Structure, Function and Process in Texts

Edited by

Lu Wang, Reinhard Köhler, Arjuna Tuzzi.

RAM-Verlag 2018

Wang, Lu
Köhler, Reinhard
Tuzzi, Arjuna

Structure, Function and Process in Texts

© Copyright 2018 by RAM-Verlag

Publisher: RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.eu

The publisher cannot be held responsible for any linguistic errors in the book: Such responsibility is only up to the authors.

ISBN: 978-3-942303-56-9

Contents

Preface

Founded in 1994, the International Quantitative Linguistics Association (IQLA) aims at promoting the development of theoretical, empirical, and applied Quantitative Linguistics (QL) and fostering an effective communication among scientists that work in this highly interdisciplinary field. This multidisciplinary background enables us to share methods and findings across the boundaries of the individual sub-disciplines, languages, and methodological areas. Quantitative linguistics cannot only be characterised as a strongly co-operative science but this property can be considered as an essential principle.

IQLA organizes on a regular basis the QUAntitative LInguistics COnference (QUALICO) that represents the most important meeting for IQLA members and for QL scholars. The first eight conferences in this series took place in: Trier (1991), Moscow (1994), Helsinki (1997), Prague (2000), Athens, Georgia (2003), Graz (2009), Belgrade (2012) and Olomouc (2014). The 9th meeting, Qualico2016, took again place in Trier. The scientific program was organized over 3 days in August 24-28 in two parallel sessions. The lecture halls were rented in the Deutsche Richterakademie (German Academy of European Law).

This book includes a selection of the papers presented at Qualico2016 from various areas of QL. The contributions highlight the richness of this branch of research and the wide range of different topics, methods and approaches available in QL.

A substantial amount of contributions are focussed on research in morphology. ***Klyshinsky, Logacheva and Nivre*** study morphological ambiguity, aiming at assisting natural language processing, by defining six types of word ambiguity and computing their distributions in nine languages. In order to measure grammatical ambiguity, ***Wang and Guo*** introduce polyfunctionality, which refers to the number of parts of speech a word can be attributed to. They suggest that a universal model can capture the distributions on data from German, Dutch, English and Chinese. ***Galieva and Nevzorova*** report an attempt to assess the morphological complexity of the Tatar language. ***Pawłowski, Topolski, Malak, Kocoń and Marcińczuk*** compare the differences between the distributions of vocabulary in the entire corpus and in its subcorpora (consisting of individual parts of speech), on data from Polish and explain the reasons. ***Vulanović and Ruff*** present a formula, based on set theory, to measure how much a linguistic system violates the One-Meaning-One-Form principle.

Two contributions to this volume involve stylometrics. ***Cortelazzo, Nadalutti, Ondelli and Tuzzi*** work on an Italian case of authorship attribution and thorough methods of

text clustering investigate the distinctive traits of Elena Ferrante's writing style with a specific focus on the presumed similarity with other novels written by Italian authors from Naples and its surroundings. The study proposed by ***Sun and Jin*** on a Japanese novel *Otome no minato*, which is published under the name of Yasunari Kawabata but disputed to be written by Tsuneko Nakazato. This study identifies the authorship by examining stylometric features such as part-of-speech bi-grams, particle bi-grams and phrase patterns and classifiers such as correspondence analysis, hierarchical cluster analysis, etc.

The contribution by ***Embleton, Uritescu and Wheeler*** deal with a dialectometric problem. They test the relationship between the linguistic distance and geographic distance on data from a Romanian dialect. Their study involves as an innovative feature, taking the popularity of telecommunication into consideration: The researchers measure geographic distance by both travel distance and travel time.

Sanada demonstrates a syntactic research on Japanese complement and adjunct. The quantitative properties of the two sentence components including frequency, length and position are investigated in various aspects.

Benešová, Faltýnek and Zámečník propose a discussion on functional explanation in quantitative linguistics, especially in synergetic linguistics. The authors focus on the philosophic reflection on epistemological and methodological bases and suggest a reconstruction of the functional model of explanation in synergetic linguistics.

The conference was evaluated both by the organisers and the participants as another successful event.

The Editors,

Lu Wang, Reinhard Köhler, Arjuna Tuzzi

Authorship Attribution and Text Clustering in Contemporary Italian Novels: Does Elena Ferrante's and Domenico Starnone's regional origin play a role?*

Michele A. Cortelazzo[1], Paolo Nadalutti[2], Stefano Ondelli[3], Arjuna Tuzzi[1]
[1] University of Padova, Italy
[2] Interdisciplinary Text Analysis Group (www.giat.org), Italy
[3] University of Trieste, Italy

Abstract
The aim of authorship attribution methods is to exploit observable textual features to identify the distinctive traits of an author's writing style. In quantitative approaches, the problem of authorship attribution is often tackled in terms of the search for a reliable procedure to measure the similarity between texts and for a consistent text clustering method. Among several proposals available in the scientific literature, this new study explores a large corpus of contemporary Italian novels and focusses on a famous case of disputed authorship. The object of the research is Elena Ferrante's work, with a specific focus on the presumed similarity with other novels written by Italian authors from Naples and its surroundings. Although we expected that the authors' geographical origin should play a role in terms of topic selection and linguistic features, results show that regional traits are not particularly relevant: of all the 10 candidates taken into consideration, the only author who can be considered "textually close" to Elena Ferrante is Domenico Starnone.

Keywords: *text clustering, content mapping, authorship attribution, contemporary Italian novels, Elena Ferrante*

Introduction

In 2013 our research group (Cortelazzo et al. 2013) published a study aimed at testing and improving the performance of Dominique Labbé's method to calculate intertextual distance (Labbé and Labbé 2001; 2007). In order to offset the impact of text size differences, we proposed a new iterative procedure based on repeated measures of the distance between text chunks of the same size. We tested the revised method on a corpus of 160 Italian novels written by 101 different authors: 33 writers contributed to the corpus with more than one work (92 novels) and 68 with only one novel. Among the latter, Elena Ferrante's *L'amore molesto* [Troubling Love] (1992) and Domenico Starnone's *Via Gemito* (2000) appeared to be very similar in terms of their mutual intertextual distance. Figure 1 shows the distribution of the average values of intertextual distance for pairs of novels written by the same author (histogram) and different authors (curve). For each novel, 10,000 word-token chunks were extracted 200 times to calculate 200 intertextual distance values (and the mean distance) for

* Address correspondence to: Prof.ssa Arjuna Tuzzi, Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata, via M. Cesarotti, 10/12, 35123 Padova, Italy. Email: arjuna.tuzzi@unipd.it

each pair of novels. Calculations involved all (grammar and content) words in text chunks. The dot and the arrow on the horizontal axis highlight the distance between Ferrante (1992) and Starnone (2000): undeniably, we are more likely to obtain this value (0.416) when measuring the distance between two novels written by the same author rather than by two different authors.

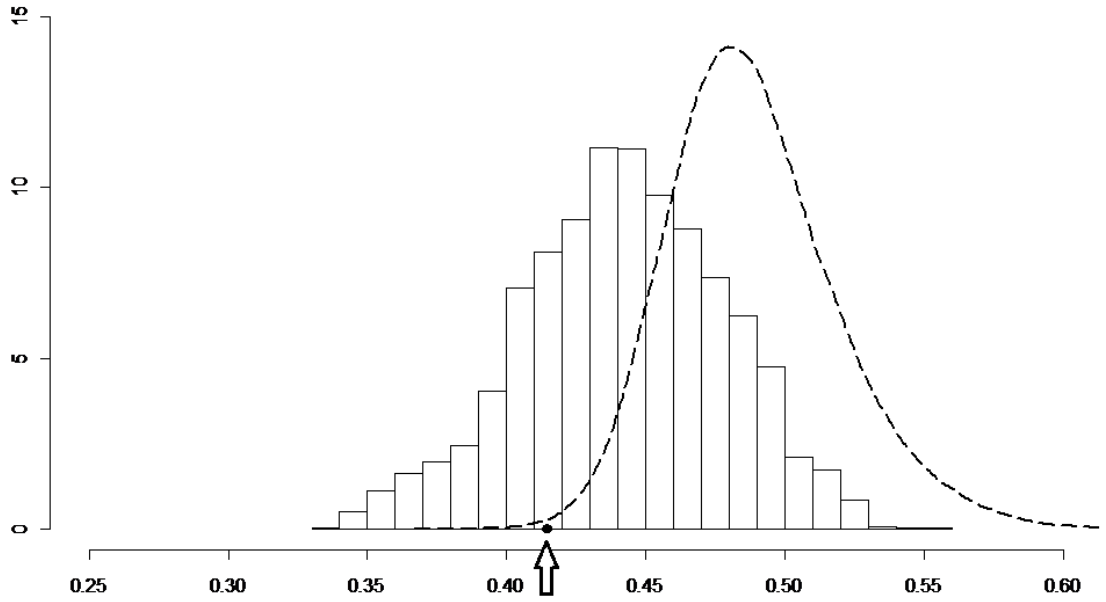

Figure 1. Corpus of 160 Italian novels (Cortelazzo et al. 2013). Distribution of intertextual distances for novels written by the same author (histogram) and different authors (curve). Calculations based on 200 replications involving text chunks of 10,000 word-tokens in length.

The objective of the research conducted in 2013 was testing the reliability of intertextual distance for the purpose of authorship attribution. Since we were comparing authors included in the corpus with more novels and authors only represented by one novel, Ferrante and Starnone were included in the latter group only to check whether pairs of different authors emerged and, among the thousands of pairs to be considered (160*159/2 = 12,720), their remarkable similarity went unnoticed. After all, we were not focussing on Elena Ferrante at all, but things changed in the following years as she started to draw the attention of the general public.

However, 25 years after the publication of her first – very successful – novel, despite the heated debate on her true identity, the Italian Academia does not seem to be particularly interested in Elena Ferrante. The relevant literature available is rather limited: mention can be made of a book edited by Russo Bullaro and Love (2016) and a number of papers, mostly published outside Italy, such as Alsop (2014); Bakopoulos (2016); Bovo-Romoeuf (2006); Ceccoli (2017); Chemotti (2009); Crispino and Vitale (2011); Deutsch (2015); Dow (2016); Reyes Ferrer (2016); Lee (2016); Milkova (2013); Mullenneaux (2007); Segnini (2017).

Elena Ferrante is a fictitious name and we still do not know who is hiding behind that pseudonym. Her pen name has become famous internationally but the mystery of her true identity remains, although some claim they have already solved it and point to Domenico Starnone, e.g. Luigi Galella (2005; 2006), Simone Gatto (2016), Rastelli (2016), Cortelazzo and Tuzzi (2017), Tuzzi and Cortelazzo (2018). Two exceptions are provided by Marco Santagata (2016) who, based on biographical data contained in her novels, concluded that Elena Ferrante actually is Marcella Marmo, a Neapolitan historian, and Claudio Gatti (2016), who identified the person who receives royalties from Ferrante's book sales as Starnone's wife, Anita Raja.

Of course, it may be claimed that Neapolitan authors setting their novels in Naples may share – at least to a certain extent – the same topics, and this may impact their mutual textual distance. Along with contents and settings, from a purely linguistic perspective geographical (or diatopic, as linguists term it) variation plays a major role in Italian (probably greater than in other European languages): in addition to occasional local dialect insertions, speakers or writers from the same region usually share a number of lexical and syntactic traits which differ from standard Italian and make them recognizable and more similar to each other. Since the Neapolitan cultural and linguistic background is considered crucial to understand Elena Ferrante's work (Benedetti, 2012; Caldwell, 2012; Cavanaugh, 2016; Falotico, 2015; Librandi, *in press*; Ricciotti, 2016), the common geographical origin of two Neapolitan authors may be responsible for the proximity of their works when pairs are formed in a corpus according to a lexically based measure of similarity such as Labbé's intertextual distance.

Starnone himself has repeatedly argued that the similarities between his novels and those written by Elena Ferrante simply mirror their common Neapolitan background. His stance is best illustrated in the last section of his book *Autobiografia erotica di Aristide Gambia* (Starnone 2011: 432):

> "Ci sono le caratteristiche regionali, - argomentai, - e, diciamo, storicosociologiche. Io e Ferrante abbiamo in comune la Campania, Napoli, gli anni Cinquanta, l'ambiente piccolo borghese, gli stessi oggetti d'epoca, la stessa eco dialettale nella frase. Per forza che qualche somiglianza c'è […].
> Certo. Galella ha mostrato che due scrittori molto diversi - uno di sesso maschile incline all'ironia e l'altra di sesso femminile incline ai sentimenti profondi - possono avere tratti in comune che dipendono dall'area dentro cui sono cresciuti. È interessante. Sono cose di cui la critica letteraria parla poco o niente, ormai. Ma non è sufficiente per costruirci un'intera pagina e segnalare addirittura il pezzo in prima."

> "Regional features come into play – I contended – as well as historical and sociological factors, so to speak. Ferrante and I both write about Campania, Naples, the 1950s, the middle classes, the same vintage props, the same dialectal overtones in utterances. Not surprisingly, similarities do emerge […].
> Of course. Galella has shown that two very different authors – a man with an inclination for humour and a woman with a preference for deep feelings – may share traits deriving from their common background. It is interesting. Literary

> critics pay little or no attention at all to such things nowadays. However, that is not enough to write a full page or even devote a front-page headline to the matter."

A new corpus of writers from Naples and Campania

This study deals with the issue of Elena Ferrante's relation to other novelists from Naples or neighbouring areas in the region of Campania. Our objective is providing evidence that the similarity between Ferrante's and Starnone's works cannot be justified only in terms of the authors' common geographical background.

The texts used for this research were extracted from a corpus (the original corpus) including 150 novels by 40 different authors (Cortelazzo and Tuzzi, 2017; Tuzzi and Cortelazzo, 2018), compiled according to criteria accounting for the most significant analyses carried out on Ferrante's works by both literary critics and journalists. The original corpus comprises novels highly representative of the Italian literary production of the last three decades in terms of both their success – as best sellers or winners of literary awards – and their literary value, according to the opinions of experts in the field of literature (critics, members of the Academia etc.). All the novels in the corpus were published in a thirty-year time span (1987-2016) and were originally written in Italian, i.e. we did not take into consideration translations. Only four novels were published before 1987: two novels by Michele Prisco (*Una spirale di nebbia*, 1966; *La provincia addormentata*, 1969), one by Dacia Maraini (*Memorie di una ladra*, 1972) and one by Marta Morazzoni (*La ragazza col turbante*, 1986). We decided to include them in the corpus to increase the number of novels by these three authors.

The original corpus comprised 39 authors and 143 texts, plus 7 novels by Elena Ferrante (their translations in English are reported in brackets): *L'amore molesto* (1992; *Troubling Love*, 2006), *I giorni dell'abbandono* (2002; *The Days of Abandonment*, 2005), *La figlia oscura* (2006; *The Lost Daughter*, 2008), *L'amica geniale Infanzia, adolescenza* (2011; *My Brilliant Friend*, 2012), *Storia del nuovo cognome. L'amica geniale volume secondo* (2012; *The Story of a New Name*, 2013), *Storia di chi fugge e di chi resta. L'amica geniale volume terzo* (2013; *Those Who Leave and Those Who Stay*, 2014), *Storia della bambina perduta. L'amica geniale volume quarto* (2014; *The Story of the Lost Child*, 2015). The last four novels are part of a tetralogy entitled *L'amica geniale*, focussing on the intertwined lives of two Neapolitan women, Raffaella Cerullo and Elena Greco. Elena Ferrante also published *La spiaggia di notte* (2007; *The Beach at Night*, 2016), a children's book, and *La Frantumaglia* (latest Italian edition 2016; *Frantumaglia. A Writer's Journey*, 2016), a collection of interviews, essays and letters; neither is included in our corpus since they are not consistent with the "novel for adult readers" as a genre.

To test whether the similarities with Domenico Starnone's books could be explained merely in terms of the common geographical origin– as pointed out by literary critics –, from the main corpus we extracted 46 novels written by eleven authors from Naples or the neighbouring areas in Campania. Our selection includes: Erri De Luca, Diego De Silva, Rossella Milone, Giuseppe Montesano, Valeria Parrella, Francesco Piccolo, Michele Prisco, Fabrizia Ramondino, Ermanno Rea,

Domenico Starnone, and, of course, Elena Ferrante, as a presumed Neapolitan writer. This new corpus is composed of 2,839,098 word tokens; the size of each novel in terms of word-tokens and word-types is reported in Table 1.

Table 1. Size of each novel in terms of word-tokens and word-types

Author year title	w-tokens	w-types
De Luca 1998 Tu mio	29387	5303
De Luca 1999 Tre cavalli	23982	4430
De Luca 2009 Il giorno prima della felicità	33148	6053
De Luca 2011 I pesci non chiudono gli occhi	21868	4896
De Silva 1999 La donna di scorta	37226	5749
De Silva 2001 Certi bambini	39682	6288
De Silva 2007 Non avevo capito niente	82546	11726
De Silva 2010 Mia suocera beve	88435	13010
De Silva 2011 Sono contrario alle emozioni	32988	7223
Ferrante 1992 L'amore molesto	42030	7665
Ferrante 2002 I giorni dell'abbandono	54356	8975
Ferrante 2006 La figlia oscura	36784	6716
Ferrante 2011 L'amica geniale. Infanzia adolescenza	97893	11482
Ferrante 2012 Storia del nuovo cognome. L'amica geniale II	142215	14447
Ferrante 2013 Storia di chi fugge e di chi resta. L'amica geniale III	122101	13693
Ferrante 2014 Storia della bambina perduta. L'amica geniale IV	140440	14742
Milone 2013 Poche parole moltissime cose	53660	7991
Milone 2015 Il silenzio del lottatore	61275	9533
Montesano 1999 Nel corpo di Napoli	75324	10539
Montesano 2003 Di questa vita menzognera	52850	8557
Parrella 2005 Per grazia ricevuta	27534	5325
Parrella 2011 Behave	8129	1879
Piccolo 1996 Storie di primogeniti e figli unici	36765	5373
Piccolo 2003 Allegro occidentale	78183	9966
Piccolo 2007 L'Italia spensierata	56356	7807
Piccolo 2008 La separazione del maschio	59883	7133
Piccolo 2010 Momenti di trascurabile felicità	27509	4853
Piccolo 2013 Il desiderio di essere come tutti	79657	9331
Piccolo 2015 Momenti di trascurabile infelicità	29625	5044
Prisco 1966 Una spirale di nebbia	112578	14411
Prisco 1969 La provincia addormentata	63815	10502
Ramondino 1995 In viaggio	55633	11996
Ramondino 1998 L'isola riflessa	44296	9296
Rea 1995 Mistero napoletano. Vita e passione…	123185	16881
Rea 2002 La dismissione	113597	15531
Rea 2012 La comunista. Due storie napoletane	23848	5645
Starnone 1987 Ex cattedra	40463	7662
Starnone 1989 Il salto con le aste	69554	11143
Starnone 1991 Fuori registro	38103	8048
Starnone 1993 Eccesso di zelo	43772	7943
Starnone 1994 Denti	47825	8555
Starnone 2000 Via Gemito	143437	16569
Starnone 2007 Prima esecuzione	41346	7900
Starnone 2011 Autobiografia erotica di Aristide Gambìa	124294	15194
Starnone 2014 Lacci	37243	6583
Starnone 2016 Scherzetto	44278	7336

Correspondence analysis

To obtain a graphical representation of all the authors and to position them on a bi-dimensional graph, a content mapping procedure was conducted by means of a classical Correspondence Analysis (CA). Based on Singular Value Decomposition (SVD) as well as Principal Component Analysis (PCA), CA exploits the information carried by the rows (word types) and columns (authors) of a two-way contingency table. CA maps the occurrences of word-types (the cells of the contingency table account for the frequencies of words) into coordinates on a multidimensional Cartesian system (Greenacre, 1984, 2007; Murtagh, 2005, 2010, 2017; Lebart et al. 1984, 1998). To map authors in terms of the similarity of their lexical profiles, the corpus was arranged in 11 subcorpora pooling together all of the novels by the same author, while a contingency table was created including the vocabulary with the words-per-author frequencies (word-types and a selection of multi-words). CA offers the opportunity to position all authors in a multidimensional space through the transformation of an appropriate distance (the χ^2 distance) into coordinates on Cartesian axes. At the same time, CA positions all word-types on a parallel multidimensional space overlapping with the space accounting for the authors.

The first two dimensions of the multidimensional space created the first Cartesian plane (Figures 2 and 3). From the graphical representation of the contents of the corpus, Elena Ferrante emerges as a peculiar author since she contributes greatly to determine the first axis, i.e. the most important dimension in terms of explained inertia. Giuseppe Montesano stands out as a significant author on the second axis. On the left-hand half-plane, we may also note that Diego De Silva and Francesco Piccolo are apparently different from Elena Ferrante. In her turn, Elena Ferrante is not completely isolated on the right-hand side of the diagram, since Domenico Starnone is located in the same region of the plane (first axis). A moderate affinity also emerges with Rossella Milone, even though the contribution of this novelist to the first axis is limited.

Compared to the other novelists from Campania, the only author displaying a clear similarity with Ferrante on our content map is Domenico Starnone. In this respect, even though all the authors considered here are liable to share a certain number of linguistic traits owing to their geographical origin, this factor alone is not sufficient to explain the persisting proximity of Elena Ferrante and Domenico Starnone. As already explained by Tuzzi and Cortelazzo (2018), the results of the analysis do not change if we replace authors with their novels.

Text clustering

In quantitative and computational approaches, Authorship Attribution methods are often dealt with in terms of how to measure the distance between two texts and how to cluster texts resembling each other according to this measure of (dis)similarity (Eder, 2016; Mikros, 2013; Rybicki et al. 2016; Savoy, 2015). From a stylometric perspective, the author's style is the effect of a combination of words, sequences of words, syllables, characters, punctuation marks and syntactic structures, i.e. all the linguistic resources the author exploits when drafting a text. Although the relevant

scientific literature includes hundreds of different proposals (Grieve, 2007; Juola, 2008; Juola and Stamatatos, 2013; Koppel et al. 2008; Rudman, 1998; Stamatatos, 2009) and new methods are still being developed, no approach can be considered the most suitable in absolute terms (Juola, 2015). In the final analysis, the researcher's pick depends heavily on the characteristics of the texts and the purpose of the analysis.

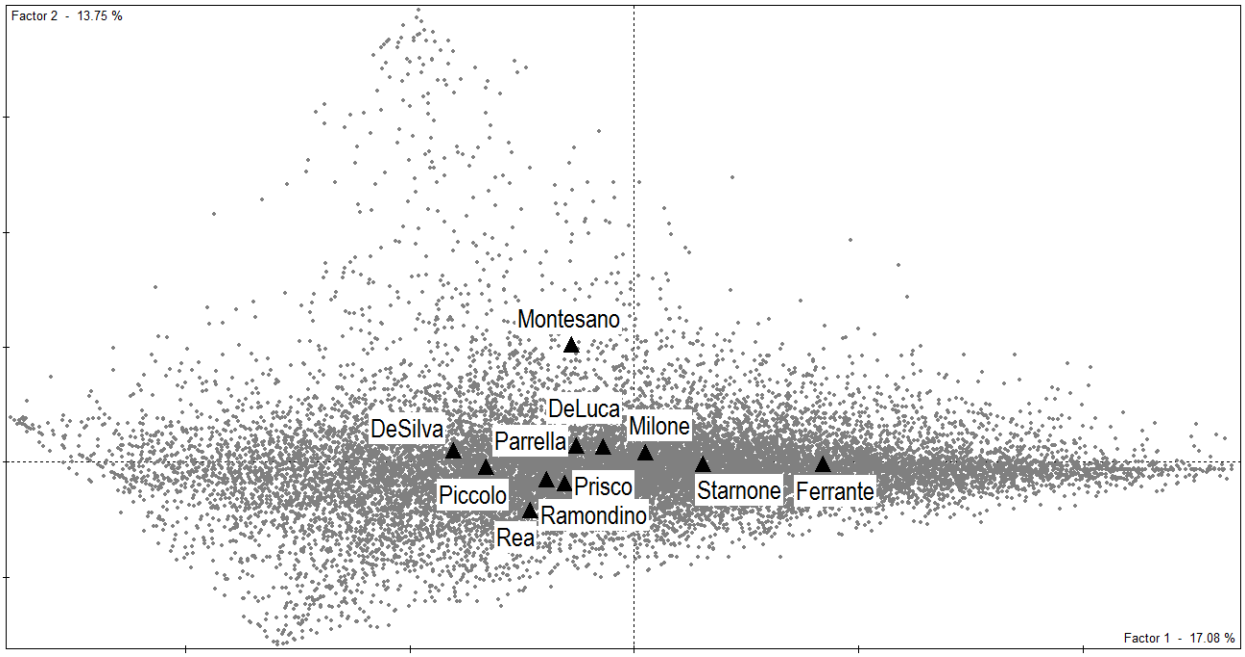

Figure 2. First plane from CA. Projection of all authors (triangles) and word-types (grey dots)

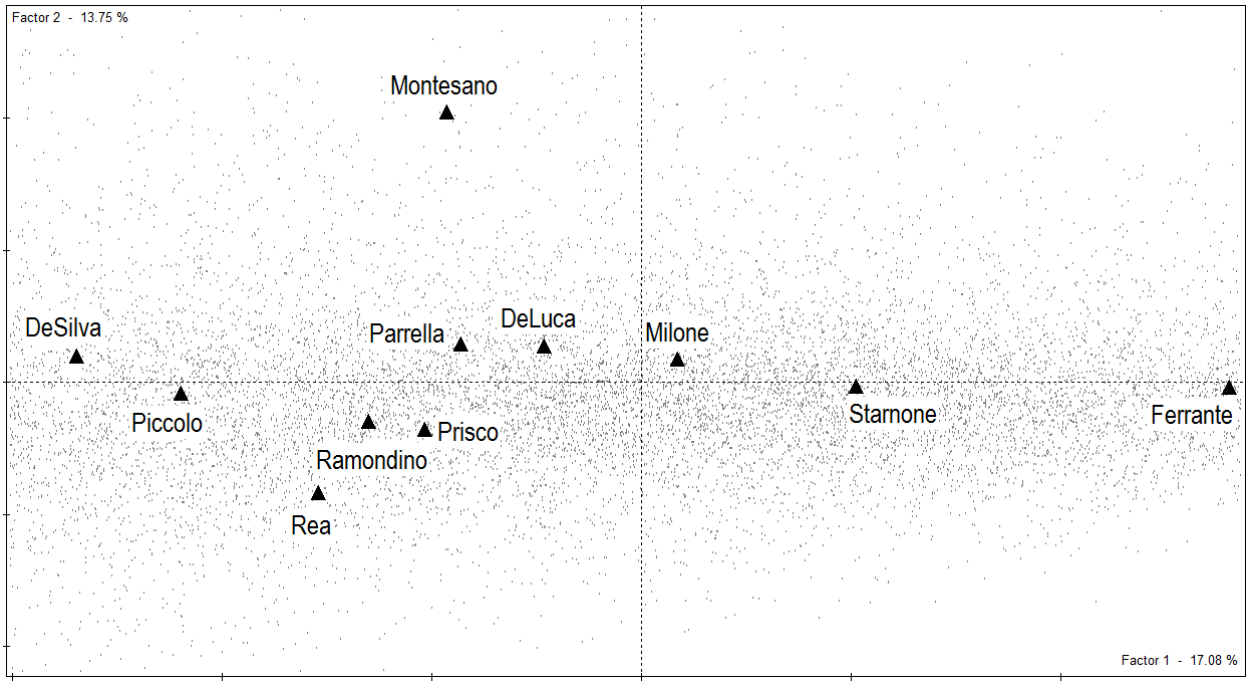

Figure 3. First plane from CA. Projection of all authors (triangles) and word-types (grey dots). Close-up.

In previous works we have endeavoured to contribute to the debate on Authorship Attribution methods by trying to improve Labbé's intertextual distance (Labbé and

Labbé, 2001, 2007), comparing the performance of different measures of similarity in clustering texts of different genres and discussing new graphic representation modes to compare the performance of different methods (Cortelazzo et al. 2012; Tuzzi, 2010). Our results show that our revised method for calculating intertextual distance ensures better results than cosine similarity (Tuzzi, 2010) and our iterative version proves even more effective when texts are long and their size varies considerably (Cortelazzo et al. 2012; 2013).

For each replication *j*, the procedure extracts a sample of text chunks (one from each text) of the same length in word tokens (*n*) and calculates the distance *d* for each pair of text-chunks. Given two texts *A* and *B* (Figure 4), the procedure takes into account the union set $V_{a \cup b}$ including the word-types of text chunks *a* and *b* (extracted, respectively, from *A* and *B*) and calculates the difference between the frequency of each word *i* in *a* and *b*. After *k* replications, the distance between *A* and *B* is the mean value $\hat{d}$ of the distances calculated on *k* pairs of the text chunks *a* and *b*.

$$\hat{d}(A,B) = \frac{\sum_{j=1}^{k} d_k}{k}$$

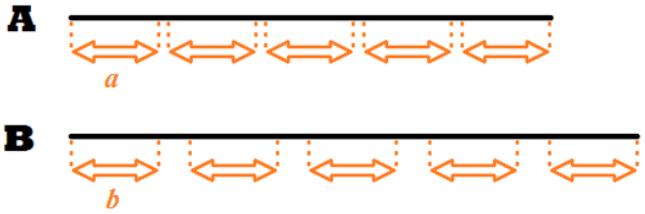

Figure 4. Two texts *A* and *B* and their text chunks *a* and *b*

In the present study, we implemented this revised version of Labbé's intertextual distance for all pairs of novels written by the authors included in the corpus, according to two procedures. The first procedure considered the whole vocabulary with $k = 300$ replications and text chunks of $n = 10{,}000$ word tokens in length. The square matrix reporting the distances between each pair of novels provides the basis for a text clustering procedure through an agglomerative hierarchical cluster algorithm with complete linkage. The second procedure involved $k = 300$ replications and text chunks of $n = 5{,}000$ word tokens, but we only considered grammar words (articles, conjunctions, prepositions and pronouns).

The results of both procedures are represented graphically by means of dendrograms (Figures 5 and 6, respectively). Figure 5 shows that all of Elena Ferrante's novels are grouped together, but they are interspersed with Starnone's recent works. The early novels by Domenico Starnone (*Ex cattedra*, *Il salto con le aste*, *Fuori registro*) are grouped together in a distinct cluster immediately connected to the clusters including the works by Diego De Silva and Francesco Piccolo (with the exception of one text by Valeria Parrella: *Behave*, 2011). Domenico Starnone's novels prove similar to each other but are split into two groups and only the books he published in the last twenty-five years (the turning point is the early 1990s) are similar

to Elena Ferrante's works. These results, illustrated in Figure 5, are based on the occurrences of both content and grammar words, therefore they account for the combined effect of the contents (conveyed by content words) and the purely linguistic style (conveyed by grammar words) of the novels.

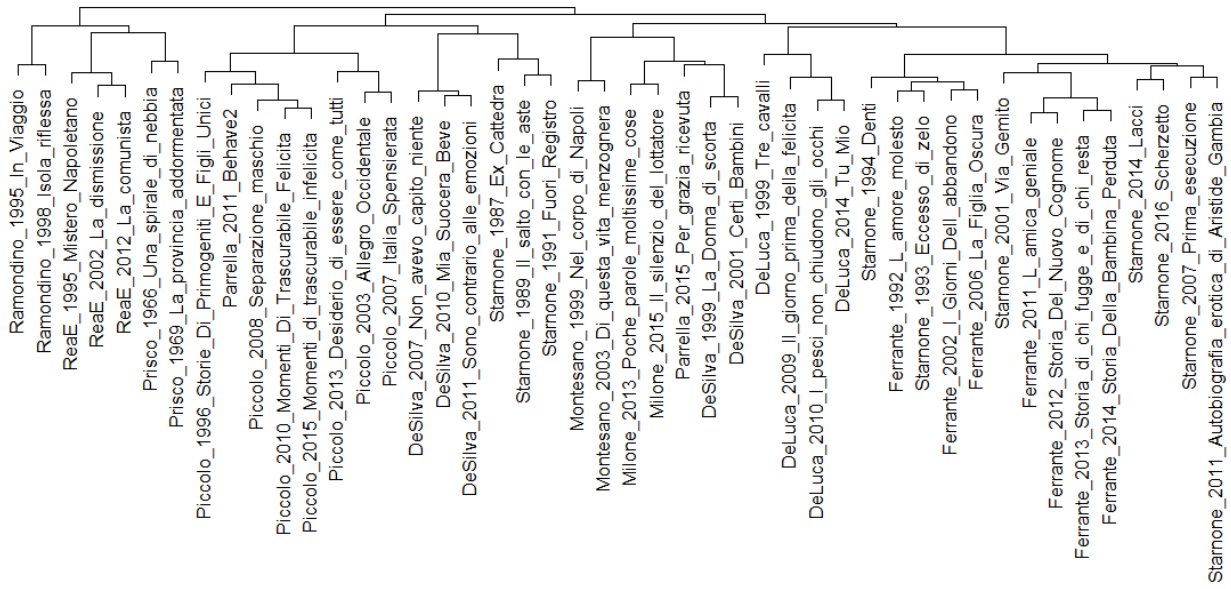

Figure 5. Dendrogram obtained through an agglomerative hierarchical cluster algorithm with complete linkage. Intertextual distance based on the whole vocabulary of 46 novels by authors from Campania.

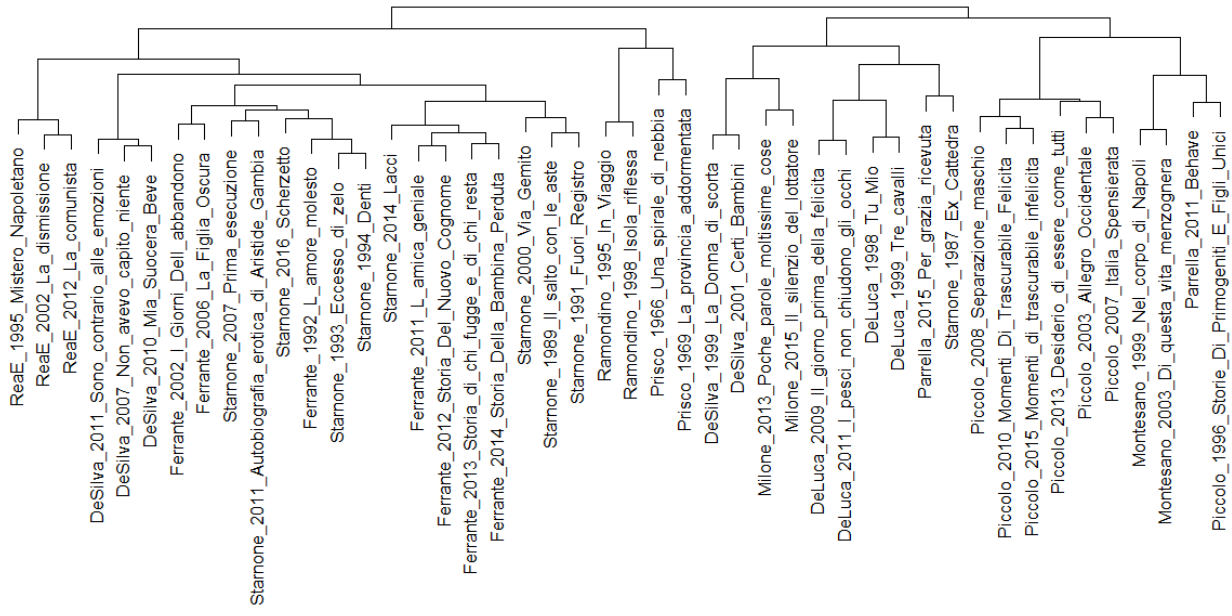

Figure 6. Dendrogram obtained through an agglomerative hierarchical cluster algorithm with complete linkage. Intertextual distance based on the grammar words of 46 novels by authors from Campania.

Figure 6 illustrates the results of the second procedure, only accounting for the occurrences of grammar words. With only one exception (Starnone's very first novel,

Ex cattedra), all the books by Domenico Starnone and Elena Ferrante are grouped together to form one cluster. Since in this case the procedure is based on grammar words, it does not capture the contents of the novels; however, the similarity between Ferrante and Starnone emerges more clearly and the distribution of texts within the cluster seems to be more dependent on the publication year.

Discussion and conclusions

Additional remarks from a qualitative viewpoint may be added to a purely quantitative analysis with reference to the use of Neapolitan dialect. If Elena Ferrante really is a Neapolitan writer setting her novels in Naples, then she should share the same social and cultural background of other Neapolitan novelists writing about similar settings and topics. Many Neapolitan writers (e.g. Giuseppe Montesano) often include Neapolitan words or expressions in their novels to add a local touch to their language. In contrast, Elena Ferrante rarely does the same (Librandi, *in press*). In her texts, she often underlines that the characters speak Neapolitan dialect, but quotes are nearly always in standard Italian. This feature is even more striking if we compare the novels with the corresponding film adaptations. For example, in the film adapted from *L'amore molesto* (Ferrante, 1992) and directed by Mario Martone (1995), the actors often speak Neapolitan and use words and expressions not contained in the book, so much that subtitles were added to make the dialogues intelligible to Italian viewers from other regions (Giani, 1995).

To a broader extent, at the lexical level Elena Ferrante does not show specific affinities with other novelists from Campania, with the notable exception – once again – of Domenico Starnone (Cortelazzo and Tuzzi, 2017; Tuzzi and Cortelazzo, 2018). This finding confirms the conclusions that may be drawn from lexically-based methods of content mapping and text clustering, i.e. that local or regional traits fail to account for the linguistic similarity between Domenico Starnone and Elena Ferrante. This similarity emerges clearly with the novels published by Starnone after 1990 and is even more apparent if only grammar words are considered, i.e. contents are disregarded and only purely stylistic features come into play.

The results of our research invalidate the hypothesis that Starnone's and Ferrante's novels appear to be similar because both authors share the same geographical, social and cultural background. If this had been the case, the comparison with a group of novelists from Campania – accounting for the regional features of contemporary Italian literature – would have produced several combinations matching and pairing different novels and authors. In contrast, the only other authors (in addition to Ferrante) who appear to be similar to Starnone are Francesco Piccolo and Diego De Silva, but their proximity only involves the books published by Starnone before 1992. After that date, the only match for Starnone is Elena Ferrante.

References

Alsop, Elizabeth (2014). Femmes Fatales: "La fascinazione di morte" in Elena Ferrante's "L'amore molesto" and "I giorni dell'abbandono", *Italica*, 91(3): 466-485.

Bakopoulos, Natalie (2016). We are always us: the boundaries of Elena Ferrante, *Michigan Quarterly Review*, 55(3): 396-419.

Benedetti, Laura (2012), Il linguaggio dell'amicizia e della città: L'amica geniale di Elena Ferrante tra continuità e cambiamento, *Quaderni d'italianistica*, 33(2): 171-187.

Bovo-Romoeuf, Martine (2006). Sensualité et obscénité dans "L'amore molesto" et "I giorni dell'abbandono" d'Elena Ferrante, *Cahiers d'études italiennes*, 5: 129-138.

Caldwell, Lesley (2012). Imagining Naples: The Senses of the City. In Bridge Gary and Watson Sophie (eds), *The New Blackwell Companion to the City*. Malden: Wiley-Blackwell, pp. 337-346.

Cavanaugh, Jillian R. (2016) Indexicalities of Language in Ferrante's Neapolitan Novels: Dialect and Italian as Markers of Social Value and Difference. In: Russo Bullaro Grace, Love Stephanie V. (eds) The Works of Elena Ferrante. Reconfiguring the margins. New York: Palgrave Macmillan, New York 45-70.

Ceccoli, Velleda C. (2017). On Being Bad *and* Good: My Brilliant Friend Muriel Dimen, *Studies in Gender and Sexuality*, 18(2): 110-114.

Chemotti, Saveria (2009). *L'inchiostro bianco. Madri e figlie nella narrativa italiana contemporanea*. Padova: Il Poligrafo.

Cortelazzo, Michele A. and Tuzzi, Arjuna (2017). Sulle tracce di Elena Ferrante: questioni di metodo e primi risultati. In Palumbo, Giuseppe (ed), *Testi, corpora, confronti interlinguistici: approcci qualitativi e quantitativi*. Trieste: EUT Edizioni Università di Trieste, pp. 11-24.

Cortelazzo, Michele A., Nadalutti, Paolo and Tuzzi, Arjuna (2013). Improving Labbé's Intertextual Distance: Testing a Revised version on a Large Corpus of Italian Literature, *Journal of Quantitative Linguistics*, 20(2): 125-152.

Cortelazzo, Michele A., Nadalutti, Paolo, Tuzzi, Arjuna (2012). Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea, in: Dister Anne, Longrée Dominique and Purnelle Gérald (eds), *JADT 2012 Actes des 11es Journées internationales d'analyse statistique des données textuelles*, LASLA – SESLA, Liège – Bruxelles, Belgio: 295-307.

Crispino, Anna Maria and Vitale, Marina eds. (2016), *Dell'ambivalenza. Dinamiche della narrazione in Elena Ferrante, Julie Otsuka e Goliarda Sapienza*. Roma: Iacobelli Editore.

Deutsch, Abigail (2015). Fiction in review: Elena Ferrante, *Yale Review*, 103(2): 158-165.

Dow, Gillian (2016). The 'biographical impulse' and pan-European women's writing. In Batchelor Jennie and Dow Gillian (eds), *Women's Writing, 1660-1830: Feminisms and Futures*. London: Palgrave Macmillan, pp. 193-213.

Eder, Maciej (2016). Rolling stylometry, *Digital Scholarship in the Humanities*, 31(3): 457-469.

Falotico, Caterina (2015). Elena Ferrante: Il ciclo dell'Amica geniale tra autobiografia, storia e metaletteratura, *Forum Italicum*, 49(1): 92-118.
Ferrante, Elena (1992). *L'amore molesto*. Roma: E/O.
Ferrante, Elena (2002). *I giorni dell'abbandono*. Roma: E/O.
Ferrante, Elena (2006). *La figlia oscura*. Roma: E/O.
Ferrante, Elena (2007). *La spiaggia di notte*. Roma: E/O.
Ferrante, Elena (2011). *L'amica geniale. Infanzia, adolescenza*. Roma: E/O.
Ferrante, Elena (2012). *Storia del nuovo cognome. L'amica geniale volume secondo*. Roma: E/O.
Ferrante, Elena (2013). *Storia di chi fugge e di chi resta. L'amica geniale volume terzo*. Roma: E/O.
Ferrante, Elena (2014). *Storia della bambina perduta. L'amica geniale volume quarto*. Roma: E/O.
Ferrante, Elena (2016). *La Frantumaglia*. Roma: E/O.
Galella, Luigi (2005). Ferrante-Starnone. Un amore molesto in via Gemito, *La Stampa*, Torino, 16 January 2005: 27.
Galella, Luigi (2006). Ferrante è Starnone. Parola di computer. *L'Unità*, Roma, 23 November 2006.
Gatti, Claudio (2016). Elena Ferrante, le «tracce» dell'autrice identificata, *Il Sole 24 Ore – Domenica*, Milano, 2 October 2016, pp. 1-2.
Gatto, Simone (2016). *Una biografia, due autofiction. Ferrante-Starnone: cancellare le tracce*, Lo Specchio di carta. Osservatorio sul romanzo italiano contemporaneo, 22 October 2016. www.lospecchiodicarta.it.
Giani, Fausto (1995), Martone al nord in napoletano con sottotitoli, La Repubblica.it (http://ricerca.repubblica.it/repubblica/archivio/repubblica/1995/05/17/martone-al-nord-in-napoletano-con-sottotitoli.html)
Greenacre, Michael J. (1984). *Theory and Application of Correspondence Analysis*. London: Academic Press.
Greenacre, Michael J. (2007), *Correspondence Analysis in Practice*. London: Chapman & Hall.
Grieve, Jack (2007). Quantitative authorship attribution: An evaluation of techniques, *Literary and Linguistic Computing* , 22(3): 251-270.
Juola, Patrick (2008). *Authorship Attribution*, vol. I(3), Foundations and Trends® in Information Retrieval, pp. 233-334.
Juola, Patrick (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions, *Digital Scholarship in the Humanities*, 30(1): 100-113.
Juola, Patrick and Stamatatos, E. (2013), Overview of the authorship identification task, *Proceedings of PAN/CLEF 2013*, Valencia, Spain.
Koppel, Moshe, Schler, Jonathan and Argamon, Shlomo (2008). Computational methods in authorship attribution, *Journal of the American Society for Information Science and Technology*, 60(1): 9-26.
Labbé, Cyril and Labbé, Dominique (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière, *Journal of Quantitative Linguistics*, 8(3): 213-231.
Labbé, Dominique (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1), 33–80.

Lebart, Ludovic, Morineau, Alain and Warwick, Kenneth M. (1984). *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*. New York: Wiley.

Lebart, Ludovic, Salem, André and Berry, Lisette (1998). *Exploring Textual Data*. Dordrecht: Kluwer Ac. Pub.

Lee, Alison (2016), Feminine Identity and Female Friendships in the 'Neapolitan' Novels of Elena Ferrante, *British Journal of Psychotherapy*, 32(4): 491-501.

Librandi, Rita (*in press*), Una lingua silenziosa: immaginare il dialetto negli scritti di Elena Ferrante. In Jamrozik, Elzbieta and Tylusińska-Kowalska, Anna (eds) *Dal monologo al polilogo: l'Italia nel mondo. Lingue, letterature e culture in contatto, Atti del Convegno (Varsavia 6-8 aprile 2017).*

Mikros, George K. (2013). Systematic stylometric differences in men and women authors: a corpus-based study. In Köhler, Reinhard and Altmann, Gabriel (eds), *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the occasion of his 70th birthday*. Lüdenscheid: RAM – Verlag, pp. 206-223.

Milkova, Stiliana (2013). Mothers, daughters, dolls: On disgust in Elena Ferrante's La figlia oscura, *Italian Culture*, 31(2): 91-109.

Mullenneaux, Lisa (2007). Burying Mother's Ghost: Elena Ferrante's "Troubling Love", *Forum Italicum*, 41(1): 246-250.

Murtagh, Fionn (2005). *Correspondence Analysis and Data Coding with Java and R*. London: Chapman & Hall/CRC.

Murtagh, Fionn (2010). The Correspondence Analysis platform for uncovering deep structure in data and information, Sixth Boole Lecture, *The Computer Journal*, 53(3): 304-315.

Murtagh, Fionn (2017). Big data scaling through metric mapping: Exploiting the remarkable simplicity of very high dimensional spaces using Correspondence Analysis. In Palumbo, Francesco, Montanari, Angela and Vichi, Maurizio (eds), *Data Science - Innovative Developments in Data Analysis and Clustering*. Cham: Springer International Publishing, pp. 295-306.

