixon's speech has proved to be the lowest in terms of lexical diversity as well as the highest value of average repeat of words.

At this stage it would be relevant to group our findings into 3 groups. The first group (A) contains 5 speeches where diversity indices are the highest. To this group we shall refer Roosevelt (diversity index = 0.43, average repeat = 2.3), Carter (0.38, 2.6), Kennedy (0.36, 2.7), Obama (0.34, 2.9) and Johnson (0.33, 3.0). Three speeches belong to 60-s and 70-s years, one speech dates back to 1945 and one belongs to 2009. The second group (B) includes the speeches, whose diversity indices are a little lower from the first group: Bush (2005) (0.32, 3.0), Eisenhower (0.32, 3.1), Clinton (0.31, 3.2), Regan (0.31, 3.2) and Harding (0.31, 3.1). Three speeches belong to 80's, 90's and 2000's years, one speech to the 50's and one speech to the 20's. The third group (C) contains the speeches with the diversity indices lower than the second group: Truman (0.30, 3.2), Bush (1989) (0.29, 3.4), Hoover (0.27, 3.6), Coolidge (0.27, 3.6) and Nixon (0.26, 3.7).

The next step of our research will be to reveal whether there is a difference between at least two groups in terms of lexical diversity scores (indices). In this case it would be relevant to have a look at the boxplot of the distribution of lexical diversity between these groups in Figure 1

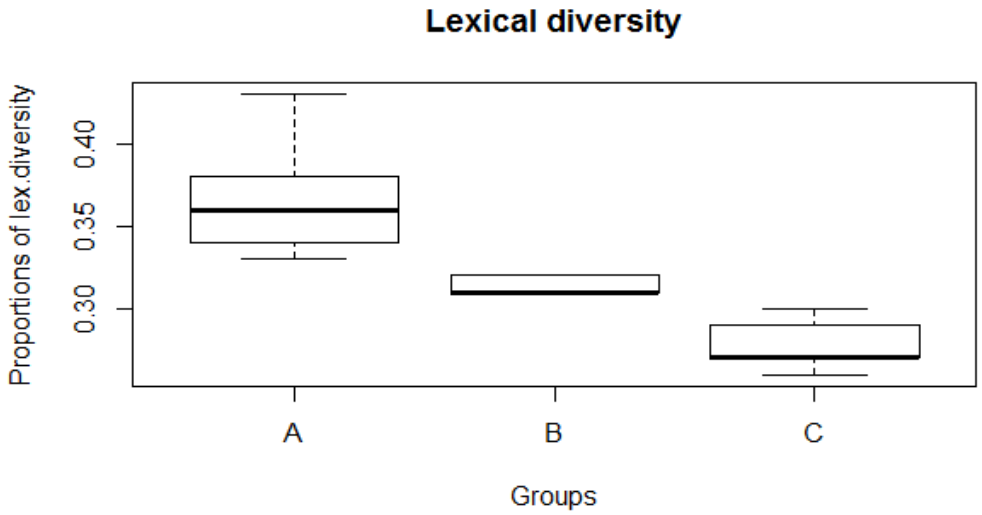


Figure 1: Lexical diversity

From Figure 1 one can see that the average means for Group A (Roosevelt, Carter, Kennedy, Obama, Johnson) is 0.36. The arithmetic means for B (Bush, 2005, Eisenhower, Clinton, Reagan, Harding) is 0.31 and 0.27 for C (Truman, Bush (1989), Hoover, Coolidge, Nixon). In order to reveal the difference between the means of these groups, we can perform the one-way parametric ANOVA. In this case our alternative and zero non-directional hypotheses are as follows:

H_1: there is at least two groups that differ in their average means.

H_0: there is no difference in the average means between groups.

Aiming to perform the one-way test, our research must meet four assumptions: independent observations in a sample, interval-scaled response variable, normality and the homogeneity of the variances. All these steps (as well as the performance of appropriate statistical tests) have been described in Book "How to do Linguistics with R: Data exploration and statistical analysis" (Levshina, 2015:176-181). The requirement of the independence of observations and interval-scaled variable are fulfilled. As far as the normality is concerned, we must perform the Shapiro Test according to Formula 1.3

$$W = \frac{(\sum_{i=1}^{n} aixi)^2}{\sum_{i=1}^{n}(x_i - mean(x))} \qquad (1.3)$$

x_i = ith smallest value of x
a_i = Shapiro constant

Using Formula 1.3 we have received the following results:

Table 2:
The significance levels for three groups

Group	p-value
A	0.54
B	0.00
C	0.48

When p-value is higher than 0.05, we can consider that the normality is met. Group A (p = .54) and B (p = .48) have met the requirement of a normal distribution. Only group B (0.00) has turned out not to meet the assumption of normality. This will have the influence on the choice of the ANOVA test.