Rastelli, Alessia (2016). Elena Ferrante, lo studio statistico richiama in causa Starnone. *Corriere della Sera*, Milano, 12 October 2016, p. 37.

Reyes Ferrer, Maria (2016). La funzione dello specchio nel romanzo *I giorni dell'abbandono* di Elena Ferrante, *Cuadernos de Filología Italiana*, 23: 221-236.

Ricciotti, Adele (2016). Un confronto tra Elena Ferrante e Anna Maria Ortese: la città di Napoli, la fuga, l'identità, Zibaldone. *Estudios italianos de La Torre del Virrey*, 4(2): 111-122.

Rudman, Joseph (1998). The state of authorship attribution studies: Some problems and solutions, *Computers and the Humanities*, 31: 351-365.

Russo Bullaro, Grace and Love, Stephanie V. (2016, eds), The Works of Elena Ferrante. Reconfiguring the Margins, New York, Palgrave Macmillan.

Rybicki, Jan, Eder, Maciej and Hoover, David (2016). Computational stylistics and text analysis. In Crompton, Constance, Lane, Richard L. and Siemens, Ray (eds), *Doing Digital Humanities*, London and New York: Routledge, pp. 123-144.

Santagata, Marco (2016). Elena Ferrante è …, *La lettura – Corriere della Sera*, Milano, 13 March 2016, p. 2 and p. 5.

Savoy, Jacques (2015). Text Clustering: An application with the State of the Union Addresses, *Journal of the American Society for Information Science and Technology*, 66(8): 1645-1654.

Segnini, Elisa (2017). Local flavour vs global readerships: The Elena Ferrante project and translatability, *Italianist*, 37(1): 100-118.

Stamatatos, Efstathios (2009). A Survey of Modern Authorship Attribution Methods, *Journal of the American Society for Information Science and Technology*, 60(3): 538-556.

Starnone, Domenico (2000). *Via Gemito*. Milano: Feltrinelli.

Starnone, Domenico (2011). *Autobiografia erotica di Aristide Gambìa*. Torino: Einaudi.

Tuzzi, Arjuna (2010), What to put in the bag? Comparing and contrasting procedures for text clustering, *Italian Journal of Applied Statistics / Statistica Applicata*, 22(1): 77-94.

Tuzzi, Arjuna, Cortelazzo Michele A. (2018), What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities* (fqx066, https://doi.org/10.1093/llc/fqx066, online 19 January 2018).

Functional Explanation in Synergetic Linguistics*

Martina Benešová, Dan Faltýnek, Lukáš Zámečník
Department of General Linguistics, Palacký University Olomouc

Abstract
The aim of this paper is to open up a debate on the nature of the functional explanation in quantitative and above all in synergetic linguistics. It also attempts to highlight considering the praxis of handling with linguistic data supported with sound and consistent metatheory of this praxis and the necessity of reconstruction of functional model of explanation in synergetic linguistics. It also discusses approaches towards the functional explanation, and offers the most efficient one.

Keywords: *quantitative linguistics, synergetic linguistics, functional explanation, deterministic laws, analytic philosophy of science*

Introduction

There were some small polemics in the history of quantitative linguistics concerned with the methodological and scientific theoretical backgrounds of this approach to linguistic phenomena (see Bunge, 1995; Meyer, 2002; Grzybek, 2006 etc.). Few years ago, there was also yet smaller polemics about the status of the law in quantitative (see Benešová, Faltýnek & Zámečník, 2015) or especially in synergetic linguistics (see Milička, 2015; Zámečník, 2014). Now we are trying again to ginger up this debate in one important aspect which is the nature of the functional explanation used in quantitative and mainly in synergetic linguistics.

Synergetic linguistics (SL) represents probably the most sophisticated quantitative linguistic approach. It is a reality which we in any case do not intend to dispute. Our comments focus on the philosophic (in tradition of analytical philosophy of science) reflection on epistemological and methodological bases of SL. We will show that SL should put aside some of exaggerated claims which try to anchor SL as a theory, which disposes a D-N model of explanation (Hempel, 1965). These requirements are unrealistic because they imply the need for provision of deterministic laws of linguistics which enable the possibility of deductive inference from these laws toward empirical facts.

First, we focus on elucidation of concepts of function and functional explanation as they have been explored in the context of analytic philosophy of science since the 50s until today. To fulfill this step, there is a longstanding demand among quantitative

* Address correspondence to: Department of General Linguistics, Faculty of Arts, Palacký University Olomouc, Křížkovského 14, 771 47; E-mail: martina.benesova@upol.cz, dan.faltynek@upol.cz, lukas.zamecnik@upol.cz * The study is a part of the collective work within the SGS_IGA project Modely mentální reprezentace slova v kontextu morfologie s přihlédnutím k jejich explanačnímu potenciálu, no. IGA_FF_2017_022.

linguists because most philosophic assumptions of QL are enshrined only in the (undeniably influential) philosophical work by Mario Bunge. Subsequently, we will show which problems there are that current SL suffers in connection with the concept of functional explanations from. And finally, we will try to outline the nature of the functional explanation of SL devoid of the synergetic principle.

Function in context of functional explanation in analytic philosophy of science[1]

The beginnings of reflection on a functional explanation originated with the founders of analytic philosophy of science Hempel and Nagel, who tried to assimilate the teleological explanation and using of the concept of function and functional explanation to establish a uniform explanatory standard across scientific disciplines. Since at least the mid-70s, it has become clear that this approach is inadequate or unavailable, which set up the classic work Functions by Larry Wright (1973). The last significant comprehensive work on this topic was presented by Peter McLaughlin (2001).

Wright asks the question what function statements mean; and he offers two answers which constitute two current traditions of understanding the concept function. The claim "The function of X is Z" may mean a) „X is there because it does Z", which defines the etiological approach which states that "F must explain how it can be that the effect produced by a kind of entity can have a causal relevance to the existence of the entity." (Garson, 2008, p. 526) The second possibility is b) "Z is a consequence (or result) of X's being there," which defines a consequentialist approach which states that "F of an entity consists in a consequence that it produces, and has nothing to do with the cause or origin of the item itself." (Garson, 2008, p. 527)

Etiological views exist in two variants, both of which generally regard to maintain a valid form of the causal nexus (without any problematical teleological backward causation). The representationalist strategy states that causal factors are internal representations (beliefs, desires) of the future effect. However, the mentalistic variant of this strategy may be only purely heuristical and metaphorical (despite the strictly theological view), and the non-mentalistic variant depends on the concept of teleonomy (Mayr), which operates with the conception of an internal program which controls the future effect.

Non-representationalist strategies state that a functional entity may be validly explained via the natural history of that entity and prominently use the biological conceptual borrowing of the natural selection. The most influential is the "selected effect view" of the function which states that "having a function means having been selected for by natural selection" (Garson, 2008, p. 533). In confrontation with the problem of disappearance of the function, it was stated that it is necessary to "identify the function with the effect for which it was selected in the recent evolutionary past**.**" (Garson, 2008, p. 536)

Consequentialist views exist in four variants, however except the first, the others are highly similar. These variant are: interest-contributions, goal-contributions, good-contributions and fitness-contributions views. The I-C view states that "the function

[1] This section is prepared on the basis of summarizing writing of the views of function and functional explanation in Garson, J. (2008) Function and Teleology, pp. 525-549.

consists in its contribution to maintaining some property of a system that is of interest to an investigator," (Garson, 2008, p. 538) and the G-C view states that "the function of a part of the system consists in its contribution to a goal of that system" (Garson, 2008, p. 539).

The I-C view sounds to be the most promising. Some primordial version of I-C may be found in Hempel (1965). It was systematically elaborated by Cummins (1975), although he named his view as "the systemic capacity" view of the function. In connection with this view, we speak about the Cummins-function (C-function). According to this view, "functions refer primarily to a distinctive style of explanation ("functional analysis"), and only secondarily to a distinctive object of study." (Garson, 2008, p. 538). This shows that it is a methodological approach that highlights the role of the researcher who chooses to ascribe the function and perform functional analysis in accordance with a certain partially chosen perspective of the research. Not quite a pleasant consequence of this approach, however, is that it is markedly pluralistic – different interests may in connection with the same system lead to mutually incompatible functional explanations. This plurality is, however, in strong contradiction with the universal character of the D-N explanation. The functional analysis enables to explain how some particular effect is produced, but not, why it exists. This also shows the incompatibility of the I-C with D-N model of explanation.

The goal-contributions view stems from the classical work by Ernst Nagel (1953, 1977), in which he tried to reduce the teleological explanation through the conception of "the goal-directed system". The goal of the system is understood as an achievement of the given value for the system variable. And a specific contribution (of a part) of the system to the specific goal is conceived as the function (of that part) of the system. Nagel's view is closely related to the concept of self-regulation. One of the limitations of this classical view is the necessity to establish the independence condition which states that controlling variables must be independently manipulable. The biggest problem of this view, however, is the problem of goal-failure. Most of philosophers of science conclude that the successful solution of this problem requires addition of an internal representation of the goal-state (via Mayr). This solution, however, leads G-C very close to representationalist view of the function.

Possibilities of functional explanation in SL

After this short summary of views of the function elaborated in philosophy of science in last decades, we may ask which view of the function may be accepted and used by SL. We think that there are three possibilities. The non-mentalistic representationalist view is possible, but only in connection with some conception of the fundamental role of internal representations which cause the functional item. However, investigation of these internal representations remains behind the possibilities of SL, and if, then only in connection with neurolinguistics. The selected effects view is possible, but also only with elaboration of the concept of natural selection and recent evolutionary past in the context of evolution of the language which SL may investigate only in tight connection with biolinguistics and psycholinguistics.

As we think the only view which is non-problematically acceptable for SL usage is the I-C view of function. SL elaborated very deeply the functional analysis of the

language system which is capable of explaining (and also to a certain extent and in statistical limitation, of predicting) a wide area of language phenomena. The actual engine that allows it is a network of interconnected economization principles (minimization of decoding effort, minimization of production effort and minimization of memory). These principles are, methodologically speaking, a very helpful package of instruments. These principles, however, were selected heuristically according to the interest of SL, they may not play the role of structural axioms in the functional explanation (as SL knows), and the only way "to deduce" them (very freely and logically incorrectly speaking) is through establishing the synergetic principle.

Problems

It seems to us that the use of the function by SL and functional explanation oscillate between etiological and consequentionalist views. For example when generally speaking about semiotic systems, Köhler et al. say that: "An explanation of existence, properties, and changes of semiotic systems is not possible without the aspect of the (dynamical) interdependence between the structure and function", (Köhler et al., 2005, p. 761) which sounds as the etiological view. In another place, it seems as the consequentionalist view of the functional analysis: "The elements under consideration have become a part of the language system, because they possess certain properties and have certain functions within the system." (Köhler et al., 2005, p. 762) However, in another place, they state that "this type of explanation /it is functional explanation in linguistics/ is a special case of the D-N explanation" (Köhler et al., 2005, p. 765) which may be possible if ever, only in some very strong reductive type of the etiological view.

SL was conceived (in the 80s) as an embryonic state of a linguistic theory, which will be in the future highly explanatory (and not only traditionally descriptive). For embryonic theories, it is very typical, and non-problematical, to base their preliminary principles on conceptual borrowing from other theories and disciplines. So is for SL, we may remind the borrowed concepts as fluctuation (from non-equilibrium thermodynamics), mutation and selection (from molecular and evolutionary biology) or self-organization (from the systems theory). However, we think that for an adult discipline, it is necessary to emancipate from this primordial borrowing and metaphors. And we think that the emancipation in the case of SL is in forsaking of the synergetic principle.

It should be noted that synergetics is today not a new subdiscipline (Köhler et al., 2005, p. 760) of the systems theory. Synergetics was new in the 80s, and in the next decade (the 90s), it was widely criticized – most comprehensibly in Achim Stephans Emergenz (1999). We think, and we may be wrong, that synergetics was a part of the holistic paradigm, which dominated in plenty of disciplines (physics not excluded) in the 70s and 80s. It is also for us very important to open the question whether the following of this paradigm is fruitful in contemporary quantitative linguistics.

The most important however is, that SL fails in its important scientific philosophical effort. We assume that SL does not satisfy the D-N model of explanation. There is a wide concensus in philosophy of science that the functional explanation via the functional analysis is not compatible with the D-N model. The functional analysis does not provide the D-N explanation because the inferences in FE are typically

unsound.[2] Soundness is of course a strong logical criterion, which in common with consistence forms an essential pillar of logically properly formed systems of inference. The D-N model stems in the very rigid syntactical view of philosophy of science (in connection with the syntactic structure of scientific theories, logic of scientific testing and intertheoretical reduction). The D-N model may be with some objections used in physics (in reconstructed classical theories), its application beyond the physical field is highly speculative.

Another important remark is connected with the type of laws which may figure in explanans of the D-N model of explanation. The wide concensus is that these are only deterministic laws. Although Hempel did not exclude other types of laws from explanans, as was made explicit at the latest in Salmon (see e.g. Salmon, 2009), for validly established D-N, it is necessary to establish a causal connection via deterministic (causal law). And as we know the typical law of quantitative linguistics is a statistical law.

Köhler et al. (2005, p. 764) report as an example of the D-N explanation the inference made from Altmann's law

$$m = AWL^{-b}$$

where "m is the number of meanings of the word, WL is the length of the word, A is the mean polysemy of words of length 1 and b is the measure of syntheticism of the language under consideration" (ibid.). This law is of course also statistical, however if we want, we may do "inferences" or "deduction" in the sense that we simply use the mathematical function (if we have AW, A, b, we may „infer" m etc.). Analogically, if we have some basic model of ideal gas – the state equation

$$pV = AT$$

where p is pressure, V volume, A specific constant and T thermodynamic temperature of ideal gas, we may „infer" for example T, if we know p, V and A. However in contrast with the SL case, in statistical physics, we may explicitly define macroscopic variables in statistical terms. We do not think that anything like that is possible in the SL case.

The highly important question is whether SL really needs a synergetic principle in any other than the metaphorical sense. The answer may be connected with the answer to another question: Why biological explanations do not start with self-organization principles when organisms are the best examples of self-organizing entities? We think that it is because this principle is too hard, general, and somehow vague. And of course, biology disposes with other subtle principles which are of better use. So it seems to us that establishing of SP in SL represents a big explanatory jump. It is understandable why it is established because only with help of this strong structural axiom/principle, it is possible to conclude that SL may propose a valid D-N explanation. However, is SP really such an axiom, which may non-problematically warrant the D-N character of functional explanation in SL?

[2] See for example Routledge Encyclopedia of Philosophy, item Functional explanation.

We think that SL (when S is meant metaphorically) may without problems prosper without SP, however of course in resignation to the D-N character of functional explanation in SL. It is not a reason for shame, as we said, also in philosophy of physics there is a wide continence to such a type of explanation (see Cartwright, 1999; Morrison, 2015 etc.). Functional explanation in SL is highly compatible with the interest-contributions view of function and may be in this line further developed.

Functional explanation for SL established

After elaboration of the analysis of the main problems of the backgrounds of the functional explanation in SL, we may now concern ourselves with the possibility of reconstruction of the structure of FL as it was presented in Köhler et al. (2005). This reconstruction is going in two main directions (with reminder of the strong solution presented in Zámečník, 2014) which are: the I-C consequentionalist view of the function in SL and N-R etiological view of the function in SL.

The explanation in synergetic linguistics originally assumes the following structure:

"(1) The system S is self-organising. For each need, it possesses mechanisms to alter its state and structure in such a way that the need is met.

(2) The needs $N_1 \ldots N_k$ *have to be met by the system.*

(3) The need N can be met by the functional equivalents $E_1 \ldots E_f \ldots E_n$.

(4) The interrelation between those functional equivalents which are able to meet the need N is given by the relation $R_N (E_{N1} \ldots E_{Nn})$.

(5) The structure of the system S can be expressed by means of the relation Q $(s_1 \ldots s_m)$ *among the elements* s_i *of the system.*

E_f *is an element of the system S with load* R_{Nf} *."* (Köhler et al., 2005, p. 765)

When seeing the structure of the functional explanation in SL, we believe this contemporary SL functional explanation seems to be the consequentionalist G-C view of function. We believe in this because of the role of principle (1), which states the synergetic principle or self-organization. Problems of this approach to the functional explanation were discussed (see above). The goal of the system is connected with the saturation of several needs. But unfortunately for SL even if the condition of independence[3] has been met, the problem of goal-failure[4] is forcing a synergetic linguist to transform this view in some sense.

In Zámečník (2014), there is recommended the extension of this SL functional explanation via adding one more special principle:

[3] See Nagel, 1977, p. 273.
[4] See Garson, 2008, pp. 540-541.

(6) "In system S enslave Ordners O … O_o elements of system s … s_m and achieves the transformation of relation $Q_1(s_1 \ldots s_m) \ldots Q_p(s_1 \ldots s_m)$." (Zámečník, 2014, p. 111)

This recommendation concerns the *reductio ad absurdum* strategy in the way it is trying to make the ontological commitments of synergetic linguists explicit when establishing SL as an explanatory and not only descriptive theory. There is also a recommendation for adopting the functional reduction – but we recognized that this assumption is for SL for several reasons non-realistic (see Zámečník, 2014, pp. 108-109).

Now in the context of analysis of uses of terms of functional explanation in analytical philosophy of science (see above), we may try to find the view which fits with SL and enables SL to connect with any other contemporary approach to language. As we said, we see two possibilities: the I-C consequentionalist view of the function in SL and the N-R etiological view of the function in SL.

The Interest-Contribution view is acceptable without other commitments when we are reconciled with the pure descriptiveness of this functional model, where functional explanation is reduced to the level of functional description. But the question is whether this is not the only sound variant of manipulation with the function which is not responsible for not acceptable ontological commitments (e.g. downward causation, teleonomy etc.).

In this case of the Interest-Contribution Consequentialist view of functional explanation, we may transform the first problematical premise of the model of functional explanation in synergetic linguistics in this way:

(1)* *The systemic capacity* (or Cummins function) of system S *enables to system S*, for each need, *the changes* in its state or structure so that the need is met.

The I-C Consequentialist view represents for us now the best variant of using functional description in synergetic linguistics in the sense it is the output of a thorough functional analysis. We are not concerned with a causal nexus or evolutionary descend, these both are beyond possibilities of our knowledge (and are not necessary for our purposes).

When we do not want to resign for the search of causal nexus and evolutionary descend of language system, we may adopt the Non-Representationalist Etiological view of the functional explanation. As in other cases, we may evade the need for representation (mental or non-mental) of the future effect (and inner states) by adopting the "selected effect" view of the function. This may be done, as we believe, due the use of Millikan's proper function. (see Reboul, 2017, p. 33)

The starting point of Millikan is the view of the language as a communication system in the strong (adaptive) sense. As stated by Reboul: "(…) Communicative system gathers behaviours produced by organisms belonging to a given species, whose proper function is to communicate information to other members of the same species." (Reboul, 2017, p. 20)

For the definition of the proper function, we may cite Millikan: "(…) for an item A to have a function F as a ‚proper function, , it is necessary (and close to sufficient) that (…) A originated as a ‚reproduction, (…) of some prior item or items that, due in

part to possession of the properties reproduced, have actually performed F in the past, and A exists because (causally, historically because) of this or these performances." (Grifiths, 1993, p. 413)

When the language is studied primarily as a communication system, and it is of course also in context of synergetic linguistics, then we have to recognize the proper function of the signal, which is probably the ability to triggering of specific response (on the part of a recipient) due to the information transfer (see Reboul, 2017, p. 39)

We have explained the role of the proper function in Millikan's view, and so we can establish the reformulation of the first premise of the model of functional explanation in synergetic linguistics:

(1)** *The proper function of system S with an evolutionary history* is that, for each need, it possesses mechanisms to alter its state and structure in such a way that the need is met.

The N-R etiological view is sympathetic according to the belief of possible explication of the causal nexus, which represents the continuum of changes going back through our evolutionary history. However we believe there are some major difficulties which stem from the general conception of the code model of the language.

Instead of conclusion: Is really language primarily a communication system?

Maybe, it is the case we are trying to consider quantitative linguistics in more conservative and old school terms, when trying to establish the right variant of explanatory model situated in the basis of SL, as e.g. Milička writes (see Milička, 2015, p. 1-2). However, our belief is that it is not enough to largely consider only the praxis of handling with linguistic data without any attempt to build sound and consistent metatheory of this praxis. So is the case with the necessity of reconstruction of functional model of explanation in synergetic linguistics.

We are aware of the fact, that the attempt to generally characterize, the way a linguist works in our quantitative branch is very difficult to almost impossible (see Grzybek, 2006). We may remember Meyer's vision, considering the possibility of establishing the new approach of research in linguistics – the complicity-driven approach:

"The ´big question´ that comes to mind here is whether a "third way" besides a traditional, qualitative understanding of the subject matter of linguistics and the inductive quantitative descriptions of contemporary QL (…) is conceivable at all. Recent contributions to the theory of complex systems suggest that qualitative-only and even functional treatments of systems may, in many scientific contexts, be both inevitable and explanatory fruitful." (Meyer, 2002, p. 77)

When we neglect some subversive visions (see Zámečník, 2014, pp. 117-119), we strongly believe this Meyer's vision is the last one, which considers some type of a general theory of the language in the sense, which should be inspiring for synergetic linguists.

Maybe the reason why all general attempts are not successful lies in the very nature of our quantitative linguistic models. This nature is expressed in terms of viewing the language as a communication system in the strong sense. This is very important especially when dealing with the N-R etiological view of functional explanation in SL because of the need to consider the language as an adaptation to communication in the natural selection of our evolutionary ancestors.

As we may find present in Reboul (2017), the paradigm shift from viewing the language as a communication system in the strong sense (the language as an adaptation to communication) to the opposing view (the language as an exaptation to communication) may build difficulties for all these models of the language which stem from the economization criteria for a successful communication.[5]

If this critique may be right, if the language primarily originates from the need to organize the conceptual space enlarged through the ability of abstraction (see Reboul, 2017, p. 63), than probably the only variant for the functional explanation in SL will remain the I-C consequentialist approach.

References

Benešová, Martina, Faltýnek, Dan, Zámečník, Lukáš (2015). Menzerath-Altmann Law in Differently Segmented Texts. In: Tuzzi, Arjuna, Benešová, Martina, Mačutek, Jan (Eds.), *Contributions to Quantitative Linguistics*. Berlin: Walter De Gruyter, pp. 27-40.

Bunge, Mario (1995). Quality, Quantity, Pseudoquantity, and Measurement in Social Science. In: *Journal of Quantitative Linguistics, 2*, pp. 1-10.

Cartwright, Nancy (1999). *The Dappled World: A Study of the Boundaries of Science.* Cambridge: Cambridge University Press.

Cummins, Robert (1975). Functional analysis. In: *Journal of Philosophy, 72*, pp. 741–65.

Garson, Justin (2008). Function and Teleology. In: *A Blackwell Companion to the Philosophy of Biology*, pp. 525-549.

Grzybek, Peter (2006). Introductory Remarks: On the Science of Language in Light of the Language of Science. In: Grzybek, Peter (Eds.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht: Springer, pp. 1-14.

Hempel, Carl Gustav (1965). The logic of functional analysis. In: Hempel, C. G. (Eds.), *Aspects of Scientific Explanation.* New York: Free Press, pp. 297-330.

Köhler, Reinhard (2005). Synergetic linguistics. In: Köhler, R., Altmann, G. & Piotrowski, G. (Eds), *Quantitative Linguistics: An International Handbook* (pp.760-774), Berlin: Walter de Gruyter.

McLaughlin, Peter (2001). *What Functions Explain: Functional Explanation and Self-Reproducing Systems.* Cambridge: Cambridge University Press.

[5] See the large astonishing critique of Milikan s view in Reboul (2017, pp. 35-42).

Meyer, Peter (2002). Laws and Theories in Quantitative Linguistics. In: *Glottometrics, 5,* pp. 62-80.

Milička, Jiří (2015). Synergetic Linguistics: Do We Need Better Explanatory Mechanism? In: *Glottotheory, 6,* pp. 1-9.

Morrison, Margaret (2015). *Reconstructing Reality: Models, Mathematics, and Simulations.* Oxford: Oxford University Press.

Nagel, Ernst (1953). Teleological explanation and teleological systems. In: Ratner, S. (ed.), *Vision and Action.* New Brunswick, NJ: Rutgers University Press, pp. 537-558.

Nagel, Ernst (1977). Teleology revisited: goal-directed processes in biology and functional explanation in biology. In: *Journal of Philosophy, 74,* pp. 261-301.

Reboul, Anne (2017). *Cognition and Communication in the Evolution of Language.* Oxford: Oxford University Press.

Stephan, Achim (1999). Emergenz: *Von der Unvorhersagbarkeit zur Selbstorganisation.* Dresden: Dresden University Press.

Wright, Larry (1973). Functions. In: *Philosophical Review, 82,* pp. 139-68.

Zámečník, Lukáš (2014). The Nature of Explanation in Synergetic Linguistics. In: *Glottotheory, 5,* pp. 101-120.

An Expanded Quantitative Study of Linguistic vs. Geographic Distance Using Romanian Dialect Data*

Sheila Embleton, Dorin Uritescu and Eric S. Wheeler
York University, Toronto, Canada

Abstract
In a previous study, we had a quantified case of a partial correlation between linguistic distance (representing dialect differences) and geographic distance, using data from the Crișana region of Romania. In this region, the geography provides more than one way of measuring distance: travel distance, and travel time between locations are not the same as the direct distance. However, that study was limited for practical purposes to only 8 geographic sites.

In this expansion of the study, we look at distances between all pairs of the 120 locations in the available dialect data. Furthermore, there are various ways of subdividing the linguistic data into subsets, such as phonetic, syntactic, etc. or even more specific selections. We examine how well different linguistic aspects correlate with one or another of the geographic distances.

While it is clear that geography cannot account for all of the dialect variation, especially in a modern world where telecommunications of all sorts override the traditional effects of physical distance, and where other factors such as culture and social structure must also play a role, the effects of geography can be measured quantitatively, and a base established against which the other factors operate.

It is perhaps worth noting that in contrast to many quantitative linguistic studies, which observe a quantified pattern and search for an explanation of the pattern, in this case we have a ready-made explanation and are searching for how much quantified observation there is to support it.

Keywords: *Crișana Region, Romanian dialects, dialectometry, quantitative dialectology, dialect variation, linguistic distance, dialect distance, Mantel test*

Introduction

This paper is about both findings in dialectology and the quantitative methods we used for getting those findings.

Linguistic distance (representing dialect differences) varies with geographic distance. Using data from field locations in the Crișana region of northwest Romania, we quantify those distances and use a statistical test to measure how well the one correlates with the other. There is a strong correlation.

In the Crișana region, the geography provides more than one way of measuring geographic distance: travel distance and travel time between locations are not the same as the direct distance. An earlier study (Embleton, Uritescu and Wheeler. 2015),

* Address correspondence to: Dr. Sheila Embleton, Department of Languages, Literatures & Linguistics, York University, 4700 Keele Street, Toronto, Ontario, CANADA M3J 1P3. E-mail: embleton@yorku.ca

limited to 8 locations, showed promising results, and so we have extended the approach to all 120 locations in our data set, and combined with a refined definition of linguistic distance, have found that travel distance and travel time have a modestly stronger correlation with linguistic distance than does direct distance.

Furthermore, there are various ways of subdividing the linguistic data into subsets, such as phonetic, syntactic, etc. or even more specific selections. We examine how well different linguistic aspects correlate with one or another of the geographic distances.

Finally, we look at how well a single correlation coefficient represents an entire data set, by taking random subsets of the data set, and correlating them to the full data set. In this case, the subsets are close to the full data set, and our method shows that.

We also look at the correlations themselves as distances in a multidimensional (MDS) space.

Linguistic distance

We use the 370 interpretive maps in our Romanian data set. An interpretive map has one or more values (interpretations of the underlying raw data) at each of the 120 field locations. Thus, for example, on map 3 which records the analyst's interpretation of one or more observations made in the field, location 101 may have a value of type 1, location 102 of type 1 and type 3, and location 103 of type 2. For map 3, the distance between 101 and 102 is 0 (because there is a match on type 1) and the distance between 102 and 103 is 1, because there is no match on any of the possible comparisons. The overall distance is the sum of such measures over the 370 maps, ranging from 0 to 370. (In early work, we allowed for multiple matches between pairs of locations, but that seems to unfairly give more weight to locations with multiple values; here, a location pair can have at most the same contribution to the distance as any other pair). The distances are represented as a distance matrix (a square 120 x 120 matrix, with 0 on the main diagonal, and symmetric entries: entry(i,j) = entry(j,i).)

Geographic distances

The direct geographic distance was calculated using the Euclidean distance between the coordinates of each field location (the coordinates being proportional to the latitude and longitude).

For travel distance and travel time, we queried Google maps (https://www.google.ca/maps/) and used the time (in minutes) and kilometres for the first (and shortest) route between each of 7140 distinct pairs of locations. (This represented a substantial amount of work; our talented research assistant, Lacramioara Oprea, was able to do over 60 queries an hour).

The distances are labelled “geo”, “km” and “min” respectively.

The Mantel test

To compare two distance matrices, the Mantel test (Legendre and Fortin. 2010; see

also http://en.wikipedia.org/wiki/Mantel_test) provides a measure of correlation (how closely does one matrix match the other) and a measure of confidence (a measure of whether this result is genuine or just a chance correspondence). Because the distances in a matrix are not independent (i.e. move one location, and a whole row of distances will change), one has to employ an appropriate test: in the Mantel test, the correlation is (naively) measured; then the values of the locations are randomly permuted a thousand times, and the correlation measured again, each time. If the measured correlation is much higher with the initial arrangement than with the random ones, it indicates that the initial correlation is significant.

The correlation coefficient (corr.) from the Mantel test will be 1.0 if the correlation is perfect, and 0.0 if there is no correlation, with typical values coming somewhere in between.

The confidence number (p-value) should be very small, if the result is strong (e.g. 0.01). Typically, we had a value of 0.001998002, which was the p-value for a matrix correlated with itself, and presumably the best value the program could provide.

Our Mantel test was a package on the R statistics system (R 2012) in the *ncf* library. (See https://cran.**r**-project.org/package=ncf)

Random samples

For the 370 interpretive maps, we used the Generalized Online Dialect Atlas (GODA) adapted to the Romanian data set, to generate the linguistic distance matrix, and 5 random subsets of the 370 maps, each random subset producing a distance matrix of about 180 items each. The selection process gave a 50/50 chance to each map to be in the subset.

Results

The distance matrices are:

- all371 All 370 interpretive maps (linguistic distance) [labelled "all371" but in practice we had only 370 maps]
- r1 to r5 Random subsets of the 370 maps (linguistic distance)
- geo Direct geographic distance
- km Distance in kilometres (from Google)
- min Distance (travel time) in minutes (from Google)

The correlation coefficients among the linguistic distances are shown in Table 1.

The full data matrix correlates 1.0 with itself, as it should by definition. As a confirmation that the software for the Mantel test is not misleading us, we also checked that mantel (m1, m2 ...) produced the same results as mantel (m2, m1 ...).

The random samples all more or less have corr. = 0.97, indicating they are very close to the full data set. (See Embleton, Uritescu, and Wheeler 2016 for a discussion of why it is important to check on how much "variance" could be in the all371 matrix).

The correlations among the geographic maps are shown in Table 2.

Table 1. Correlation coefficients among linguistic distances

	all371
all371	1.0
r4	0.9764655
r5	0.9751735
r1	0.9687155
r3	0.9662242
r2	0.9658761
Average r1 to r5	0.97049096
Standard deviation	0.0050068888

Table 2. Correlations among geographic maps

	km	min
geo	0.9611154	0.9285799
km		0.9798781

The geographic maps are similar, but not the same. As we see below, the differences in the geographic maps show up as differences in the correlation between linguistic and geographic distances.

The correlation coefficients for the geographic vs. linguistic matrices are:

Table 3. Correlation coefficients for geographic vs. linguistic matrices

	geo	km	min
all371	0.7708616	0.8009089	0.7954661
r1	0.7124076	0.7427141	0.7281619
r2	0.6907492	0.73135	0.7277309
r3	0.7070289	0.736329	0.7265886
r4	0.7600945	0.7897172	0.782036
r5	0.7826928	0.8063899	0.8048785
Average r1 to r5	0.7305946	0.76130004	0.75387918
Standard deviation r1 to r5	0.0348087105	0.0306810214	0.0331168354

Generally, linguistic distance is correlated to geographic distance with corr.= 0.73 to 0.76. This is a strong correlation, but clearly there are other factors involved in some way. It is not the same as the 0.97 correlation inherent in the linguistic matrices themselves.

If we list the km and min values as a ratio of the geo value, we get:

Table 4. Values of km and min as ratio of geo value

	geo	km	min
all371	1	1.039	1.032
r1	1	1.043	1.022
r2	1	1.059	1.054
r3	1	1.041	1.028
r4	1	1.039	1.029
r5	1	1.030	1.028
Average r1 to r5	1	1.042	1.032

We see that using the travel distance and travel time, we get an improvement across all the samples, but only of a modest amount (2% to 6%). However, given the similarity of the different geographic matrices, it would be unreasonable to expect more, and the use of km and min is an improvement.

We can speculate that on a more local scale, the differences between direct distance and travel distance or time would be more pronounced, whereas on a broader scale (e.g. continent-wide) the differences might be insignificant, with a corresponding difference for the match of linguistic and geographic distances.

Correlation as distance

It is possible to take the correlation coefficients and change them into distances by subtracting them each from 1.0. In this way, matrices with a high correlation have a small distance, and smaller correlations have a larger distance. With multidimensional scaling (MDS; in R the useful function is “cmdscale”), the matrices can then be placed on a graph reflecting these distances. (See Figure 1).

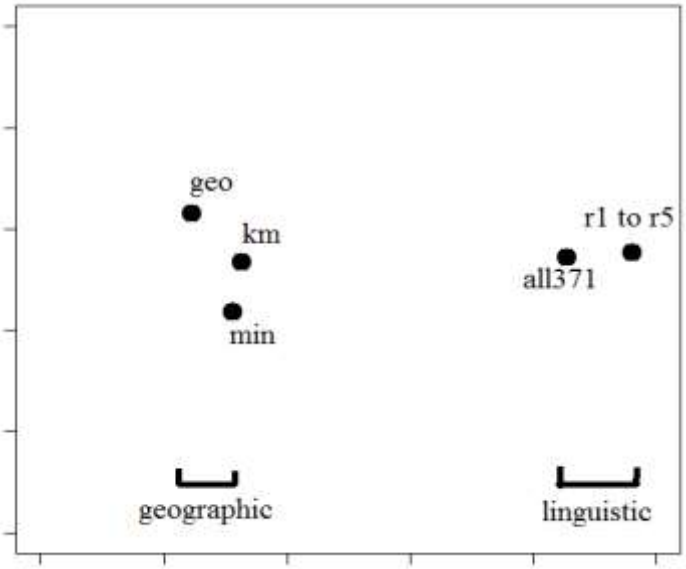

Figure. 1: Distance matrices for 370 maps, spread by correlation distance = (1-correlation coefficient)

The scales and dimensions are arbitrary, but reflect the relative correlation of the matrices. The average of the random samples (r1 to r5) is not as close to the linguistic matrices as the full data (all371) is, and the travel distance (km) and travel time (min) are somewhat closer to the geographic distance than the direct distance (geo) is. Furthermore, the linguistic matrices (all371, r1 to r5) are bunched together, and so are the three geographic distances, so that it is reasonable to talk about the differences between geographic and linguistic distance (without specifying which matrix in particular).

Functional subsets

On the linguistic side, it is possible to take subsets, not randomly, but based on linguistic function. Here, we take 143 (of the 370) maps that are about lexical items (as opposed to phonetic, morphophonemic, morphological or syntactic) and repeat the processes done for the full data set.
We used the same geographic matrices and added:

- lex 143 maps oriented to lexical items
- lex r1 to lex r5 randomly chosen subsets of lex, with 59 to 80 maps each.

The correlation coefficients among the linguistic distances are:

Table 5. Correlation coefficients among linguistic distances

	lex
lex	1.0
lex r2	0.9153888
lex r3	0.9275218
lex r4	0.9335443
lex r5	0.9431725
lex r1	0.9542546
Average r1 to r5	0.9347764
Standard deviation	0.014831732

In the case of “lex”, the random samples are not as tightly clustered around the full matrix as they were for “all371”, but they are still reasonable close. We might attribute the greater spread to the fact that sample size is much smaller.

The correlation coefficients for the geographic vs. linguistic matrices are:

Table 6. Correlation coefficients for geographic vs. linguistic matrices

	geo	km	min
lex	0.7990865	0.8089702	0.7948557
lex r2	0.7228087	0.7289443	0.7189263
lex r3	0.7230399	0.7488993	0.7503728
lex r4	0.7528244	0.7553345	0.7379978
lex r5	0.7555094	0.7718498	0.8048785
lex r1	0.7638457	0.7833471	0.7780887
Average lex r1 to r5	0.74360562	0.757675	0.75805282
Standard deviation	0.0172731436	0.0188038594	0.0302858115

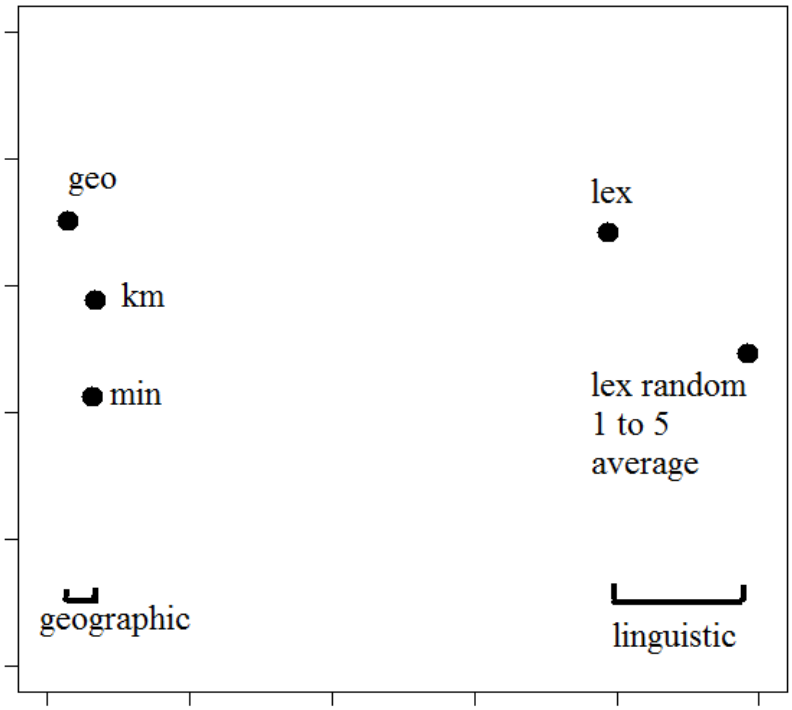

Figure 2: Distance matrices for 143 lexical maps, spread by correlation distance = (1-correlation coefficient)

Here, we get much the same results as for the all371 data set. Again, converting the correlation coefficients into distances, and viewing them in an MDS plot (see Figure 2), we see that the three geographic distances have similar relationships to the linguistic distances, with the km and min distances generally modestly better than the direct distances (though not always), and the random samples (represented by their average) not as close as the full lex data set. If we extend the same approach to other subsets, and create the MDS picture, we get Figure 3.

Here, we see the geographic distances on the right side (orientation and scale are

all relative to the data sets chosen), the morphophonemic and lexical data sets close to the geographic points, the phonetic data sets further away, and the morphological data set in isolation.

Such a visualization of what can also be represented by a table of correlation coefficients (not listed here) invites one to speculate on why one set correlates better than another with geography. Is it the case that people are more conscious of lexical forms (e.g. “spigot” vs “tap”) than say morphological variants (“ox”/ “oxen” / “oxes”), and so more consciously make the differences with distance travelled? We cannot say from the data here, but the visualization does raise the questions.

Visual presentation of statistical “facts” can be an important part of presenting,

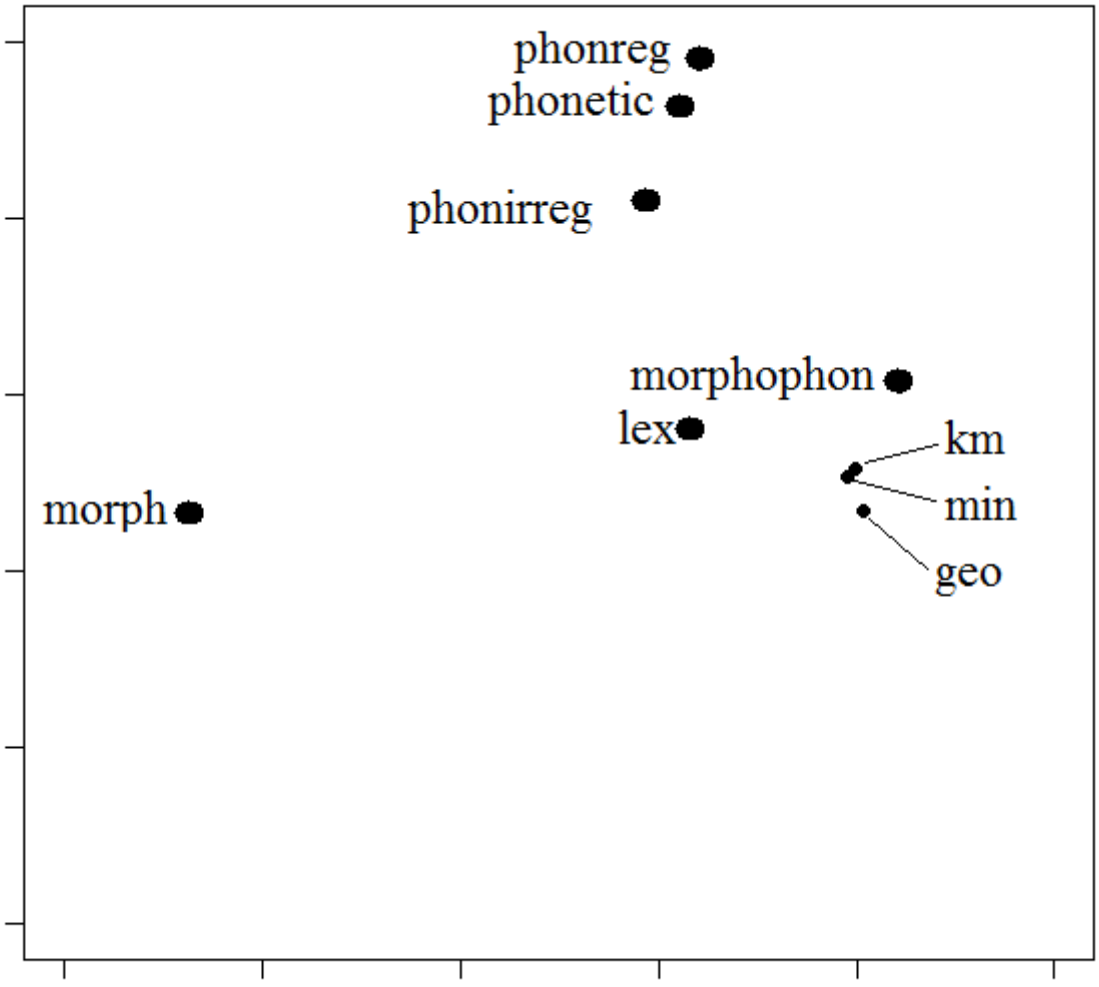

Figure 3: Distance matrices for various subsets of the linguistic data, spread by correlation distance = (1-correlation coefficient)

understanding, and confirming statistical results.

Conclusions

The hypothesis that dialect distance corresponds to geographic distance is amply supported by the evidence of the Romanian case. With a correlation coefficient of about 0.8 (and a very small p-value of 0.001998002), we not only see a good correlation, but we also see that there must be other factors at work, and can conclude that geography is not everything in this case.

Innovative approaches to defining geographic distance, using travel distance and travel time prove helpful, but only modestly, because there is only a modest difference among the various geographic distance matrices. It invites testing in a situation where more radical differences exist.

Tests for variation within the linguistic data sets, by processing random subsets, indicate that the linguistic data is fairly homogenous, and the results we observed are not merely based on the selection of linguistic data. Likewise, when we purposefully segment the linguistic data to focus on one type, we find that data is fairly homogenous, and that the functional subset has a similar relationship to the geographic distances as the full data set did.

NOTE: *For an MDS analysis of the contribution of different dialect features to the dialect structure of the Crişana dialect, based on both analytic (raw data) maps and on the 370 interpretive maps used in this paper, see Embleton, Uritescu, Wheeler 2013.*

References

Embleton, Sheila; Uritescu, Dorin & Wheeler, Eric S. (2016). *Play with the Data*! In: Kelih, Emmerich; Knight, Róisín; Mačutek, Ján & Wilson, Andrew (eds.), *Issues in Quantitative Linguistics*, vol. 4: *Festschrift for Reinhard Köhler on the occassion of his 65th birthday*, Lüdenscheid, RAM-Verlag, p. 128–134. (Studies in Quantitative Linguistics, 23).

Embleton, Sheila; Uritescu, Dorin & Wheeler, Eric S. (2015). Exploring Linguistic vs Geographic Distance Quantitatively. Presentation to VIII. Congress of the International Society for Dialectology and Geolinguistics (SIDG) 14 - 18 September 2015, Famagusta, Turkish Republic of North Cyprus.

Embleton, Sheila; Uritescu, Dorin & Wheeler, Eric S. (2013). *Continuum et fragmentation géolinguistiques d'après l'Atlas linguistique de la Crişana en ligne*, In: Casanova, Emili Herrero & Calvo, Cesáreo Rigual (eds.), *Actes del 26é Congrés de Lingüística i Filologia Romàniques*, vol. VI, Berlin, de Gruyter, p. 119–129.

Legendre, Pierre & Fortin, Marie-Josée (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecological Resources*. 10. 831-844. doi: 10.1111/j.1755-0998.2010.02866.x

R. 2012. R version 2.15.2 (2012-10-26) -- "Trick or Treat". The R Foundation for Statistical Computing. ISBN 3-900051-07-0. Platform: x86_64-w64-mingw32/x64 (64-bit)

Statistical Distributions of Parts of Speech Frequencies in Polish. Big Data Analysis*

Adam Pawłowski[1], *Krzysztof Topolski*[1], *Piotr Malak*[1],
Jan Kocoń[2], *Michał Marcińczuk*[2]
[1]*University of Wrocław*
[2]*Wrocław University of Science and Technology*

Abstract
In the present paper statistical distributions of vocabulary in a big corpus of Polish and in its sub-corpora are estimated and compared. A hypothesis is tested which states that statistical distributions in the entire corpus and in its sub-corpora are different with regard to type and/or parameter values (in case of similar models). The research was carried out on the material of a balanced corpus of Polish using NLP programming tools prepared in the framework of the CLARIN-PL consortium.

The results prove that only some parts of speech (adverbs and interjections), if they are automatically extracted in a separate corpus, have well determined statistical distributions. The models fitted were respectively inverse Gauss-Poisson and Zipf-Mandelbrot distribution. Other POS classes did not pass multidimensional chi2 test. For function words Zipf-Mandelbrot was fitted and for named entities – finite Zipf-Mandelbrot model. One reason for this is that corpora are objects created by engineers or linguists to serve various practical purposes, but they do not represent adequately all properties of texts created by humans. Another reason is that POS in big corpora are not coherent classes of lexemes because they are defined using different, sometimes contradictory criteria.

Keywords: *statistical distributions, Zipf-Mandelbrot distribution, Gauss-Poisson distribution, parts of speech, big data, text corpus, Polish*

1. Research objectives and hypotheses

The objective of this article is to examine selected quantitative characteristics of certain natural language lexical subsystems on the basis of corpus data. For the purpose of the study, such subsystems were defined as the basic parts of speech (or POS – noun, verb, adjective, adverb, participle, interjection, numeral, functions words), as well as certain classes derived on the basis of semantic criteria, namely common nouns and one-word named entities (vaguely corresponding to proper names). These word classes were identified and named in antiquity by Dionysius Thrax (circa 100 BC). They have been codified for centuries in Indo-European language grammar books and are commonly used in teaching curricula. Therefore they are recognizable to users and are perceived, above all by people living in Indo-European language usage regions, as being natural, even universal. However, these subsystems are not identified according to clear criteria, hence any attempt to analyze them in corpus research must inevitably

* Address correspondence to: Prof. Adam Pawłowski, Uniwersytet Wrocławski, Instytut Informacji Naukowej i Bibliotekoznawstwa, pl. Uniwersytecki 9/13, 50-137 Wrocław, Poland. Email: adam.pawlowski@uwr.edu.pl

deal with inconsistencies. For example, there are 35 lexical categories (with particular inflections) defined in the Polish Language National Corpus that do not correlate with categories defined by the educational canon (Przepiórkowski, Bańko, Górski, et al.: 62-67). POS definitions are essentially a peculiar combination of syntactic and semantic criteria, but above all remain heavily influenced by a centuries-old didactic tradition that serves only to classify vocabulary and not to elucidate its internal structure. The resulting status quo is that certain classes may overlap (for example, verbs and adjectives in the form of a past participle, verbal and nominal gerunds); there is also a significant number of units that do not have a precise categorization. The upshot of this inconsistency and lack of classificatory criteria means that lexical categories are only seemingly natural or obvious and, in fact, constitute valuable objects of corpus research and analysis.

Due to the fact that the above-mentioned division of the lexical system into POS subsystems has proven to be exceptionally impervious to changes in language theory and methods, we have based our research on the assumption (hypotheses) that, despite difficulties with automatic segmentation of text into relevant lexical units, sub-corpora consisting of individual parts of speech should have specific properties, also in the light of methods of lexical statistics. This specificity may be the result, inter alia, of the fact that individual parts of speech serve different functions in a sentence (discourse), and thus possess particular distributive (compositional) characteristics. We assume, therefore, that the study of the statistical distribution of POS frequency on a major corpus of texts may indirectly reflect these specific features.

An innovative feature of the approach presented in this study is the fact that it takes into account categories of named entities as opposed to common nouns, as well as entities not recognized by the morphosyntactic tagger. This type of study has never been undertaken on the basis of the Polish language because there were no tools available for automatic selection of this semantic category (named entities) from large corpora, and manual examination was far too labour intensive. Research based on these classes of vocabulary must be regarded as the exploration of entirely new issues, making it difficult to offer any plausible hypotheses. Only in the case of named entities can one suppose that their distribution is similar to that of nouns, as named entities are functionally nominal units.

From among the many available tools of quantitative study of vocabulary we have chosen statistical distribution of lexeme frequency as a means of verifying the above-stated hypothesis. In particular, two types of information will be generated for each particular sub-corpus of POS and of named entities: the best distribution type, and parameter values of the model in the case of matching the same distribution type. However, one cannot reject the possibility that any plausible statistical distribution fits the data in a statistically satisfactory way. Actually many specialists in quantitative linguistics (e.g. Reinhard Köhler) claim that corpora are a sort of artificial objects, which do not have properties comparable to the authorial texts, created by humans as coherent and closed entities. Large corpora should thus display a rather complicated mixture of statistical features, typical for different individual styles. This is particularly important here, as big corpora of Polish are processed. Despite these objections we decided to carry out a corpus analysis of Polish and corroborate or reject the above hypotheses. Actually big data analysis of the Polish language is potentially an appropriate method

for extracting and synthesizing information (if it exists!) as it brings the researcher closer to discovering properties that can be regarded as points of reference for more specific research projects that focus, for example, on small corpora, authorial texts or stylistic varieties.

It should be noted that the essence of corpus research and, at the same time, that which distinguishes corpus linguistics from normative linguistics, is the necessity of carrying out an analysis that is all-encompassing. All encountered entities must be processed or included by some means, and not simply those that fall within the adopted paradigms, categories or other pre-defined schemes. The traditional approach to linguistic analysis, based on small (e.g. authorial) corpora or solely on intuition and selected examples of use, does not take into account entities that for some reason are regarded as difficult, e.g., partly incorrect, colloquial (conversational) or occasional (and thus not found in any dictionary), acronyms (in a concise form in texts, but extended in speech), alphanumeric combinations, etc. Automatic methods have not brought about any major change in this situation. At the time when the digital text more and more dominates over speech, languages continue to be dynamic systems that are constantly changing together with the communities that use them. And they do not become more rational or correct then they were in the past. From this pragmatic perspective on data follows the assumption that the results obtained in a study of real texts will always contain some percentage of error[1], which in the best case may be minimised by referring to the law of large numbers.

2. Methodology and corpus

The research method applied in this project includes two basic modules: the preparatory phase and modelling. The resulting model is then subject to interpretation. The preparatory phase consists of the following stages: transformation of the corpus to lemmatized form, morphosyntactic annotation, extraction of sub-corpora, and generation of so-called frequency spectra (called also blind frequency lists).

Typically two ways of summarizing the word frequency counts are considered in the quantitative linguistic literature. The first one is a rank frequency distribution which is obtained when in the sample of size N the frequency $f(k,N)$ of the k-th type is presented as a function of its rank k. The ranks are assigned to frequencies in such a way that $f(k,N) \geq f(k+1,N)$, for all $k \geq 1$. The second way to present frequencies is a grouped frequency distribution or the frequency spectrum, $V(m,N)$, $m \geq 1$, which is defined as the number of types with frequency m in a sample of N tokens. Formally we may write:

[1] For example, in the corpus that is the subject of this research project, the number of units recognised as tokens with a frequency higher than 3 amounted to 296 021 749, while the number of types amounted to 921 142. However, the number of tokens ignored by the morphosyntactic tagger in the text totalled 12 054 564 (approx. 4%), while the number of types ignored amounted to 337 423, which is nearly 37%. In the case of the complete corpus, which took into account the frequency of 2 and 1, the number of ignored types exceeded 50%, and 10% of the tokens. Data curation improved the quality of the result, of course, but errors are virtually impossible to entirely avoid.

$$V(m,N) = \sum_{i=1}^{V(N)} I\{f(i,N) = m\}$$

where $V(N) = \sum_m V(m,N)$ and the indicator function $I\{x\}$ is equal 1 if expression *x* is true, and zero otherwise.

Table 1. The rank frequency list is presented in the left table and in the right table there is the corresponding frequency spectrum

Rank	**Lemma**	**Frq.**	**m**	**V(m)**
1.	a	9	1	7
2.	b	8	2	5
3.	c	5	3	4
4.	ab	5	5	2
5.	ba	3	8	1
6.	ca	3	9	1
7.	caab	3		
8.	acaa	3		
9.	caba	2		
10.	baca	2		
11.	aacc	2		
12.	ccaa	2		
13.	aaab	2		
14.	bbbc	1		
15.	bbba	1		
16.	cccb	1		
17.	cbcb	1		
18.	abca	1		
19.	acba	1		
20.	acca	1		

There is a natural connection between rank-frequency distribution and frequency spectrum:

$$V(m,N) = \sum_i I\{f(i,N) \geq m\} - \sum_i I\{f(i,N \geq m+1)\}.$$

Assume that a rank–frequency distribution is described by Zipf-Mandelbrot law [M] of the form:

$$f(i,N) = \frac{K}{(i+b)^a},$$

where $a > 1$ and $b \geq 1$ are parameters and K is the normalizing constant.

Notice that $I\{f(i,N) \geq m\} = 1$, if and only if $\left(\frac{K}{m}\right)^{1/a} - b \geq i$, which implies the following equation:

$$V(m,N) = \left(\frac{K}{m}\right)^{1/a} - b - \left[\left(\frac{K}{m+1}\right)^{1/a} - b\right] \quad \text{or} \quad V(m,N) = K^{1/a}\left(\frac{1}{m^{1/a}} - \frac{1}{(m+1)^{1/a}}\right).$$

It may happen that for some high frequency occurrences the number of words with such frequency is zero. In table 1 such situation happens for m = 4 and 6. This fact must be taken into account during the procedure of fitting the distribution to data.

The Sichel model uses generalized inverse Gauss-Poisson distribution as a description of word probabilities (Sichel, 1975). Let us recall that the probability density function for generalized inverse Gauss-Poisson distribution has the following form:

$$g(x) = Mx^{a-1}\exp(-\frac{x}{c} - \frac{b^2c}{4\mathrm{x}})$$

The normalizing constant M is of the form $M = \frac{(2/bc)^{a+1}}{K_{a+1}(b)}$, where $K_a$denotes the modified Bessel function of the second kind of order a. The maximum likelihood estimators of the generalized inverse Gauss-Poisson distribution parameters are described in (Sichel, 1982).

Taking into account a somewhat wider spectrum of publications on quantitative linguistics, it should be noted that many researchers are interested in rank-frequency distributions, and not in frequency spectra used in this study. The approach based on ranks is, however, burdened with a certain error, which consists in the fact that interrelationships are constructed between the values belonging to a rank scale (or so called ordinal scale), composed of subsequent integer numbers, and the frequency expressed de facto in real numbers (although absolute lexeme frequencies are integers, they can be easily transformed into relative values or relative probabilities of occurrence, which are real numbers). Frequency spectra do not have this drawback as rank scales are not applied.

As far as input data are concerned, we used one of the largest corpora of Polish created by G4.19 Research Group at Wrocław University of Science and Technology, called KGR10. It contains more than 4,3 billion tokens and covers texts from a wide range of domains like: blogs, science, stenographic recordings, news, journalism, books, parliamentary transcripts and the content of Polish websites listed in DMOZ (a multilingual open-content directory of World Wide Web links). All texts come from the second half of the 20th century and represent the modern Polish language. Due to component licenses, KGR10 corpus is not publicly available.

The KGR10 corpus was pre-processed by a set of basic tools for text processing, i.e. WCRFT – a morphosyntactic tagger for Polish (Radziszewski, 2013), Liner 2 – a tool for named entity recognition for Polish (Marcińczuk et al., 2017) and Polem – a lemmatizer for multi-word expressions and named entities for Polish (Marcińczuk, 2017). WCRFT was used to split the texts into sequences of tokens and to assign morphological tags for each of them (according to the NKJP tagset (Przepiórkowski, 2011)). Liner 2 was used to identify occurrences of named entities – sequences of tokens which should be treated as single units, not as sets of independent tokens. As the WCRFT tagger provides lemmas on the level of single tokens we needed to lemmatize

the multi-word named entities separately. To lemmatize the named entities we used the Polem tool.

In both cases, the system does not recognize multi-word units, and does not include semantic information that distinguishes polysemous meanings or homonyms. This in fact does not deviate from the standards of big data research in quantitative linguistics, because even the current natural language processing technology does not allow for such distinctions. Distribution, selection and fitting were conducted in such a manner that, first of all, an effort was made to select one distribution type that fits all the sub-corpora being examined, and then its parameters were compared. If this type of distribution proved to be inadequate for some subset, then another was chosen that was a better fit.

It was assumed that fitting of different distributions for specific POS would definitively confirm the proposed hypothesis regarding the specificity of each subset of POS and named entities. It was also assumed that the above hypothesis will be confirmed by the presence of the distributions of the same type, but with significantly different parameter values. If, on the other hand, some (or all) of the tested subsets are fit to the same distribution with very similar characteristics, the hypothesis that assumes specific quantitative properties of subsets of the parts of speech will be disproven.

Finally an extremely negative hypothesis was also considered that among available statistical distributions described in the literature on quantitative linguistics, there will be no model fitting data in a statistically satisfactory way. It should be stated here that graphical data presentation may be misleading in that it displays apparently similar lines, but actual values do not pass appropriate quality tests (e.g. chi2).

The parametric models of the frequency distributions which we considered in this paper are log-normal, generalized inverse Gauss-Poisson and generalized Zipf's. These classes of distribution laws are proposed in relation to word frequency distribution by Herdan (1960), Sichel (1975) and Orlov / Chitashvili (1983) respectively and discussed in detail by Harald Baayen (2001). We evaluated goodness-of-fit of the models following the procedure presented in Baayen (*ibid*.).

Let $V(N)$ denote the observed vocabulary size at sample size N and let $V(m,N)$ denote the number of types with frequency m in a sample of N tokens. For the vector $\boldsymbol{x} = (V(N), V(1,N), \ldots, V(k,N))$, the following goodness-of fit statistic is evaluated:

$$X^2_{N,k} = (\boldsymbol{x} - \boldsymbol{m})^T \sum^{-1} (\boldsymbol{x} - \boldsymbol{m}),$$

where $\boldsymbol{m} = (E[V(N)], E[V(1,N)], \ldots, E[V(k,N)])$ is the vector of expectation and $\sum^{-1}$ is the corresponding covariance matrix, both computed under evaluated model word frequency distribution.

The parameters of the model are estimated by comparing sample value of $V(1,N)$ and $V(N)$ with $E[V(N)]$ respectively. For the model with r parameters, statistics $X^2_{N,k}$ has χ^2_{k+1-r} distribution. The details of described test procedure are described in (Baayen 2001).

3. State of research

To the best of our knowledge a big data study of statistical distributions of specific parts of speech in the Polish language has never been previously carried out. Neither have such research projects been found regarding other languages, apart from very few exceptions.

It is indicative of this state of affairs that, in the book *Problems in Quantitative Linguistics* (Strauss, Fan, & Altmann 2008), which is a collection of interesting problems in quantitative linguistics in need of resolutions, the question of modelling word class distributions is broached. The authors propose the following claim as a working hypothesis: "The rank-frequency distributions of different word classes abide by the same probability distribution." (*ibid*. 27). The authors also propose the following: „Count different word classes (noun, verbs, adjectives, adverbs,...) separately in a text. If there are ambiguous cases, decide ad hoc to which class a word belongs. Then fit the same probability distribution to all empirical distributions, e.g Zipf's truncated zeta distribution $P_x = C/x^a$ $(x = 1,2,3, \dots, n)$ where C is the normalizing constant and $n = x_{max}$. Examine the behaviour of the parameter a. Is it equal in all cases or are there differences?" (*ibid*.) Actually the authors do not explain, why different word classes should behave in this way, neither the notion of a word-class is defined.

There are also studies which model the distribution of occurrences of different grammatical classes in the text, i.e. frequencies of nouns, verbs, adjectives etc. (and not particular lexeme frequencies) against their ranks. Examples of this type of analyses can be found in the "invitation" to quantitative linguistics by Karl Heinz Best and Otto Rottmann (2017: 64-66).

This scarcity of works does not mean, however, that the researchers totally ignored the information potential flowing from the quantitative analysis of the participation of the parts of speech in the texts. Their main concern was, however, stylometric and material work focused on building statistical linguistic profiles of specific authors. As Bernard Moreaux affirmed, „La fréquence des 'parties du discours' concerne moins directement la syntaxe car ces classes sont hétérogènes puisqu'elles sont fondées sur des critères à la fois morphologiques, syntaxiques et même, dans le cadre de la traditionnelle, sémantique. Leur hétérogénéité même les rend surtout utiles dans l'étude de cet 'objet' hétérogène qu'est le style." (Moreaux, 1982: 308). Such studies were generally carried out on relatively short texts, often authorial works, and not on large corpora. It is also quite characteristic that the researchers did not estimate statistical distributions, limiting themselves to examination of the frequency of POS occurrence in the entire text. A good example of the earliest stylometric research of this kind are the pioneering computer-assisted quantitative analyses of ancient era texts published by the staff of the Laboratoire d'Analyse Statistique des Langues Anciennes at the University of Liège during the early 1960s. Some of the problems approached at that time are of interest also today and, despite rather rudimentary methodology, a few of the studies are worth mentioning at this point.

Describing the chronology of the works of Seneca, Anna Nicolova-Burova used occurrence of POS frequency to distinguish between dialogue texts and descriptive texts. The former demonstrated, as could be expected, a higher turnout of verbs, while the descriptive texts demonstrated a high frequency of nominal classes (Nicolova-

Burova 1975:20). This criterion is fairly common sense, and also appears in other works on the subject of stylostatistics (cf. Kamińska, 1990: 84–85). For instance Gilles Maloney discussed the frequency of parts of speech in a study of Euripides' "Electra" (Maloney, 1970). In subsequent years Charles Muller applied POS distribution in his works on Molière (e.g. Muller 1979). This approach is also often found in the British linguistics, which since the late nineteenth century has always evinced a preference for practical applications of statistical methods as opposed to a search for universal or theoretical models. Researchers have always gladly taken on the problem of authorship attribution of great cultural texts. Carrington B. Williams, among others, has recommended analysing specific word classes by looking at POS in terms of average length and frequency (Williams, 1970: 41–51, 64–72). Calculating POS numerical parameters was also a common practice in frequency dictionaries of various languages, published often in the 20th century.

The approach applied in this article differs from the previous ones, as described above, in many aspects because it uses the frequency spectra method, it is based on automatic text annotation, and, last but not least, it is carried out on a corpus which is unprecedented in terms of size.

4. Summary results

After analyzing the data, it appeared that the best fit was obtained with two models of word frequency distribution: generalized Zipf's and Sichel's. Estimations of the distribution functions for particular POS are presented below.

Nouns

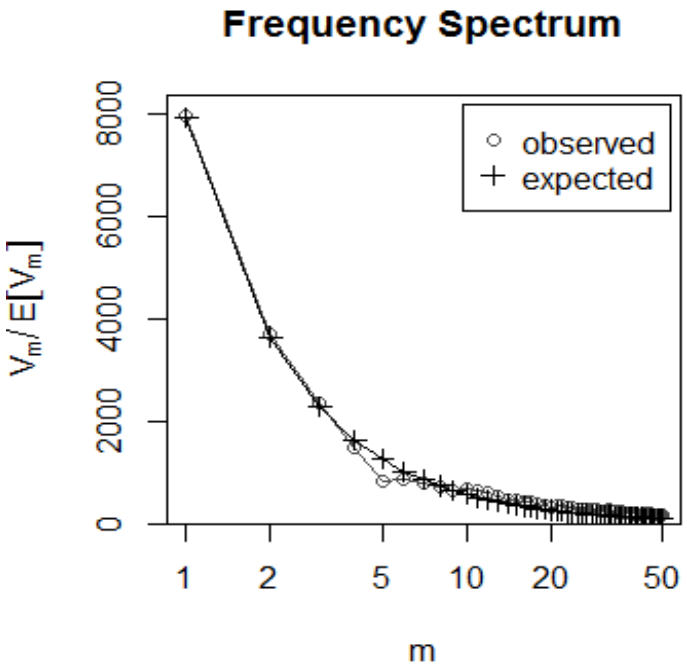

Figure 1a. The frequency spectrum for nouns (circles), the finite Zipf-Mandelbrot fit (solid line and crosses).

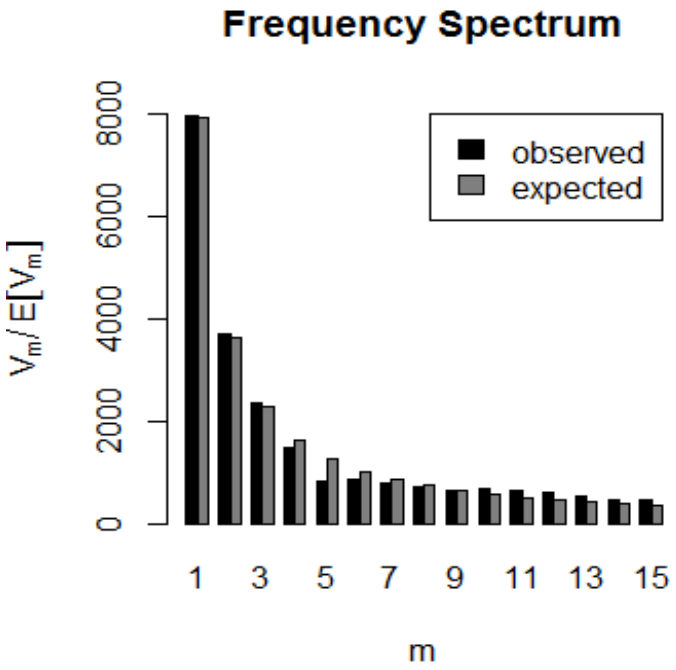

Figure 1b. Bar plot for first 15 spectrum elements.

Verbs

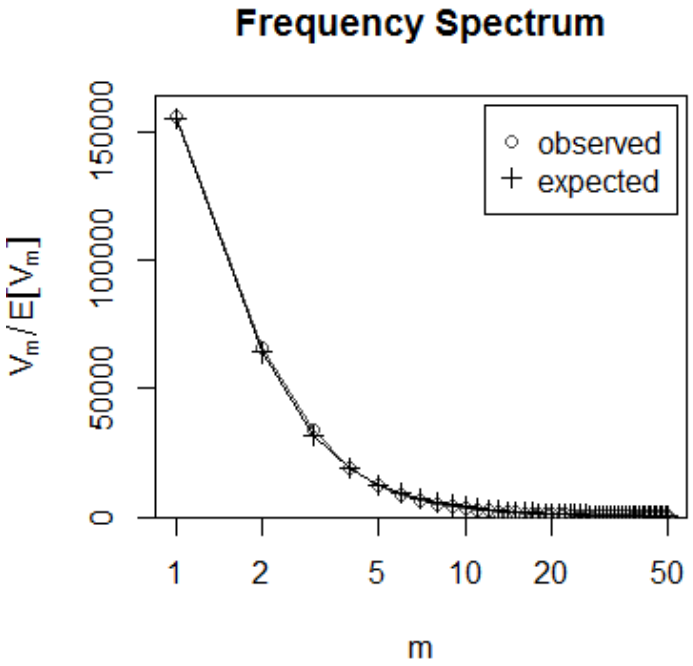

Figure 2a. The frequency spectrum for verbs (circles), the finite Zipf-Mandelbrot fit (solid line and crosses).

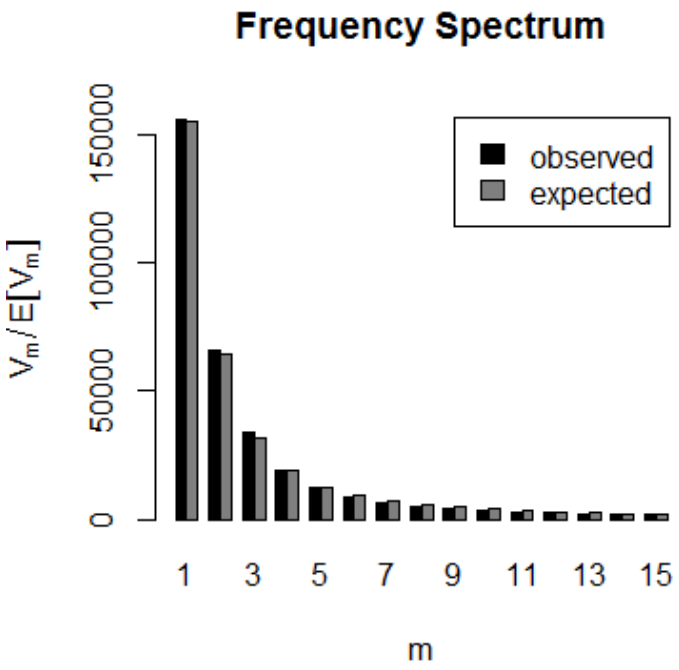

Figure 2b. Bar plot for first 15 spectrum elements.

Adjectives

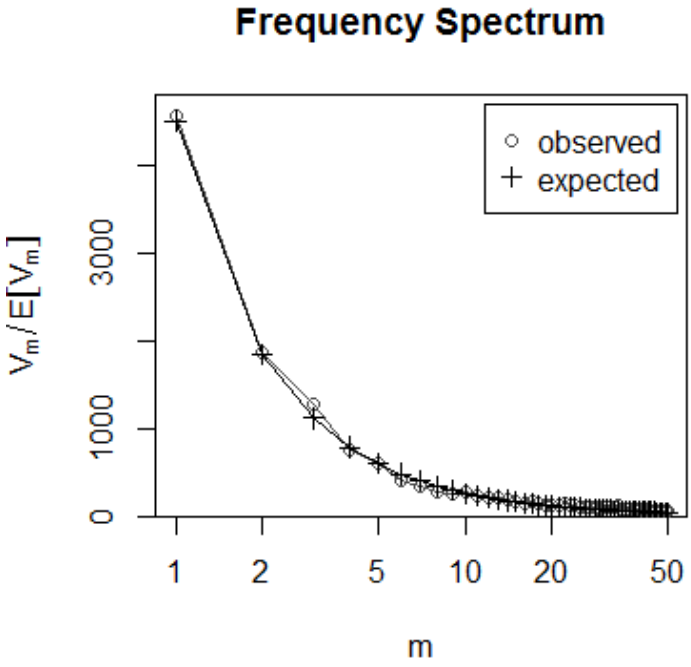

Figure 3a. The frequency spectrum for adjectives (circles), the Zipf-Mandelbrot fit (solid line and crosses).

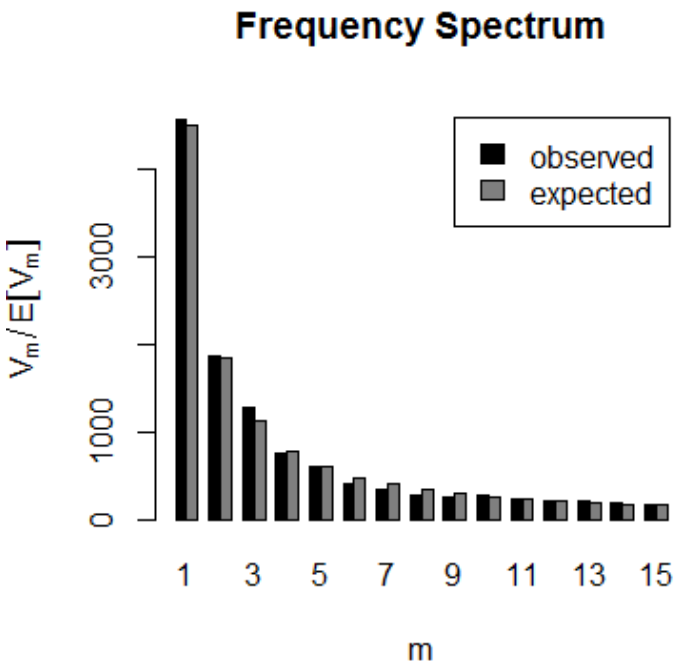

Figure 3b. Bar plot for first 15 spectrum elements.

Adverbs

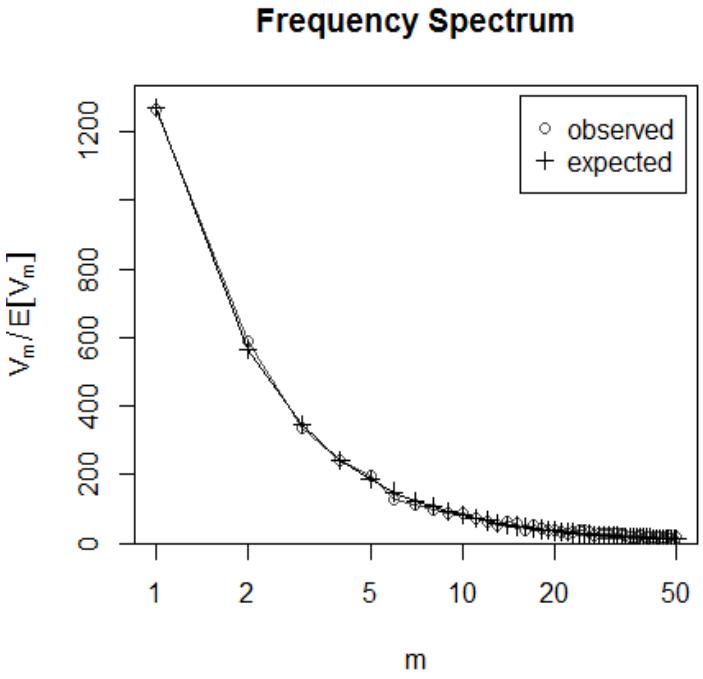

Figure 4a. The frequency spectrum for adverbs (circles), the generalized-Gauss-Poisson fit (solid line and crosses).

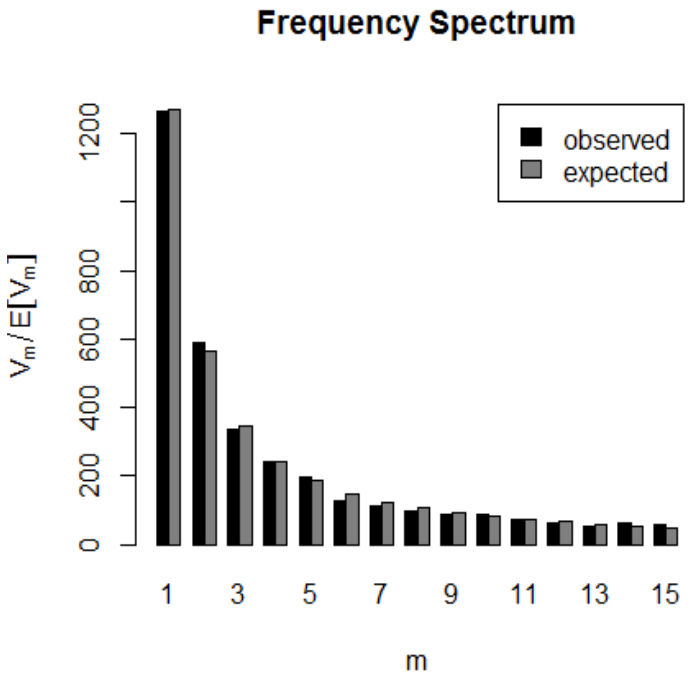

Figure 4b. Bar plot for first 15 spectrum elements.

Numerals

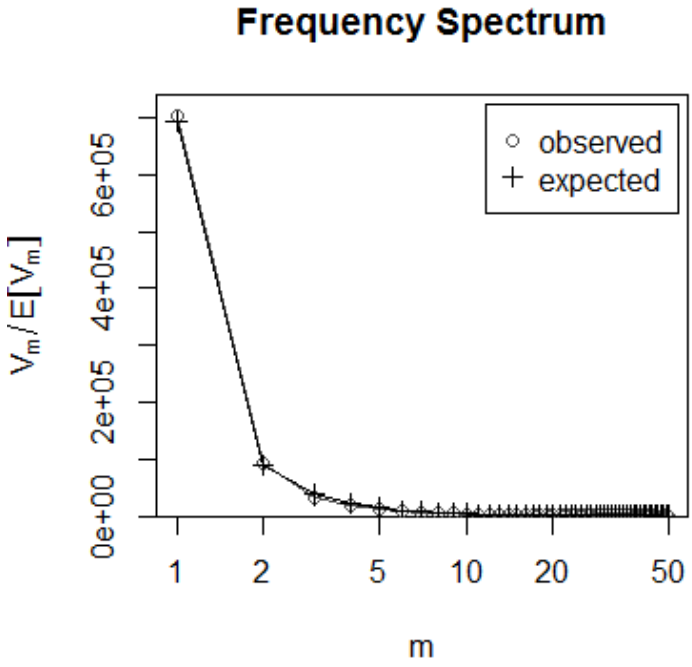

Figure 5a. The frequency spectrum for numerals (circles), the Zipf-Mandelbrot fit (solid line and crosses).

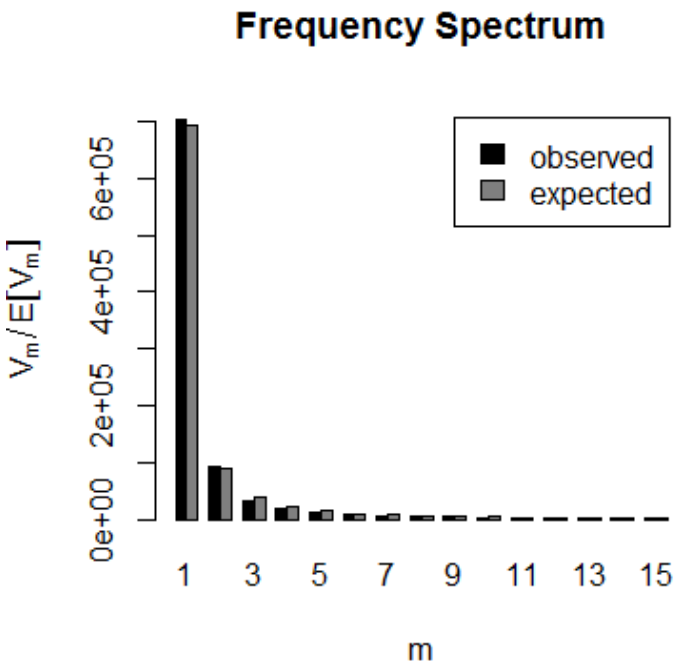

Figure 5b. Bar plot for first 15 spectrum elements.

Interjections

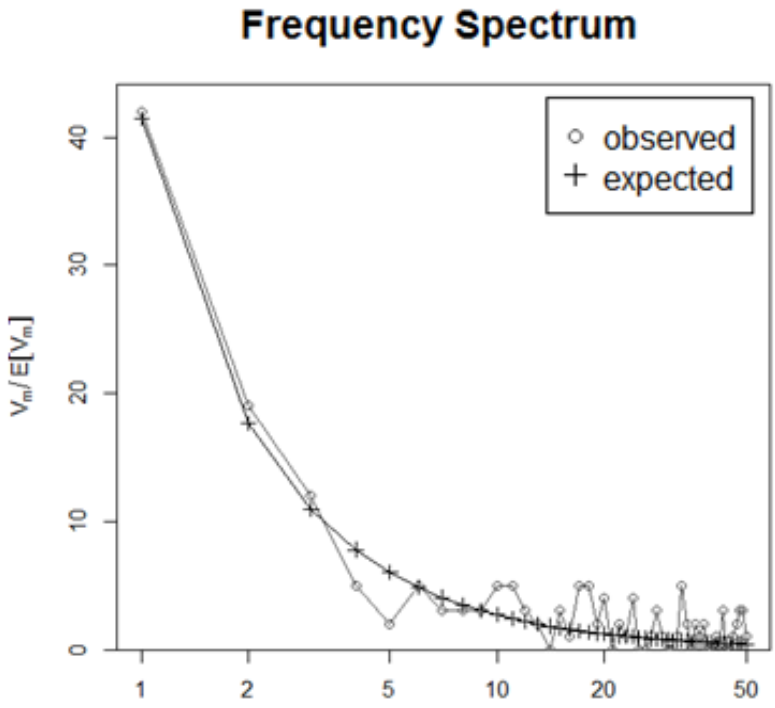

Figure 6a. The frequency spectrum for interjections. (circles), the Zipf-Mandelbrot fit (solid line and crosses).

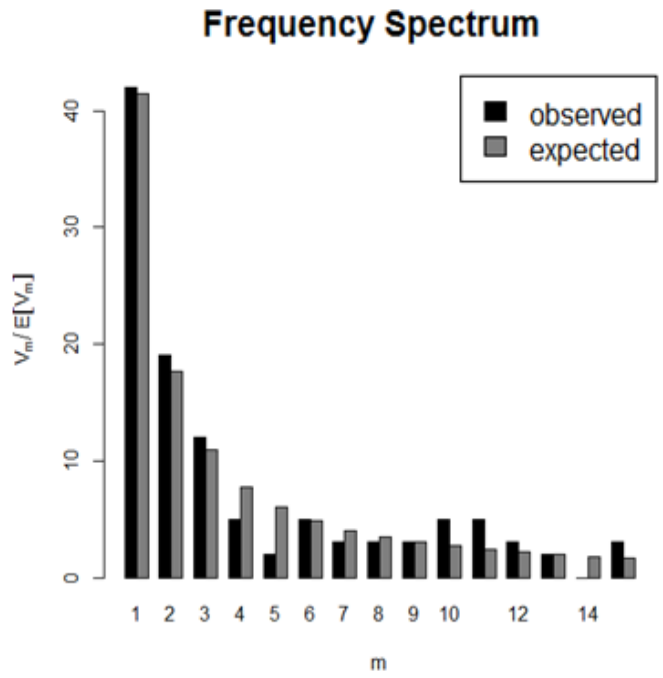

Figure 6b. Bar plot for first 15 spectrum elements.

Unrecognised units

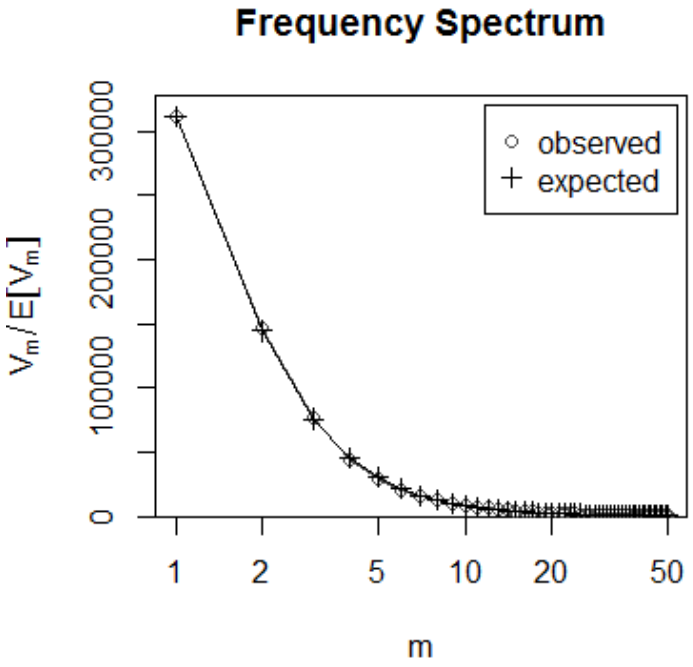

Figure 7a. The frequency spectrum for adverbs (circles), the generalized-Gauss-Poisson fit (solid line and crosses).

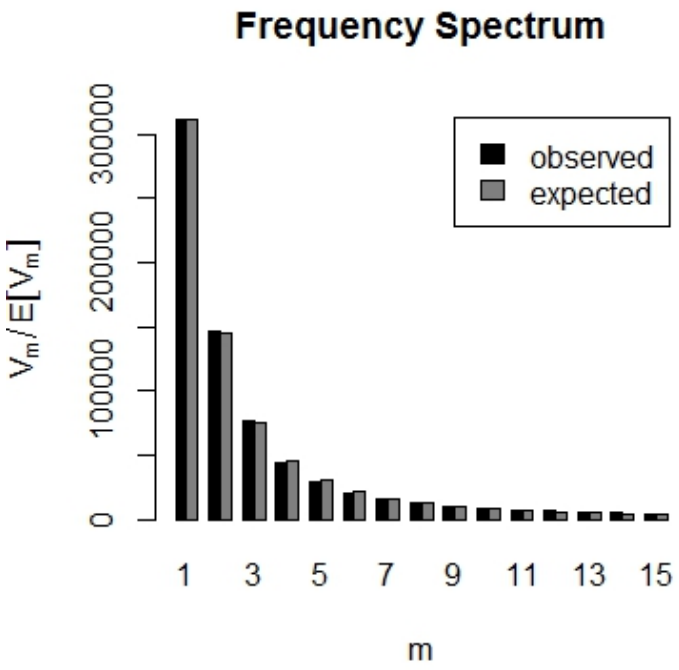

Figure 7b. Bar plot for first 15 spectrum elements.

Function words

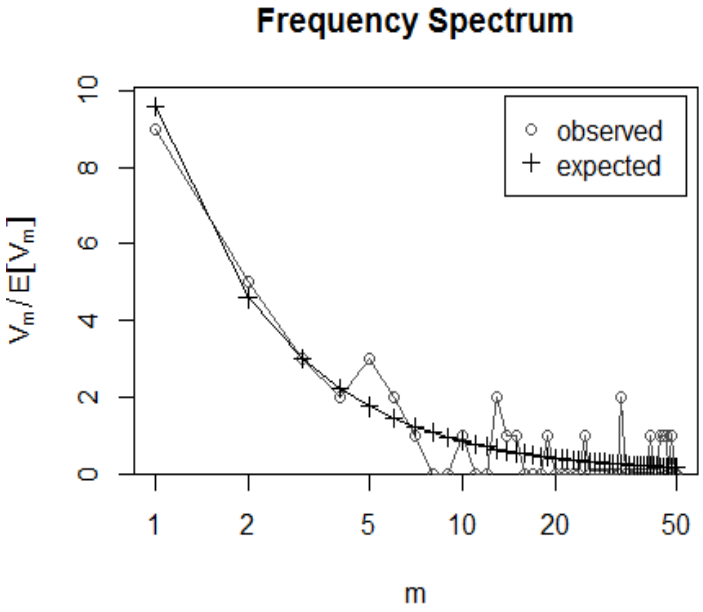

Figure 8a. The frequency spectrum for grammatical (circles), the Zipf-Mandelbrot fit (solid line and crosses).

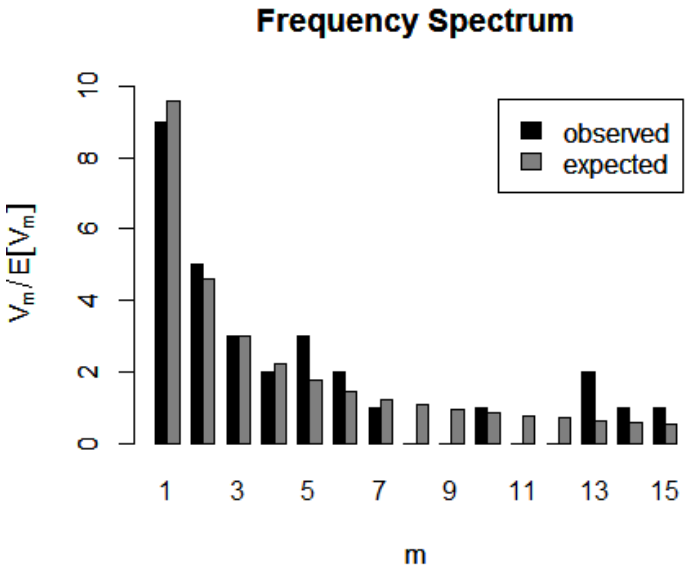

Figure 8b. Bar plot for first 15 spectrum elements.

Named entities

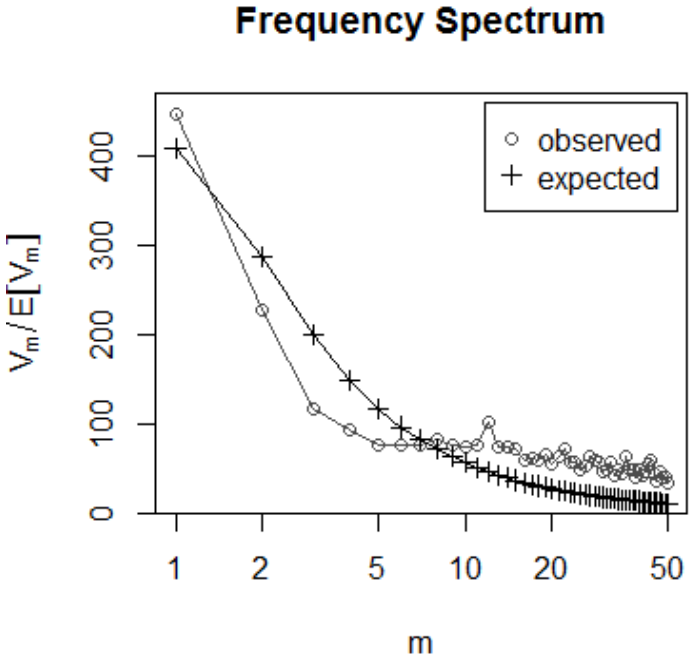

Figure 9a. The frequency spectrum for named entities (circles), the finite Zipf-Mandelbrot fit (solid line and crosses).

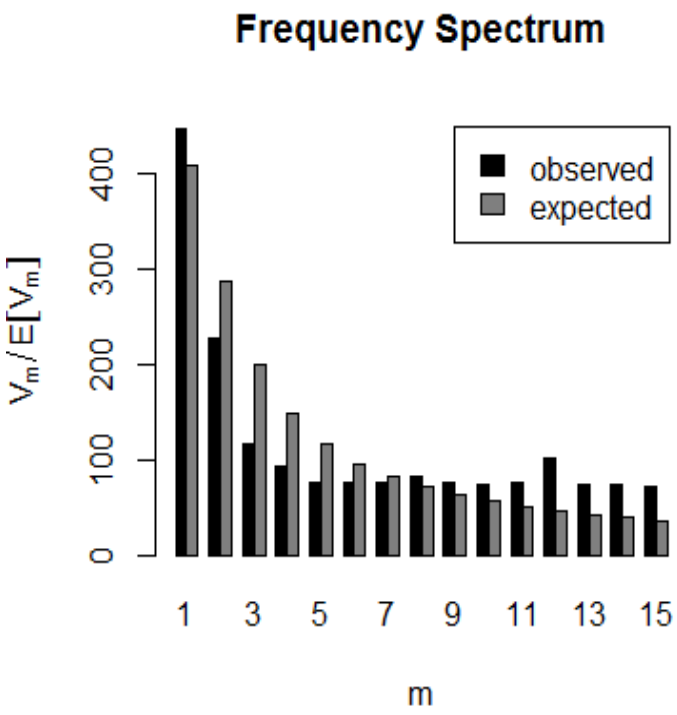

Figure 9b. Bar plot for first 15 spectrum elements

After estimating the model we proceeded to the evaluation of goodness-of-fit of the results obtained. Calculations were performed with the help of *ZipfR* package (Evert & Baroni, 2007). The best fit for POS is given in the Table 2 where fZM, ZM and GIGM denote Zipf-Mandelbrot, finite Zipf-Mandelbrot and generalized inverse Gauss-Poisson distribution respectively. Out of all the POS sub-corpora analysed only adverbs, function words and interjections passed multivariate chi-square tests and can be considered as statistically satisfactory (values of test statistics are indicated in Table 3). Other POS sub-corpora failed this test, despite being visually correct (values of test statistics for nouns, verbs, adjectives, adverbs, numerals and unrecognized units are indicated in Tab. 4). In the case of nouns it was even impossible to fit generalized Gauss-Poisson model with reasonable parameters. Interestingly enough, the visual inspection of the fits presented in figures 1-3, 5 and 7 may be considered as reasonable. Human perception and statistical tests are, however, different from one another. Chi-square test is very sensitive to large values (of frequency spectrum), much more sensitive than human eye, and therefore it produces negative results even when graphical presentation looks satisfactory (Grotjahn &Altmann, 1993).