The last assumption concerns the homogeneity of variances or homoscedacity, which is calculated by means of the Bartlett-Test according to Formula 1.4:

$$x^2 = \frac{2.303}{c} * [(N_{ges} - p)*\ln(MS_{within}) - \sum_{j=1}^{p}(nj - 1) * \ln(s^2_{j})] \qquad (1.4)$$

with df = p-1 where

$$c = 1+ \frac{1}{3*(p-1)} * (\sum_{j=1}^{p} \frac{1}{nj-1} - \frac{1}{N_{ges-1}}) \qquad (1.5)$$

with
N_{ges}: the total number of all research items
n_j : the number of items in the j^{th} group;
p: the number of groups;
MS_{within}: the means square sum within the groups;
s^2_j: the variance within the groups;

The Bartlett-Test has shown that the variances are homogeneous with $x^2 = 4.195$, p = 0.12. This means that the requirement of homogeneity is met.

After checking the necessary assumptions for the ANOVA, we can test our alternative and zero hypotheses. In this case, it is worth mentioning that there is only one violation in the normal distribution. This presupposes the usage of the Kruskal-Wallis one-way ANOVA according to Formula 1.5:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{Ri^2}{ni} - 3(n+1) \qquad (1.6)$$

n – the sample size;
k – the groups in the data;
R_i – the sum of ranks for group *i*;
H – the value of the Kruskal-Wallis one-way ANOVA;

The Kruskal-Wallis Test has turned out to be statistically significant with x^2 (2) = 12.63, p = .001. As the significance level is lower than p = .05, we can observe a significant difference in at least two groups in terms of lexical diversity. The zero hypothesis is rejected and the alternative hypothesis is accepted.

The next step of our research is to reveal between which groups one there is a significant difference with the help of the Tukey Honest Significant Difference (HSD) test (Formula 1.6).

$$T = \frac{M1-M2}{\sqrt{\frac{MSw}{n}}} \qquad (1.7)$$

M are the group means;
n is the total number of observations in a group;
MSw is the Mean Square within;
T is the Tukey's value;

The Tukey HSD test has been computed in our research in the statistical program R-Studio and we have received the following the values in Table 3:

Table 3
The results of the Honest Significant Differences (HSD)

	difference	**lwr**	**upr**	**p.adj**
B-A	-0.054	-0.096	-0.011	0.013
C-A	-0.090	-0.132	-0.047	0.000
C-B	-0.036	-0.078	0.006	0.097

From Table 3 one can see that there are differences in group means of B-A (p = .013) and B-A (p = .000). The greatest difference is between groups C-A. This is well illustrated in Figure 1. It is a negative difference in so far as the mean lexical diversity in C is higher than in A. This difference (C-A) is also statistically significant considering a small p-value in the last column of Table 3. The result for B-A are similar. The negative difference between B-A (-0.054) shows that the mean value of the diversity index variable of B is smaller than A. This difference has proved to be statistically significant. The difference between C and B is also negative, but not statistically significant. Moreover, the columns **lwr** is the lower end points of the interval, **upr** is the upper limit. These intervals can be visualized in Figure 2:

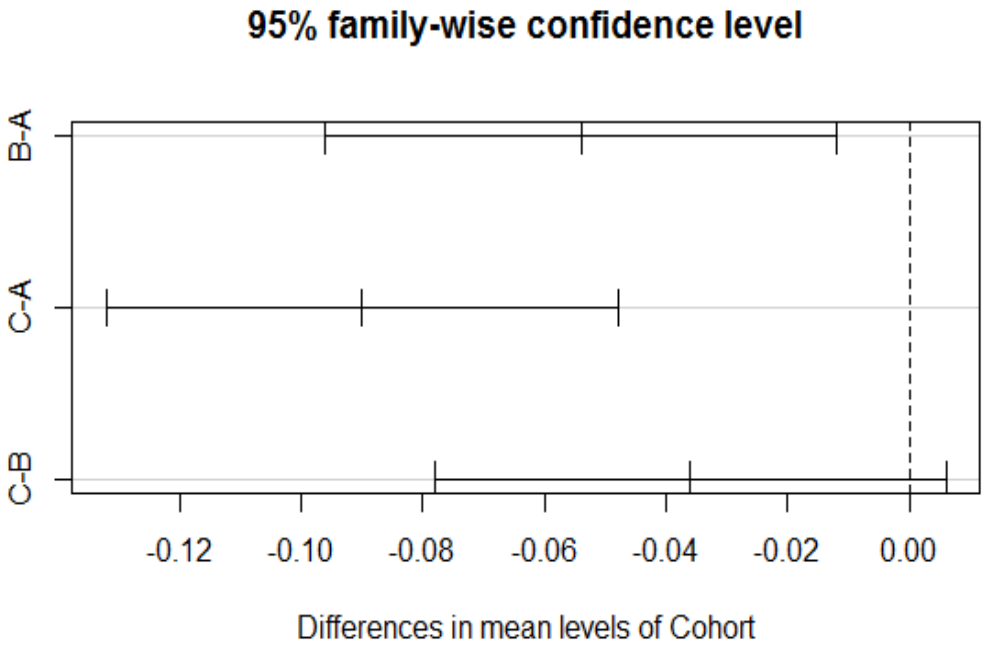


Figure 2. Confidence intervals of differences between group means

We can conclude here that the lexical richness of the USA Presidents of three analysed groups is different for two groups (B-A, C-A). We have found in our research the diversity indices in the inauguration speeches of the USA presidents. The highest diversity index and the lowest words' repeats have proved to be in the speech by Roosevelt (1945). It is possible to notice that his speech dates back to the end of the World War. Moreover, this president was elected more than for two terms. The lowest diversity index is found for Nixon's inauguration speech (1973), who was known to be the only one to resign from the president's position. We have decided to group these diversity indices in three groups according to their frequency. In

Group A we refer 5 speeches with high diversity indices. The next group (B) contain the speeches with middle frequencies and the last group (C) – with the lowest frequencies. We intended to reveal whether there is any difference in the lexical richness of the inauguration speeches by 15 USA presidents. The difference has been detected between B-A (middle-high frequent indices and low-high indices). This allows us concluding that lexical richness is quite different in the USA presidents. It would be interesting for the further research to investigate the other quantitative features as hapax legomena, exception indices, concentration indices and other features as well as compare them.

REFERENCES

Buk, Solomija. (2008). *Osnovy statystychnoji lingvistyky*. Lviv: LNU imeni Ivana Franka.

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

https://en.wikisource.org/wiki/Portal:Inaugural_Speeches_by_United_States_Presidents

Glottometrics 44, 2019, 94-99

Script Complexity in Indian Languages

Panchanan Mohanty, India
Ioan-Iovitz Popescu, Romania
Gabriel Altmann, Germany

Abstract. The present article shows a method for computing script complexity known from other articles and presents the computing in seven Indian languages. At the same time it is a hint: if the scripts should be unified, one should choose the simplest script. The simplest script means the choice of the simplest letters, not the whole script. In print, letters are chosen because of the need for ornamentality; but for schools the simplest forms should be used.

Keywords: *Script, complexity, Odia, Bengali, Devanagari, Gujarati, Telugu, Tamil, Gurumukhi*

Complexity is a concept used in all developed sciences. In linguistics any entity can be scaled according to this property, beginning with writing systems and ending somewhere in semantics and text science. The views may be quite different. One may distinguish local or global complexity, e.g. phoneme "as such" is global, its realizations in speech are local; the global complexity corresponds to Riedl's (1975) "norm", the local complexity is rather the "use" or "realization". Further, one may consider complex an entity having many parts, e.g. a sentence with many clauses is surely more complex that a simple sentence; a letter having many strokes is in any case more complex than a letter with only few. A word may be phonetically simple but semantically very complex, e.g. the English "and" having a simple form but more than 40 grammatical and semantic functions. Complexity may also mean a multilevel hierarchy in a system and also the kinds of links between entities, parts or levels. This aspects is, of course, the most challenging enterprise because it may show the inner structure of dynamic systems, chaotic systems, communication systems, etc.

It must be emphasized just at the beginning that complexity is no "inherent" property of things, it is the property of our concepts with which we try to capture the reality and find orientation in it. The same holds of all other properties which help us to identify and characterize the objects and find their reflections in our mind. Some concepts are important for their theoretical impact, other ones for their practical aspect. Script complexity is important for learning to read and write. But here not only the script complexity itself is important, also the relation of written form to the words they symbolize. In some languages, one performs – from time to time – a writing reform, e.g, in German, Slovak, Indonesian, other languages adhere to classical form and get thereby ever more hieroglyphic, e.g. English, French or Irish.