5. Conclusions and discussion

The aim of the study was to check the relevance of the quantitative approach, when applied to the centuries-old problem of division of lexemes into parts of speech. Using a big corpus of Polish, we tried to determine statistical distributions that could provide a reference model for studies on small authorial texts or corpora that represent the spe-

cific language styles, genres or varieties. It was determined that content words display different spectra compared to function words. Interestingly, content words behave in a different way too, but, regarding mere parameter values of respective models (GIGP, ZM, fZM), it was difficult to discover some convincing pattern. It is probably due to the fact that corpora are mixtures of styles and genres where specific lexemes have various syntactic functions and distribution in a sentence. The hypothesis posed at the beginning, which assumed that there are significant statistical differences between POS was thus rejected. On the other hand, our study showed that common and proper nouns display fundamental differences, probably due to their different semantic status. Quite surprisingly, the "error" class of unrecognized units displayed a pattern that was not abnormally different from regular POS patterns[2].

Table 2. Word class distributions – summary of results.

POS	**Distribution**	**Par. *a***	**Par. *B***	**Par. *c***
Noun	fZM	0.134363	4.448E-10	0.000465
Verb	fZM	0.635685	3.990E-10	0.524694
Adjective	ZM	0.183516	0.001464	-
Adverb	GIGP	-0.185368	0.001082	0.005377
Numerals	ZM	0.742752	1.025647	-
Interjections	ZM	0.146975	0.085470	-
Named entities	fZM	0.059084	5.952E-08	0.002642
Gramaticals	ZM	-0.185368	0.146975	-
Unrecognized units	GIGP	-0.8500646	0.007458	0.002243

Table 3. Result summary of multidimensional chi-square test for adverbs and interjections (positive results)

	Adverb	**Interjection**	**Gramaticals**
Parameter	**GIGP model**	**ZM model**	**ZM model**
a	-0.185	0.147	0.046
b	0.001	0.085	0.212
c	0.005	-	-
CHI^2	12.46	5.040	1.167
df	13	4	3
p	0.491	0.283	0.983

[2] The same is true for “kubliki”, another class of recognised and repetitive but unqualified POS.

Table 4. Result summary of multidimensional chi-square test – negative results[3].

Zipf-Mandelbrot model:

	Nouns	**Verbs**	**Adjectives**	**Adverbs**	**Interjections**	**Numerals**	**Named entities**	**Grammaticals**	**Unrecognized**
a	0.132	0.541	**0.184**	0.219	**0.147**	**0.743**	0.052	**0.045**	0.609
b	0.0005	0.039	**0.0014**	0.008	**0.085**	**1.025**	0.003	**0.212**	0.0003
X^2	722.89	50937	**117.45**	91.37	**-**	**8921**	1414	**0.166**	436560
df	14	14	**14**	14	**5.039**	**14**	14	**3**	14
p	34E-145	0	**2E-18**	2E-13	**4**	**0**	1.E-293	**0.983**	0

Finite size Zipf-Mandelbrot model:

	Nouns	**Verbs**	**Adjectives**	**Adverbs**	**Interjections**	**Numerals**	**Named entities**	**Grammaticals**	**Unrecognized**
a	**0.134**	**0.636**	0.184	0.230	0.147	0.743	**0.059**	0.473	0.789
Lower cutoff	**4E-10**	**4E-10**	3E-22	1.E-08	2.334	7.E-25	**5.9E-08**	2E-09	2E-08
Upper cutoff	**0.0005**	**0.525**	0.0015	0.0081	0.085	1.0293	**0.00264**	0.209	0.0008
X^2	**602.48**	**2534**	117.45	15.3	5.04	8926	**762.9**	1.129	3724
df	**13**	**13**	13	13	3	13	**13**	3	13
p	**2E-120**	**0**	6.E-19	0.28	0.16	0	**1E-154**	0.770	0

Generalized inverse Gauss-Poisson model:

	Nouns	**Verbs**	**Adjectives**	**Adverbs**	**Interjections**	**Numerals**	**Named entities**	**Grammaticals**	**Unrecognized**
a	-	-0.615	-0.148	**-0.185**	-4.3E-06	-0.675	-	-	**-0.850**
b	-	4E-05	0.0012	**0.0011**	1.5E-11	1.3E-13	-	-	**0.0076**
c	-	0.3702	0.0009	**0.0054**	0.028	0.1442	-	-	**0.0022**
X^2	-	5624	3164	**12.456**	9.5	13539	-	-	**373**
df	-	13	13	**13**	4	13	-	-	**13**
P	-	0	0	**0.4906**	0.05	0	-	-	**1.E-71**

[3] Values in bold correspond to the best model of distribution fitted.

This study allowed us, however, to reach at least partly our objective of determining the reference distributions for specific POS. For adverbs it was generalized inverse Gauss-Poisson model and for interjections – Zipf-Mandelbrot model. Other classes did not pass multidimensional chi2 test. For function words we obtained Zipf-Mandelbrot model, and for named entities finite Zipf-Mandelbrot model.

All the above observations suggest that statistical analysis of vocabulary should take into account semantic criteria to a much greater extent than is currently the case. What lies behind subsets distinguished on the basis of statistical distributions research (content words, function words, named entities) is their meaning reference, and not just their syntactic function in a text. Another interesting conclusion, demonstrated empirically, is that great corpora are specific and difficult objects of modelling. They are indeed artificial objects, created by engineers or linguists for practical purposes, contrary to the "natural" authorial texts, serving communicative or aesthetic purposes. Nonetheless an attempt at describing large collections of texts is justified in that they apparently seem to simulate the utopian concept of general language.

Does the result achieved signify that traditional divisions of parts of speech are in some sense incorrect or invalid? Of course not. Their function is completely different: they assist in effective first language acquisition, contribute to maintaining stability of "didactic order," and, to a lesser degree, facilitate the teaching of foreign languages.

References

Baayen, R. Harald (2001). *Word Frequency Distributions.* Kluwer Academic Publisher, Dordrecht, Boston, London.

Best, Karl Heinz & Rottmann, Otto (2017). *Quantitative Linguistics. An Invitation*. 3rd ed. Lüdenscheid: RAM-Verlag.

Evert, Stefan & Baroni, Marco (2007). *zipfR*: Word frequency distributions in R. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*. Prague, Czech Republic.

Grotjahn, Rudiger & Altmann Gabriel (1993). Modeling the distribution of word length: some methodological problems. In: Reinhard Koehler, Burghard B. Rieger (ed.), *Contributions to quantitative linguistics.* Kluwer, Dordrecht, 141-153.

Herdan, Gustav (1960). *Type-Token Mathematics.* The Hague: Mouton.

Kamińska, Irena (1990). *Różnice leksykalne między stylami funkcjonalnymi polszczyzny pisanej*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

Maloney, Gilles (1970). La frequence et l'ordre des formes verbales dans l'oeuvre de Thucydide. *RELO* VI, 1970, 87–109. Cited after: http://promethee.philo.ulg.-ac.be/RISSHpdf/annee1970/03/GMaloney.pdf

Mandelbrot, Benoit (1962). On the theory of word frequencies and on related Markovian models of discourse. In: R. Jacobson (ed.), *Structure of Language and its Mathematical Aspects, Proceedings of Symposium in Applied Mathematics.* Vol. XII, AMS, Providence, Rode Island, 1962, 190–219.

Marcińczuk, Michał (2017). Lemmatization of Multi-word Common Noun Phrases and Named Entities in Polish. In: *Proceedings of the 11th International*

Conference on "Recent Advances in Natural Language Processing", RANLP 2017, 2-8 September, 2017, Varna, Bulgaria.

Marcińczuk, Michał; Oleksy, Marcin & Kocoń, Jan (2017). Liner 2 – a Generic Framework for Named Entity Recognition. In: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Valencia, Spain, 4 April 2017. Association for Computational Linguistics, 86–91.

Muller, Charles (1979). Sur quelques scènes de Molière, essai d'un indice du style familier. In: Ch. Muller (ed.), *Langue françise et linguistique quantitative*. Genève: Slatkine, 107–116.

Nicolova-Burova, Anna (1975). On the chronology of Seneca's philosophical dialogues. *RELO* 2, 1975, 1-30. Cited after: http://web.philo.ulg.ac.be-/rissh/1975-2/

Orlov, Jurij K. & Chitashvili, Revas Y. (1983). Generalized Z-distribution generating the well-known "rank-distributions", *Bulletin of the Academy of Sciences, Georgia,* 110.2, 268–272.

Przepiórkowski, Adam (2011). *Tagset* (NKJP), Cited after: http://nkjp.pl/poliqarp/help/ense2.html

Radziszewski, Adam (2013). A Tiered CRF Tagger for Polish. In: Robert Bembenik (ed.), *Intelligent Tools for Building a Scientific Information Platform*. Berlin etc.: Springer, 215–230.

Sichel, Herbert S. (1975). On a distribution law for word frequencies, *Journal of the American Statistical Association*, 70, 542–547.

Sichel, Herbert S. (1982). Asymptotic efficiency of the three methods of estimation for the inverse gaussian-Poisson distribution, *Biometrika*, 69, 467–472.

Strauss, Udo; Fan, Fengxiang & Altmann, Gabriel (2008). *Problems in Quantitative Linguistics* 1. (2nd ed.) Lüdenscheid: RAM-Verlag.

Williams, Carrington B. (1970). *Style and vocabulary: numerical studies*. London: Griffin.

Polyfunctionality Studies in German, Dutch, English and Chinese*

Lu Wang[1, 2], *Yahui Guo*[1]
[1] *School of Foreign languages, Dalian Maritime University, Dalian, China*
[2] *Computational Linguistics and Digital Humanities, University of Trier, Trier, Germany*

Abstract
In many languages, there are grammatically ambiguous words. For example, *list* is a noun in "*a list of candidates*", but a verb in "*I have listed four reasons below*"; and *tear* is a noun in "*a tear rolled down her cheek*" but a verb in "*She began to tear them into small pieces*". Since such words have more than one grammatical function, we call them polyfunctional words. The number of grammatical functions (part of speech) a word has is called polyfunctionality.

This paper focuses on the polyfunctionality distribution, which is left almost uninvestigated so far. We assume that polyfunctionality is lawfully distributed. Therefore, we attempt to find a unified model to capture such distributions in various languages. In this study, the Celex dictionary is adopted to extract data from German, Dutch and English; and the Chinese data is derived from The *Modern Chinese Dictionary* (5th Edition). Given that polyfunctionality is the kind of spectrum aspect of linguistic property, we tested many proper models. Finally, the Waring distribution is found to capture all the data and perform excellent goodness-of-fitting results.

Further, the polyfunctional words form part-of-speech patterns. As shown in the above example, the part-of-speech pattern of both *list* and *tear* is noun and verb, which polyfunctionality is 2. In this way, part-of-speech patterns also form a polyfunctionality distribution. Again, we expect to find a model to capture such distributions. The fitting results show that the data of all the 4 languages abide by the binomial distribution.

Keywords: *polyfunctionality, parts of speech, Chinese*

1. Introduction

Each word belonged to one and only one part of speech in the moment of its creation. Then, some words are expanded to new parts of speech by way of adopting affixes, e.g. Zahl (noun), zahlen (verb), and zahlreich (adjective) in German, or without any change ("conversion"), e.g. 左右 (noun, verb and adverb) in Chinese and run (noun and verb) in English. The latter words, which have more than one part of speech, are called poly-functional words. The quantitative concept "polyfunctionality" means the number of parts of speech a polyfunctional word possesses. "Part-of-speech pattern" refers to the

* Address correspondence to: Lu Wang, Computerlinguistik und Digital Humanities, University of Trier, 54286 Trier, Germany. Email: wanglu-chn@hotmail.com

specific parts of speech of a polyfunctional word.

This paper focuses on polyfunctionality and investigates polyfunctionality distributions on the levels of words and part-of-speech patterns on data from the German, Dutch, English and Chinese languages.

2. Data

In this study, the Celex dictionary is adopted to extract data of from German, Dutch and English; and the Chinese data is derived from *The Modern Chinese Dictionary* (5th Edition). Some words in German, Dutch and English have the first character capitalized. For example, the German word "schreiben" is a verb; while "Schreiben" is a noun. However, the capitalized form "*S''* and the lower form "*s''* are merely allo-forms of the same character. Therefore, despite the two words are listed as different entries in a dictionary, we do not discriminate them in this study. The word "schreiben" is considered as a polyfunctional word, whose polyfunctionality is equal to 2 and its part-of-speech pattern is "noun and verb".

3. Results

3.1 Polyfunctionality distribution

Parts of speech as any nominal entities form two kinds of distribution: the rank distribution (where the variable is the rank associated with the corresponding number of words) and the spectrum distribution i.e. polyfunctionality distribution (obtained by counting the polyfunctional words with the given number of parts of speech). Most of the previous works on parts of speech studied the rank-frequency distribution, while polyfunctionality is not thoroughly investigated yet.

The first report on a polyfunctionality distribution is given in Fan and Altmann (2008) on the basis of 165 randomly selected English words, which are shown to fit with the Shenton-Skee-geometric distribution. In a later research on data from *The Modern Chinese Dictionary (5th edition)* and consisting of over 50000 words, the Waring distribution, which is used to capture spectrum linguistic properties such as polysemy and frequency, is selected and performs a perfect result (Wang, 2016). To further explore the regularity of polyfunctionality, we test the Waring distribution together with other models that have been shown to capture spectrum distributions on dictionary data of from the German, Dutch, English and Chinese languages. Considering all the four languages, the Waring distribution still performs best, as shown in Table 1 and Figure 1, 2, 3 and 4.

3.2 Polyfunctionality and Part-of-speech patterns

Polyfunctional words form part-of-speech patterns. For example, the Dutch word "het" is an article (determiner) in the first sentence but a pronoun in the second sentence. Therefore, its part-of-speech pattern is noun and verb (N, V). Abbreviations of parts of

speech are listed in Appendix.

Het eerste liedje dat ik schreef, was in het Engels, het tweede in het Chinees. Het was de honing in het water die het zoet maakte, het was de olie die het ziek maakte.

Table 1. Fitting the Waring distribution to the polyfunctionality data.

Waring											
German			**Dutch**			**English**			**Chinese**		
x[i]	**f[i]**	**NP[i]**	**x[i]**	**f[i]**	**NP[i]**	**x[i]**	**f[i]**	**NP[i]**	**x[i]**	**f[i]**	**NP[i]**
1	50003	50003	1	119078	119078	1	40469	40573.55	1	47414	47414
2	627	625.57	2	1599	1600.08	2	4804	4569.5	2	3299	3285.55
3	25	27.23	3	103	101	3	367	514.63	3	364	380.71
4	3	2.21	4	10	10.91	4	68	57.96	4	60	60.16
			5	2	2.02	5	14	6.53	5	17	11.81
						6	1	0.83	6	2	3.77
b = 29.4583 n = 0.3859			b = 17.604 n = 0.2534			b = 4405873.113 n = 559177.6934			b = 16.6657 n = 1.3153		
X^2 = 0.4718 P(X^2) = 0.4922 DF = 1 C = 0.0000 R^2 = 1.0000			X^2 = 0.1167 P(X^2) = 0.9433 DF = 2 C = 0.0000 R^2 = 1.0000			X^2 = 64.3363 P(X^2) = 0 DF = 2 C = 0.0014 R^2 = 0.9999			X^2 = 3.8983 P(X^2) = 0.2727 DF = 3 C = 0.0001 R^2 = 1.0000		

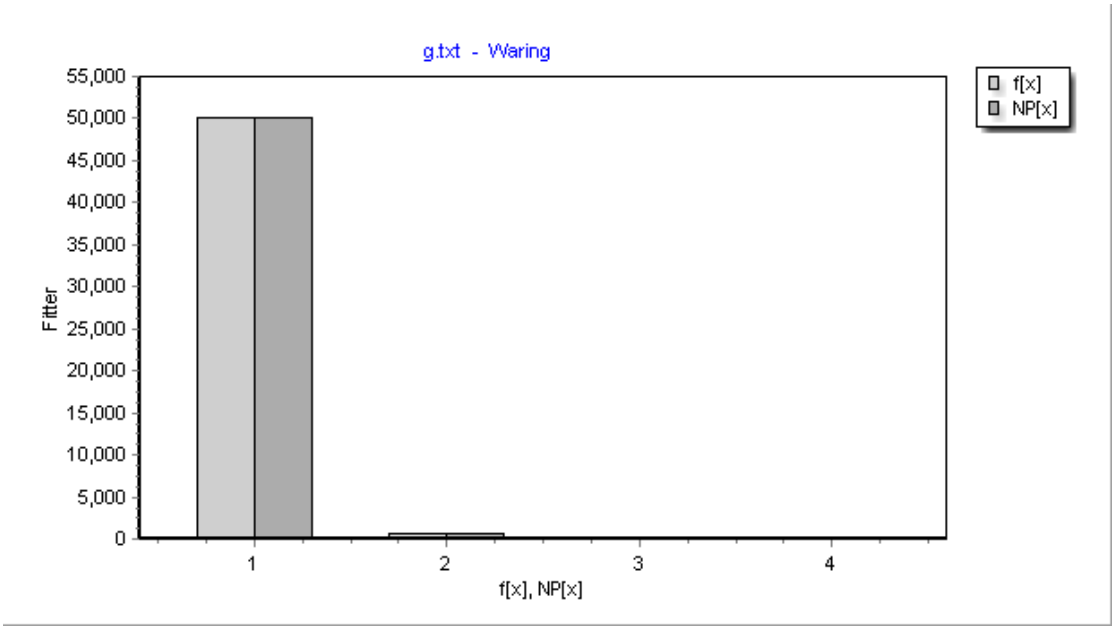

Figure 1. Fitting the Waring distribution to the polyfunctionality data of the German language.

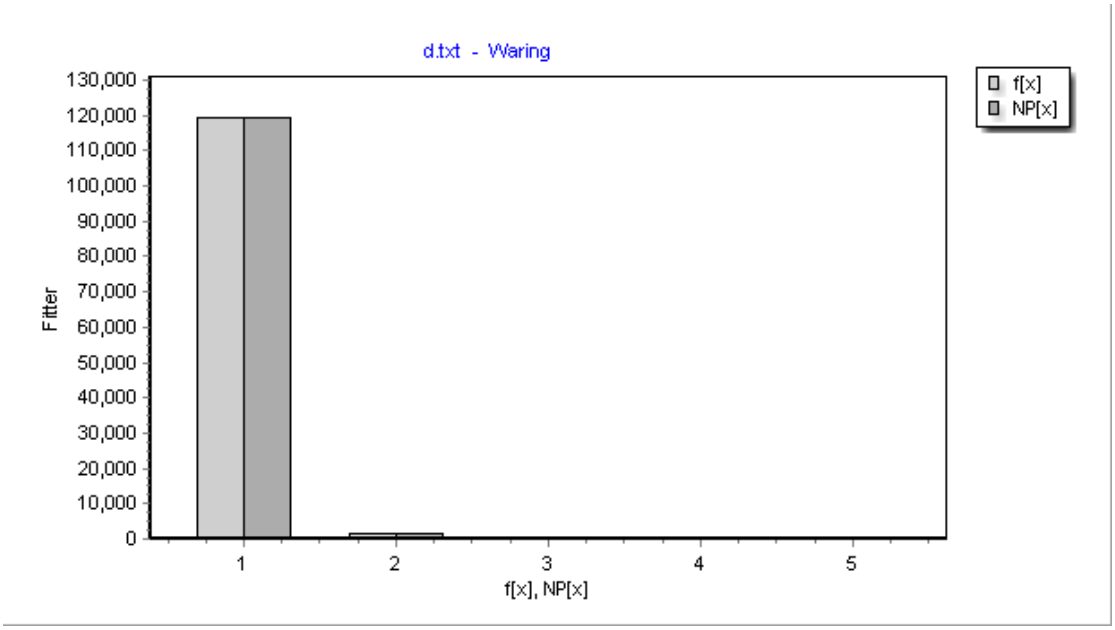

Figure 2. Fitting the Waring distribution to the polyfunctionality data of the Dutch language.

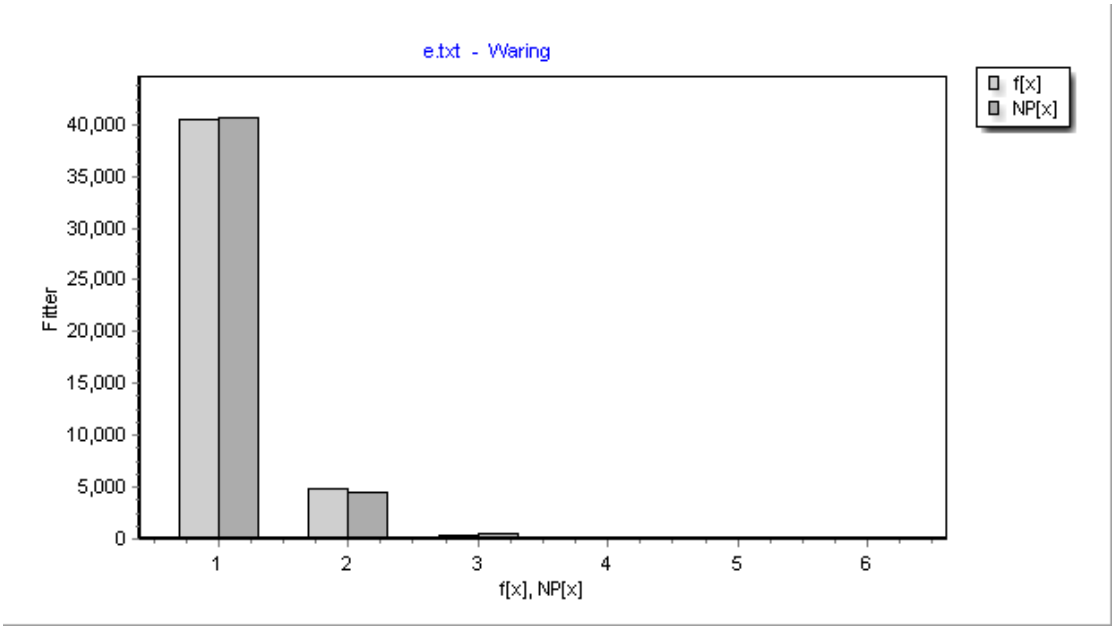

Figure 3. Fitting the Waring distribution to the polyfunctionality data of the English language.

The German word “sein” is a noun, a verb and a pronoun in the following sentences successively. The part-of-speech pattern is noun, verb and pronoun (N, V, PRON).

Er denkt über das menschliche Sein nach.

Es kann nicht wahr sein.

Sein Vater ist Richter.

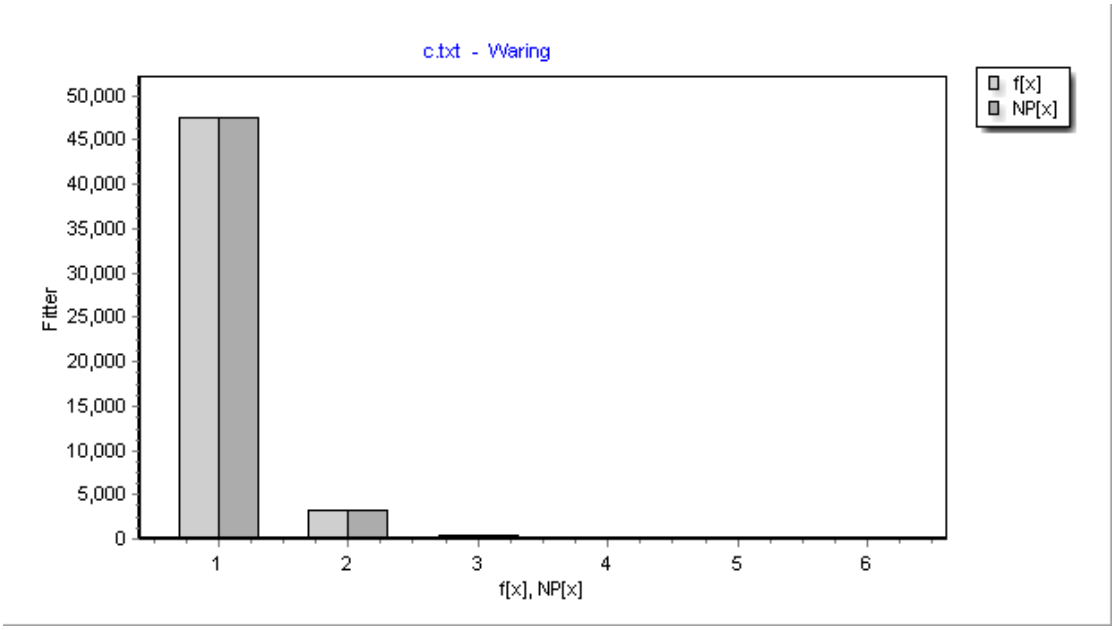

Figure 4. Fitting the Waring distribution to the polyfunctionality data of the Chinese language.

The Chinese word "就" is a verb, an adverb, a preposition and a conjunction respectively in the following sentences. Its pattern is verb, adverb, preposition and conjunction (V, ADV, PREP, C).

花生仁儿就酒。

他要是不来，我就去找他。

就工作经验来说，他比别人要丰富些。

你就送来，我也不要。

The English word "like" in the following sentences is a noun, a verb, an adjective, an adverb, a preposition and a conjunction. Its part-of-speech is noun, verb, adjective, adverb, preposition and conjunction (N, V, A, ADV, PREP, C).

I've never seen the like of it!
Do you like fish?
They are not twins, but they are very like.
It'll rain this afternoon, as like as not.
I, like everyone else, had read these stories in the press.
She acts like she owns the place.

Table 2, 3, 4 and 5 list the specific part-of-speech patterns (POS patterns) with the corresponding polyfunctionality (PF). As can be seen from the tables, the patterns have polyfunctionality as a property, in analogy to words. In this way, part-of-speech patterns

also form a polyfunctionality distribution.

Table 2. Part-of-speech patterns in German language

PF	**POS patterns**				**Total patterns**
2	N, A N, V N, ADV ADV, PREP ADV, C	A, ADV N, NUM A, V N, I N, PRON	A, PREP N, PREP N, C A, NUM PRON, ADV	PREP, C PRON, C NUM, V ART, PRON	19
3	N, ADV, PREP A, ADV, PREP N, A, ADV ADV, PREP, C	NUM, ART, ADV N, V, PRON N, PRON, ADV N, PREP, C	N, NUM, ADV N, ADV, I N, A, V N, A, I	N, A, PRON N, A, PREP N, A, NUM A, ADV, C	16
4	N, ADV, PREP, C ADV, PREP, C, I				2

Table 3. Part-of-speech patterns in Dutch language

PF	**POS patterns**				**Total patterns**
2	N, A N, V N, ADV A, V EXP, N N, I N, NUM ADV, PREP	N, PRON EXP, A A, I A, PREP A, NUM ADV, C A, ADV PRON, ADV	ADV, I PREP, C N, PREP N, C EXP, I EXP, ADV A, C NUM, ADV	EXP, V V, ADV NUM, PRON EXP, ART ART, PRON	29
3	EXP, N, A N, A, NUM N, A, I N, A, V N, A, ADV N, NUM, V N, PRON, C	N, A, PREP N, ADV, I N, ADV, C N, A, C EXP, N, I ADV, C, I A, ADV, I	EXP, N, ADV A, PRON, ADV N, V, PRON A, PRON, I A, PREP, C N, ART, ADV N, ADV, PREP	N, NUM, PRON ADV, PREP, C A, NUM, PRON A, ADV, PREP NUM, PRON, ADV NUM, ART, PRON N, PRON, ADV	28
4	N, ADV, PREP, I N, ADV, PREP, C N, ADV, C, I N, A, NUM, V N, A, ADV, I		EXP, N, ADV, PREP N, A, ADV, C EXP, NUM, PRON, ADV EXP, N, ADV, C A, NUM, PRON, ADV		10
5	N, NUM, PRON, ADV, C		N, A, NUM, ART, PRON		2

Table 4. Part-of-speech patterns in English language

PF	POS patterns				Total patterns
2	N, V N, ADJ ADJ, ADV V, ADJ N, NUM N, ADV N, INT	ADV, PRON ADV, PREP N, PRON ADJ, PREP N, PREP ADV, C V, INT	V, ADV PREP, C ADV, INT ADJ, PRON V, PRON N, C ADJ, INT	V, PREP V, C PRON, PREP PRON, ART N, ART ADV, ART	27
3	N, V, ADJ N, V, ADV N, V, INT N, V, PREP N, V, NUM V, PREP, C N, ADV, C	N, ADJ, ADV N, ADJ, INT N, ADV, INT N, V, PRON N, PREP, C V, ADJ, ADV N, ADV, NUM	N, ADV, PREP ADJ, ADV, C N, ADJ, PREP N, ADJ, PRON ADV, PREP, C N, ADV, PRON ADV, PRON, C	V, ADJ, PRON N, NUM, PRON ADJ, ADV, PREP ADJ, ADV, PRON	25
4	N, V, ADJ, ADV N, ADJ, ADV, PREP N, V, ADJ, INT N, ADV, PRON, C N, ADJ, ADV, INT V, ADJ, ADV, PREP N, V, PREP, C		N, V, ADV, NUM N, V, ADV, INT ADJ, ADV, PREP, C N, ADV, PRON, INT N, ADJ, ADV, PRON N, ADJ, PRON, PREP ADV, PRON, C, INT		14
5	N, V, ADJ, ADV, INT N, V, ADJ, ADV, PREP N, V, ADJ, ADV, PRON		N, ADV, PRON, PREP, C N, ADJ, ADV, PRON, PREP		5
6	N, V, ADJ, ADV, PREP, C				1

We expect to find a model to capture such distributions of the four languages. The Chinese data is reported to fit the Positive binomial distribution best (Wang, 2016). In this study, the four languages belonging to different linguistic types all abide by binomial distribution. The binomial distribution consists of fewer parameters than the positive binomial distribution and performs perfect results, as shown in Table 6 and Figure 5, 6, 7 and 8. Therefore, we consider the binomial distribution as a better unified model.

4. Discussion and conclusion

The present paper studied polyfunctionality based on dictionary data of the German, Dutch, English and Chinese languages. We get the following conclusions:

(1) Like other properties of words, polyfunctionality is distributed lawfully. The distributions in the four languages abide by the Waring distribution.

(2) The polyfunctionality distributions formed by part-of-speech patterns are perfectly captured by binomial distribution.

Table 5. Part-of-speech patterns in Chinese language.

PF	**POS patterns**				**Total patterns**
2	N, V N, A V, A V, C N, C V, I N, I A, C N, QUA A, I	V, ADV A, ADV V, QUA N, NUM ADV, C V, PREP N, ADV V, AUX N, AUX C, PREP	N, ONOM ONOM, I N, PREP ADV, PRON V, ONOM N, PRON V, PRON AUX, I ADV, NUM C, PRON	ADV, AUX V, NUM PRON, AUX PRON, NUM QUA, AUX ONOM, AUX ADV, QUA ADV, PREP ADV, ONOM A, ONOM	40
3	N, V, A V, A, C N, V, C V, A, I N, A, QUA N, ADV, C N, V, NUM N, V, QUA N, V, ADV N, A, ADV	V, A, ADV N, V, PREP N, C, PRON N, V, ONOM N, C, PREP V, A, AUX N, V, AUX V, A, PREP N, A, NUM N, ADV, QUA	N, ADV, PRON N, PRON, AUX V, ADV, PREP N, ADV, NUM ADV, C, PRON N, QUA, AUX N, ADV, PREP V, QUA, PREP V, A, ONOM N, A, PRON	N, QUA, NUM V, PREP, AUX N, PREP, AUX N, AUX, NUM V, ADV, QUA N, ADV, AUX ADV, C, AUX ONOM, AUX, I V, ONOM, AUX	39
4	N, V, QUA, PREP N, V, ADV, PREP N, V, ADV, AUX N, V, QUA, ONOM N, V, ONOM, PREP N, V, PRON, NUM N, V, PREP, AUX V, ADV, PREP, AUX N, ADV, C, PRON		N, V, A, ADV N, V, A, QUA N, V, ADV, C N, QUA, C, AUX N, V, ADV, QUA V, ADV, C, PREP N, V, ONOM, I N, V, QUA, C N, V, A, PREP	N, V, A, PRON N, A, QUA, PREP N, V, A, C N, A, ADV, QUA N, A, ADV, PRON N, A, ADV, C N, C, PRON, AUX N, ADV, C, AUX N, V, C, PREP	27
5	N, V, A, ADV, QUA N, V, ADV, C, PREP N, V, A, ADV, PREP N, V, QUA, C, PREP N, V, QUA, PREP, AUX N, V, QUA, C, AUX N, V, A, QUA, PRON		N, V, ADV, QUA, PREP N, V, ADV, PREP, AUX N, V, A, QUA, PREP N, V, A, ADV, PRON N, V, A, ADV, NUM N, ADV, QUA, PRON, PREP		13
6	N, V, A, QUA, C, PREP		N, V, A, ADV, C, PREP		2

Table 6. Fitting the Binomial distribution to the polyfunctionality of pattern data.

Binomial											
German			**Dutch**			**English**			**Chinese**		
x[i]	**f[i]**	**NP[i]**	**x[i]**	**f[i]**	**NP[i]**	**x[i]**	**f[i]**	**NP[i]**	**x[i]**	**f[i]**	**NP[i]**
1	9	9.09	1	11	11.47	1	10	10.86	1	12	13.18
2	19	19.55	2	29	28.68	2	27	25.45	2	40	36.10
3	16	14.01	3	28	26.89	3	25	25.57	3	39	42.38
4	2	3.35	4	10	11.21	4	14	14.27	4	27	27.64
			5	2	1.75	5	5	4.78	5	13	10.81
						6	1	1.07	6	2	2.89
n = 3.0000 p = 0.4174			n = 4.0000 p = 0.3847			n = 7.0000 p = 0.2509			n = 7.0000 p = 0.2812		
X^2 = 0.8410 $P(X^2)$ = 0.3591 DF = 1 C = 0.0183 R^2 = 0.9737			X^2 = 0.2338 $P(X^2)$ = 0.8897 DF = 2 C = 0.0029 R^2 = 0.9953			X^2 = 0.1949 $P(X^2)$ = 0.9784 DF = 3 C = 0.0024 R^2 = 0.9940			X^2 = 1.5259 $P(X^2)$ = 0.6763 DF = 3 C = 0.0115 R^2 = 0.9746		

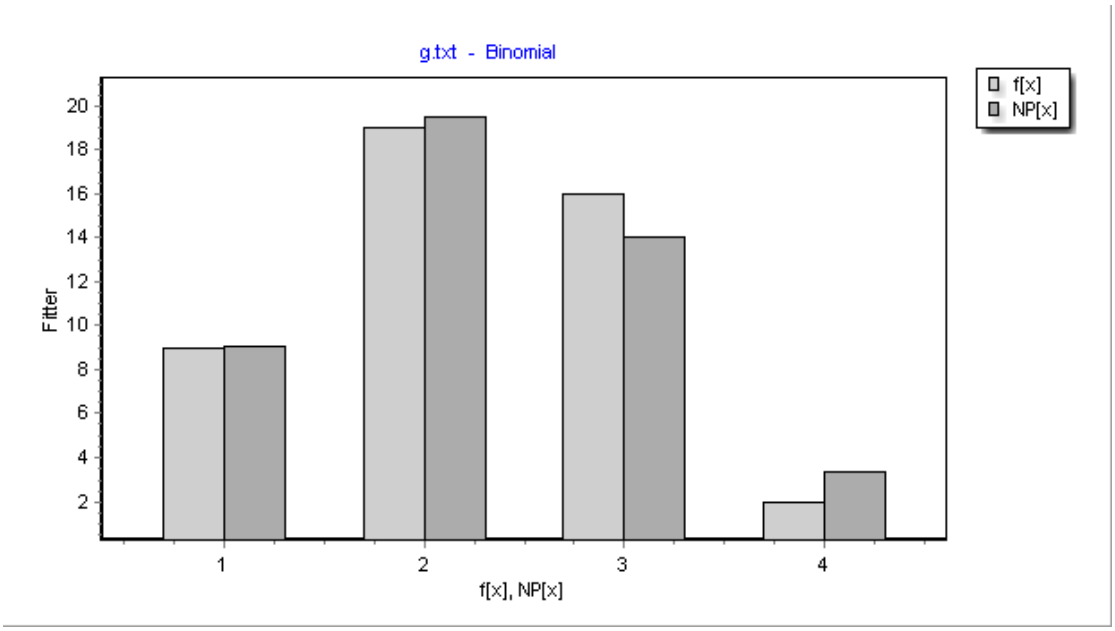

Figure 5. Fitting the Binomial distribution to the polyfunctionality distribution of patterns of the German language.

We choose the above models because they yield excellent results on the data from all the four languages and they involve less parameters than others. However, whether they are also the best choice for other languages is still a question, since four languages is a small sample compared with the 7000 odd known languages in the world. Therefore, to find better models (or more general models), data from more languages need to be tested.

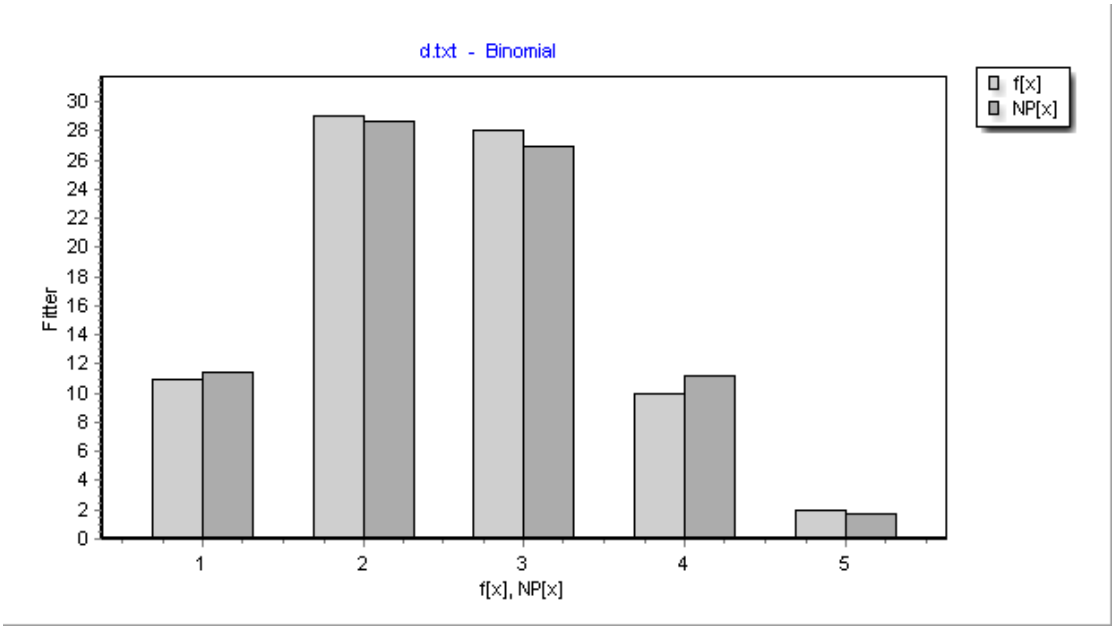

Figure 6. Fitting the Binomial distribution to the polyfunctionality distribution of patterns of the Dutch language.

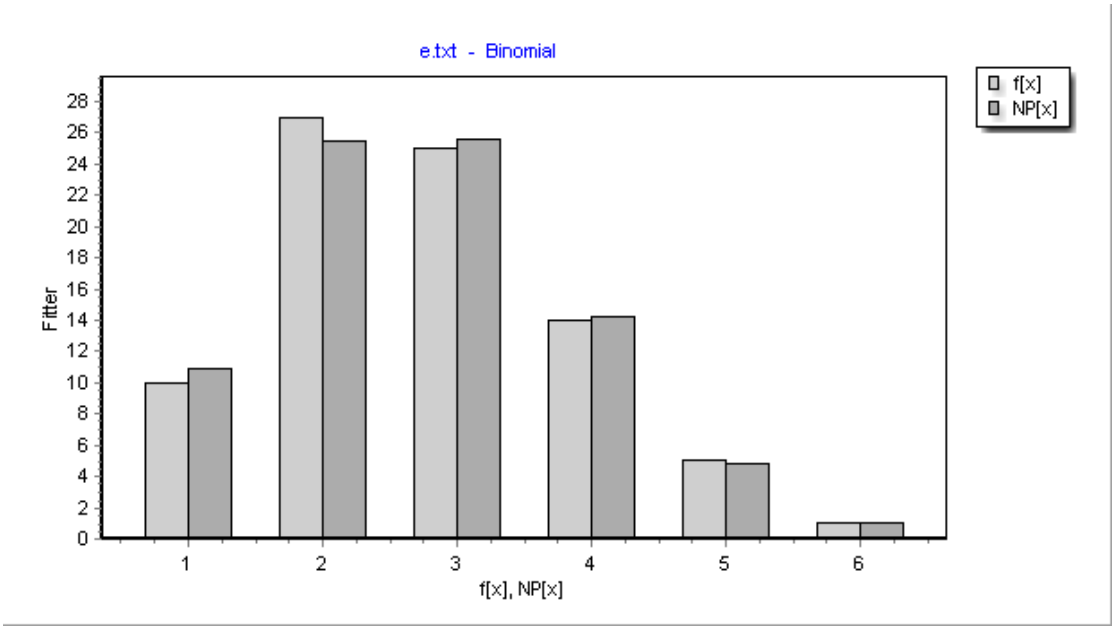

Figure 7. Fitting the Binomial distribution to the polyfunctionality distribution of patterns of the English language.

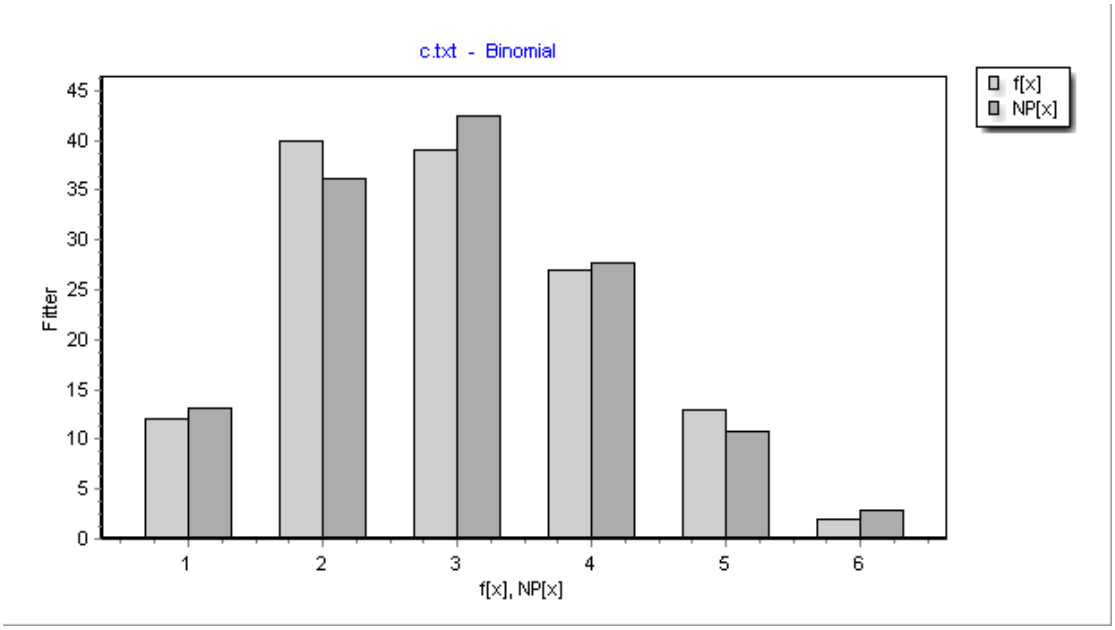

Figure 8. Fitting the Binomial distribution to the polyfunctionality distribution of patterns of the Chinese language.

References

Fan, Fengxiang & Altmann, Gabriel (2008). On meaning diversification in English. *Glottometrics 17*, 69-81.

Köhler, Reinhard (2005). Gegenstand und Arbeitsweise der Quantitativen Linguistik [Aims and methods of Quantitative Linguistics]. In: Köhler, Reinhard; Altmann, Gabriel & Piotrowski, Rajmund (Eds.), *Quantitative Linguistics. An International Handbook*, (pp.1-16), Berlin: de Gruyter.

Wang, Lu (2016). Parts of speech studies in Chinese. In: *Journal of Quantitative Linguistics 23 (3)*, 235-255.

Appendix

Part-of-speech abbreviations of English, German and Dutch are the same as used in Celex dictionary; that of Chinese are translations from *Modern Chinese Dictionary* (5 Edition).

Abbreviations	**English**	**German**	**Dutch**	**Chinese**
N	Noun	Noun	Noun	Noun
A	Adjective	Adjective	Adjective	Adjective
NUM	Quantifier/ Numeral	Quantifier/ Numeral	Quantifier/ Numeral	Numeral
V	Verb	Verb	Verb	Verb
ART	Article	Article	Article	
PRON	Pronoun	Pronoun	Pronoun	Pronoun
ADV	Adverb	Adverb	Adverb	Adverb
PREP	Preposition	Preposition	Preposition	Preposition

C	Conjunction	Conjunction	Conjunction	Conjunction
I	Interjection	Interjection	Interjection	Interjection
EXP			Expression	
QUA				Quantifier
ONOM				Onomatopoia
AUX				Auxiliary

Measuring the Degree of Violation of the One-Meaning–One-Form Principle*

Relja Vulanović, Oliver Ruff
Department of Mathematical Sciences, Kent State University at Stark
6000 Frank Ave NW, North Canton, Ohio, USA

Abstract
The situation when there is no one-to-one correspondence between linguistic forms and their meaning is described mathematically and a formula is proposed to measure the deviation from the one-meaning–one-form principle. Application examples are provided. They include subject and object marking and part-of-speech systems.

Keywords: *One-Meaning–One-Form Principle, mathematical relation, bijection, subject and object marking, parts-of-speech system.*

1. Introduction

According to Miestamo (2008), the linguistic principle 'One-Meaning–One-Form' (from now on, the 'Principle') has been known under this name since Anttila's (1972) book. Speaking of the Principle in his book (p. 181), Anttila states: "A maximally efficient system avoids polysemy (forms with many [related] meanings, especially if these occur in the same semantic sphere) and homophony, two (unrelated) meanings getting the same form. […] Avoidance of homophony and polysemy provide clear evidence of this mental force 'one meaning, one form'." The book gives many examples of the manifestations of the Principle. Nevertheless, examples of the Principle being violated also abound. It is therefore of interest to be able to measure how much a linguistic system departs from the Principle. This kind of measure can become a component of the measure of grammar complexity or grammar efficiency. The Principle is proposed in (Miestamo, 2008) as one of the criteria for measuring absolute (theory-oriented, objective, as opposed to user-oriented) language complexity (see also the whole volume Miestamo et al., 2008). Formal descriptions of grammar complexity and efficiency can be found, e.g., in (Vulanović, 2003, 2007).

We approach the question of how to measure the degree of violation of the Principle in a formal, mathematical way. In Section 2 we consider general relations between two sets in contrast to the situations when a bijection can be established between them. This results in a relatively simple formula, called the *basic formula*, for measuring the departure from the Principle. The basic formula is generalized by introducing weights for the elements of the two sets considered. Subject-object marking types are used in Section 3 as an elementary linguistic example to illustrate an application of the basic formula. More complicated linguistic structures, viz. parts-of-speech systems in the sense of (Hengeveld, 1992), are discussed in Section 4, where

* Address correspondence to: Relja Vulanović, Department of Mathematical Sciences, Kent State University at Stark, 6000 Frank Ave NW, North Canton, Ohio, USA. E-mail: rvulanov@kent.edu.

we show that the weighted formula with appropriate weights can be used to achieve a better measure of the violation of the Principle. Several different weight systems are considered and compared. Section 5 offers some concluding remarks.

2. Mathematical description

Let X and Y be two finite non-empty sets and let ρ be a relation between the elements in X and those in Y, $\rho \subseteq X \times Y$, such that every element of X and every element of Y appear in at least one ordered pair in ρ. The set X can be interpreted as the set of meanings, and Y as the set of forms, or the other way around.

If for each $x \in X$ there exists exactly one $y \in Y$ so that $(x, y) \in \rho$, then ρ is a function with domain X and range Y (a function onto Y). If additionally, $|X| = |Y|$, where $|A|$ denotes the number of elements in a finite set A, then ρ is a bijection (a one-to-one correspondence) between X and Y. We want to find a reasonable measure, denoted by $\mu(\rho)$, of how far a general relation ρ is from a bijection. The measure should be flexible enough to allow for different applications in linguistics. Moreover, it should be invariant under the transformation which interchanges the sets X and Y, that is, ρ should satisfy the following property:

Property 1. $\mu(\rho) = \mu(\rho^{-1})$, where $\rho^{-1} = \{(y, x) : (x, y) \in \rho\}$.

For each $y \in Y$, let us define $\nu_X(y)$ as the number of elements in X that are paired up with y,

$$\nu_X(y) = |\{x \in X : (x, y) \in \rho\}|.$$

Similarly, let $\nu_Y(x)$ be the number of elements in Y that are paired up with $x \in X$,

$$\nu_Y(x) = |\{y \in Y : (x, y) \in \rho\}|.$$

We can now define a set of one-to-one pairs in ρ,

$$B = \{(x, y) \in \rho : \nu_X(y) = \nu_Y(x) = 1\}.$$

We have that $|B| \leq |\rho|$. The formula

$$\mu(\rho) = \mu_*(\rho) := |\rho| - |B| \tag{1}$$

is then a simple way to measure how different from a bijection the relation ρ is. In the application we are interested in, this indicates the degree of violation of the Principle. Note that μ_* satisfies Property 1. Moreover, if ρ is a bijection, then there is a one-to-one correspondence between X and Y and $\rho = B$, giving $\mu_*(\rho) = 0$. When ρ is not a

bijection, then $\mu_*(\rho) > 0$. This kind of a measure is suitable when the degree of violation of the Principle is a term (an additive component, as opposed to a factor – a multiplicative component) in the formula for measuring grammar complexity. However, if the formula consists of terms, each term has to be assigned a weight which indicates the relative importance of the component in question. Such weights may be hard to determine, and it is more practical to have a formula which consists of factors, the degree of violation of the Principle being one of them. Then, just one weight, a scaling coefficient, suffices (Vulanović, 2007). Moreover, μ_* is not suitable for comparing two different relations via a relative measure. This is because 0 is one of the possible values of this measure. Therefore, it would be better to have $\mu(\rho)$ satisfying

Property 2. (a) $\mu(\rho) = 1$ when ρ is a bijection,
(b) $\mu(\rho) > 1$ when ρ is not a bijection.

To make $\mu(\rho)$ compatible with the measure of relative grammar complexity (ibid.), we should also require the following property, which (1) does not satisfy:

Property 3. $\mu(\rho)$ should be a measure relative to the size of $|X|$ and $|Y|$. It should be inversely proportional to $|X|$ and $|Y|$.

At the same time, we should keep $\mu(\rho)$ simple and preserve the linearity present in (1):

Property 4. $\mu(\rho)$ should be a linear function of both $|\rho|$ and $|B|$. It should increase when $|\rho|$ increases or when $|B|$ decreases.

A formula that meets Properties 1-4 is

$$\mu(\rho) = \mu_\theta(\rho) := \frac{(1+\theta)|\rho| - \theta|B|}{\min\{|X|, |Y|\}}, \tag{2}$$

where θ is a parameter, $\theta > 0$. All properties are easy to verify, except perhaps Property 2(b). This property holds true because if ρ is not a bijection, then we have that $|\rho|$ is strictly greater than $|B|$ and at least one of $|X|$ or $|Y|$. Therefore,

$$\mu_\theta(\rho) > \frac{(1+\theta)|\rho| - \theta|\rho|}{\min\{|X|, |Y|\}} = \frac{|\rho|}{\min\{|X|, |Y|\}} > 1.$$

We refer to the formula (2) as the *basic formula* and we work with it in Section 3. However, in Section 4, we show that a generalization of this may be better suited for some applications.

Consider an arbitrary set A with n elements. If the elements are regarded to be of different importance, we may want to assign different weights to them. The weights are chosen to be positive numbers $w_1, w_2, \ldots, w_n$, which can be normalized so that $\min_i w_i = 1$. Then the quantity

$$\|A\| := w_1 + w_2 + \Lambda + w_n$$

can be used instead of the simple $|A| = n$. In this sense, (2) can be generalized to the weighted formula (3),

$$\hat{\mu}_\theta(\rho) = c \cdot \frac{(1+\theta)\|\rho\| - \theta\|B\|}{\min\{\|X\|, \|Y\|\}}, \quad \theta > 0, \tag{3}$$

where c is a positive scaling coefficient which ensures that $\hat{\mu}_\theta(\rho) = 1$ when ρ is a bijection so that Property 2(a) still holds true. Note that each occurrence of $\|\cdot\|$ in (3) depends on the set that $\|\cdot\|$ is applied to. Different weights may be used for the elements of X, as opposed to the elements of Y, so $\|X\|$ and $\|Y\|$ may be calculated differently, and $\|\rho\|$ may use yet another selection of weights. However, the weights in $\|B\|$ must use the corresponding weights from $\|\rho\|$ since $B \subseteq \rho$. Because the weights in $\|\rho\|$, $\|X\|$, and $\|Y\|$ may be different, Property 2(a) cannot be guaranteed in general without a scaling coefficient c. However, if $\|\rho\|$, $\|X\|$, and $\|Y\|$ are defined so that

$$\|\rho\| = \|X\| \text{ or } \|\rho\| = \|Y\| \tag{4a}$$

and

$$\|X\| = \|Y\| \text{ when } \rho \text{ is a bijection,} \tag{4b}$$

then $c = 1$.

As for Property 2(b), consider the case when ρ is not a bijection, so that $\|B\| < \|\rho\|$. Then we get from (3) that

$$\hat{\mu}_\theta(\rho) > c \cdot \frac{\|\rho\|}{\min\{\|X\|, \|Y\|\}}.$$

If (4) is satisfied, then it immediately follows that $\hat{\mu}_\theta(\rho) > 1$. Otherwise, the quotient $\frac{\|\rho\|}{\min\{\|X\|, \|Y\|\}}$ is expected to have the smallest value when ρ is a bijection and, since c is the reciprocal of this value, we get $\hat{\mu}_\theta(\rho) > 1$.

The formula (3) may lose Property 1, but if different weights are used for *X* and *Y*, the two sets are not treated equally (which may be for a good reason) and then the preservation of Property 1 cannot be justified.

3. The basic formula: subject and object marking

Most of the examples in Anttila (1972) that illustrate the violation of the Principle are specific cases found in the vocabulary (pp. 181-184). Since we measure the departure from the Principle keeping in mind that this can become a component of the measure of grammar complexity/efficiency, our interest here is in examples of morpho-syntactic nature. In this section, we consider different situations how languages mark the subject and object in transitive sentences.

Six different subject-object marking types are considered in Table 1, which also shows the values of the components needed to calculate μ_θ using formula (2), the value of μ_θ itself, as well as the special case $\theta = 1$, when (2) reduces to

$$\mu_1(\rho) = \frac{2|\rho| - |B|}{\min\{|X|, |Y|\}}. \tag{5}$$

Table 1. Subject-object marking types and the corresponding measures μ_θ and μ_1.

Type	S	O	$\lvert\rho\rvert$	$\lvert B\rvert$	$\min\{\lvert X\rvert, \lvert Y\rvert\}$	μ_θ	μ_1
1	N	N	2	0	1	$2(1+\theta)$	4
2	Nom	Acc	2	2	2	1	1
3	N, Nom	N, Acc	4	0	2	$2(1+\theta)$	4
4	An	An, In	3	0	2	$1.5(1+\theta)$	3
5	Nom	Acc, In	3	1	2	$1.5+\theta$	2.5
6	An, Nom	An, Acc, In	5	0	2	$2.5(1+\theta)$	5

In Table 1, S stands for the Subject and O for the Object. The set of meanings is *X* = {S, O} and *Y* is the set of different nominal classes or forms used to convey the meaning of S or O. Therefore, $\min\{|X|, |Y|\} \le 2$. This minimum can equal 1 only if $|Y| = 1$, which is the case in Type 1 marking, where only one nominal class is used. The class is simply Noun, denoted by N. Then, ρ = {(S, N), (O, N)} and there is no one-to-one pair. This explains why $|\rho| = 2$ and $|B| = 0$. Type 1 is exemplified by languages like English.

Type 2 is like in Latin, where Nom(inative) is used to convey S and Acc(usative) to convey O. The relation ρ is now a bijection and $\mu_\theta(\rho) = 1$. However, Type 3 models the situation in Latin better because there is a class of nouns which does not distinguish between Nom and Acc (this is also the situation in many Slavic

languages). In this case, $\rho = \{(S, N), (S, Nom), (O, N), (O, Acc)\}$ and $|\rho| = 4$, $|B| = 0$. It is interesting to note that the measure μ_θ is equal for marking types 1 and 3.

Types 4 and 5 model the marking in languages with the category of animacy. An(imate) nouns can have both S and O roles, but the In(animate) ones can only be objects. For instance, this is the case in the Algonquian language Cree (Wolfart and Carroll, 1981) and the Uto-Aztecan language Luiseño (Steele, 1978). In Luiseño, however, animate nouns have both Nom and Acc forms, which is modeled by Type 5.

Finally, Type 6 is hypothetical. It models the case when there are two classes of animate nouns, those that have Nom and Acc, and those that do not. This is where the departure from the Principle is greatest.

4. The weighted formula: parts-of-speech systems

In this section, we apply the weighted formula (3) to Hengeveld's (1992) parts-of-speech (PoS) systems (see also Hengeveld, Rijkhoff, and Siewierska, 2004; Hengeveld and van Lier, 2010).

Hengeveld's approach to PoS systems is based on the four propositional functions (syntactic slots) shown in Table 2. Some languages have less than four propositional functions. Table 3, in which *l* stands for the number of propositional functions, indicates what combinations are possible. The table also shows what propositional functions have to be conveyed by simple intransitive sentences in dependence on the existing propositional functions. In general, except for $l = 1$, the heads P and R should be identifiable in any sentence, whereas the modifying functions p and r are optional.

Table 2. The four propositional functions

	Head	Modifier
Predicate phrase	P	p
Referential phrase	R	r

Table 3. Possible combinations of propositional functions

l	Propositional functions in a PoS system	Propositional functions conveyed by sentences
4	{P, R, r, p}	{P, R}, {P, R, r}, {P, R, p}, {P, R, r, p}
3	{P, R, r}	{P, R}, {P, R, r}
3	{P, R, p}	{P, R}, {P, R, p}
2	{P, R}	{P, R}
1	{P}	{P}

The propositional functions existing in a PoS system constitute the set *X* of the meanings that need to be conveyed. Based on the number of occurrences of the propositional functions in the last two columns of Table 3, it can be concluded that

different importance should be given to different propositional functions. This means that they should carry different weights in the measure $\|X\|$. Let α, β, γ, and δ be the weights for P, R, r, and p, respectively. We use two weight systems, W1 and W2,

W1: $\alpha = 2.5,\ \beta = 2,\ \gamma = \delta = 1,$

W2: $\alpha = 3\frac{1}{3},\ \beta = 3,\ \gamma = \delta = 1.$

The system W1 is based on the counts of how many times each propositional function appears in the second column of Table 3. Analogously, W2 uses the counts in the third column. In both cases, the weights are scaled down so that the smallest ones equal 1. In conclusion, the general formulas for $\|X\|$ are

$\|X\| = \alpha + \beta + l - 2$ if l = 2, 3, 4 and $\|X\| = \alpha$ if l = 1.

Table 4. Word classes for l = 4 and their weights

Word class	P	R	r	p	Simple weight	Accrued weight
Verbs	V	-	-	-	α	α
Nouns	-	N	-	-	β	β
Adjectives	-	-	a	-	1	1
Manner adverbs[1]	-	-	-	m	1	1
Heads	H	H	-	-	$\alpha+\beta$	$2\alpha+\beta$
Predicatives	~~P~~	-	-	~~P~~	$\alpha+1$	$2\alpha+1$
Nominals	-	~~N~~	~~N~~	-	$\beta+1$	$2\beta+1$
Modifiers	-	-	M	M	2	3
*	X_1	-	X_1	-	$\alpha+1$	$2\alpha+1$
*	-	X_2	-	X_2	$\beta+1$	$2\beta+1$
Non-verbs	-	Λ	Λ	Λ	$\beta+2$	$3\beta+3$
*Non-nouns	Z	-	Z	Z	$\alpha+2$	$3\alpha+3$
*	X_3	X_3	X_3	-	$\alpha+\beta+1$	$3\alpha+2\beta+1$
*	X_4	X_4	-	X_4	$\alpha+\beta+1$	$3\alpha+2\beta+1$
Contentives	C	C	C	C	$\alpha+\beta+2$	$4\alpha+3\beta+3$

The propositional functions are fulfilled by different word classes. Each word class is defined by the propositional functions it can have. In many languages, there are word classes that have more than one propositional function, which means that the Principle is violated in that case. Table 4 shows how all theoretically possible word classes and their corresponding weights are defined when l = 4 (the definitions are

[1] Adverbs other than manner adverbs are not considered because they typically modify not the head of the predicate phrase, but the whole sentence.

analogous for $l < 4$). Unattested word classes are denoted by an asterisk. All but one unattested class are left without a name.

Two kinds of weights are considered in Table 4. The simple weight is just the sum of the weights of the propositional functions fulfilled by the word class. The accrued weights are calculated as follows. The propositional functions that a word class has are considered one by one in the P-R-r-p order. The first function contributes its own weight (in the system W1or W2) to the total word-class weight. Each next propositional function adds its own weight to the weight assigned to the preceding existing function. Consider non-verbs for instance. We have β assigned to the R function, $1+\beta$ to r, and $1+(1+\beta)=2+\beta$ to p. This results in $\beta+(1+\beta)+(2+\beta)=3\beta+3$ for the accrued weight of Λ. The weights calculated this way introduce a greater penalty for the word classes that have more propositional functions.

Table 5. Measures of the degree of violation of the Principle using the simple word-class weights

l	*k*	PoS-system type	The basic formula (5)	Formula (6) with W1 and simple word-class weights	Formula (6) with W2 and simple word-class weights
4	4	VNam	1	1	1
	3	VNMM	2	1.308	1.240
		V~~NN~~m/VX_2aX_2	2	1.462	1.480
		~~Þ~~Na~~Þ~~/X_1NX_1m	2	1.538	1.520
		HHam	2	1.692	1.760
	2	VΛΛΛ	3.5	1.615	1.600
		ZNZZ	3.5	1.692	1.640
		$X_4X_4aX_4/X_3X_3X_3$m	3.5	1.846	1.880
		HHMM	4	2	2
		~~ÞNNÞ~~/$X_1X_2X_1X_2$	4	2	2
	1	CCCC	8	2	2
3	3	VNa/VNm	1	1	1
	2	V~~NN~~/VX_2X_2	2.5	1.546	1.546
		~~Þ~~N~~Þ~~/X_1NX_1	2.5	1.636	1.591
		HHa/HHm	2.5	1.818	1.864
	1	$X_3X_3X_3/X_4X_4X_4$	6	2	2
2	2	VN	1	1	1
	1	HH	4	2	2
1	1	V	1	1	1

In a PoS system, Y is the set of the word classes used, and $\|Y\|$ can be defined as the sum of their weights.[2] Then, it is always the case that $\|Y\| \geq \|X\|$ and $\|Y\| = \|X\|$ if ρ is a bijection. We also define $\|\rho\| = \|Y\|$. Therefore, the condition (4) is satisfied and the scaling coefficient c is equal to 1. Moreover, in what follows, we consider the simplest choice of θ, $\theta = 1$. Then, the formula (3) becomes

$$\hat{\mu}_1(\rho) = \frac{2\|\rho\| - \|B\|}{\|X\|}. \tag{6}$$

In Tables 5 and 6, we present all theoretically possible PoS systems. They are coded by the existing word classes, which are listed in the P-R-r-p order of the propositional functions they convey. For instance, in the PoS system denoted as VΛΛΛ, V has the role of P and the consecutive Λ's the roles of R, r, and p. The number of word classes in the PoS system is denoted by k, $k \leq l$. Whenever $k = l$, ρ is a bijection.

Table 6. Measures of the degree of violation of the Principle using the accrued word-class weights

l	k	PoS-system type	Formula (6) with W1 and accrued word-class weights	Formula (6) with W2 and accrued word-class weights
4	4	VNam	1	1
	3	VNMM	1.615	1.480
		V~~NN~~m/VX_2aX_2	2.077	2.200
		~~Þ~~Na~~Þ~~/X_1NX_1m	2.308	2.320
		HHam	2.462	2.560
	2	VΛΛΛ	3.154	3.280
		ZNZZ	3.538	3.480
		$X_4X_4aX_4/X_3X_3X_3m$	4.000	4.200
		HHMM	3.077	3.040
		~~ÞNN~~Þ/$X_1X_2X_1X_2$	3.385	3.520
	1	CCCC	5.846	6.080
3	3	VNa/VNm	1	1
	2	V~~NN~~/VX_2X_2	2.273	2.364
		~~Þ~~N~~Þ~~/X_1NX_1	2.545	2.500
		HHa/HHm	2.727	2.773
	1	$X_3X_3X_3/X_4X_4X_4$	4.545	4.636
2	2	VN	1	1
	1	HH	3.111	3.053
1	1	V	1	1

[2] Because of this definition, the weighted formula (3) does not reduce in general to the basic formula (2) when all weights are set equal to 1.

Table 5 shows the measures of the degree of violation of the Principle when both the basic formula (5) and the weighted formula (6) are used, the latter with the simple word-class weights in each W1 and W2 weight systems. We can see that the basic formula is too simplistic since the values do not vary much, if at all, within the same PoS-system type. This justifies the need for the weighted formula and, indeed, the formula (6) provides a greater variation of values. With the simple word-class weights, we have that $\|Y\| = \|X\|$ and the formula (6) simplifies further to $\hat{\mu}_1(\rho) = 2 - \frac{\|B\|}{\|X\|}$. Therefore, 2 is the greatest value in the last two columns in Table 5, which is achieved whenever $\|B\| = 0$. This result is problematic, particularly for the four PoS-system types with $l = 3$. It seems like the HHMM system, for instance, is closer to a bijection than CCCC, but the formula does not indicate this. Hence, a further modification of the weights is called for and this is why the accrued word-class weights are introduced. They are considered in Table 6, combined again with each weight system W1 or W2. The variation of the values is the greatest in Table 6 and the measure for the CCCC system shows correctly that this system is farthest away from a bijection.

In general, it can be concluded that the two weight systems, W1 and W2, provide similar results.

5. Conclusion

We have proposed a formula for measuring how much a general relation between two finite sets is different from a bijection between the same sets. The formula can be used in linguistics as a measure of the departure from the One-Meaning–One-Form Principle of Anttila (1972). When constructing the formula, we had certain desirable properties in mind, particularly motivated by possible applications to grammar complexity or efficiency, as discussed in Vulanović (2003, 2007).

Two versions of the formula are considered in the paper, a basic formula and a weighted one, with appropriate linguistics examples illustrating their applications. The weighted formula is applied to parts-of-speech systems in the sense of Hengeveld (1992). Grammar efficiency of parts-of-speech systems is discussed in Vulanović (2008, 2009), for instance. The formula for calculating grammar efficiency includes a factor called *parsing ratio*, which requires the number of parses of all possible simple sentences. This is obtained through a relatively cumbersome combinatorial analysis. On the other hand, the formula presented here is much simpler and produces results which are in a general agreement with the corresponding values of grammar efficiency. The greater the departure of the parts-of-speech system from the Principle, the lower its grammar efficiency (the greater its complexity). This was one of the motivations for considering the topic of this paper. A more detailed analysis of the relationship between the measure proposed here and grammar efficiency is yet to be performed. It remains to be seen whether such a measure can entirely replace the formula for grammar efficiency or just the parsing ratio.

References

Anttila, Raimo (1972). *An Introduction to Historical and Comparative Linguistics*. New York: Macmillan.

Hengeveld, Kees (1992). Parts of Speech. In: Fortescue, Michael & Harder, Peter & Kristoffersen, Lars (eds.), *Layered Structure and Reference in Functional Perspective*. Amsterdam/Philadelphia: John Benjamins, 29-55.

Hengeveld, Kees; Rijkhoff, Jan & Siewierska, Anna (2004). Parts-of-Speech Systems and Word Order. *Journal of Linguistics, 40,* 527-570.

Hengeveld, Kees & van Lier, Eva (2010). An Implicational Map of Parts of Speech. *Linguistic Discovery, 8,* 129-156.

Miestamo, Matti (2008). Grammatical Complexity in a Cross-Linguistic Perspective. In: Miestamo, Matti; Sinnemäki, Kaius & Karlsson, Fred (eds.) (2008), 23-41.

Miestamo, Matti; Sinnemäki, Kaius & Karlsson, Fred (eds.) (2008). *Language Complexity: Typology, Contact, Change.* Amsterdam: Benjamins.

Steele, Susan (1978). Word Order Variation: A Typological Study. In: Greenberg, Joseph H. (ed.), *Universals of Human Language*. Stanford: Stanford University Press, 587-623.

Vulanović, Relja (2003). Grammar Efficiency and Complexity. *Grammars, 6,* 127-144.

Vulanović, Relja (2007). On Measuring Language Complexity as Relative to the Conveyed Linguistic Information. *SKY Journal of Linguistics, 20,* 399-427.

Vulanović, Relja (2008). A Mathematical Analysis of Parts-of-Speech Systems. *Glottometrics, 17,* 51-65.

Vulanović, Relja (2009). Efficiency of Flexible Parts-of-Speech Systems. In: Reinhard Köhler (ed.), *Issues in Quantitative Linguistics* (*Studies in Quantitative Linguistics, 5*). Lüdenscheid: RAM-Verlag, 136-157.

Wolfart, H. Christoph & Carroll, Janet F. (1981). *Meet Cree: A Guide to the Cree Language*. Lincoln: University of Nebraska Press.

Quantitative Interrelations of Properties of Complement and Adjunct*

Haruko Sanada
Rissho University, Japan

Abstract
The present study focuses on (1) the proportion of complements and adjuncts in the clause, (2) relationships between complement and adjunct numbers and the position of clauses in the sentence, and (3) relationships between complement and adjunct lengths and the position of clauses in the sentence, and (4) differences between complement and adjunct lengths in the non-embedded clause and those in the embedded clause. The number of adjuncts is not related to the number of complements, but both are related to the position of the clause in the sentence and related to the type of clause, whether it is non-embedded or embedded.

Keywords: *valency, sentence structure, length, frequency, complement, adjunct, position in the sentence, Japanese, Synergetic Linguistics.*

1. Aim of the paper and background of the problem

This study looks at the naturalness of the amount and length of linguistic entities from an empirical point of view. Every language has a certain amount or a certain length of linguistic entities as common knowledge among native speakers[1]. For example, In English or in German the amount of adjuncts in a clause has no theoretical or grammatical limit. Japanese has no strict obligatory complements like a subject or a direct object. The maximum number of complements depends on the verb, but the number of complements can be less if redundancy is to be avoided. However, in Japanese there is a flexible rule on length or amount of linguistic entities in the sentence, from the point of view of naturalness. Valency dictionaries of Japanese[2] show individual verbs and their grammatically correct complements. However, valency dictionaries do not show information on the frequency of

* Address correspondence to: Haruko Sanada, Rissho University, 4-2-16, Osaki, Shinagawaku, Tokyo 141-8602, Japan. E-mail address: hsanada@ris.ac.jp

[1] Mark Twain mentioned long compound words in German in his text "The Awful German Language" (Twain 1880). Such German words seem to be unusual for him as an English speaker.

[2] Ishiwata & Ogino (1983), Koizumi et al. (2000), Rickmeyer (2008), and IPAL developed by Information Technology Promotion Agency are well known as the valency dictionary of Japanese.

complements or the number of complements that are common in a sentence.

We can deduce that such numbers or lengths are decided by linguistic property, which obeys Köhler's synergetic linguistic system (Köhler 1999, 2012). The results will also be applicable to artificial text producing systems or the second language education where knowing such common numbers or common lengths in the language is extremely useful.

The present study focuses on (1) the relationship of proportions between frequencies of complements and adjuncts, (2) relationships of complement or adjunct numbers in the clause with the position of the clause in the sentence, (3) relationships of complement or adjunct length in the clause with the position of the clause in the sentence, and (4) relationships of complement or adjunct properties with the type of clause, whether it is an embedded clause or non-embedded clause.

From our former study (Sanada, to appear) with sentences including the verb "meet", we obtained a distribution of the length of clauses in arguments (linguistic entities of which clauses consist, i.e. complements, adjuncts, and a predicate), which obeys the Positive-Cohen Poisson distribution (see Figure 1). We can observe that the length of the clause, i.e. the number of arguments, has a certain physical upper limit.

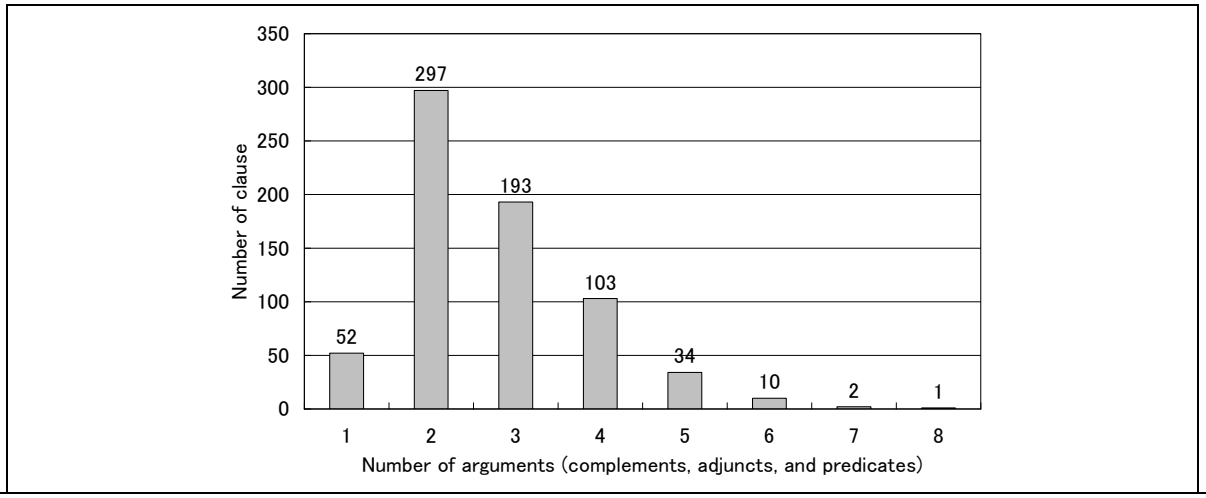

Figure 1. Distributions of the number of arguments (complements, adjuncts, and predicates) in 692 clauses of 243 sentences (Sanada, to appear)

We also obtained a distribution of the complement and adjunct types (see Figure 2.) (Sanada, 2016)[3]. In the case of the verb "meet", three complements - a subject, a direct object and a place - are registered with the form of nouns and a postposition for the case to the valency dictionaries and the valency data-base that we employed.

[3] We refer to Čech & Uhlířová (2014) who categorized adverbials in Czech into 13 groups.

For most subjects and objects, complements are supplied to the clause. However, it can be observed that for other types, e.g. complements for the place or adjuncts with the form of nouns and a postposition or adverb, various words are supplied to the clause. However, a proportion of complements and adjuncts in the clause were not investigated.

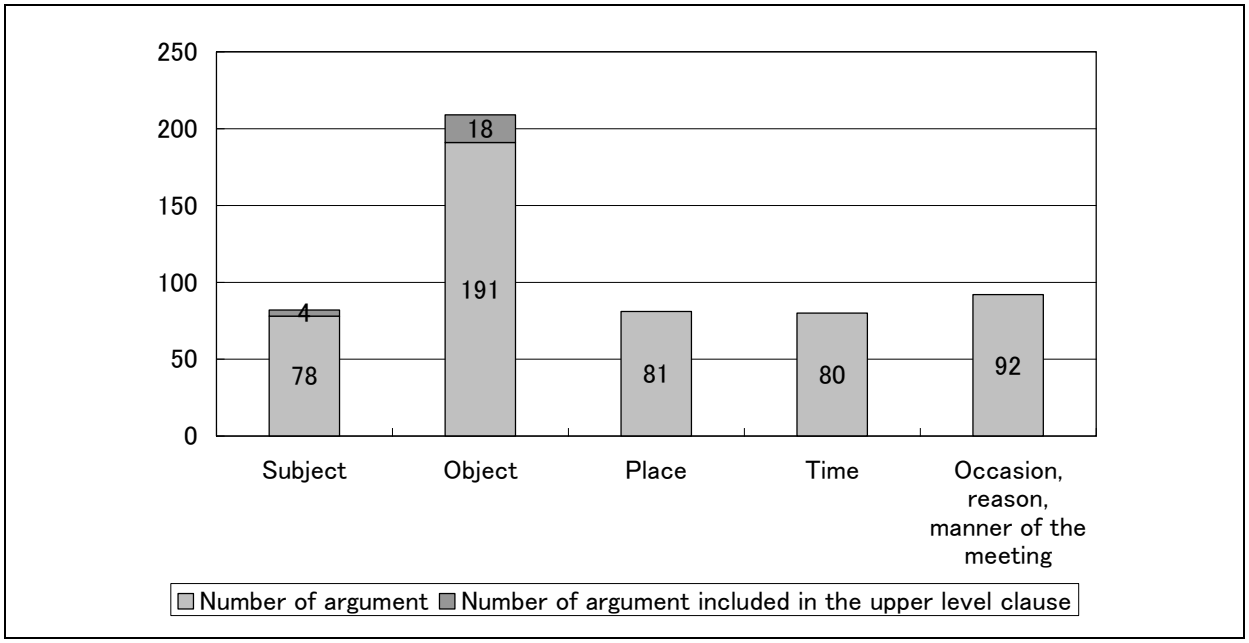

Figure 2. Complement and adjunct types from 243 clauses containing the verb "meet" (Some clauses have several arguments in the same type) (Sanada 2016)[4]

We also already obtained negative functions of the number of arguments in the clause and the position of the clause in the sentence (Figure 3[5] and vice versa) (Sanada, to appear). Their regression curves are:

$$y = 3.1853\, x^{-0.2471} \qquad (1)$$

with a correlation coefficient of $R^2 = 0.9627$ when y is an average of the number of arguments in the clause and x is the position, and

$$y = 3.2097 x^{-0.4734} \qquad (2)$$

with $R^2 = 0.9534$ when y is an average of the position and x is the number of arguments. For the length of arguments, we obtained negative functions between the

[4] A few numbers are corrected.
[5] Hereafter white circles in the figures are excluded from regression curves because a number of data points are less than 10.

length of arguments in morphemes and the position of the clause in the sentence (Figure 4 and vice versa) (Sanada, to appear). Their regression curves are:

$$y = 10.375\, x^{-0.4147} \qquad (3)$$

with $R^2 = 0.9916$ when y is an average of the argument length in morphemes and x is the position, and

$$y = 3.9226x^{-0.3227} \qquad (4)$$

with $R^2 = 0.8551$ when y is an average of the position and x is the argument length in morphemes. However, it has not yet been investigated how the number or the length of linguistic entities in the clause, i.e. complements and adjuncts, affects these negative functions.

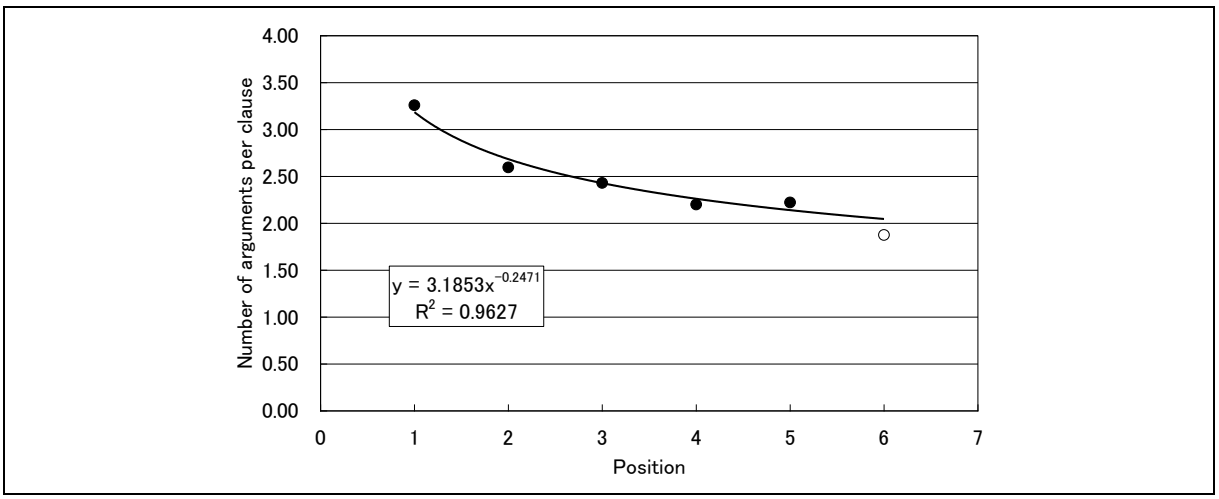

Figure 3. Relationships between the average number of arguments in the clause (y) and the position of the clause in the sentence (x) with 692 clauses (x=1 as the beginning of the sentence) (Sanada, to appear)

From the above background, the present study investigates the following four problems:

1. Proportions of complements and number of adjuncts in the clause.

2. Relationships between the number of arguments in the clauses and the position of clauses in the sentence.

3. Relationships between the length of arguments in morphemes in the clauses and the position of clauses in the sentence.

4. Difference between the number or length of arguments in the non-embedded and those in the embedded clause.

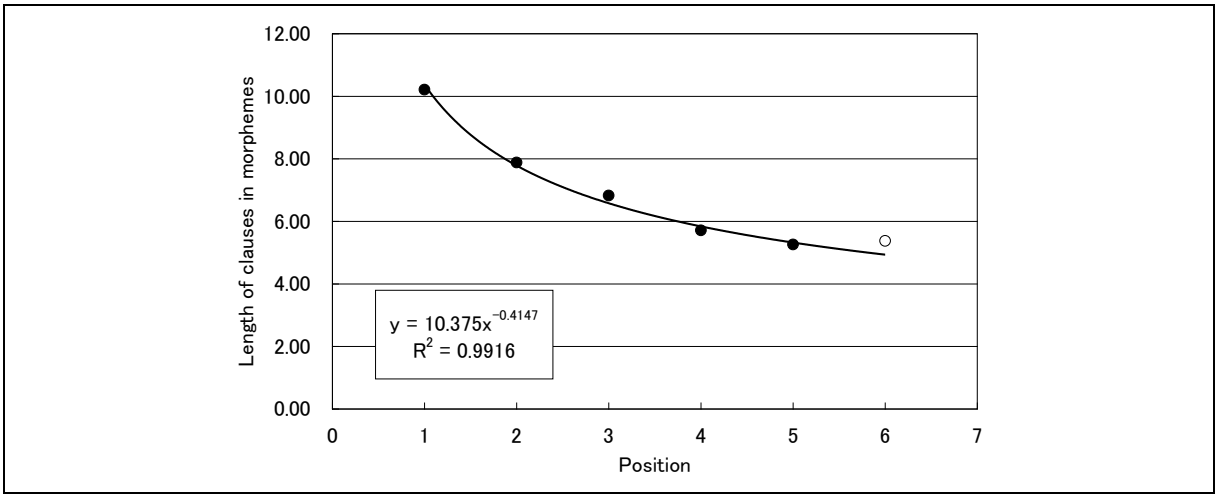

Figure 4. Relationships between the average length of clauses in morphemes (*y*) and the position of the clause in the sentence (*x*) with 692 clauses (*x*=1 as the beginning of the sentence) (Sanada, to appear)

2. Hypotheses

We set up the following hypotheses from the above problems.

Hypothesis 1. The number of adjuncts is a function of the number of complements.

Hypothesis 2. The clause that appears later in the sentence has **less** complements and adjuncts than a clause that appears earlier.

Hypothesis 3. The clause that appears later in the sentence has **shorter** complements and adjuncts than clauses that appear earlier.

Hypothesis 4. The Embedded clause has **less** complements and adjuncts than the non-embedded clause has, or the embedded clause has **shorter** complements and adjuncts than the non-embedded clause has.

3. Descriptions of data and grammatical definitions

We regard the present study as one in the series of our valency studies, and we employed the Japanese valency data-base (Ogino et al. 2003) the same as the one employed in our previous studies (Sanada 2012, 2014, 2015, 2016, to appear). In the previous studies 240 sentences containing the verb "meet" were extracted from the valency data-base. These extracted sentences also include many other verbs because each sentence has one or more predicates. 3 of the 240 sentences have 2 predicates with the verb "meet".

We used the Japanese morphological analyzer *Mecab* (Graduate Schools of Informatics in Kyoto University et al.) and the electronic dictionary *Unidic*

(National Institute for Japanese Language and Linguistics) for the extracted sentences. The software shows a boundary of the "short unit" as a morpheme. Errors were corrected by hand. Of the 240 sentences, 692 clauses, 1,889 arguments and 5,626 morphemes were obtained.

In the present study we mainly employed the 243 clauses that contain the verb "meet". The clauses have 765 arguments and 2,365 morphemes.

Our former study (Sanada 2016, to appear) took four linguistics levels, i.e. sentence, clause, argument and morpheme. The clause must have one predicate for each[6], and consist of complements, adjuncts, and a predicate. The present study follows these definitions for consistency of the study.

For the position of the clause in the sentence, we take the beginning of the clause. In the case that the last clause in the sentence is divided by an embedded clause, the beginning of the first half is taken as the basis of data, and the second half is not considered. Figure 5 shows a model of the position of clauses in the sentence and the embedded clause.

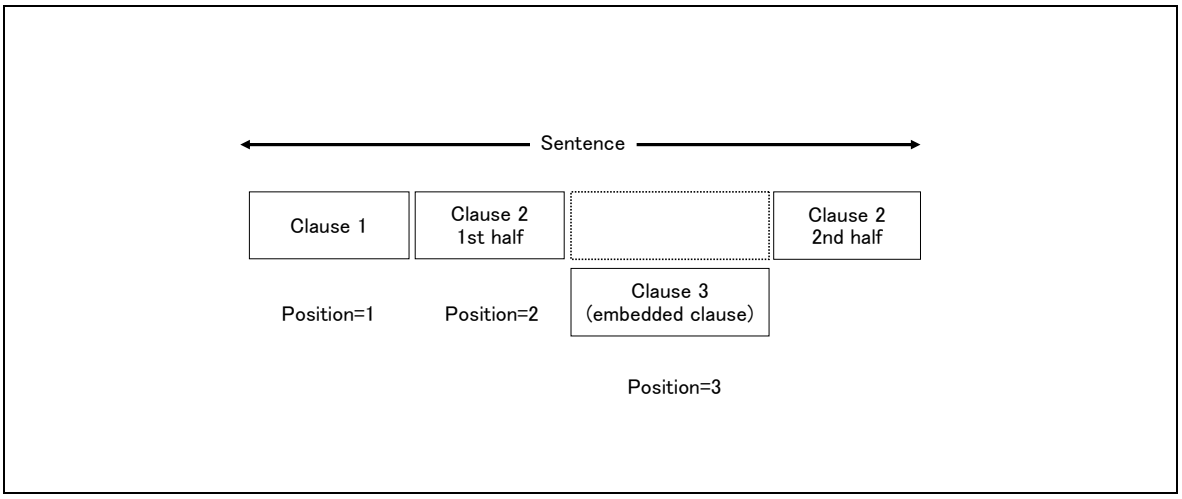

Figure 5. The position of the clause and the embedded clause

The rules used for counting the data are shown here with examples, which were also shown in our former study. A space in the example shows a morpheme boundary. A single slash mark (/) and a double slash mark (//) show the boundary of arguments and boundaries of the clauses respectively. The number of clauses, arguments and morphemes of the example follow its English equivalents. The numbers shown with "ID" after the example indicate a sentence number in the data-base.

(1) Abbreviations and grammatical remarks in the examples

Abbreviations or grammatical remarks used in the examples are as follows:

[6] In 2 of the 692 clauses, a predicate is missing because the sentences are grammatically incorrect.

ATTR=attributive, CONNECT=connective form, CONTINUOUS=continuous aspect, COPULA, GEN=genitive, INS=instrument, LOC=location, NOM=nominalized form, OBJ=object, PAST=past tense form, PERFECT, PP=postposition, PROG=progressive form, SUBJ=subject[7], TOPIC.

(2) Definitions of the sentence and the morpheme

The "sentence" in Japanese is optically clear as it has a sign at the end. The "morpheme" is a topic that is still being discussed. We employ the definition of the "short unit" as a morpheme, which was developed by the National Institute of Japanese Language and Linguistics (National Language Research Institute 1964).

(3) Definition of the clause

The clause is also a topic that is still being discussed in Japanese linguistics. Minami (1974, 1993) analyzed grammatically important types of Japanese clauses[8]. Here, referring to his model, we studied our data using quantitative and empirical analysis. In the present study we defined it as a linguistic unit that has a predicate on the surface of the sentence.

(4) Definition of the argument

The argument is defined as the level between the clause and the morpheme. We regard the predicate and grammatical elements that are linked to the predicate as arguments in the clause. All elements in the clause are employed as arguments. Attributive elements, i.e. a noun and a postposition were treated as a part of the argument (see underlined words in Example 1). Among the arguments, those tagged in the Japanese valency data-base (Ogino et al. 2003) were defined as the complement, the rest of the arguments, except the predicate, are defined as the adjunct. Conjunctions, except postpositions, which belong to the clause were also regarded as a member of the clause in the present study, and categorized into an adjunct. This definition should be discussed in our future studies.

Example 1:

<u>Yogo no</u> Ishikawa Keiko sensei wa,/ chugaku 3 nen no shojo no hahaoya ni/ at ta.
(ID: JCO0217129)

[7] The Japanese postposition "*ga*" is a subject marker and "*wa*" is a topic marker. However, "*wa*" also works as a subject marker.

[8] "Minami's model" is also employed in software of the National Institute of Japanese Language and Linguistics to find the boundaries of some types of clauses. However, his model does not cover all types of Japanese clauses in the corpus because it focuses on grammatical and semantic aspects.

[nurse-teacher-ATTR Mrs. Keio Ishikawa-SUBJ/ junior high school 3rd grade-ATTR girl-GEN mother-OBJ/ meet-PAST]
(Mrs. Ishikawa Keiko of a nurse-teacher met a mother of the girl in the 3rd grade of the junior high school.)
Clause=1, Argument = 3, Morpheme=16.

(5) The sub clause and the embedded clause

Japanese has no marker of the relative pronoun for sub clauses (see the underlined beginning of the sentence in Example 2) or embedded clauses that divide the upper level clause into two parts (see the underlined words in Example 3.).

Example 2:
Watashi ga/ at ta// chiji no hotondo wa,/ chiji shitsu ni/ nihon no ningyo ya okimono wo/ oi te i ta. (ID: JCO0138531)
[I-SUBJ/ meet-PAST// prefectural governors-ATTR most-TOPIC/ prefectural governor office-LOC/ Japan-ATTR dolls or ornamental objects-OBJ/ display-CONNECT- CONTINUOUS-PAST]
(Most of prefectural governors whom I met displayed Japanese dolls or ornaments in their office.)
Clause=2, Argument = 6, Morpheme=21.

Example 3:
Henshu cho shitsu de/ nan do ka/ at ta ga,// itsu mo/ taatorunekku no seetaa ni// zakkuri shi ta// sebiro wo/ haot te i ta. (ID: JCO0209028)
[chief editor room-LOC/ several times/ meet-PAST-CONNECT,// always/ turtleneck-ATTR sweater-CONNECT// roughly weaved-PAST// jacket-OBJ/ was wearing (PROG-PAST)]
(I met him for several times in the office of the chief editor, and he always wore a sweater with a turtleneck and a jacket which roughly weaved.
Clause=3, Argument = 8, Morpheme=25.

For more detailed definitions, e.g. definitions related to the predicate or other special cases, Sanada (2016) should be referred to.

4. Results and interpretations

4.1. Relationship between the number of complements and the number of adjuncts

Here we investigated the following hypothesis:

Hypothesis 1. The number of adjuncts is a function of the number of complements.