The study of script complexity is important also from the cultural-historical point of view. In general, we may state that complexity decreases, though in some languages it is quite constant. In dead languages nothing changes any more but in living languages one should perform a reform of orthography.

For some national states in which different scripts are used like in India the problem of complexity is lethal. There is a number of multilingual speakers who must, unfortunately, learn not only the other language but also its script. Besides, learning to write English is

perhaps the greatest difficulty leading to many simplifications. One can find them in the e-mails of English speakers from different parts of the world.

When analyzing the complexity of a script, one must consider merely the not compound forms. For Chinese, one should analyze only the individual parts of signs, whose inventory is restricted, not the complete signs and their combinations whose inventory is potentially infinite. If a language uses written syllables, as e.g. the Japanese *katakana* or *hiragana*, then all syllables must be analyzed.

The simplest way of measuring script complexity is the use of Altmann's system (2004) distinguishing three categories: dots, straight lines and arches. An arch begins somewhere and ends at a point of inflection hence a non-straight line can consist of several parts.

There are also three kinds of connections between these categories: continuous connection between two parts of a curve (like e.g. in O), crisp connection where a line touches another line (like e.g. in T), or crossing (like e.g. in +, X, etc.). The scaling is shown below. Separate scores can be added to signs if they are filled. For example, ► can be considered a dot but also a triangle with three sides, three contacts and a filling. Signs of this kind are usual in hieroglyphic scripts. They seem to be simple but their complexity degree depends on our evaluation.

Form	**Point of any size**	**Straight line of any size and direction**	**Arch of any size and direction**[1]
Value	1	2	3
Examples	• ▪ ►	−/ | \	⁀ ⁔ ⁀ ⁀ ⁔ ∩ ∪ ⊃ ⊂
Contact	**Continuous**	**Crisp**	**Crossing**
Value	1	2	3
Examples	O ~	┐┘ F ┬┴〈∠	× + ≠

This scaling can be extended by taking into account also the positions of components in the sign, e.g. low – mid – high, or left – mid – right, but usually this is not necessary. It depends on the kind of script whether one considers further scales.

Any script, even hieroglyphs, can be evaluated using these two scales (cf. e.g. Hegenbarth-Reichard, Altmann 2008; Sanada, Altmann 2008). In order to show some examples, consider the Latin "N" (printed in Arial): it has three straight lines (3*2) and two crisp connections (2*2) yielding 6 + 4 = 10. The letter "O" consists of two arches (2*3) and two continuous connections (2*1), hence 6 + 2 = 8.

For practical purposes not only sign simplicity is important. If the correspondence "sound vs. sign" is not 1:1, then learners get difficulties and the learning of the written form of the language becomes a problem. In extreme cases, the writing of each word must be learned separately. This is not better in English than in Chinese.

Finally, the distinctivity of signs (cf. Antić, Altmann 2005) plays a very important role. The signs may be simple but if they are not sufficiently distinguished from other signs, the writing gets ineffective. This concerns, of course, the written, not the printed form of script: in printed version, all signs have their prescribed form but in handwriting everybody has his problems, especially with the script of other persons. In Japanese, one must learn two sets of signs: the official version and the slightly simplified handwritten version whose distinctivity is diminished.

[1] Eden (1961) and Eden and Halle (1961) distinguish *hook*, *arch* and *loop*, but for the purposes of measuring complexity they need not be distinguished, as their complexity is equal.

In modern times, writing is not a sign of cultural development of a nation but a sign of smaller or greater learning effort for school-children. On the other hand, a change of the script, e.g. unification of several scripts or taking-over of another script, would mean increased effort with learning to understand the own past culture.[2] But this evil is smaller because it is a normal way of evolution. The Egyptians use today the Arabic writing system, the Germans took over the Latin script, etc. As of India, the unification of scripts which developed from a common source would mean a much smaller effort for children who are to a great extent bi- or trilingual and in the school they must learn also the Latin script.

In order to evoke some ideas we compute the complexity of 7 Indian writing systems. The primary data consist of the alphabets of these seven scheduled and major Indian languages- Bengali, Devanagari (for Hindi), Gujarati, Gurmukhi (for Panjabi), Odia (earlier Oriya), Tamil, and Telugu. We will compare the printed forms and use the simplest version, which means a version without serifs, i.e. Baraha fonts. Each serif would increase the complexity by a straight line (= 1) and a crisp contact (= 2), i.e. each serif has an additional weight of 3.

As can be seen, there are differences which can be expressed in various ways. One can compute the usual moments, the repeat rate, Shannon's entropy, or one can choose a two-dimensional characterization using Ord's criterion (1972). Considering the Oriya script, which has already been characterized (cf. Mohanty 2007), we obtain

mean = 22.7556
m_2 = 39.9180
m_3 = 81.9714
I = 1.7542
S = 2.053

where $I = m_2/mean$, $S = m_3/m_2$, and m_2 and m_3 are the central moments. All values of the script systems are presented in Table 1.

Table 1
Ord's criterion of script complexity

Language	mean	m_2	m_3	I	S
Odia	22.7556	39.9180	81.9714	1.75	2.05
Bengali	20.7674	28.9227	45.4935	1.39	1.57
Devanagari	19.9778	28.8217	97.8992	1.44	3.40
Gujarati	14.7111	34.1165	249.8362	2.32	7.32
Telugu	20.6889	32.6588	139.0005	1.58	4.26
Tamil	22.7500	81.1875	366.4419	3.57	4.51
Gurumukhi	21.2326	45.0157	235.3253	2.12	5.23

The place of a script system in the (I,S)-space is presented in Figure 1. As can be seen, Gujarati has the simplest script signalized by the mean, and Odia has the most complex script.

[2] This would be the case especially in Chinese: today, South-Chinese, North-Chinese and Japanese can communicate with one-another writing the signs on their palms, though their languages are quite different.

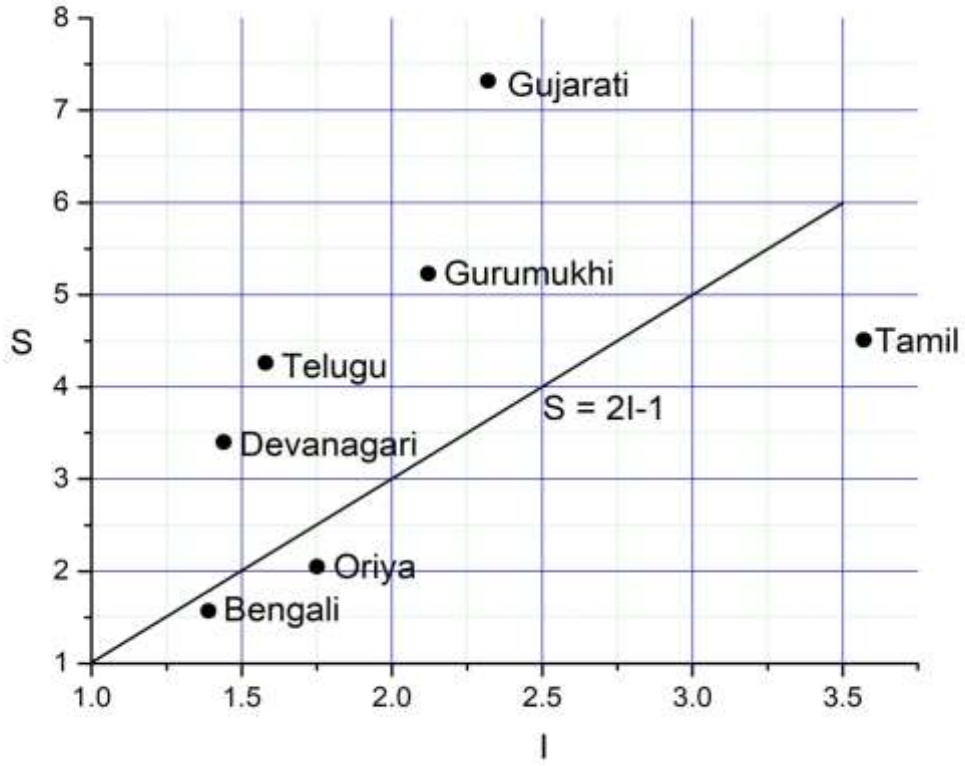


Figure 1. Ord's criterion for 7 Indian script systems

Another characterization of script complexity is the finding of a common model – a distribution or a function – expressing the course of complexities. To this end we consider the smallest and the greatest complexity value and find minC = 7 and maxC = 42. We construct 5 classes, namely 7-13, 14-20, 21-27, 28-34, 35-42 and call them 1,2,3,4,5, leaving the highest class to be a little greater than the other ones. In Tamil, we insert the isolated complexity 6 into class 7-13, and the isolated class 43 into class 35-42.Counting the complexities for Odia, we obtain the results presented in Table 3. As can be seen, the curve is concave and slightly asymmetric. There are many functions which can be successfully used for capturing the given sequence. However, in Gujarati we see the decreasing function for which a concave model is not adequate. That means, some Indian writing systems strive for simplicity as displayed by the monotonic decrease of complexities. Since we want to find a common model for all, we use the Lorentzian function in which the parameter *b* shows a kind of trend. For Odia one can see the results in Table 2, for the other systems in Table 3 and 4.