The relationship between the number of complements (x) and the number of adjuncts (y) is shown in Figure 6. The number of data points is also shown in the figure with the size of each circle. We did not take averages of y because this section focuses on the proportion of two variables. We obtained a regression function between the number of adjuncts (y) and the number of complements (x) from all data points as follows:

$$y = 0.0053\, x^{3.9636} + 0.6542 \qquad (5)$$

with their correlation coefficient (R^2) 0.0176 which is very low. The correlation between two variables is 0.056, which is also very low. Therefore, the frequency of adjuncts is almost stable regardless of the frequency of complements.

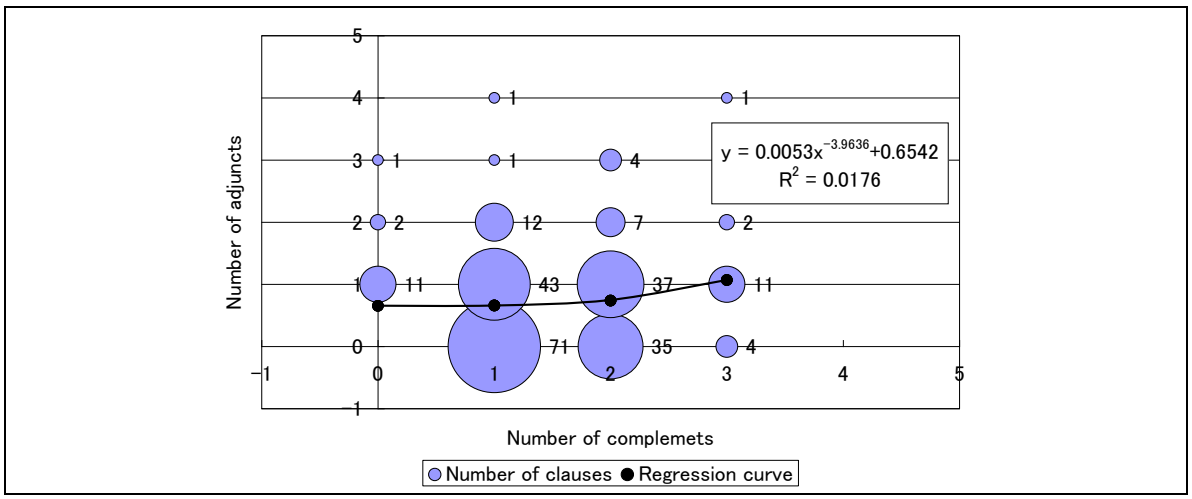

Figure 6. Number of adjuncts (y) and number of complements (x) with a regression curve (243 clauses including a verb "meet")

By observing the regression curve, it becomes clear that the number of adjuncts (the average is 0.72) is mostly equal to or less than the number of complements (the average is 1.43). We conducted a t-test for the averages of 243 pairs with a one sided

test as we expected that the number of complements must be more than the number of adjuncts. The value of *t*-statistics is 10.699, and the critical value is 2.255 with 243 degrees of freedom at 2.5% (5%* 1/2) level. The null hypothesis is rejected and the number of adjuncts is significantly less than the number of complements. It can be interpreted that the complement has higher priority than the adjunct in the clause.

4.2. Relationships between the number of complements or adjuncts and the position

In this section we investigated the following hypothesis:

Hypothesis 2. The clause that appears later in the sentence has **less** complements and adjuncts than a clause that appears earlier.

First we confirmed that there is a negative function with 243 clauses that include the verb "meet", the same as the case with 692 clauses from 240 sentences[9]that include the verb "meet". The figure is shown as Figure 7b, and Figure 3 is also shown as Figure 7a again for comparison. We obtained a regression function

$$y = 3.4176x^{-0.2249} \qquad (6)$$

with $R^2 = 0.9274$ when y is the number of arguments and x is the position, and a regression function for the reverse relationship is:

$$y = 3.2377x^{-0.615} \qquad (7)$$

with $R^2 = 0.9771$ when y is the position and x is the number of arguments. It can be assumed that the tendency of the relationships is almost the same even though the number of clauses is less.

[9] 3 of 240 sentences have 2 clauses for each, which include a verb "meet" in their predicate. Therefore 243 clauses that include the verb "meet" are employed for the present study.

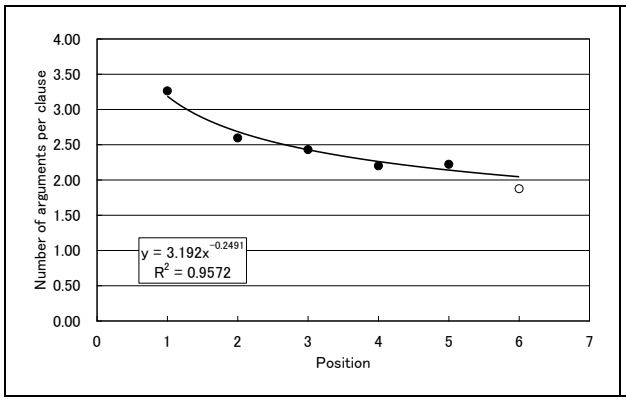

Figure 7a (=Figure 3). Relationships between the average number of arguments in the clause (*y*) and the position of the clause in the sentence (*x*) with 692 clauses (*x*=1 as the beginning of the sentence)

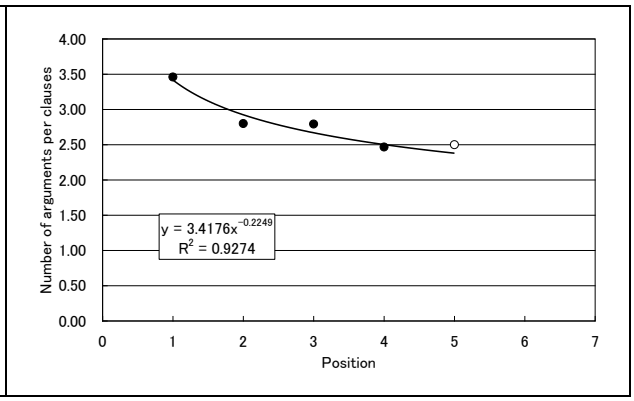

Figure 7b. Relationships between the average number of arguments in the clause (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Now we tested relationships between the number of complements or adjuncts and the position. There are 348 complements and 174 adjuncts in 243 clauses including a verb "meet". The results are shown in Figure 8a and Figure 8b, and their data and regression functions are shown in Table 1a and Table 1b respectively.

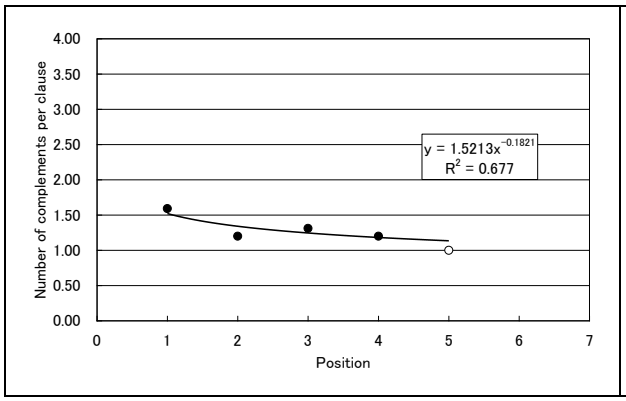

Figure 8a. Relationships between the average number of complements per clause (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

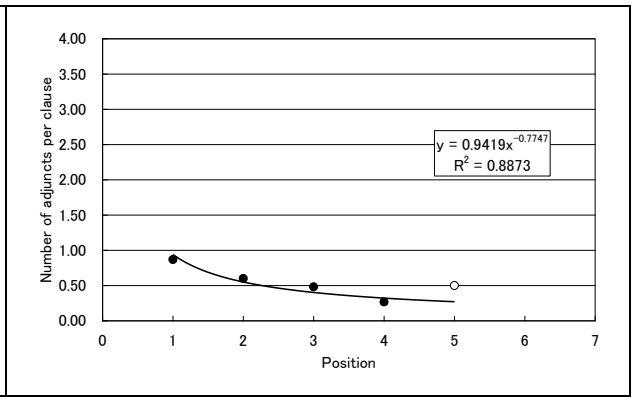

Figure 8b. Relationships between the average of adjuncts per clause (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Table 1a. Relationships between the average number of complements per clause (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Position of the clause in the sentence (*x*) (x=1 as the beginning of the sentence)	Average number of complements per clause (*y*)	Number of data points
1	1.59	137
2	1.20	60
3	1.31	29
4	1.20	15
5	1.00 (*)	2
Total		243
Regression function $y = 1.5213x^{-0.1821}$	$R^2 = 0.677$	(8)
Regression functions for the reverse relationship		
$y= 0.1806\ x^{1.156965} + 0.1990893$	$R^2 = 0.0475$	(9)
$y = -0.2262x + 1.9574$	$R^2 = 0.8824$	(10)

(*) Data is excluded from the regression curve.

Table 1b. Relationships between the average number of adjuncts per clause (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Position of the clause in the sentence (*x*) (x=1 as the beginning of the sentence)	Average number of adjuncts per clause (*y*)	Number of data points
1	0.87	137
2	0.60	60
3	0.48	29
4	0.27	15
5	0.50 (*)	2
Total		243
Regression function $y = 0.9419x^{-0.7747}$	$R^2 = 0.8873$	(11)
Regression functions for the reverse relationship		
$y = 0.1302\ x^{0.6031} + 0.0903$	$R^2 = 0.0521$	(12)
$y = -0.2988x + 1.9162$	$R^2 = 0.972$	(13)

(*) Data is excluded from the regression curve.

The values of R^2 can be regarded as being enough, meaning that the hypothesis is

supported, i.e. the clause appearing later in the sentence has **less** complements and adjuncts than a clause appearing earlier.

4.3. Relationship between the length of complements or adjuncts and the position

Here we investigated the following hypothesis:

Hypothesis 3. The clause that appears later in the sentence has **shorter** complements and adjuncts than a clause that appears earlier.

The same as in the previous section, we confirmed first that there is a negative function with 243 clauses that include the verb "meet" the same as the case with 692 clauses from 240 sentences which include the verb "meet". The figure is shown as Figure 9b, and Figure 4 is also shown again as Figure 9a for comparison. We obtained the following regression function

$$y = 10.963x^{-0.3456} \qquad (14)$$

with $R^2 = 0.8424$ when y is the length of arguments in morphemes and x is the position, and the regression function for the reverse relationship is:

$$y = 3.1667x^{-0.2926} \qquad (15)$$

with $R^2 = 0.3857$ when y is the position and x is the length of arguments in morphemes. For the relationship of the length of arguments (y) and the position (x), the negative function fits the data even though the number of clauses is less.

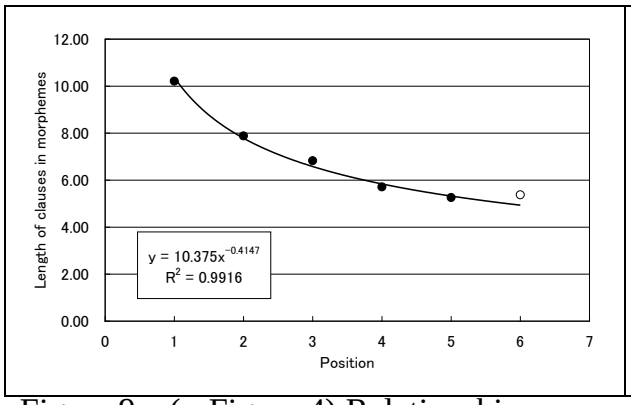

Figure 9a. (= Figure 4) Relationships between the average length of clauses in morphemes (y) and the position of the clause in the sentence (x) with 692 clauses (x=1 as the beginning of the sentence)

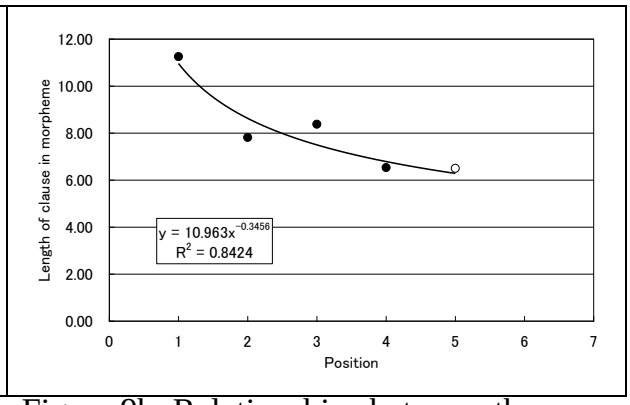

Figure 9b. Relationships between the average length of clauses in morphemes (y) and the position of the clause in the sentence (x) with 243 "meet" clauses (x=1 as the beginning of the sentence)

For the next step relationships of argument lengths in morphemes with the position of the clause in the sentence were investigated for 692 clauses and 243 "meet" clauses respectively. The results are shown in Figure 10a and Figure 10b, and their data and regression functions are shown in Table 2a and Table 2b respectively. We observed that both of the regression curves in Figure 10a and Figure 10b are slightly decreasing, but flatter than curves of the whole clause length (Figure 9a and Figure 9b), and that the tendencies of Figure 10a and Figure 10b are almost the same as each other even though the number of clauses is less.

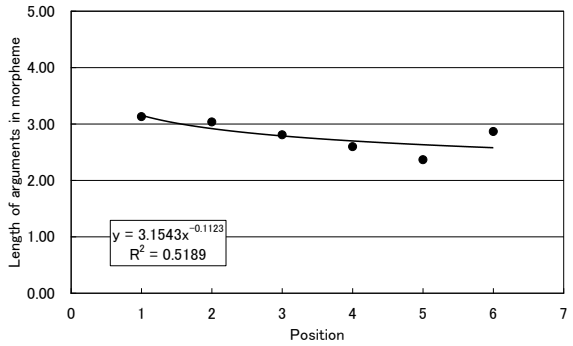

Figure 10a. Relationships between the average length of arguments in morphemes (y) and the position of the clause in the sentence (x) with 692 clauses (x=1 as the beginning of the sentence)

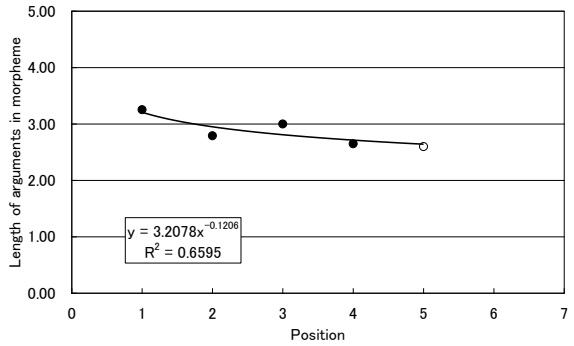

Figure 10b. Relationships between the average length of arguments in morphemes (y) and the position of the clause in the sentence (x) with 243 "meet" clauses (x=1 as the beginning of the sentence)

Table 2a. Relationships between the average length of arguments in morphemes (y) and the position of the clause in the sentence (x) with 692 clauses (x=1 as the beginning of the sentence)

Position of the clause in the sentence (x) (x=1 as the beginning of the sentence)	Average length of arguments in morphemes (y)	Number of data points
1	3.13	782
2	3.04	545
3	2.81	333
4	2.60	154
5	2.37	60
6	2.87	15
Total		1889
Regression function $y = 3.1571x^{-0.113}$	$R^2 = 0.5219$	(16)
Regression functions for the reverse relationship $y= 2.2424\ x^{-0.1082}$	$R^2 = 0.5947$	(17)

Table 2b. Relationships between the average length of arguments in morphemes (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Position of the clause in the sentence (*x*) (x=1 as the beginning of the sentence)	Average length of arguments in morpheme (*y*)	Number of data points
1	3.25	474
2	2.79	168
3	3.00	81
4	2.65	37
5	2.60 (*)	5
Total		765
Regression function $y = 3.2078x^{-0.1206}$	$R^2 = 0.6595$	(18)
Regression functions for the reverse relationship $y = 1.8149x^{-0.1412}$	$R^2 = 0.6584$	(19)

(*) Data is excluded from the regression curve.

Then relationships of the length of complements or adjuncts with the position were investigated. The data used is the same as in the previous section, i.e. 348 complements and 174 adjuncts in 243 clauses including the verb "meet". The results are shown in Figure 11a and Figure 11b, and their data and regression functions are shown in Table 3a and Table 3b respectively.

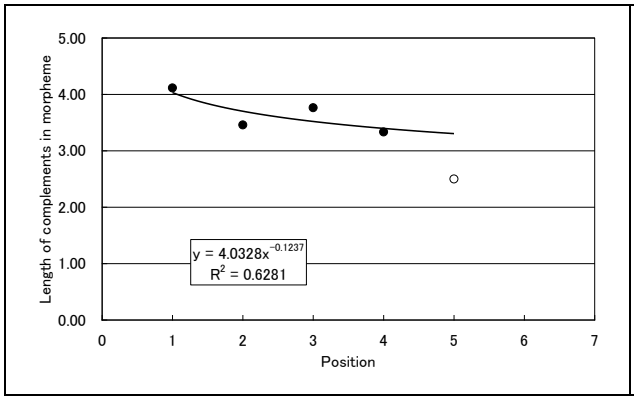

Figure 11a. Relationships between the average length of complements in morphemes (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

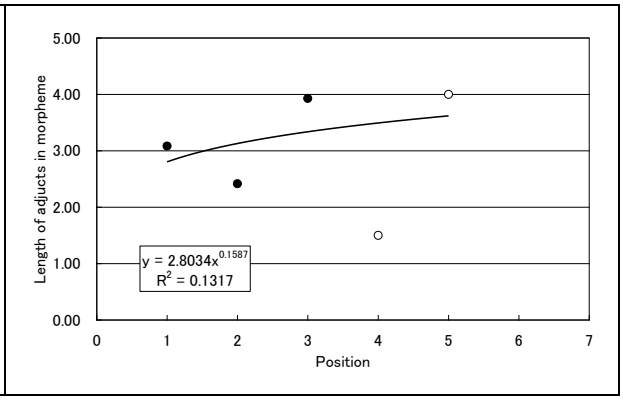

Figure 11b. Relationships between the average length of adjuncts in morphemes (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Table 3a. Relationships between the average length of complements in morphemes (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Position of the clause in the sentence (*x*) (x=1 as the beginning of the sentence)	Average length of complements in morpheme (*y*)	Number of data points
1	4.11	218
2	3.46	72
3	3.76	38
4	3.33	18
5	2.50 (*)	2
Total		348
Regression function $y = 4.0328x^{-0.1237}$	$R^2 = 0.6281$	(20)
Regression functions for the reverse relationship $y= 1.9562\ x^{-0.157}$	$R^2 = 0.3796$	(21)

(*) Data is excluded from the regression curve.

Table 3b. Relationships between the average length of adjuncts in morphemes (*y*) and the position of the clause in the sentence (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the sentence)

Position of the clause in the sentence (*x*) (x=1 as the beginning of the sentence)	Average length of adjuncts in morpheme (*y*)	Number of data points
1	3.08	119
2	2.42	36
3	3.93	14
4	1.50 (*)	4
5	4.00 (*)	1
Total		174
Regression function $y = 2.8034x^{0.1587}$	$R^2 = 0.1317$	(22)
Regression functions for the reverse relationship $y = 1.546\ x^{-0.077}$	$R^2 = 0.3862$	(23)

(*)Data is excluded from the regression curve.

The values of R^2 for the regression functions are smaller than 0.60, excepting the relationship between the average length of complements in morphemes (*y*) and the position of the clause in the sentence (*x*) (Figure 11a and Table 3a). Therefore our hypothesis, i.e. the clause appearing later in the sentence has **shorter** complements

and adjuncts than a clause appearing earlier, is not supported except for the case of the complements.

4.4. Numbers or length of complements or adjuncts in non-embedded/embedded clauses

A *t*-test from our former study (Sanada 2016) showed that the number of arguments per clause in the embedded clause is significantly less than those in the non-embedded clause (see Table 4). The distributions of the number of arguments are shown in Figure 12a.

Now we can confirm that the length of arguments in morphemes in the embedded clause is significantly lower than those in the non-embedded clause by means of a one-sided *t*-test. From our former study we expected that the length in the embedded clause must be shorter than those in the non-embedded clause. The number of non-embedded clauses is 563, and the number of embedded clauses is 129. The value of *t*-statistics is 5.323, and the critical value is 2.332 with 690 degrees of freedom at 0.5% (1%* 1/2) level. The null hypothesis is rejected and the length in the embedded clause is significantly shorter than the length in the non-embedded clause. The result of the *t*-test is shown in Table 4, and the distributions are shown in Figure 12b.

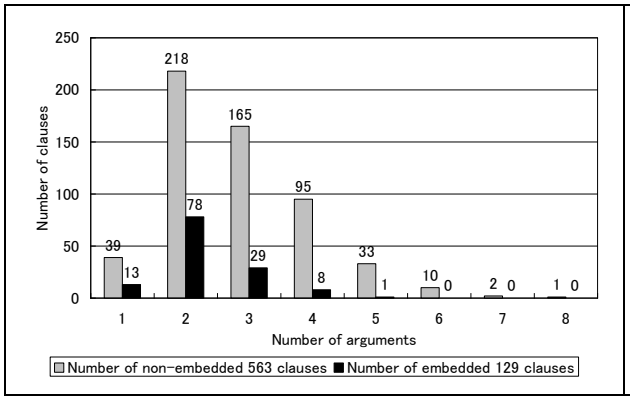

Figure 12a. Distributions of the number of complements and adjuncts in the non-embedded/ embedded clause (692 clauses)

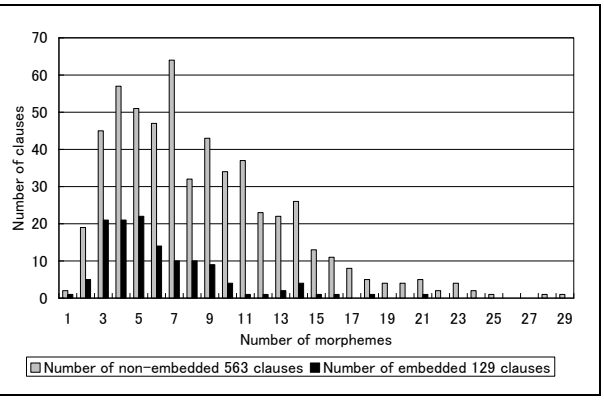

Figure 12b. Distributions of the length of complements and adjuncts in the non-embedded/ embedded clause (692 clauses)

Then employing 243 clauses that include the verb "meet", we investigated the difference between the number or length of complements and adjuncts in the embedded clause and those in the non-embedded clause. Distributions of the number of complements and adjuncts are shown in Figure 13a, and distributions of length of

complements and adjuncts in morphemes are shown in Figure 13b.

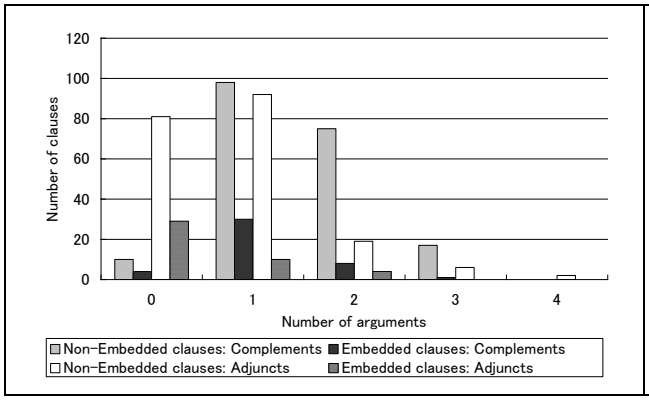

Figure 13a. Distribution of the number of arguments in non-embedded/ embedded "meet" clauses (243 clauses)

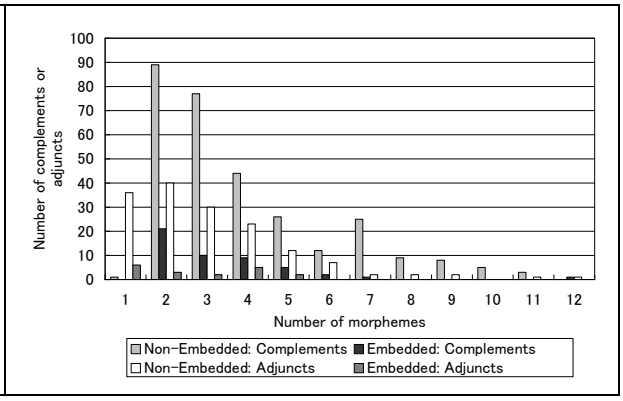

Figure 13b. Distribution of the length of arguments in morphemes in non-embedded/ embedded "meet" clauses (243 clauses)

We conducted one-sided *t*-tests for the number of complements and the number of adjuncts in the non-embedded and the embedded clauses respectively. The values of *t*-statistics are 3.007 for the complements and 2.718 for the adjuncts. The critical value is 2.342 with 241 degrees of freedom at 0.5% (1%* 1/2) level for both the number of complements and the number of adjuncts. The null hypothesis is rejected and the number of complements and the number of adjuncts in the embedded clauses are significantly smaller than those in the non-embedded clauses.

We also conducted one-sided *t*-tests for the length of complements and the length of adjuncts in the non-embedded and the embedded clauses respectively. The values of *t*-statistics are 1.928 for the complements and 0.726 for the adjuncts. The critical value is 1.649 with 346 degrees of freedom at 2.5% (5%* 1/2) level for the length of complements, and 1.654 with 172 degrees of freedom at 2.5% level for the length of adjuncts. In the case of complement length, the null hypothesis is rejected and those in the embedded clauses are significantly shorter than those in the non-embedded clauses. However, in the case of adjunct length, the null hypothesis is not rejected.

The summary of the *t*-test results is shown in Table 4.

Table 4. Summary of *t*-test results for the numbers or the length in the non-embedded / embedded clauses

Type of data	Number of samples	Values of *t*-statistics/ Critical value	Degrees of freedom/ Level (one side)	Figure of distributions
Number of arguments in 692 clauses	Non-embedded: 563 clauses Embedded: 129 clauses	5.3565** 2.332	690, 0.5%	Figure 12a
Length of arguments in 692 clauses	Non-embedded: 563 clauses Embedded: 129 clauses	5.2323**, 2.332	690, 0.5%	Figure 12b
Number of complements in 243 clauses	Non-embedded: 200 clauses Embedded: 43 clauses	3.007**, 2.342	241, 0.5%	Figure 13a
Number of adjuncts in 243 clauses	Non-embedded: 200 clauses Embedded: 43 clauses	2.718**, 2.342	241, 0. 5%	Figure 13a
Length of complements in 243 clauses	Non-embedded: 299 complements Embedded: 49 complements	1.928*, 1.649	346, 2.5%	Figure 13b
Length of adjuncts in 243 clauses	Non-embedded: 156 adjuncts Embedded: 18 adjuncts	0.726, 1.654	172, 2.5%	Figure 13b

(**) It is significant at 1% level. (*) It is significant at 5% level.

5. Discussion and conclusions

From our former results we obtained a distribution of arguments, and we observed that there is a negative relationship between the number of arguments and the position of the clause in the sentence. Therefore the interrelations of properties in the clause are our present topic. In the present study we investigated the proportion of the number of complements and the number of adjuncts in the clause, relationships between the number or the length of complements and adjuncts and the position in the sentence, and the difference of properties in non-embedded/ embedded clauses.

The following conclusions were obtained:

1. The number of adjuncts is almost stable regardless of the number of the complements. The average number of adjuncts is significantly less than the average number of complements.

2. Relationships between the length of clauses containing the verb "meet" (i.e. length in the arguments or in the morphemes) and the position of clauses in the sentence have a decreasing function. Namely, the clause is longer in the beginning of the sentence, and according to the position of the clause closing to the end, the shorter the clause in the arguments or in the morphemes. This tendency is the same as relationships between sentences with various verbs.
3. To make the clause shorter according to the position of the clause closing to the end, two options are possible. One is to make number of arguments less, the other one is to make arguments themselves shorter in morpheme. If the number of adjuncts is relatively stable, the length of adjuncts in morpheme must be less. Comparing to the functions of the length in morpheme, functions of number of complements or adjuncts fit to curves better. It can be interpreted that to omit arguments from the clause and make less numbers of arguments is preferable to making arguments themselves shorter in morpheme.
4. Complements in the embedded clause are significantly shorter or less than ones in the non-embedded clause. Adjuncts in the embedded clause are significantly less than ones in the non-embedded clause. However, the length of adjuncts in the embedded clauses and one in the non-embedded clause are not significantly different from each other.

We need more data with various verbs to confirm these characteristics, which we shall investigate in the future.

References

Čech, Radek & Uhlířová, Ludmila. (2014). Adverbials in Czech: Models for their frequency distribution. In: Uhlířová, Ludmila; Altmann, Gabriel; Čech, Radek; & Mačutek, Jan. (eds.) (2014). *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag, pp.45-59.

Grzybek, Peter. (ed.) (2006). *Contributions to the Science of Text and Language: Word Length Studies and Related Issues.* Dordrecht: Springer.

Helbig, Gerhard & Schenkel, Wolfgang. (1969, 1983). *Wörterbuch zur Valenz und Distribution deutscher Verben.* Leipzig: Bibliographisches Institut. (Reprint by De Gruyter)

Ishiwata, Toshio & Ogino, Takano. (1983). Nihongo Yogen no Ketsugoka [Valency of declinable words in Japanese]. In: Mizutani, Shizuo. (ed.), *Bunpo to Imi 1* (Grammar and meaning 1), Tokyo: Asakura Shoten, pp.226-272.

Köhler, Reinhard. (1999). Syntactic Structures: Properties and Interrelations. In:

Journal of Quantitative Linguistics, 6 (1), pp.46-57.

Köhler, Reinhard. (2012). *Quantitative Syntax Analysis.* Berlin: Mouton De Gruyter.

Koizumi, Tamotsu; Funaki, Michio; Honda, Kyoji; Nitta, Yoshio & Tsukamoto, Hideki. (2000). *Nihongo Kihon Doshi Yoho Jiten* [Dictionary of usage of basic verbs in Japanese]. Tokyo: Taishukan Shoten.

Minami, Fujio. (1974). *Gendai nihongo no kozo* [The structure of the present Japanese]. Tokyo: Taishukan.

Minami, Fujio. (1993). *Gendai nihongo bunpo no rinkaku* [The outline of the grammar of the present Japanese]. Tokyo: Taishukan.

National Language Research Institute. (1964). *Gendai Zasshi 90shu no Yogo Yoji: Dai3bunsatsu: Bunseki* [Vocabulary and Chinese Characters in Ninety Magazines of Today: vol.3: Analysis of Results]. Tokyo: Shuei Shuppan.

Ogino, Takano; Kobayashi, Masahiro & Isahara, Hitoshi. (2003). *Nihongo Doshi no Ketsugoka* [Verb valency in Japanese]. Tokyo: Sanseido.

Rickmeyer, Jens. (2008). *Kleines japanisches Valenzlexikon. 2nd edition.* Hamburg: Helmut Buske Verlag.

Sanada, Haruko. (2012). Joshi no Shiyo Dosu to Ketsugoka ni Kansuru Keiyoteki Bunseki Hoho no Kento [Quantitative approach to frequency data of Japanese postpositions and valency]. In: *Rissho Daigaku Keizaigaku Kiho* [The quarterly report of economics of Rissho University], vol. 62, no. 2, pp. 1-35.

Sanada, Haruko. (2014). The choice of postpositions of the subject and the ellipsis of the subject in Japanese. In: Uhlířová, L.; Altmann, G.; Čech, R. & Mačutek, J. (eds.) *Empirical Approaches to Text and Language Analysis*. Lüdenschied: RAM-Verlag, pp. 190-206.

Sanada, Haruko. (2015). A co-occurrence and an order of valency in Japanese sentences. In: Tuzzi, A.; Mačutek, J. & Benešová, M. (eds.) *Recent Contributions to Quantitative Linguistics.* Berlin: Walter de Gruyter, pp. 139-152.

Sanada, Haruko. (2016). The Menzerath-Altmann law and sentence structure. *Journal of Quantitative Linguistics, vol. 23-3,* pp. 256-277.

Sanada, Haruko. (To appear). Quantitative aspects of the clause: the length, the position and the depth of the clause. *Journal of Quantitative Linguistics.*

Twain, Mark. (1880, 1977). A Tramp abroad. Hartford, Conn.: American Pub. Co.; London: Chatto & Windus. (Reprinted by Michigan: Scholarly Press).

Tesnière, Lucien. (1959, 1988). *Éléments de Syntaxe Structurale. 2nd edition.* Paris: Klincksieck.

Software and digital dictionaries

Altmann, Gabriel. Fitter, version 3.1.3.0. Lüdenschied: RAM-Verlag.

Graduate Schools of Informatics in Kyoto University; NTT Communication Science Laboratories. (2008). Morphological analyzer: *MeCab*, version 0.97. (https://code.google.com/p/mecab/)

Information Technology Promotion Agency. (2007). Digital dictionary for the natural language processing: *IPAL (verbs, adjectives and nouns)*, GSK edition. (http://www.gsk.or.jp/en/catalog/)

National Institute for Japanese Language and Linguistics. (2008). Digital dictionary for the natural language processing: *UniDic*, version 1.3.9. (http://www.ninjal.ac.jp/corpus_center/unidic/)

Remarks

This work was supported by the Alexander von Humboldt Foundation [grant number 2014-2016].

Multilingual Quantitative Analysis of Morphological Ambiguity*

Eduard Klyshinsky[1], *Varvara Logacheva*[1], *Joakim Nivre*[2]
[1] *Keldysh Institute of Applied Mathematics, Russian Academy of Science*
[2] *Uppsala University, Department of Linguistics and Philology*

Abstract
The article deals with the morphological ambiguity. We define six types of word ambiguity and compute their distributions in nine languages. These distributions vary significantly depending on the number of grammatical features and other properties of a language. We have found that the distribution for more frequent lexis varies rarer one. We state that designing a high-quality POS-tagger for highly-inflected languages is more difficult than for languages like English.

Keywords: *morphological ambiguity, ambiguity classes, words' frequencies.*

1. Introduction

Nowadays the most promising techniques in natural language processing (NLP) are the techniques related to statistical analysis of the data, since more and more texts are becoming available. It has become possible not only to describe language phenomena, but also to define their relative frequency in a language. This possibility is crucial for NLP. Almost all layers of language are organized as follows: there is a reasonable number of rules and regularities, and a large number of exceptions to these rules. Those exceptions are usually too numerous and each of them often needs a separate manual declaration when creating a computational model of language. Such exceptions are difficult to incorporate into text analysis applications. However, statistical analysis can define the frequency of the exceptions, and therefore can help make a decision on whether those descriptions are possible to omit without significant loss in quality. In addition, the observation of a big number of extreme cases can help in finding an alternative system in them.

During the process of an NLP application development we have faced the lack of statistical information about languages, e.g. frequencies of parts of speech in different languages, the amount of ambiguous words and the most common types of ambiguity. Most of this information is easy to obtain. However, to the best of our knowledge, there is no comprehensive statistical digest of a language or a group of languages. Such information is reported in small chunks in different NLP papers as a side effect of the preparatory research. Projects like the World Atlas of Language Structures Online (WALS)[1] collect information on different languages, but it is rather qualitative

* Address correspondence to: Eduard Kyshinsky, Keldysh Institute of Applied Mathematics, Miusskaya sq.. 4, 125047, Moscow, Russia. E-mail: klyshinsky@itas.miem.edu.ru; varvara.logacheva@gmail.com; joakim.nivre@lingfil.uu.se
[1] http://wals.info

than quantitative. Nevertheless, we believe that various NLP applications can benefit from statistical information on languages.

2. Problem statement

One of the applications of such a collection of statistical data on a number of languages could be cross-lingual comparison of the performance of various NLP applications. Nowadays the results of part-of-speech (POS) tagging for different languages are not comparable. Therefore, we cannot tell if bad performance of a POS-tagger is due to the lack of training data, or the flaws of a method, or some properties of the language that bound the quality to some threshold. The statistical information on the amount of ambiguity in a language could give us a clue to that problem.

Another issue in contemporary NLP which can be tackled by this research is the transferability of methods and techniques to new languages. Although many of developed methods are declared to be language-independent, they are rarely tested on more than a limited number of languages, and there is often no analysis of limitations of the methods with respect to the properties of a language. Unfortunately, examples of such limitations can be often found when adapting POS-taggers originally designed for English to highly inflected languages (e.g. Russian). Protopopova and Bocharov (2013) describe the adaptation of Brill's (1995) POS-tagger to Russian: although the method itself is shown to be suitable, the feature set designed for English cannot capture the rules for Russian, it needs to be extended with grammatical features (number, gender, case, etc.). Sharoff and Nivre (2011) demonstrate an analogous situation: the TnT tagger developed by Brants (2000) achieves reasonable quality in Russian only with an extended tag set. The analysis of statistical information can help us decide if a method is applicable to languages with some set of grammatical properties. We may also be able to define what amount of training data is needed for different languages to achieve a certain level of performance.

Various NLP applications may require statistical information on any level: from distribution of phonemes to characteristics of entire texts. However, we decided to start with research on the distribution of POS-tags, because POS-tagging is the initial stage of text analysis in many NLP applications, e.g. parsing, information extraction. More generally, this paper deals with the differences in languages with respect to morphological properties including not only POS-tags but also grammatical features and lemmas. The issue we would like to examine is the issue of ambiguous words, i.e. words whose surface form does not uniquely determine the word’s lemma, part of speech or grammatical features, but offers several variants. This phenomenon in language has made POS-tagging a non-trivial task which still cannot be solved with 100% accuracy. The problem of ambiguous words exists in all languages, but the causes can vary. For example, English is known for its polysemy, so a large number of content words can belong to different parts of speech, which requires special techniques for POS-tagging. On the other hand, in highly inflected languages the amount of part-of-speech ambiguity is lower, but they have a problem of distinguishing different forms of one word. These differences require specific approaches in POS-tagging and parsing depending on the features of a language. This fact is

generally understood by researchers, but there is no quantitative information that could be directly applied to this problem.

We try to fill this gap by examining the distribution of words in corpora among different classes of ambiguity divided by such parameters as ambiguity by lemma, by part of speech, and by grammatical features of the word. The results of the experiments show some crucial differences between the considered languages, which are: English, French, Spanish, Portuguese, Italian, German, Russian, Polish, and Turkish.

The rest of the paper is organized as follows. Section 3 provides a brief review of related work in this area. Section 4 introduces six types of ambiguity. Section 5 describes the tools and corpora used. Section 6 contains the results of our experiments. Section 7 provides a discussion of the results and some ideas about their usefulness.

3. Related work

Most of the papers that deal with the problem of morphological ambiguity usually give information on a single language. In most cases, the topic of such research is an individual issue such as grammatical case distribution in texts of a given genre (Kopotev, 2008; Lyashevskaya, 2013) or development of a frequency dictionary for a given language (Bolshakov et al., 2002). The comparison of ambiguity in different languages is performed using empirical methods based only on the researcher's intuition. However, statistical language processing needs numerical evaluation of phenomena. For example, Fabricz (1986) demonstrates the influence of POS-ambiguity on the performance of machine translation systems, Krovetz (1997) shows that the resolution of POS-ambiguity improves the information retrieval quality.

Some statistical information on individual languages is reported as a motivation for the development of language-specific tools. For example, Hajič (1998) gives information on part-of-speech ambiguity in Czech. The authors introduce the term "ambiguity class", which is used to represent the set of ambiguous categories of one word form, for example, a word like *process*, which can be a noun as well as a verb or an adjective, is assigned to the ambiguity class POS_{NVA}. The notion of ambiguity class is further used in some works on development of morphological taggers with disambiguation. Such works contain statistical data as a starting point of their research. Some statistical data (the amount of ambiguous tokens, the average number of tags per token) is given by Oravecz and Dienes (2002) for Hungarian and English, and by Tufiş (2000) for Romanian. Pinnis and Goba (2011) also consider ambiguity classes, they study their size and properties for Estonian, Latvian, and Lithuanian words.

Some works on discourse give another type of statistical information, for example the parts of speech distribution in texts (Li, 2012). Since this data can vary significantly throughout different genres or even for different authors, it is used as a basis for further research. A brief comparison of ambiguity in the Russian and English languages can be found in work by Klyshinsky et al. (2013), however, the comparison of two languages is not enough for a conclusion. Therefore, in the current research we include more languages from different language groups, and give a formal description of ambiguity classes. The PhD thesis of Pertsova (2007) contains background for further research on ambiguity patterns in different languages. Unfortunately, the author does not provide enough numerical results for comparison using the information from

the abovementioned WALS project instead. By and large, the existing research does not give any scale for comparison of the level and type of ambiguity in different languages. Our research is intended to bridge this gap.

4. Method of analysis

Let us represent text T as a sequence of tokens from the vocabulary *V:* $T = <w_1, w_2, \ldots, w_n>$, where $w_i \in V$ is a word of the text. Note that the vocabulary *V* contains only word tokens, i.e. we do not consider punctuation marks and other non-word tokens like numbers.

The morphological analysis of a word token *w* includes its lemma, its part of speech, and its grammatical features. The list of grammatical features varies depending on the language and the word's part of speech. Let us define a word form *v* as a tuple $v=<l, \pi, \mu>$, where *l* is the lemma of this word form, π is its part of speech (POS) and μ is its set of grammatical features. The result of morphological analysis of a word *w* is a set of word forms $\varphi(w) = \{v_1, v_2, \ldots v_k\}$, where each v_i is a distinct word form. If the word *w* is not present in the vocabulary, $\varphi(w) = \emptyset$ (i.e. $k = 0$), otherwise $k > 0$. If $\varphi(w)$ contains more than one word form ($k > 1$), the word *w* is ambiguous. Many NLP applications that use the results of POS-tagging need this ambiguity to be resolved as they need every word to belong to a particular part of speech and have only one set of features. However, some types of ambiguity are easier to resolve than others, and some tasks allow the use of ambiguous results of morphological analysis.

We distinguish six types of ambiguity depending on what parameters of word forms differ. The description of the types is given in Table 1.

Let us consider the ambiguity types in more detail. All of them are technically equal in the sense that if a word belongs to any of the ambiguity types, we have more than one possible analysis for it and therefore cannot be sure about its part-of-speech tag and morphological features. However, we have an intuition that some ambiguity types are easier to deal with than others.

For example, if a word is **ambiguous only by features** we already have the correct information on its part of speech. Without considering contextual information we still would not be able to identify its role in the sentence, but this role is already bounded to some subset of roles applicable to the given part of speech. In fact, many NLP tasks that consider bag-of-words models (i.e. information retrieval, sentiment analysis) do not care about the morphological information. Part-of-speech information can be important for them, because it can change the word's meaning, while noun case or verb tense are not always meaningful features. Thus, some applications could make use of the ambiguous analysis of a word.

On the other hand, **POS-ambiguous** words give us no clue about their role in any syntactic or semantic structures that could be present in a sentence. In order to find any information about them we should apply some homonymy resolution rules to the ambiguous cases. These rules are usually extracted from a labelled dataset, but they are likely to make many errors unless the labelled dataset is very big. The majority of languages do not have such resources.

Table 1. Ambiguity Classes. POS = part of speech, LEM = lemma, GF = grammatical features. 0 means that the corresponding parameter matches for different variants of morphological analysis, 1 means that it has multiple values

	POS	**LEM**	**GF**	**Description**
1	0	0	0	**Unambiguous** (word has one possible analysis)
2	0	0	1	**Ambiguous by features**: A word has coinciding forms with different grammatical features. **Example:** The German verb 'wohnen' ('to live') has the same form 'wohnen' for the infinitive, the present tense 1st person plural, present tense 3rd person plural, present tense 2nd person polite form.
3	1	0	–	**Ambiguous by part of speech**: A word has coinciding forms with different parts of speech but the same lemma; features are not comparable. **Example:** The English word 'close' can be a noun, a verb, or an adjective.
4	0	1	0/1	**Ambiguous by lemma:** Different words of the same part of speech that occasionally have coinciding form; features may be the same or not. **Examples with same features:** The Russian word 'смели' [smeli] can be a form of the verb 'сметь' [smet'] ('dare') or a form of the verb 'смести' [smesti] ('wipe off'). In both cases the verb is in the form of past tense, plural, 3rd person. **Example with different features: T**he Russian word 'вина' [vina] can be the noun 'вина' [vina] ('guilt') in singular, nominative case, or the noun 'вино' [vino] ('wine') in plural, nominative or accusative cases
5	1	1	–	**Ambiguous by part of speech and lemma**: This class includes different words of different parts of speech that occasionally have coinciding forms; features are not comparable. **Example:** The French 'est' can be the noun 'est' ('east') or the verb 'être' ('to be'), in present tense, singular, 3rd person.
6	–	–	–	**Out-of-vocabulary words**: Not contained in the vocabulary of the tagger.

5. Data and tools

The analysis has been performed for the following languages: English, French, Spanish, Portuguese, Italian, German, Russian, Polish, and Turkish. For this purpose we used the existing POS-taggers (see Table 2). We did not predict the tags and grammatical features for unknown words, they were put into the **out-of-vocabulary words** class.

Previous experiments have shown that the result depends on the style of the analyzed corpora (see (Klyshinsky et al., 2013)). In order to eliminate the influence of

text style, the experiments for all languages were conducted on texts from the same domain, namely news texts. Table 2 outlines the statistics of the corpora used.

In order to make the cross-lingual comparison more accurate and eliminate the differences between texts we also performed the comparison on a multilingual parallel corpus. We used the News Commentary corpus issued for the Workshop on Statistical Machine Translation for English, French, German, and Spanish.[2]

Every POS-tagger we used has a language-specific tagset. We examined these tagsets and found that they differ significantly. For uniformity reasons, we normalized all tag sets reducing all syntactic features or moving grammatical properties from the tag to the set of grammatical features. For some languages (e.g. Spanish and Italian) the normalized tagsets do not differ from the initial ones. Tagsets for several languages are presented in Table 3. The second row shows the number of grammatical features used in the language. For the detailed description of tagsets, see the documentation of the listed tools.

Table 2. POS-taggers and corpora.

Language	**POS-tagger**	**Vocabulary size (lexemes)**	**Corpus**	**Corpus size (words)**
English	Extended AOT vocabulary[3]	105 000	Reuters	300 mln
French	Morphalu[4]	68 000	Le Parisien	43.1 mln
Spanish	FreeLing[5]	76 000	Abc.es	15.2 mln
Italian	FreeLing	40 000	Corriere dela Serra	9.1 mln
German	FreeLing	155 000	Die Zeit	8.5 mln
Russian	Extended AOT vocabulary	167 000	Lenta.ru	32.4 mln
Polish	Morfologik[6]	> 400 000	Different sources	21.2 mln
Portuguese	Freeling	110 000	Expresso	41.6 mln
Turkish	Extended AOT vocabulary	64 000	Haberler.com	28.6 mln
Finnish	en.wiktionary.org	69500	Helsingin Sanomat	10.5 mln

6. Experiments

6.1. Distribution across homonymy types

We performed a series of experiments for the chosen languages. The results of these experiments are presented in Table 4 in numerical form and in Figure 1 in visual form. The proportion of words with a single morphological analysis variant ranges from 30% to 50% for the majority of languages with Polish showing a much lower rate of unambiguous words of 20%. Four highly inflected languages (Russian, Polish,

[2] http://statmt.org/wmt13/training-parallel-nc-v8.tgz, about 4 mln word tokens.
[3] http://aot.ru (Sokirko and Toldova, 2004)
[4] http://www.cnrtl.fr/lexiques/morphalou/
[5] http://devel.cpl.upc.edu/freeling/downloads?order=time&desc=1 (Padró and Stanilovsky, 2012)
[6] http://morfologik.blogspot.ru/2013/02/morfologik-20-rc2.html

German and Turkish) demonstrate a high level of ambiguity by grammatical features (about 25-40% versus 0-5% in other languages). The class distribution for the Italian and Spanish languages are quite close and demonstrate a high level of unambiguous, POS-ambiguous and POS- and lemma-ambiguous words and a low level of grammatical feature ambiguity. The distributions for English and French have different shapes in comparison with other languages. The English language demonstrates an extremely high percentage of POS-ambiguous words (about 50% versus 5-25% in other languages). Finally, the Finnish language demonstrates extremely high amount of out-of-vocabulary words because of high rate of compounds in the text.

We also compared the distributions of words among ambiguity types in the monolingual corpora, where all the texts were originally written in the language, and corresponding parts of parallel corpora, where texts were presumably translated from English. French and Spanish show the same distributions in monolingual and parallel corpora. However, the analysis of these differences is beyond the scope of our research.

6.2. POS-tagging without homonymy resolution

In section 4 we stated that we could make use of ambiguous POS-tagging for some of words. The described experiments allow us to estimate the percentage of the tagging outputs that could be used with no homonymy resolution.

Table 3. Tag sets for different languages

Russian	**English**	**Spanish**	**Italian**	**German**	**French**	**Polish**	**Parts of Speech**
12	9	19	10	4	5	12	
adj	adj	ADJ	ADJ	Adjektive	adjective	adjective	adjective
adv	adv	ADV	ADV	Adverbien	adverb	adverb	adverb
conj	conj	CONJ	CONJ	Konjunktionen		conjuction	conjuction
	article	DET	DET	Artikel	functionWord		determiner
interj	int	INTERJ	INTERJ	Interjektionen	interjection	interjection	interjection
noun	noun	NOUN	NOUN	Nomina	commonNoun	noun	noun
prep	prep	PREP	PREP	Adpositionen	functionWord	preposition	preposition
pers_pron	pronoun	PRON	PRON	Pronomina		pronoun	pronoun
verb	verb	VERB	VERB	Verben	verb	verb	verb
particle	part			Partikeln		particle	particle
card_num	card			Kardinal-zahlen		numeral	cardinal number
deepr							adverbial participle
dem_pron	pn_dem						demonstrative pronoun
ord_num	ord						ordinal number
participle						participle	participle
poss_pron	pn_poss						possessive pronoun

Table 4. The ambiguity distribution for different languages

	Unambiguous	Ambiguous by features	POS ambiguous	Ambiguous by lemma and features	POS and lemma ambiguous	Out-of-vocabulary
Russian	48.28%	27.68%	5.26%	4.67%	9.92%	4.38%
Polish	18.94%	39.13%	13.49%	7.23%	18.44%	2.78%
English	38.87%	2.79%	50.35%	0.32%	0.69%	7.65%
English NC	39,77%	3,17%	46,78%	0,30%	6,88%	3,05%
Italian	41.66%	0.14%	13.36%	0.98%	23.10%	20.77%
French	51.21%	3.69%	7.40%	7.96%	19.59%	10.15%
French NC	54.54%	2.80%	8.21%	9.54%	19,83%	5.08%
Portuguese	39.14%	5.80%	16.78%	0.95%	25.16%	12.17%
Spanish	33.00%	0.09%	21.48%	0.47%	28.97%	15.98%
Spanish NC	35.53%	0.05%	21.51%	0.33%	29.65%	12.93%
German	13,39%	25,54%	21,66%	1,53%	24,89%	12,99%
German NC	14,24%	25,71%	23,55%	1,08%	29,69%	5,73%
Turkish	28.26%	27.10%	7.19%	2.03%	16.45%	18.98%
Finnish	45,82%	0,42%	2,82%	1,56%	2,87%	46,50%

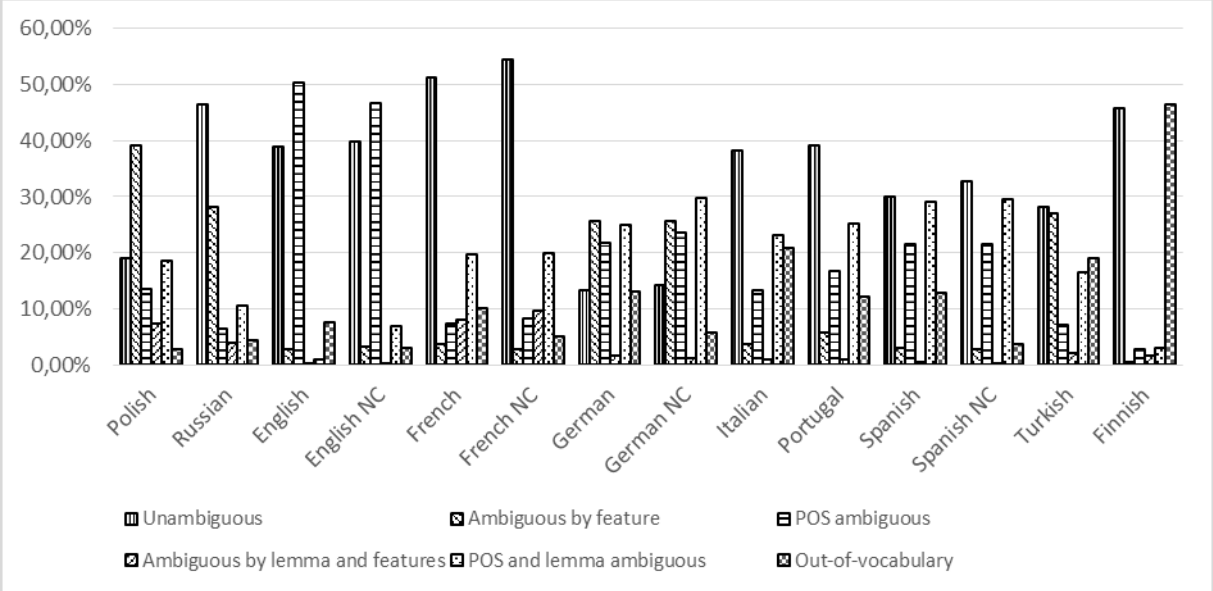

Figure 1. The ambiguity distribution for different languages.

As it was explained, some NLP applications can benefit from partial outputs of POS-taggers, if this partial output contains part-of-speech tags. In order to be included into the partial output a POS-tag has to be defined unambiguously. This is possible for words which are unambiguous or POS-unambiguous. The number of such words in a corpus can be defined as the sum of **unambiguous** words, words which are **ambiguous by features** or **ambiguous by lemma**. Figure 2 demonstrates the percentage of words which can be used without disambiguation throughout the considered corpora

(shown by solid polygonal curve). Note that the languages with larger number of grammatical features have larger percentage of such words.

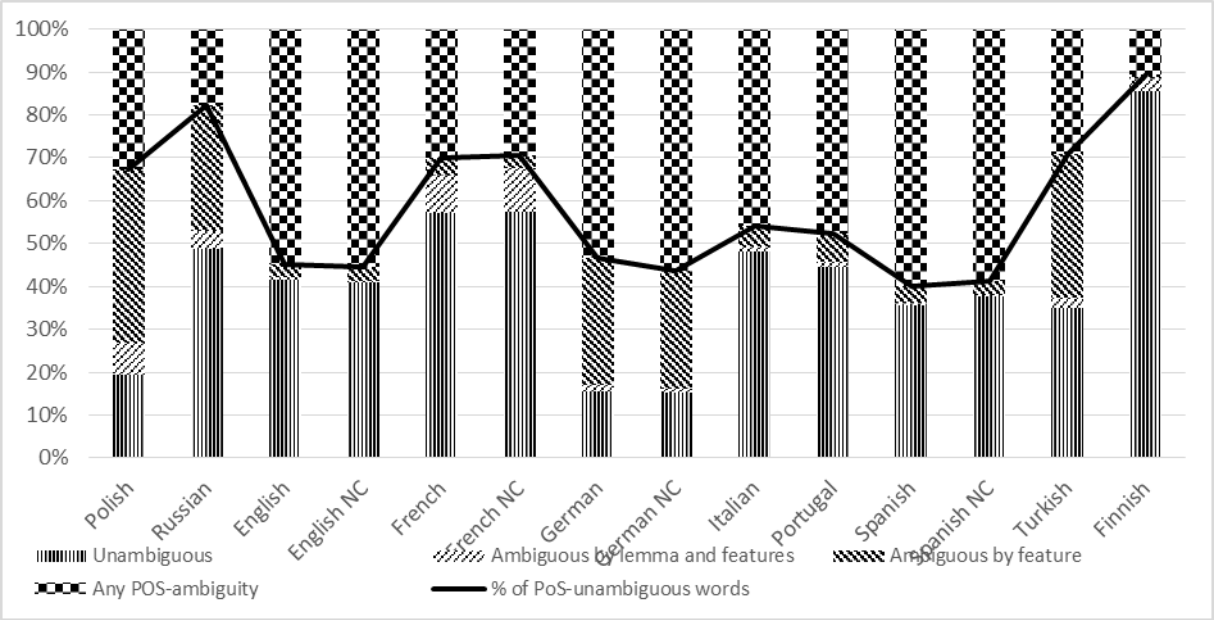

Figure 2. Percentage of words can be used without POS disambiguation.

6.3. Additional experiments

It is worth noticing that the resources and tools we used for different languages are extremely heterogeneous: the sizes of vocabularies and the number of tags alter from one language to another. That means that the discovered differences in the ambiguity distribution could be explained by the diversity of resources, and not by properties of languages. In order to exclude the dependence of results on various parameters of the data and POS-taggers we conducted a series of experiments.

The first hypothesis we need to reject is that the differences among the distributions for the considered languages are explained by the varying size of vocabulary. In order to check if this is true or not we showed how the distribution of words among the ambiguity classes changes when changing the size of the vocabulary. The experiments were conducted for Russian, English, French, Italian, Portuguese, and Spanish. We sorted all the words in the POS-taggers' vocabularies in descending order of frequency in the corpus. Then we computed distributions of words using the information about only a fraction of the vocabulary. We experimented with the most frequent 1000, 3000, and 5000 words. The graph in Figure 3 shows that the changing size of vocabulary does not affect the main trends of word distributions, i.e. Russian still has relatively high percentage of words that are ambiguous by grammatical features, while the most common ambiguity type in English is POS ambiguity. Thus, we can claim that the most frequent ambiguity types that differ significantly for different languages are the properties of those languages.

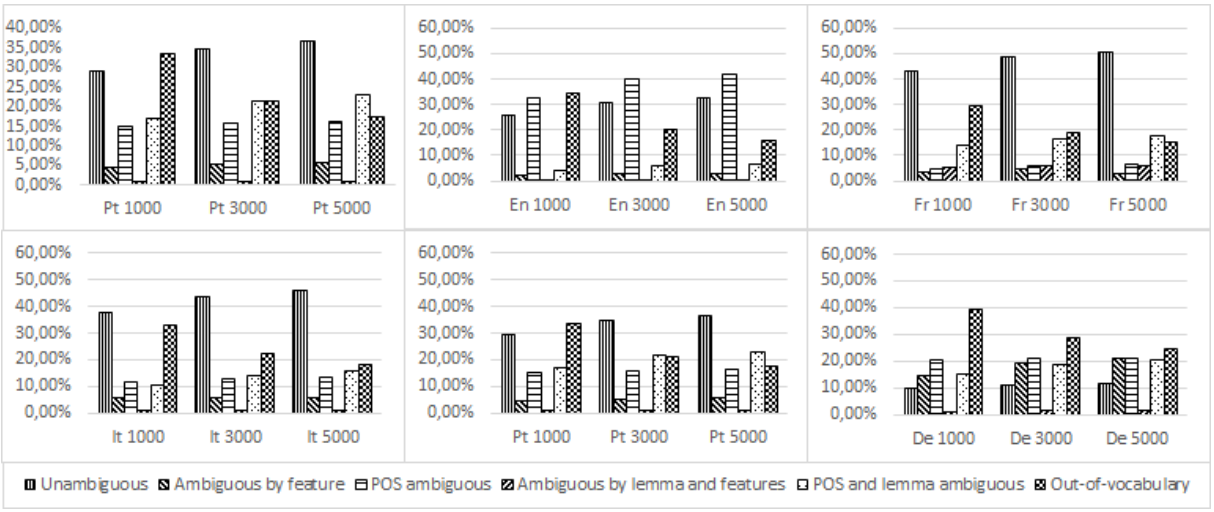

Figure 3. Influence of the lexical vocabulary size on resulting distribution.

We also experimented separately with every thousand out of the most frequent ten thousand words for Russian, Polish, Turkish, French, English, Spanish, Italian, and Portuguese. The graph in Figure 4 shows that the changing frequency of words influences the resulting distribution for all selected languages. The core vocabulary of a language (the most frequent words, denoted by the left bar for every language) has more PoS-ambiguous words than less frequent subset of vocabulary. The French core vocabulary contains more words which are ambiguous by lemma and features. While in Russian, Polish, and Turkish the amount of unambiguous words keeps constant, the rest of languages increase this number for less frequent words. The Finnish language demonstrates the extremely low rate of ambiguous words for every interval.

We also checked if the distribution was affected by the tag set used. In order to dispose of this assumption we repeated the experiment eliminating the differences between Russian and English tag sets, namely, replacing all personal pronouns with nouns, all possessive pronouns with adjectives and all gerunds and participles with verb in the Russian tag set. The changes in the resulting distributions were around 0.1% because quite many pronouns in the Russian language have a unique lemma and all gerunds and participles have specific inflections.

We examined the dependence of the distribution on the grammatical feature list: does the size of grammatical features inventory influence the distribution? In order to check this we sequentially removed such features as animacy, case, person and number from the results of morphological analysis for Russian. The deletion of the animacy feature changes numbers by 0.005%, the deletion of the person feature does not change the distribution at all. The deletion of a feature leads to the decreased percentage of ambiguity by grammatical features and the increased number of unambiguous words. The results are shown in Figure 5. The distributions for English and Polish are added there for comparison: we can see that despite the reduction of grammatical features the distributions for Russian do not come closer to any distributions for other languages.

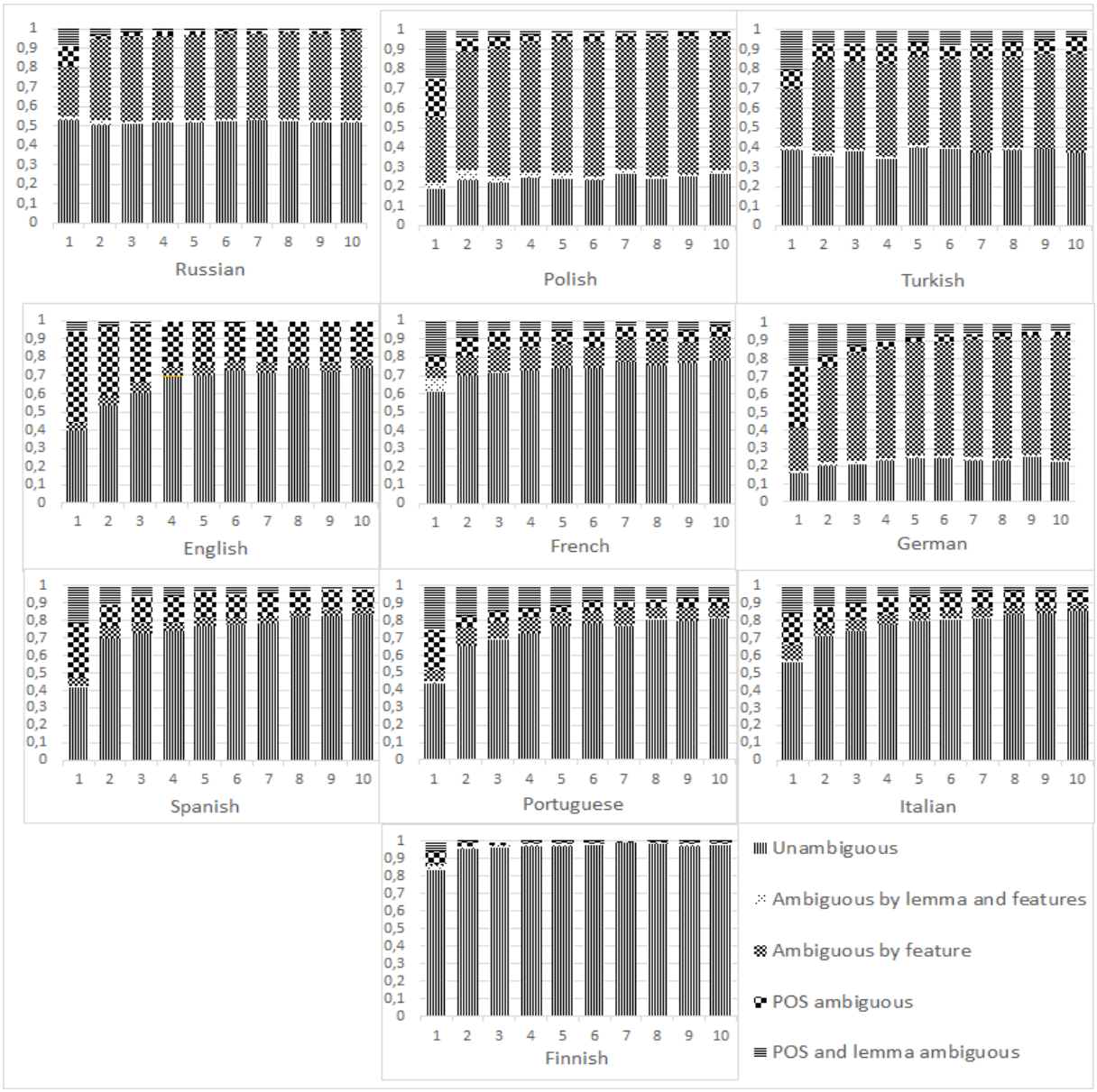

Figure 4. Influence of the words frequency on resulting distribution.

Finally, we examined how the distribution depends on the tagger we use. Different taggers can vary in the size of vocabulary, tag set and core principles of filling the vocabulary. We changed FreeLing dictionary to TreeTagger for the German language. The TreeTagger's vocabulary is shorter than the FreeLing's one. In addition, the TreeTagger German tag set contains no grammatical features, excluding the ones included in PoS tag. For example, TreeTagger tag set contains two tags for the adjective depending on its role in a sentence (ADJA for attributive adjective and ADJD for predicative ones). By contrast, the FreeLing German tag set contains only one tag for adjectives (adj) and other properties are reflected by the grammatical features of the word. For uniformity reasons, we considered two TreeTagger tags (e.g., ADJA and ADJD) as one (ADJ) that has one extra grammatical feature.

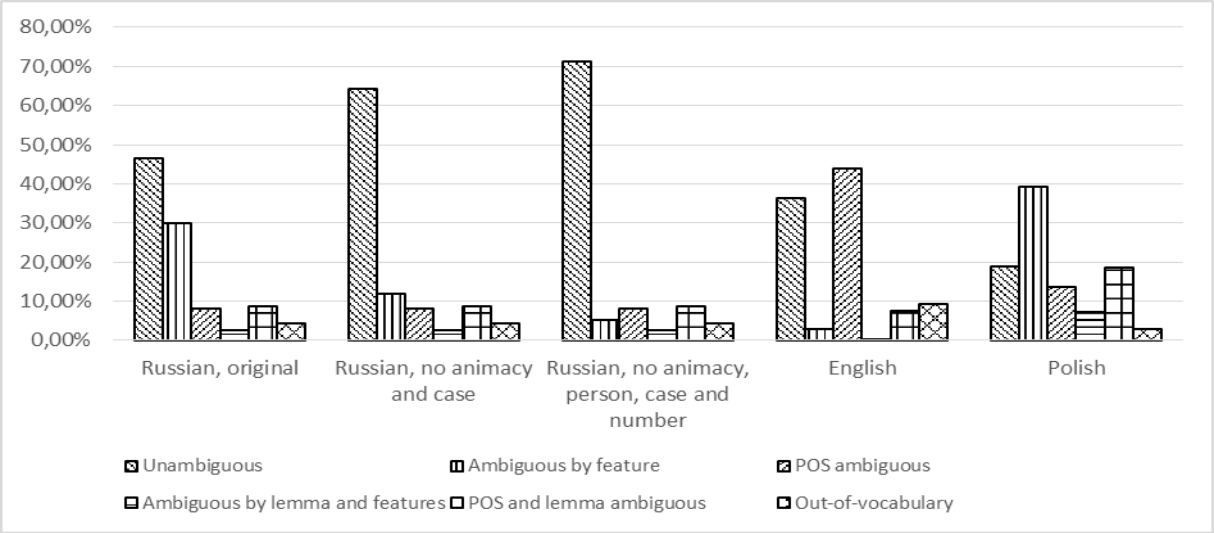

Figure 5. Influence of the set of grammatical features on resulting distribution.

We compared the results calculated for Die Zeit and the News Commentary corpus; the results are shown in Table 5 in numerical form and in Figure 6 in visual form. The difference between the distributions is crucial. However, it can be explained by the combination of three factors: reduction of the set of grammatical features, decrease of the vocabulary, and (probably) the fact that the authors of TreeTagger tried to eliminate words that are ambiguous by PoS and lemma. As it was shown before, reducing the set of grammatical feature decreases the number of words that are ambiguous by features and increases the number of unambiguous words. The rates of PoS-ambiguous word are comparable. Finally, words that are ambiguous by PoS and lemma were not included in the vocabulary.

6.4. Experiments: conclusion

We would like to note that every analyzed parameter influences the distribution among the types of ambiguity. Nevertheless, none of the tried variations leads to any significant changes in the shape of distribution: none of the distributions became similar to another one. The Russian language has larger amount of unambiguous tokens than English in any setting, while the English language has a large percentage of POS-ambiguous tokens, although the sizes of the vocabularies for English, Russian, German, and Portuguese are comparable. We should also note that the reduction of grammatical features reduces the size of **ambiguous by features** class in highly-inflected languages (Russian and German) when the parameters that create the ambiguity are not considered. However, we cannot tell if there are any regularities in such shifts. The research on this topic is left for future work.

The represented analysis is not comprehensive. Moreover, the figures we report were calculated using texts of one style, namely the news, since the previous research has shown that the distribution differs for texts of other styles (see (Klyshinsky et al., 2013)). On the other hand, our experiments showed that the size of vocabulary does not influence the distributions. Furthermore, we can state that the 5000 most frequent words define the shape of the distribution so that less frequent words are unable to change it significantly. We can also state that the core vocabularies of all considered

languages have more PoS-ambiguous words that the rest of the language. The research on this topic is left for future work as well.

Table 5. The ambiguity distribution for different vocabularies in German.

	Unambig.	**Ambig. by features**	**POS ambig.**	**Ambig. by lemma and features**	**POS and lemma ambig.**	**Out-of-vocabulary**
German FreeLing	13,39%	25,54%	21,66%	1,53%	24,89%	12,99%
German TreeTagger	33.27%	4.51%	22.40%	1.36%	13.07%	25.39%
German NC FreeLing	14,24%	25,71%	23,55%	1,08%	29,69%	5,73%
German NC TreeTagger	44.53%	9.35%	23.91%	0.91%	5.26%	16.03%

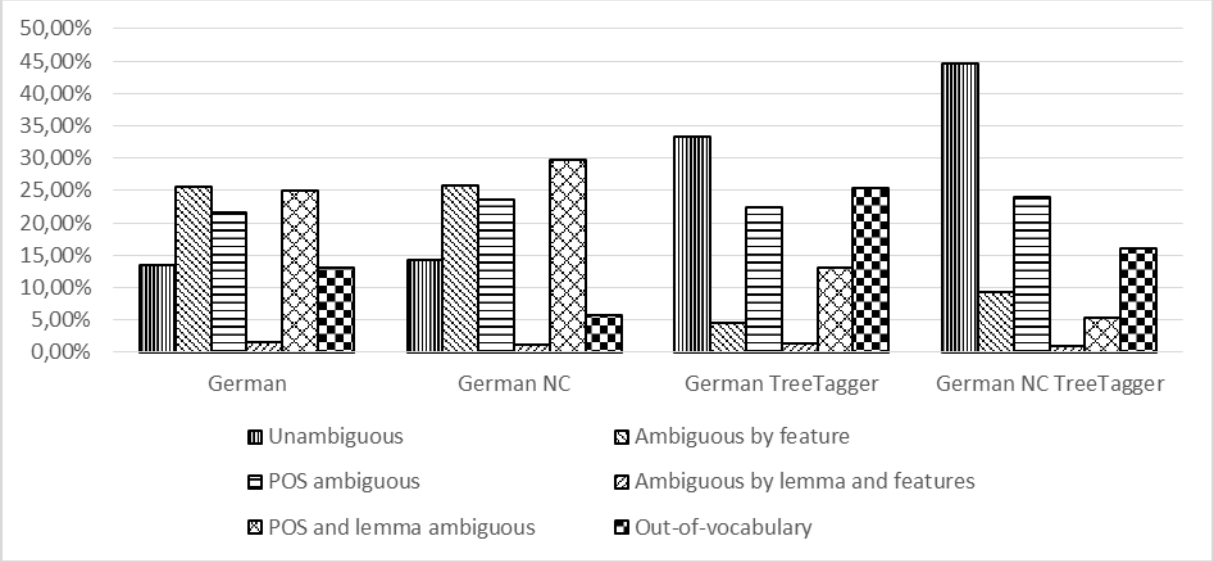

Figure 6. Influence of the selected dictionary on resulting distribution.

The shape of distribution is dependent on the vocabulary at hand. Depending on the task, researchers use one of the two approaches: eliminating grammatical features or using a complete tag set. Choosing one of them will influence the resulting distribution of words among the types of ambiguity.

7. Discussion and future work

The results of our experiments show some significant differences between the considered languages. The reasons behind these differences are a topic for another research and we will not analyze them here. However, let us just look at the resulting numbers. We have not found any correlation between the distribution of words among ambiguity types and the number of grammatical features (excluding the fact that the number of unambiguous word tokens is increasing while we are cutting grammatical features). Likewise, we did not see any correlation among distributions while we were varying the size of corpora, tag sets or vocabularies. Thus, we can conclude that the

distribution of words among the types of ambiguity is an inherent property of the considered languages.

We drew an important conclusion on the usefulness of ambiguous POS-tagger output: if the part of speech itself is defined correctly, we can use this fact during shallow parsing or words disambiguation. Our experiments demonstrate that considering the rank of a word or distribution among types of ambiguity allows to slightly increase the precision of disambiguation for POS-tagging and morphological analysis (see (Rysakov, 2015)). It has been noticed in earlier works (Hajič, 1998, Pinnis, 2011) that there are sets of part-of-speech tags that often create ambiguities: e.g. there is a class of words that can be classified as nouns as well as verbs, but nouns are quite rarely confused with prepositions. These ambiguity classes are apparently language-specific: adjectives can be often confused with nouns in English, but there are less examples of such ambiguity in Russian.

Our results allow evaluating differences between languages, their vocabularies and characteristics of words' ambiguity. The introduced methods provide a numerical evaluation of such properties and helps understand limitations of modern POS-tagging methods.

ACKNOWLEDGEMENTS

The authors would like to thank Elena Iagunova for fruitful discussions.

References

Bolshakov, Igor; Galicia-Haro, Sofia & Gelbukh, Alexander (2002). Quantitative Comparison of Homonymy in Spanish EuroWordNet and Traditional Dictionaries. In: *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 2276*.

Brants, Thorsten (2000). TnT - a statistical part-of-speech tagger. In: *Proc. of 6th Applied Natural Language Processing Conference* (pp. 224 – 231). Seattle.

Brill, Eric (1995). Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. In: *Proc. of the Third Workshop on Very Large Corpora*. Cambridge, MA, USA.

Fábricz, Karoly (1986). Particle Homonymy and Machine Translation. In: *Proc. of International Conference on Computational Linguistics* (pp. 59 – 61).

Hajič, Jan & Vidová-Hladká, Barbora (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: *Proc. of the COLING-ACL Conference* (pp. 483 – 490). Montreal, Canada.

Klyshinsky, Eduard; Kochetkova, Natalia; Litvinov, Maxim & Vadim Maximov (2011). Method of POS-disambiguation Using Information about Words Co-occurrence (For Russian). In: *Proc. of GSCL'2011* (pp. 191 – 195) University of Hamburg, Germany, September 2011.

Klyshinsky, Eduard; Kochetkova, Natalia; Mansurova, Oksana; Iagunova, Elena; Maximov, Vadim & Karpik, Olesya (2013). *Формирование модели сочетаемости слов русского языка и исследование ее свойств* [Development of Russian Subcategorization Frames and its Properties Investigation]. Preprints of

Keldysh IAM #41. Moscow: Keldysh IAM. Retrieved from http://library.keldysh.ru/preprint.asp?id=2013-41

Kopotev, Mikhail (2008). *К построению частотной грамматики русского языка: падежная система по корпусным данным* [Towards the frequency grammar of Russian: corpus evidence on the grammatical case system] In: Mustayoki A.; Kopotev M.V.; Birjulin L.A. & Protasova E.Ju. (eds.), Инструментарий русистики: корпусные подходы [Instruments of Russian linguistics: corpus approach] (pp. 136 – 150). Helsinki.

Krovetz, Robert (1997). Homonymy and Polysemy in Information Retrieval. In: *Proc. of EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics archive* (pp. 72 – 79).

Li, Ying; Yu, Yue & Fung, Pascale (2012). A Mandarin-English Code-Switching Corpus. In: *Proc. of Eighth International Conference on Language Resources and Evaluation (LREC-2012).*

Lyashevskaya, Olga (2013). *Frequency Dictionary of Inflectional Paradigms: Core Russian Vocabulary*. Preprints of HSE, Series: Humanity, WP BRP 35/HUM/2013. Retrieved from http://www.hse.ru/data/2013/06/27/1285976210/35HUM2013.pdf

Oravecz, Csaba & Dienes, Peter (2002). Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: *Proc. of Third Int. Conf. on Language Resources and Evaluation (LREC'02).*

Padró, Lluis & Stanilovsky, Evgeny (2012). FreeLing 3.0: Towards Wider Multilinguality. In: *Proc. of the Language Resources and Evaluation Conference (LREC'12)* ELRA. Istanbul, Turkey.

Pertsova, Katya (2007). *Learning Form-Meaning Mappings in Presence of Homonymy: a linguistically motivated model of learning inflection* (PhD Thesis in Linguistics). Retrieved from http://linguistics.ucla.edu/people/grads/pertsova/pertsovaThesis.pdf

Pinnis, Marcis & Goba, Karlis (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In: *Proc. of Systems and Frameworks for Computational Morphology - Second International Workshop (SFCM 2011)*. Zurich, Switzerland.

Protopopova, Ekaterina & Bocharov, Victor (2013). Unsupervised learning of part-of-speech disambiguation rules. In: *Proc. of Computational Linguistics and Intellectual Technologies (Dialog 2013)*. Bekasovo, Russia.

Rysakov, Svyatoslav (2015). How to kill homonymy? In: *System Administrator Magazine*, October 2015, pp. 92-95.

Sharoff, Serge & Nivre, Joakim (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: *Proc. of Computational Linguistics and Intellectual Technologies (Dialog 2011)*. Bekasovo, Russia.

Schmid, Helmut (1995). Improvements in Part-of-Speech Tagging with an Application to German. In: *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Sokirko, Aleksey & Toldova, Svetlana (2004). *Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка.* [Comparing a stochastic tagger based on Hidden Markov Model

with a Rule-based tagger for Russian] In: *Proc. of Corpus Linguistics* – 2004, Saint-Petersburg.

Tufiş, Dan (2000). Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In: *Proceedings of Second International Conference on Language Resources and Evaluation*. Athens.

Collaborative Writing of Otome no minato*

Hao Sun[1], *Mingzhe Jin*[2]
[1]*Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan*
[2]*Faculty of Culture and Information Science, Doshisha University, Kyoto, Japan*

Abstract
Otome no minato, a novel published under the name of Yasunari Kawabata, was suspected to have been written by the ghostwriter Tsuneko Nakazato. In this paper, we identify the actual author of *Otome no minato* by means of authorship attribution. Part of speech (POS) bi-grams, particle bi-grams, and phrase patterns are used as stylometric features. Correspondence analysis (CA), hierarchical cluster analysis (HCA), Adaptive Boosting (AdaBoost), High-Dimensional Discriminant Analysis (HDDA), Logic Model Tree (LMT), Random Forest (RF), and Support Vector Machine (SVM) are used as classifiers. Our results suggested that it is highly possible that *Otome no minato* was collaboratively written by Yasunari Kawabata and Tsuneko Nakazato.

Keywords: *Yasunari Kawabata; ghostwriter; authorship attribution; machine learning*

1. Introduction

Yasunari Kawabata (1899-1972) was a famous Japanese novelist who won the Nobel Prize in Literature in 1968. Representative works by him include *Snow Country*, *The Old Capital*, *The Sound of the Mountain*, and *House of the Sleeping Beauties*. Kawabata suffered from anxiety after the loss of all his close relatives. He began to take sleeping pills in large quantities after his anxiety worsened in the 1960s, and was hospitalized due to his sleeping pill addiction in 1962. Tragically, Kawabata chose to end his life in 1972.

It has been suspected that some of Kawabata's novels were written by ghost-writers. It was suspected that *Otome no minato*, a novel, was written by Tsuneko Nakazato even though it is included in Kawabata's newest collected writings, published in 1984. The possible ghostwriter, Tsuneko Nakazato was a famous Japanese female novelist who received one of the most important prizes, the Akutagawa Prize, in 1939.

To date, some compelling evidence has suggested that *Otome no minato* was not written by Yasunari Kawabata. The most powerful evidence is correspondence concerning instructions about the writing of *Otome no minato* between Kawabata and Nakazato. In a letter from Nakazato to Kawabata, Nakazato wrote: 'Nowadays readers buy *Syojyo no tomo* because *Otome no minato* is serialized in it. It is my pleasure to

* Address correspondence to: Hao Sun, ph.D, Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan, 1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu 610-0394. Email: Sonnkou1985@gmail.com

write this novel'. In his reply, Kawabata wrote to Nakazato: 'I have received the manuscript of (*Otome no minato*)'. Atsushi Koyano, a writer and literary critic, mentioned in his book that *Otome no minato* could have been written by Nakazato (Atsushi, 2013). However, the above evidence only suggests that Nakazato may have been one of the writers of *Otome no minato*, rather than the only one. Unlike previous studies, we consider the ghostwriter problem of *Otome no minato* as a collaborative one. That is, Kawabata could have revised some parts of the manuscript because he was the writing adviser to Nakazato. The key point of the *Otome no minato* ghostwriter problem is the degree of revision carried out by Kawabata. This novel should be recognized as written by Nakazato if the manuscript was not rewritten by Kawabata at all. Nonetheless, it should be treated as Kawabata's work if he rewrote all the chapters of *Otome no minato*. Another possibility is that Kawabata rewrote some parts of *Otome no minato*. In this case, we need to decide which parts were rewritten to decide the attribution of the novel. However, we do not know which parts of *Otome no minato* belong to whom because the novel has never been discussed in parts. In this paper, we chunk *Otome no minato* into ten categories because it is naturally divided into ten chapters. We discuss the attribution problem of each chapter of *Otome no minato* in terms of authorship attribution.

The collaborative writing problem has attracted researchers' interest for a long time. Computer-assisted stylometric methods were first introduced to the well-known 'The Federalist Paper' case in 1964 (Mosteller and Wallace, 1964). Rybicki et al. (2014) described the collaborative authorship problem of Joseph Conrad and Ford Madox. They used most frequent words (MFW) as a stylometric feature and bootstrap consensus tree with Burrow's 'Delta' distance as attribution methods. Their results showed not only that the suspected work was written jointly by Joseph Conrad and Ford Madox Frd, but that Ford wrote a sizeable fragment of *Nostromo*. Gladwin et al. (2015) discussed the joint writing problem of *The Loved Dead*, which was contested by family members of Clifford Martin Eddy, Jr. and Sunand Tryambak Joshi. In their study, besides Burrow's 'Delta', principal component analysis (PCA) of function words was applied to discuss the collaborative writing problem. Kestemont et al. (2015) also applied PCA to analyze the collaborative authorship problem of Hildegard of Bingen and Guibert of Gembloux. Crombez and Cassiers (2017) adapted the traditional stylometric methods on a theatre script to discuss a collaborative writing problem.

We achieved two purposes which have not been discussed in previous studies. One is to verify whether *Otome no minato* was jointly written by Kawabata and Nakazato. The other is to find which parts the two authors are responsible for. We discuss the attribution of all ten chapters of *Otome no minato* as regards authorship attribution. Unlike all the previous studies mentioned above, we discuss the collaborative writing problem of *Otome no minato* by applying the integrated classification algorithm, which is an ensemble learning approach proposed to avoid the collision in classification results between each pair of stylometric features and classifiers (Jin, 2014).

This paper is organized as follows. First, in section two, we list the necessary corpora. Then, in section three, we describe the three stylometric features used in this study. In section four, we present the five classifiers applied in this study. In section

five, we introduce the results generated from the integrated classification algorithm. In section six, we conclude this study and suggest possible future works.

2. Corpus establishment

We established two corpora for the present study. One is the training corpus of 40 novels, 20 of which are selected from the collaborative writings of Kawabata and the other 20 from the collaborative writing of Nakazato. The other is the test corpus, which contains all ten chapters of *Otome no minato*. We excluded all the conversations in the corpora. All the novels written by Kawabata and Nakazato were published around 1937, the year in which *Otome no minato* was published. Tables 1-2 list all the selected novels of Kawabata and Nakazato.

Table 1. Novels by Yasunari Kawabata

Year of publication	Novel
1925	*Shiroimangetsu*
1926	*Izunoodoriko*
1926	*Bunkadaigakusow*
1927	*Harukeshiki*
1928	*Shishanosyo*
1929	*Onsenyado*
1930	*Haritogarasutokir*
1932	*Jyojyouka*
1933	*Kinjyu*
1935	*Yukiguni*
1940	*Hokuronotegami*
1940	*Tsubamenodojyo*
1940	*Onanoyume*
1940	*Fusyofuwa*
1940	*Yorunosaikoro*
1945	*Saikonsha*
1946	*Saikai*
1947	*Yume*
1948	*Sorihashi*
1949	*Amenohi*

Table 2. Novels by Tsuneko Nakazato

Year of publication	Novel
1932	*Houmatsu*
1932	*Roji*
1933	*Masuku*
1936	*Hanaama*
1936	*Jiyuga*
1937	*Shukufuku*
1937	*Fumimisubito*
1937	*Kegawa*
1937	*Seiyokan*
1937	*Hanabi*
1937	*Jyuka*
1938	*Morinonaka*
1938	*Nobara*
1938	*Noriaibasha*
1938	*Nikkoshitsu*
1939	*Tenkoku*
1939	*Bansankai*
1939	*Ajisai*
1939	*Kujyaku*
1940	*Rojyo*

3. Stylometric features

Stylometric features are features which contain information concerning personal writing style. Five categories of stylometric features, namely Lexical, Character, Syntactic, Semantic and Application-specific, exist in the proposed thousands of stylometric features (Rudman, 1998; Stamatatos, 2009). Among them, an n-gram profile was described to perform better than others in authorship attribution tasks (Grieve, 2007). In this study, we selected Part of speech (POS) bi-grams, particle bi-grams, and phrase patterns as stylometric features.

POS n-grams are a widely used stylometric feature in authorship attribution studies (Zhao & Zobel, 2007). This stylometric feature also achieved high accuracy in Japanese authorship attribution tasks (Jin, 2014). We used a Japanese POS tagger (MeCab) to split Japanese sentences into words and to add POS to the words. Two adjacent POS are extracted as a POS tag bi-grams feature.

Particles are mostly used POS in Japanese. Particle bi-grams are some of the most effective stylometric features in Japanese authorship attribution studies (Jin, 2002). In this paper, particle bi-grams refer to the sequence of two adjacent particles.

Phrases are the smallest units that a sentence can be divided into before the parts became unnatural. We used a Japanese parser (CaboCha) to separate Japanese sentences into phrases. A phrase pattern is a combination of the original form of the inherent particles and symbols and the POS of the other materials in the same phrase (Jin, 2013).

4. Experiments

We show the classification result of each chapter of *Otome no minato*, and the integrated classification result under the majority rule in this section.

4.1 Correspondence analysis

Firstly, we adopted correspondence analysis to POS bi-grams, particle bi-grams, and phrase patterns. Correspondence analysis is a well-known multivariate statistical analysis method which is conceptually similar to principle component analysis. Correspondence analysis applies categorical data, the frequency of which is often summarized in a two-way table or contingency table. The method computes a set of row scores and column scores, which is used to form a plot in two-dimensional space. The points that are close together on the plot indicate similar samples or variables relation. Riba and Ginebra (2005) employed correspondence analysis to discuss the authorship of Tirant lo Blanc (1460-1464). Correspondence analysis was also used to attribute the author of threating letters of the famous "Glico-Morinaga" criminal case occurred in Japan (Zaitsu and Jin, 2016).

The result of correspondence analysis on POS bi-grams is shown in Figure 1. The 20 novels of Kawabata and Nakazato are visualized with the initial letter 'K' and 'N', separately. The 10 chapters of *Otome no minato* are visualized with the initial letter 'O'. In Figure 1, we cannot perceive a clear boundary between the novels of Kawabata (K1~K20) and Nakazato (N1~N20). The ten chapters of *Otome no minato* (O1~O20) are positioned away from the Kawabata and Nakazato's group, which means *Otome no minato*'s writing style of is different from Kawaba and Nakazato's.

The result of correspondence analysis on particle bi-grams is shown in Figure 2. In Figure 2, Kawabata's novels (K1~K20) are concentrated at the top while Nakazato's novels (N1~N20) are positioned at the bottom. Ten chapters of *Otome no minato* (O1~O10) are plotted in the middle of Kawabata and Nakazato's novels.

The result of correspondence analysis on phrase patterns is shown in Figure 3. In Figure 3, novels of Kawabata and Nakazato are divided across the first dimension. Ten chapters of *Otome no minato* are plotted on the right side.

The analysis of correspondence analysis on POS bi-grams, particle bi-grams, and phrase patterns reflect that the style of *Otome no minato* is different from Kawabata and Nakazato's.

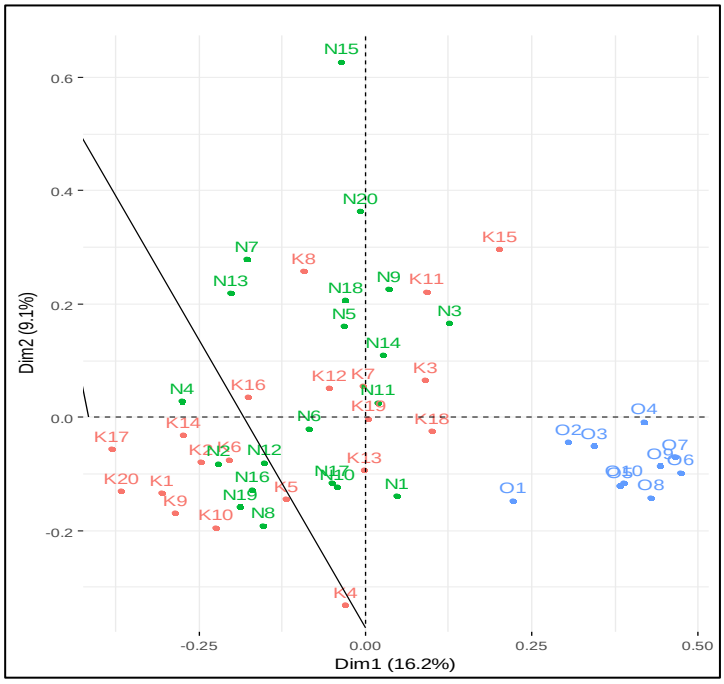

Figure 1. The biplot of correspondence analysis on POS bi-grams

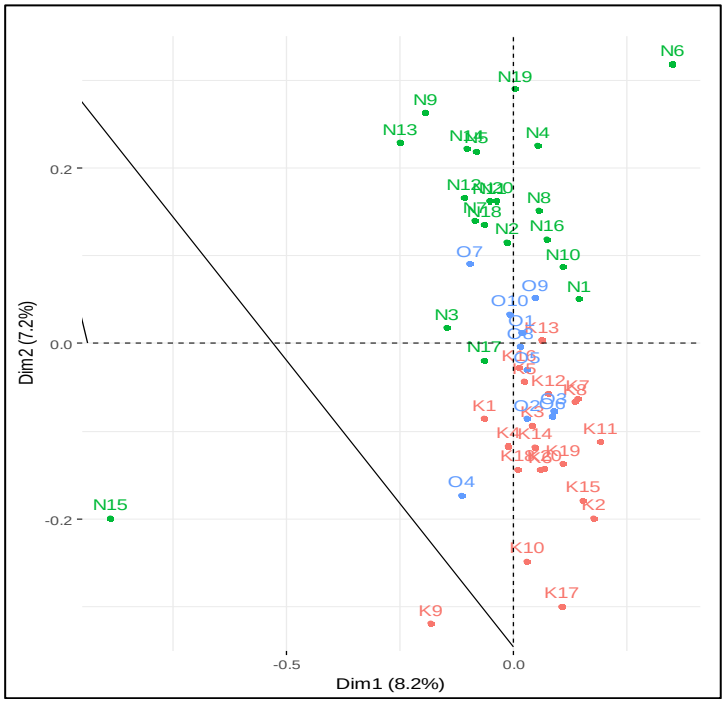

Figure 2. The biplot of correspondence analysis on particle bi-grams

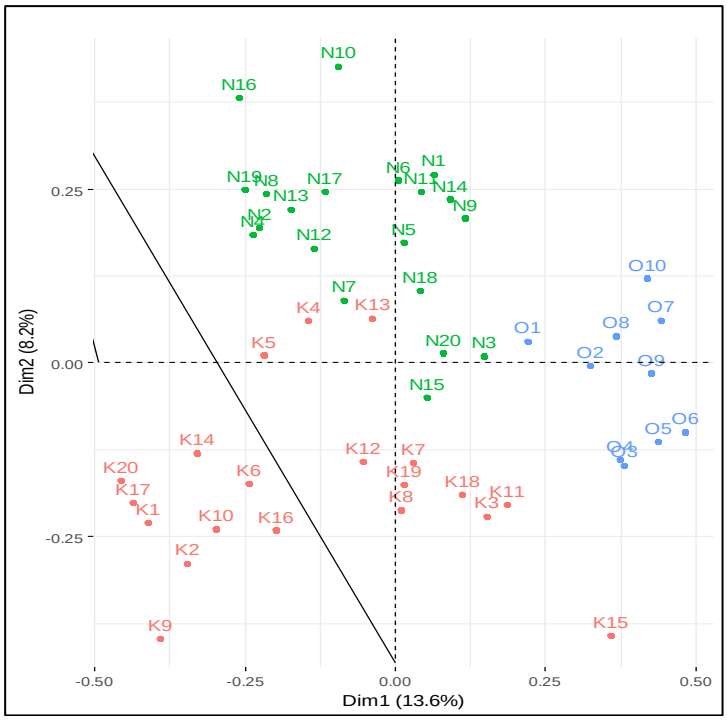

Figure 3. The biplot of correspondence analysis on phrase patterns

4.2 Hierarchical cluster analysis

Secondly, we apply hierarchal clustering analysis to POS bi-grams, particle bi-grams, and phrase patterns. Hierarchical cluster analysis is an unsupervised classification method, which combines similar samples into one cluster. The method generates a hierarchical tree structure called a dendrogram, which reflects the data grouping situation. Hierarchical cluster analyses have various applications in the authorship analysis field. Eder demonstrated that hierarchical cluster analysis is one of the powerful methods in stylometry visualization (Eder, 2017). Aljumily tried to resolve the "Shakespeare authorship question". The result of the hierarchical cluster analysis reveals that some disputed works which were claimed as Shakespeare's, seem to have not been written by him (Aljumily, 2015). We use ward's method and KLD distance for hierarchal clustering analysis. The formula for calculating KLD distance is shown below.

$$\mathrm{KLD} = \sqrt{\frac{1}{2}\sum \left(x_i \log \frac{2x_i}{x_i + y_i} + y_i \log \frac{2y_i}{x_i + y_i}\right)} \qquad (2.1)$$

The result of the hierarchical cluster analysis on character bi-grams is shown in Figure 4. From Figure 4, we can see all the novels are clustered into three clusters. Ten chapters of *Otome no minato* are in the left cluster. All the twenty novels of Nakazato are in the middle cluster. All the twenty novels of Kawabata are in the right cluster.