Table 2
A model for the Odia complexities

Complexity	Classes	Number of signs	Theor
7-13	1	3	4.31
14-20	2	13	12.49
21-27	3	20	20.23
28-34	4	8	7.13
35-42	5	2	2.84
a = 21.8808, b = 2.7520, c = 0.8673, R^2 = 0.9844			

Table 3
Other systems

Classes	Bengali	Devanagari	Gujarati	Telugu	Tamil	Gurumukhi
1	4	5	25	4	4	4
2	15	32	14	20	9	17
3	19	12	4	17	8	15
4	5	6	2	3	4	5
5	-	-	-	1	3	2

Table 4
Fitting the Lorentzian function to complexities

Class	Bengali	Devanagari	Gujarati	Telugu	Tamil	Gurumukhi
1	4.08	5.83	25.00	3.64	4.47	4.69
2	14.98	31.92	13.99	20.03	8.67	16.84
3	19.01	12.54	4.11	16.95	8.24	15.19
4	4.94	3.16	1.80	3.31	4.14	4.30
5	-	-	-	1.28	2.10	1.78
	a = 25.9226 b = 2.5863 c = 0.6861 R^2 = 0.999	a = 38.8254 b = 2.2432 c = 0.5227 R^2 = 0.9808	a = 28.8295 b = 1.2753 c = 0.7036 R^2 = 0.9998	a = 39.9788 b = 2.4613 c = 0.4623 R^2 = 0.9990	a = 9.6285 b = 2.4474 c = 1.3486 R^2 = 0.9583	a = 23.5768 b = 2.4597 c = 0.7270 R^2 = 0.9943

As can be seen, the Indian scripts follow a common model of complexity. The Lorantzian function is defined as

$$y = \frac{a}{1+\left(\frac{x-b}{c}\right)^2}$$

However, one could apply other functions, too, but it depends on the other scripts which must be first evaluated. The above result is satisfactory.

Our aim was to show that a special property of script – its complexity, as defined by us – is not a chaotic property but behaves by abiding to some laws. If a script develops then it surely develops not only towards simplicity which is a requirement of writers, but also towards distinctivity which is a requirement of readers. The functions applied to capture these two requirements should contain at least these two aspects We presented the Lorentzian function which can be derived from the general theory (cf. Wimmer, Altmann 2005). Historically, each scripts contains remnants which change very slowly and one must add some boundary conditions. For toher properties one should consult special books (cf. Altmann, Fan 2008)

REFERENCES

Altmann, G. (2004). Script complexity. *Glottometrics* 8, 68-73.

Altmann, G., Fan, F. (2008). *Analyses of Script. Properties of Characters and Writing Systems.* Berlin/New York: Mouton de Gruyter

Antić, G., Altmann, G. (2005). On letter distinctivity. *Glottometrics 9, 46-53.*

Hegenbarth-Reichard, I., Altmann, G. (2008). On the decrease of complexity from hieroglyphs to hieratic symbols. In: Altmann, G., Fan, F. (eds.), *Analyses of script: 101-110.* Berlin: de Gruyter.

Mohanty, P. (2007). On script complexity and the Oriya script. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text.* Berlin, New York: Mouton de Gruyter.

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

Riedl, R. (1975). *Die Ordnung des Lebendigen*. Hamburg: Paul Parey.

Sanada, H., Altmann, G. (2008). On two simplifications of the Japanese writing systems. In: Altmann, G., Zadorozhna, I., Matskulyak, Y. (eds.), *Problems of General, Germanic and Slavic Linguistics: 493-502.* Chernovcy: Books-XXI.

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.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language.* 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis.* 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1.* 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences*. 2015, IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5.* 2015, III+146 pp.
22. P. Zörnig et al. *Positional occurrences in texts: Weighted Consensus Strings.* 2016. II+179 pp.

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.