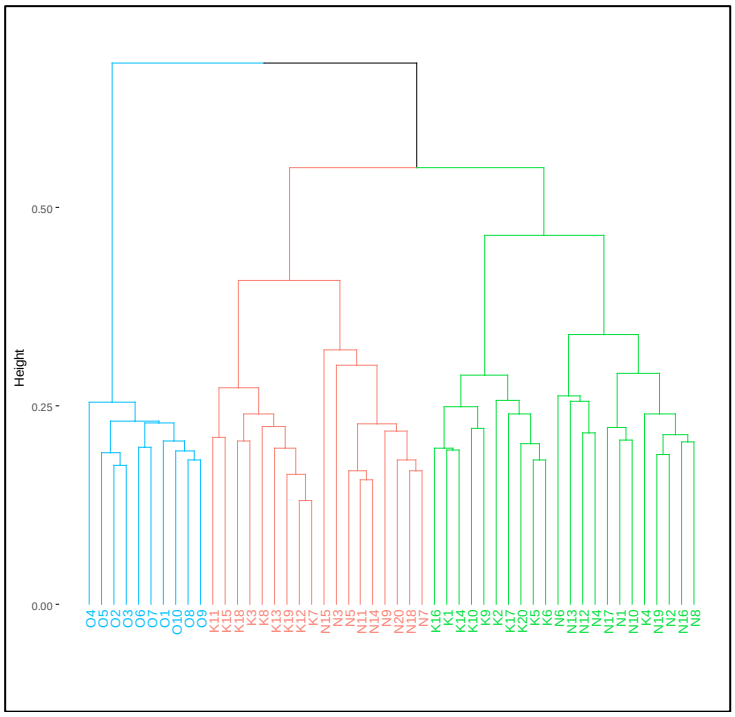

Figure 4. The dendrogram of hierarchical cluster analysis on POS bi-grams

The result of the hierarchical cluster analysis on particle bi-grams is shown in Figure 5. In Figure 5, ten chapters of *Otome no minato*, Kawabata's novels, and parts of Nakazato's novels are in the left cluster. Parts of Nakazato are in the right cluster.

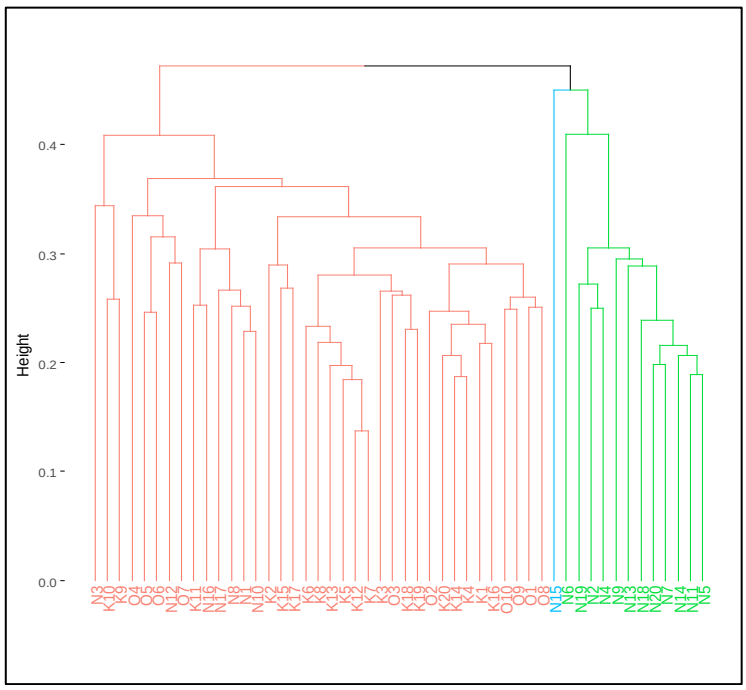

Figure 5. The dendrogram of hierarchical cluster analysis on particle bi-grams

The result of the hierarchical cluster analysis on phrase patterns is shown in Figure 6. From Figure 6, the same as character bi-grams and POS bi-grams, all the novels are clustered into three clusters. Ten chapters of *Otome no minato* are in the left cluster

while all the twenty novels of Nakazato are in the middle cluster. All the twenty novels of Kawabata are in the right cluster.

According to the analysis of correspondence analysis on POS bi-grams and phrase patterns. We can see that style of *Otome no minato* is different from Kawabata and Nakazato's.

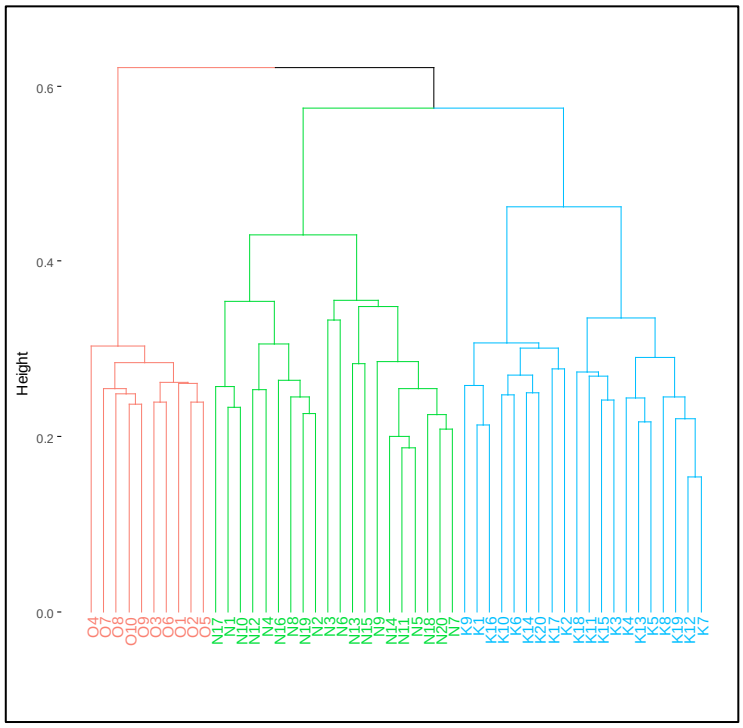

Figure 6. The dendrogram of hierarchical cluster analysis on phrase patterns

4.3 Classification results

We show the classification result of each chapter of *Otome no minato*, and the integrated classification result under the majority rule in this section.

4.3.1 Result of POS bi-grams

The classification results on POS bi-grams are described in table 4. In AdaBoost, all the chapters of *Otome no minato* were classified as Nakazato. In HDDA, chapters 1, 2, 4, and 8 were classified as Kawabata; chapters 3, 5, 6, 7, 9, and 10 were classified as Nakazato. In LMT, all the chapters were classified as Nakazato. In RF, chapters 1, 2, 3, 5, 6, 7, 8, 9, and 10 were classified as Nakazato, and chapter 4 was classified as Kawabata. The voting result under majority rule revealed that all the chapters belong to Tsuneko Nakazato.

4.3.2 Result of particle bi-grams

The classification results on POS bi-grams are described in table 4. In AdaBoost, chapters 2, 3, 4, 5, 6, 8, 9, and 10 were classified as Kawabata; chapters 1, 7 were classified as Nakazato. In HDDA, chapters 1, 2, 3, 5, 6, and 9 were classified as

Kawabata; chapters 4, 7, 8, and 10 were classified as Nakazato. In LMT, chapters 1, 2, 3, 5, 6, 8, 9, and 10 were classified as Kawabata; chapters 4 and 7 were classified as Nakazato. In RF, chapters 2, 3, 4, 6, 8, and 9 were classified as Kawabata; chapters 1, 5, 7, and 10 were classified as Nakazato. In SVM, chapters 1, 2, 3, 5, 6, 9, and 10 were classified as Kawabata; chapters 4, 7, and 8 were classified as Nakazato. The voting result under majority rule revealed that chapters 1, 2, 3, 5, 6, 8, and 9 belong to Kawabata; chapters 4 and 8 belong to Nakazato.

Table 3. Integrated result of POS bi-grams

chapters	AdaBoost	HDDA	LMT	RF	SVM	majority
1	N	N	N	N	N	N
2	N	N	N	N	N	N
3	N	N	N	N	K	N
4	K	K	N	K	N	N
5	N	K	N	K	N	N
6	K	K	N	K	K	K
7	K	K	N	N	N	N
8	N	K	N	N	N	N
9	K	K	N	N	N	N
10	N	K	K	N	N	N

K: Yasunari Kawabata, N: Tsuneko Nakazato

Table 4. Integrated results of POS tag bi-grams

chapters	AdaBoost	HDDA	LMT	RF	SVM	majority
1	N	K	K	N	K	K
2	K	K	K	K	K	K
3	K	K	K	K	K	K
4	K	N	N	K	N	N
5	K	K	K	N	K	K
6	K	K	K	K	K	K
7	N	N	N	N	N	N
8	K	N	K	K	N	K
9	K	K	K	K	K	K
10	K	N	K	N	K	K

K: Yasunari Kawabata, N: Tsuneko Nakazato

4.3.3 Result of phrase patterns

The classification results of phrase patterns are indicated in table 5. In AdaBoost, chapters 1, 3, 4, 5, 6, and 7 were classified as Kawabata, and chapters 2, 8, 9, and 10 were classified as Nakazato. In HDDA, chapters 1, 2, 7, 8, 9, and 10 were classified as Nakazato, and chapters 3, 4, 5, and 6 were classified as Kawabata. In LMT, chapters 1, 2, 3, 4, 5, 6, 7, and 9 were classified as Kawabata, and chapters 8 and 10 were classified as Nakazato. In RF, chapters 1, 2, 5, 7, 8, 9, and 10 were classified as

Nakazato, and chapters 3, 4, and 6 were classified as Kawabata. The voting result under majority rule revealed that chapters 1, 7, 8, 9, and 10 belong to Nakazato, and chapters 2, 3, 4, 5, and 6 belong to Kawabata.

Table 5. Integrated results of phrase patterns

chapters	AdaBoost	HDDA	LMT	RF	SVM	majority
1	N	N	N	N	N	N
2	N	N	N	N	N	N
3	K	K	N	K	K	K
4	N	K	K	K	K	K
5	K	K	K	K	K	K
6	K	K	K	K	N	K
7	N	N	K	N	N	N
8	N	K	K	N	N	N
9	K	K	K	K	N	K
10	N	N	N	N	N	N

K: Yasunari Kawabata, N: Tsuneko Nakazato

4.3.4 Integrated result of the three stylometric features

Finally, we combine the three integrated classification results to get the final classification result of *Otome no minato*. The majority result indicated that chapters 1, 7, 8, 9, and 10 were classified as Nakazato. Chapters 2, 3, 4, 5, and 6 were classified as Kawabata.

Table 6. Integrated results of the three stylometric features

chapters	character bi-grams	POS bi-grams	phrase patterns	majority
1	K	N	N	N
2	K	N	K	K
3	K	N	K	K
4	K	N	K	K
5	K	N	K	K
6	K	N	K	K
7	K	N	N	N
8	K	N	N	N
9	K	N	N	N
10	K	N	N	N

K: Yasunari Kawabata, N: Tsuneko Nakazato

5. Conclusion

In this paper, we focused on the suspicion that Yasunari Kawabata's novel *Otome no minato* was ghostwritten. We identified the actual author of this novel by means of an integrated classification algorithm, which integrates the classification result of three

stylometric features and five classifiers. Our integrated result suggests that the writing style of chapters 1, 7, 8, 9, and 10 of *Otome no minato* are close to that of Tsuneko Nakazato, while chapters 2, 3, 4, 5, and 6 are close to that of Yasunari Kawabata. The result reveals that *Otome no minato* should be recognized as a collaborative work by Yasunari Kawabata and Tsuneko Nakazato.

References

Aljumily, Refat (2015). Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods to "Shakespeare Authorship Question". *Social Sciences*, 4, 758-799.

Crombez, Thomas & Cassiers, Edith (2017). Postdramatic Methods of Adaptation in the Age of Digital Collaborative Writing. *Digital Scholarship in the Humanities*, 32(1), 17-35.

Eder, Maciej (2017). Visualization in Stylomerty: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities*, 32(1), 50-64.

Gladwin, Alexander A. G.; Lavin, Matthew J. & Look, Daniel M. (2017). Stylometry and Collaborative Authorship: Eddy, Lovecraft, and 'The Loved Dead'. *Digital Scholarship in the Humanities*, 32(1), 123-140.

Grieve, Jack (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251-270.

Jin, Mingzhe (2013). Authorship identification based on phrase patterns, *The Japanese Journal of Behaviormetrics*, 40(1), 35-46.

Jin, Mingzhe (2014). Using Integrated classification algorithm to identify a text's author. *The Japanese Journal of Behaviormetrics*, 41(1), 35-46.

Kestemont, Mike; Moens, Sara & Deploige, Jeroen (2015). Collaborative authorship in the twelfth century: A stylometric study of Hildegrad of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities*, 30(2), 199-222.

Koyano, Atushi (2013). *Kawabata Yasunari den - Soumen no hito*. Chuoukoron-shinsha Press.

Mosteller, Frederick & Wallace, David L. (1964). *Inference and Disputed Authorship: The Federalist*, Addison Wesley.

Rybichi, Jan; Hoover, David & Kestemont, Mike (2014). Collaborative authorship: Conrad, Ford and Rolling Delta. *Literary and Linguistic Computing*, 29(3), 422-431.

Rudman, Joseph (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, 351-365.

Stamatatos, Efstathios (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.

Wataru, Zaitsu & Jin, Mingzhe (2015). Identifying the author of illegal documents through text mining. *Japanese Journal of Forensic Science and Technology*, 20(1), 1-14.

Zhao, Ying & Zobel, Justin (2007). Searching with style: Authorship attribution in class literature. In: *Proceedings of the Thirtieth Australasian Computer Science Conference.* 59-68.

Corpus-Based Assessment of Linguistic Complexity of the Tatar Language: Methodology and Preliminary Results*

Alfiya Galieva, Olga Nevzorova
Tatarstan Academy of Sciences, Kazan, Russia

Abstract
The main objective of the paper is to discuss the theoretical construct of linguistic complexity and to accommodate it to the Tatar language data. Tatar belongs to the Turkic family; the distinctive feature of the members of the latter is agglutinative structure with rich morphology.

To assess the linguistic complexity of the Tatar language we sketch out some features that are relevant for natural language processing. We proceed from the assumption that linguistic complexity becomes apparent in parameters that can be measured, so quantitative properties of languages are essential for describing it. We consider the statistical distribution of grammatical categories on corpus data and emphasize the complex phenomenon of Tatar morphology.

The paper demonstrates that Tatar inflectional affixes tend to be universal for different parts of speech, all the allomorphs are conditioned phonetically, and irregular affixes do not exist. We also establish the frequency of allomorphs and give some explanations concerning the asymmetry of distribution of allomorphs. Particular attention is paid to nominal affixes, and we substantiate statistically that Tatar case affixes and the comparative affix are not bound to the unique grammatical class of the stem, but may be joined to a broad range of stems, including verbal ones.

Keywords: linguistic complexity, Tatar agglutinative morphology, Tatar inflectional affix, allomorph, Tatar comparative affix, Tatar case affix

1. Introduction

Linguistic complexity is currently one of debatable notions in linguistics, and there are different ways to understanding complexity depending on linguistic domains, theoretical frameworks and researchers' aims. We proceed from the assumption that linguistic complexity becomes apparent in parameters that can be measured, so quantitative properties of languages are essential for describing it.

Some of basic quantitative characteristics of a language that are related to its grammatical structure, even if not all, may be retrieved from grammatically annotated corpora. So linguistic corpora, if their system of grammatical annotation is commensurate and comparable, may provide objective data about grammatical complexity of languages. A precise formal characterization of comparable structures within and across languages can provide a complexity ranking for them (Hawkins, 2014), and the

* Address correspondence to: Alfiya Galieva, Olga Nevzorova， Tatarstan Academy of Sciences, Kazan, Russia。 E-mail: amgalieva@gmail.com

methodology should be developed on particular language data.

The task of calculating the frequency of grammatical categories and distribution of allomorphs in real texts is the first step in the assessment of linguistic complexity. The paper represents an endeavour to assess the morphological complexity of the Tatar language as a first approximation, by analyzing the distribution of certain grammatical categories on corpus data.

The remainder of this paper is organized as follows. Section 2 gives a brief overview of related works, Section 3 outlines main grammatical features of the Tatar language, Section 4 is aimed at detecting morphological complexity of Tatar basing on quantitative corpus data. Finally, Section 5 lists the conclusions that can be derived from this research.

Quantitative linguistic data reflect the current state of the Tatar National corpus (http://tugantel.tatar). Its current volume is about 100 million word forms. The data in the tables below were obtained from this corpus by corpus searches.

2. Related work

During the last decades specialized monographs (Miestamo et al., 2008; Sampson et al., 2009; Newmeyer & Preston, 2014) and journal papers (Bane, 2008; Becerra-Bonache & Jimenes-Lopez, 2015, Blache, 2011; Nevzorova et al., 2017; Palotti, 2009, etc.) that were devoted to the issue of language complexity were published. Although the concept of linguistic complexity seems intuitively clear, it has scarcely undergone formalization and analysis. Researchers regard different criteria and parameters of linguistic complexity and get different, even opposite results for the same language.

The dissertation *Linguistic Complexity. The Influence of Social Change on Verbal Inflection* by W. Kusters (Kusters, 2003) was one of the first studies focused on the influence of extralinguistic factors on the internal structure of a language. The author studies verbal inflection in certain languages (Arabic, Scandinavian, Quechua, and Swahili) and argues that a large number of non-native speakers of a language, social cohesion within a speech community, and enlargement of external contacts can lead to decreasing the complexity of verbal inflection.

In Dahl (2004) methodologically significant delimitation of a number of essential concepts such as *complexity, cost, difficult*y and *demandingness* is represented. According to this point of view, *complexity* is a theoretical construct aimed at determining an 'objective' parameter of a language, which would be important for language processing, that must not be related to a user or an agent. The notions of *cost* and *difficulty* are relevant for adult language learners. Cost implies essentially "the amount of resources — in terms of energy, money or anything else — that an agent spends in order to achieve some goal" (Dahl 2004: 39). High cost does not necessarily imply high degree of complexity - the relationship between these phenomena is not direct. "Difficulty is a notion that primarily applies to tasks, and is always relative to an agent: it is easy or difficult for someone" (Dahl 2004: 39). Demandingness is a link between complexity and difficulty: for instance, acquiring a human language natively is certainly demanding (only human children seem to fulfil the requirements), but it does not necessarily follow that children find it difficult (Dahl 2004: 39-40).

The paper by P. Juola (2008) discusses some definitions proposed in literature,

and shows how complexity can be assessed in various frameworks. The author focuses on mathematical and psychological aspects of complexity, and attempts to validate available complexity measurements. One more work in this field (Newmeyer & Preston, 2014) covers the discussions on measuring grammatical complexity from a variety of viewpoints like formal linguistics or studies of the brain.

The crucial question that remains relevant is how to actually measure and compare the degree of linguistic complexity. The programmatic sketches for measureing complexity may range from what might be called 'grammar-based' to those that might be called 'user-based'. The former approach focuses on elements of grammars and counts the amount of structural dependencies, irregularity, and so on. The latter considers complexity in terms of emphasizing the language difficulty for the user, distinguishing the first-language acquirer, the second-language acquirer, or the adult user (Newmeyer & Preston, 2014: 7).

We may say that the topic of complexity of languages has different dimensions and nowadays attracts a great deal of interest. Researchers maintain that language complexity may be regarded and evaluated on different levels: of the language as a whole, and of its separate layers; thus parameterization of linguistic complexity needs further research, and work results must be considered in the general theory of language.

The notion of complexity is conceptualized and defined differently in different domains. To specify this notion we are to take into consideration peculiarities of the internal organization of the system, its evolution, interaction with the external world, etc. We are to realise that the actual diversity of internal relations of a complex object is not easy to merely describe and parameterize, but also to discover in many cases. That is essential for such a multidimensional phenomenon as language.

Assessment of linguistic complexity supposes search for objectively evaluating and finding comparable criteria. To determine the absolute value of complexity many researchers (Dahl, 2004; Juola, 2008) use a categorical apparatus of information theory, for example, Kolmogorov complexity that may be defined as a way of measuring the amount of information in a given string - as the length of the shortest possible algorithm required to describe/generate that string (Juola, 2008). Because of practical uncomputability and nonapplicability of Kolmogorov complexity for linguistic phenomena, P. Juola applies a purely technical expedient and considers the file compression method as an attempt to approximate this kind of complexity within a tractable formal framework (Juola, 2008). Researchers maintain that formulation of complexity and complexity metrics should be based on the number and variety of the parts of the grammatical description and interactions between them (Dahl, 2004; Miestamo, 2008; Sinnemäki, 2014).

J. McWhorter, assessing linguistic complexity, relies upon the assumption that an area of grammar is more complex than the same area in another grammar to the extent that it encompasses more overt distinctions and/or rules than another grammar (McWhorter, 2001). This assumption is deployed in the following way:

1 A phonemic inventory is more complex to the extent that it has more marked members.

2. A syntax is more complex than another to the extent that it requires processing more rules, such as asymmetries between matrix and subordinate clauses.

3. A grammar is more complex than another to the extent that it gives overt and grammaticalized expressions to more fine-grained semantic and/or pragmatic distinctions than another.

4. Inflectional morphology renders a grammar more complex than another one in most cases (McWhorter, 2001: 135 - 137).

Another important delimitation of complexity is that any complexity measurement needs to focus on local, rather than global, complexity, as suggested by M. Miestamo (2008). The former concerns the complexity of individual grammatical domains of a language, while the latter considers its overall complexity, including phonological, morphological, syntactic and semantic systems.

Reduction to a common denominator of great variety of grammatical phenomena of different languages remains an insoluble problem; nevertheless the first steps in this direction may be made by means of automatic text processing.

P. Blache proposes a computational model, making it possible to give a quantitative evaluation of grammatical complexity, taking into account all the different parameters concerning the issue. The model takes advantage of a constraint based representation, displaying different types of linguistic information separately. His approach involves a precise identification of what kind of constructions are complex and, as a side effect, what is difficult to process. This model can then serve as a basis for a comparison of different languages in terms of complexity (Blache, 2011).

From our viewpoint, among the factors that influence the grammatical complexity of a language of agglutinative structure, like Tatar, the following may be viewed:

- universality of means for expressing grammatical categories (or its absence);
- variety of grammatical categories of different types;
- regularity of means of expression of a grammatical category (absence of exception to the rules);
- length of affixal chains (average length of affixal chains for each part of speech);
- potential interconversion of parts of speech;
- degree of linguistic redundancy, etc.

Some of these characteristics of a language, if not all, may be retrieved from grammatically annotated corpora. So linguistic corpora, if their system of grammatical annotation is commensurate and comparable, may provide objective data on grammatical complexity of languages.

3. Tatar agglutinative morphology

The Tatar language is related to the Turkic family and is characterized by rich agglutinative morphology. The basic way of word formation and inflection is progressive affixal agglutination when a new unit is built by consecutive addition of regular and clear-cut monosyllabic derivational and inflectional affixes to the stem. Affixal agglutination provides unified morphological means for forming derivatives within the same grammatical class of words as well as for changing the part-of-speech characteristic of the word and for turning it into another lexical or grammatical class. The boundaries between the affixes within the word form are distinct and transparent,

and the affixal joint in many cases coincides with the syllabication (Guzev & Burykin, 2007).

In Turkic languages the order of affixes is rigidly determined, for example, noun word forms have the following morphological structure: stem <plural> <possessive> <case> <modality>:

bala - lar-ıbız-ga-dır
child -PL, POSS_1PL, DIR, PROB
'maybe, to our children'

Each added affix tends to modify the whole preceding stem. Nouns have no classifying categories, like grammatical gender or animacy. The plural of nouns is formed by joining the affix *-La*r to the stem; the same plural affix is used to nominalize adjectives and to form the 3rd person of verbs in plural. Possessive affixes are used to express the person and the number of possessors.

The affixes in the affix chain are contextually unambiguous. There is one type of declension and, with certain proviso, conjugation, so only one set of affixes for each grammatical class is used (though there is a set allomorphs for each affix).

The Tatar verb has no aspect, but is characterized by tense, mood and can have the negative form.

Numerals are used with the following singular noun, which simplifies the grammatical structure of nominal phrases:

ber bala 'one child'
ike bala 'two children'

Tatar morphology is regular and predictable in many respects, and there is little or no fusion between the stem and the affixes (affixes do not change the stem in any way as it may happen in other languages, e.g. in Russian when an inflection often causes palatalization of stem's final consonant (examples like Russian *вода* 'water' (NOM)- *воде* (DAT) 'to water' are impossible).

Examples of the word inflections *урам* 'street, outdoors' below represent some salient features of Tatar morphology:

(1a) урам
'street, outdoors'

(1b) урам-да
street-LOC
'in the street, outdoors'

(1c) урам-да-гы
street-LOC, ATTR
'that who is outdoors'

(1d) урам-да-гы-лар
street-LOC, ATTR, PL
'those who are outdoors'

(1e) урам-да-гы-лар-га

street-LOC, ATTR, PL, DIR
'to those who are outdoors'

So a Tatar agglutinative word form is built by adding standard, mostly unambiguous, affixes to the stem, with the order of affixes and phonetic changes of affixes rigidly determined, and with affix boundaries clear-cut. Nevertheless an attempt to build a paradigm of an individual word shows that this paradigm is extremely complicated and divaricate, consisting of words formed by a stem and a great number of inflectional affixes.

Another typologically relevant feature of the Tatar language is absence of clear-cut borders between inflection and derivation, since the same affixes in different positions may function both as inflectional and derivational.

The most important phonetic feature of Turkic languages is progressive vowel harmony, so the nature of vowels of a stem predetermines the choice of allomorphs of the derivational and inflectional affixes.

4. Morphological complexity of Tatar: using corpus data

In this section we will show some features of Tatar morphology from the linguistic complexity viewpoint, and provide linguistic data from the Tatar National corpus.

4.1. Universality of means for expressing grammatical categories

In Tatar most word forms of different parts of speech have the same affixes, for example, the standard Plural affix is used both for noun, adjective and verb stems (3PL):

(2a) Bala-lar bara-lar.
Child-PL go-PRES, PL
'Children go'

(2b) Bala-lar matur-lar.
Child – PL beautiful-PL
'Children are beautiful'

Table 1. represents the distribution of Plural affix depending on grammatical class of the word stem.

Table 1. Distribution of Plural affix depending on the grammatical class of the word stem.

Part of speech	**Singular**	**Plural**
noun	40,250,374	7,516,467
adjective	2,469,776	570,801
verb	839,389	1,464,261

Case affixes, as well as Plural and possessive ones, are used to inflect nouns (and in Tatar grammars the grammatical case is described as a grammatical category of nouns) (Zakiev, 1993). Besides, case affixes may be used in nominalization of words of other parts of speech – adjectives, numerals, some verbal forms, etc. Consider the following examples with affix -GA – the affix of the Directive (Dative) case:

urman-ga bara
forest-DIR go-PRES
‘(he/she) goes to the forest'

Adjectives in Tatar are grammatically interchangeable and do not join inflectional affixes (excluding the comparative affix):

ak kiyem-när
white cloth-PL
'white cloths'

Joining case affixes, they turn into nominalized adjectives, and are used as nouns:
Akka manu
white-DIR paint-VN
'to paint in white (colour)'
In the same way certain verb forms may be nominalized:
Abıy kaytkan-ga şatmın.
Elder brother return-PST_IND, DIR glad-1SG
'I am glad that the elder brother has returned'.

The following example shows that nominalization of numerals is realised in the same way (3b):

(3a) un keşe
ten man
'ten persons'

(3b) Un-nan ber-ne al.
ten-ABL one-ACC take-away
'Take away one from ten'.

The same case affixes are used for different parts of speech – nouns, verbal derivatives, adjectives, numerals. Table 2 represents the distribution of case affixes depending on the stem type.

Table 2. Distribution of case affixes depending on the stem type

Case	**Part of speech of stem**			
	N	**V**	**ADJ**	**NUM**
GEN	2,353,774	258,853	87,622	22,515
DIR	4,480,058	798,613	466,618	119,488
ACC	3,859,725	890,644	428,033	36,936

ABL	1,574,473	317,259	339,393	44,146
LOC	4,599,572	795,389	376,201	48,150

So case affixes may function as special operators that turn words of other grammatical classes (adjectives, verbs, numerals, etc.) into nouns (or, rather, noun-like structures denoting concepts or situations). Plural and possessive affixes (separately or combined with case affixes) may function the same way (for turning words into nouns and noun-like structures).

The same way joining personal affixes of verbs (1-3 persons SG and PL) to nouns and adjectives one can get verb-like forms of nouns and adjectives (nouns and adjectives functioning as predicates).

Another particular feature of Tatar morphology involves the comparative affix -rAK. Prototypical comparatives denote the increase in quality, quantity, or relation expressed by adjectives or adverbs. The comparative affix in Tatar also joins to:

- certain forms of verbs (converbs and participles);
- certain forms of pronouns;
- certain forms of nouns in indirect cases.

Table 3 represents the distribution of word classes containing the comparative affix.

Table 3. Grammatical types of stems joining the comparative affix

Stem type	**Number of word forms**	**Number of unique word forms**
ADJ +COMP	238,431	1,504
N +COMP	13,872	936
V +COMP	4,704	855
PN +COMP	684	20

In particular, the comparative affix -rAK joins the case forms denoting spatial relations – the Directive, Ablative, and Locative. See examples below:

koyaş-ka-rak
sun-DIR, COMP
'to (the place) where (there is) more sunlight'

koyaş-ta-rak
sun-LOC, COMP
'in (the place) where (there is) more sunlight'

koyaş-tan-rak
sun-ABL, COMP
'from (the place) where (there is) more sunlight'

Cases expressing spatial relations permit a gradation of a quality (LOC+COMP) or a gradual change of location of the object (DIR + COMP; ABL + COMP). See the

number of each combination of affixes in Table 4.

So the Tatar language has special grammatical mechanisms (not taking into account its derivative potential) for converting nouns and verbs into each other in a sentence. Nominal affixes (case and possessive affixes) may be used as operators of nominalization, and personal affixes of verbs - for turning nouns and adjectives into predicates.

Table 4. Case forms of nouns joining the comparative affix.

Case	**Total number of word forms**	**Number of unique word forms**
DIR	4,419	307
ABL	202	28
LOC	331	22

Different parts of speech share the affixal inventory, and that leads to blurring boundaries between morphological classes of words.

Also the same comparative affix may be joined to words of a different grammatical class for expressing different gradations and intensity of a quality.

The existence of a set of common affixes for different parts of speech simplifies the language system because the number of elements is reduced; nevertheless the same reason considerably complicates the rules for describing functioning words in a sentence.

4.2. Variety of grammatical categories of different types

The system of morphological annotation of the Tatar National Corpus (Suleymanov et al. 2013; Galieva et. al. 2016) may give a general idea of the plenty of morphological categories in Tatar.

In the current version of grammatical annotation of the Corpus 11 tags for parts of speech and 75 tags for inflectional affixes are used.

Because of variety of inflectional affixes of different types and combinations of items in the affixal chain the potential number of word forms is uncertain (the number of word forms has pragmatic rather than grammatical constraints and depends on the word meaning), which leads to inapplicability of the term paradigm in the traditional sense. For example, the word *agaç* ‘tree’ is used 41,124 times in the Corpus, and has 84 unique word forms that are detected for the lemma *agaç*. Table 5 enables one to see the distribution of noun word forms of different types in the Tatar National Corpus and to assess the average length of affixal chains for nouns.

Variety of grammatical categories of different types leads to a great abundance of individual word forms in real use, including items consisting of long affixal chains, which increases the degree of linguistic complexity of Tatar. Nevertheless rigid rules of combining affixes and unambiguousness of affixes decrease the linguistic complexity.

4.3. Distribution of case affixes: types of allomorphs

As a result of progressive vowel harmony the quality of second and subsequent syllables is determined by the quality of the first or preceding syllable, and certain researchers claim that second and subsequent syllables of Turkic languages may contain three phonemes (or morphophonemes) only - a, ı, u (Guzev & Burykin, 2007).

The impact of the phonetic and morphological factors on the choice of allomorphs of Tatar case affixes were considered in (Galieva et al., 2015).

The Tatar language exhibits complete stem invariance; depending on the phonetic aspect of the stem, affixes of two vowel types are used - front vowels and back vowels. Table 6 represents the distribution of phonological variations for case affixes.

Table 5. Distribution of some grammatical forms of the word *agaç* ‘tree’, according to corpus data.

Structure of the wordform	**Example**	**English translation**	**Number of wordforms of *agaç***
stem (Absolute case)	agaç	'tree'	20,357
N+PL	agaç-lar	'trees'	5,516
N+POSS_1SG	agaç-ım	'my tree'	8
N+PL+POSS_1SG	agaç-lar-ım	'my trees'	1
N+LOC	agaç-ta	'on the tree'	76
N+PL+LOC	agaç-lar-da	'on the trees'	55
N+PL+LOC+ATTR	agaç-lar-da-gı	‘that on the tree’	10
N+POSS_1SG+ LOC	agaç-ım-da	'on my tree'	1
Total number			41,124

Table 6. Distribution of allomorphs of case affixes

Case	**Back vowel affixes**	**Front vowel affixes**
Genitive	1,776,992	1,308,121
Directive	924,942	730,608
Accusative	453,434	379,437
Ablative	388,313	273,607
Locative	917,664	590,617

Quality consonants in case affixes are also determined by the character of the last sound of the stem: in the Directive, Ablative and Locative. Table 7 demonstrates the number of case affixes depending on the consonant nature.

Table 7. Distribution of case affixes depending on consonant quality

Case	**Allomorph containing voiced consonant**	**Allomorph containing voiceless consonant**	**The ratio of the number of affixes containing voiced and voiceless consonants**
Directive	736,586	333,289	2,21
Ablative	230,972	137,427	1,68
Locative	1,160,788	259,817	4,47

The asymmetry between the amount of voiced consonant affix variations and voiceless consonant ones may be explained by the fact that:

- - the use of possessive affixes preceding case affixes influence the choice of a voiced consonant in case affixes;
- - most verbal derivatives in the active voice also phonetically require a voiced consonant in case affixes.

As a result, in the Directive, the number of voiced consonants is 2,21 times more than that of voiceless consonants; in the Locative, the number of voiced consonants is 4 times more than that of voiceless ones. The additional complicating factor that influences allomorph choice are possessive affixes POSS_1SG, POSS_2SG, and POSS_3. See Table 8 (in brackets the number of unique word forms is given).

Table 8. Distribution of case allomorphs depending on preceding possessive affixes

Case	**Standard case affixes**	**Case affixes after POSS_1SG, POSS_2SG**	**Number of case affixes after POSS_1SG, POSS_2SG**	**Case affixes after POSS_3 affix**	**Number of case affixes after POSS_3 affix**
DIR	-ga / -gä -ka / -kä	-a / -ä	42,036 (3,340)	-na / -nä	515,812 (11,680)
ACC	-nı/-ne	-nı/-ne	67,460 (6,397)	-n	697,083 (18,083)
ABL	-dan / -dän -tan / -tän	-nan / -nän	14,663 (1,743)	-nnan / -nnän	214,461 (7,559)
LOC	-da / -dä -ta / -tä	-da / -dä	18,819 (1,470)	-nda / -ndä	635,037 (6,969)

There are similar allomorphs for other affixes. Although Tatar allomorphs raise the level of linguistic complexity by multiplying the number of affix variants, they facilitate pronunciation to the speaker and word recognition to the listener by splitting up word boundaries.

5. Conclusion

The sets of grammatical categories in languages may vary significantly, and quantitative analysis enables researchers to reveal the essence of the language by examining the structures and properties we can observe in real texts. Our work is the first and a very preliminary study aimed at developing methodology for assessing linguistic complexity of the Tatar language using corpus data. We made a first attempt to elucidate certain typological features of the Tatar language from the viewpoint of linguistic complexity and to use the data of corpus grammatical annotation to measure the linguistic complexity of Tatar.

The paper presented the distribution of nominal inflectional affixes in Tatar on the data of the Tatar National Corpus. We showed that Tatar case affixes, the plural and comparative affix are not bound to the unique grammatical class of the stem, but may be joined to a broad range of stems. We also demonstrated the frequency of allomorphs and gave some explanations concerning the asymmetry of their distribution.

Different parts of speech share the affixal inventory, which leads to blurring boundaries between the morphological classes of words. The Tatar language has special grammatical mechanisms (not taking into account its potential of derivation) for converting nouns and verbs into each other in a sentence. Nominal affixes (case and possessive affixes) may be used as operators of nominalization, and personal affixes of verbs are often applied for turning nouns and adjectives into predicates. Also the same comparative affix may join words of a different grammatical class for expressing different gradation and intensity of a quality.

The existence of the set of common affixes for different parts of speech simplifies the language system because the number of elements is reduced; nevertheless the same fact considerably complicates the rules for describing functioning words in a sentence. So potential interconversion of parts of speech is rather high.

The Tatar language has one type of declension and conjugation in which one set of affixes is used only. Regularity of means of expression of a grammatical category and absence of exception to the rules simplifies the language system.

The choice of allomorphs follows rigid phonetic laws, and the set of allomorphs facilitates the process of speaking. The choice of allomorphs is influenced by morphological factors in a very low degree.

List of abbreviations

ABL - Ablative case,
ACC - Accusative case,
ADJ - Adjective,
ATR - Attributive,
COMP - Comparative,
DIR - Directive case,
GEN - Genitive,
LOC - Locative,
N - Noun,
PL - Plural,

PN - Pronoun,
POSS - Possessive,
PRESS - Present,
PROB - Probabilitive,
PST_IND - Past Indefinite,
SG - Singular,
V - Verb,
VN - Verbal noun.

References

Bane, Max (2008). Quantifying and Measuring Morphological Complexity. In: *Proceedings of the 26th West Coast Conference on Formal Linguistics*, pp. 69 -76.

Becerra-Bonache, Leonor & Jimenes-Lopez, Maria Dolores (2015). A Grammatical Inference Model for Measuring Language Complexity In: *Advances in Computational Intelligence. 13th International Work Conference on Artificial Neural Networks*, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part I, pp. 4 - 17.

Berdichevsky, Alexander (2012). Language Complexity [Yazykovaya slozhnost]. In: *Voprosy Jazykoznanija* (Topics in the study of language). Vol. 5, pp. 101 - 124. (In Russian).

Blache, Philippe (2011). A computational Model for Linguistic Complexity. In: *Frontiers in Artificial Intelligence and Applications*, 228: 155 - 167.

Dahl, Östen. (2004). *The Growth and Maintenance of Linguistic Complexity*. John Benjamins Publishing, Amsterdam.

Galieva, Alfiya; Khakimov, Bulat & Gatiatullin, Ayrat (2016). On the Way to the Relevant Grammatical Tagset for the Tatar National Corpus. In: *EPiC Series in Language and Linguistics.* Vol. 1, CILC2016, 8th International Conference on Corpus Linguistics, pp. 121 - 129.

Galieva, Alfiya; Nevzorova, Olga & Suleymanov, Dzhavdet (2015). Statistic Distribution of Some Grammatical Categories of the Tatar Language over Corpus Data. In: *Proceedings of the International Conference "Turkic Languages Processing" TurkLang - 2015*. Academy of Sciences of the Republic of Tatarstan Press, Kazan, pp. 396 - 408.

Gil, David (2008). How Complex are Isolating Languages? In: Miestamo, M., Sinnemäki, K. and Karlsson, F. (eds). Language Complexity: Typology, Contact, Change. Vol. 94. John Benjamins Publishing, Amsterdam, pp. 109 - 131.

Guzev, Viktor & Burykin, Alexey (2007). Common Structural Features of Agglutinative Languages [Obshchiye stroevye osobennocti agglutinativnykh yazykov]. In: *Acta linguistica Petropolitana. Proceedings of ILR RAS.* Vol. 3: 1, pp. 109 – 117. (In Russian).

Hawkins, John (2014). Major Contributions from Formal Linguistics to the Complexity Debate. In: Newmeyer F. J. & Preston L. B. (eds.) (2014).

Measuring Grammatical Complexity. Oxford University Press, Oxford, pp. 14 - 36.

Juola, Patrick (2008). Assessing Linguistic Complexity. In: Miestamo, Matti, Sinnemäki, Kaius & Karlsson, Fred (eds.) (2008). *Language Complexity: Typology, Contact, Change.* Vol. 94. John Benjamins Publishing, Amsterdam, pp. 89 - 108.

Kusters, Wouter (2003). *Linguistic Complexity: The Influence of Social Change on Verbal Inflection.* LOT, Netherlands Graduate School of Linguistics, Utrecht.

McWhorter, John. (2001). The World's Simplest Grammars are Creole Grammars. In: *Linguistic Typology.* Vol. 5, pp. 125 - 166.

Miestamo, Matti; Sinnemäki, Kaius & Karlsson, Fred. (eds). (2008). *Language Complexity: Typology, Contact, Change.* Vol. 94. John Benjamins Publishing, Amsterdam.

Nevzorova, Olga; Galieva, Alfiya & Nevzorov, Vladimir (2017). Toward Measuring Linguistic Complexity: Grammatical Homonymy in the Russian Language. In: *International Journal "Information Theories and Applications"*. Vol. 24, pp. 127 - 140.

Newmeyer, Frederick J. & Preston Laurel B. (eds.) (2014). *Measuring Grammatical Complexity*. Oxford University Press, Oxford.

Nichols, Johanna (2009). Linguistic Complexity: A Comprehensive Definition and Survey. In: Sampson Geoffrey, Gil David & Trudgill Peter (eds.) (2009). *Language Complexity as an Evolving Variable.* Oxford University Press, Oxford, pp. 110 - 125.

Nöth, Winfried (1995). *Handbook of Semiotics.* Indiana University Press.

Pallotti, Gabriele (2009). CAF: Defining, Refining and Differentiating Constructs. In: *Applied Linguistics 30(4),* pp. 590 - 601.

Sampson, Geoffrey; Gil, David & Trudgill Peter. (eds.) (2009). *Language complexity as an Evolving Variable.* Oxford University Press, Oxford.

Sinnemäki, Kaius (2014). Complexity Trade-offs: a Case Study. In: Newmeyer Frederick J. & Preston LaureL B. (eds.) *Measuring Grammatical Complexity.* Oxford University press, Oxford, pp. 179 - 201.

Suleymanov, Dzhavdet; Nevzorova, Olga; Gatiatullin, Ayrat; Gilmullin, Rinat & Khakimov, Bulat (2013). National Corpus of the Tatar Language "Tugan Tel": Grammatical Annotation and Implementation. In: *Procedia - Social and Behavioral Sciences.* Vol. 95, pp. 68 - 74.

Tatar National Corpus. URL: http://tugantel.tatar/?lang=en

Zakiev, Mirfatykh (ed.) (1993). *Tatar Grammar.* Tatar Publishing House, Kazan.

QUALICO 2016 in Trier. Conference Report

Reinhard Köhler, Trier

The call had a good echo: 59 Submissions were received, of which 6 were rejected and 53 accepted. The papers were selected one by one by at least three of the members of the program committee. 46 papers were finally presented. 57 researchers attended the conference, although 65 were expected. As often, some of the colleagues did not receive their visas in time or became sick etc.

The scientific program was divided into two parallel sections. Each participant had a time slot of 20 minutes for her/his presentation and 10 minutes for discussion. The program was subdivided and grouped into three or two presentations, which made 8 talks per day. As usual, the presentations were chaired by experienced colleagues. Additionally, information desks and book tables were available at any time.

After the opening of the conference and the welcome addresses (Arjuna Tuzzi, IQLA president and the vice-president of the University of Trier), the following papers were delivered:

Miroslav Kubát and Radek Čech, University of Ostrava, Czech Republic: ***Dynamic development of vocabulary richness of text***

Reinhard Köhler, University of Trier: ***Text characterization using syntactic motifs***

Makoto Yamazaki, National Institute for Japanese Language and Linguistics, Japan: ***Coherence and quantitative measures of text***

Heng Chen & Xinying Chen, Zhejiang University, China: ***Do L-motifs and F-motifs co-evolve with word length?***

Peter Grzybek, University of Graz, Austria: ***Word length research: Coming to rest. Or not?***

Jingqi Yan, Zhejiang University, China: ***The rank-frequency distribution of part-of-speech motif and dependency relation motif in the deaf learners' compositions***

Hermann Moisl, Newcastle University, U.K.: ***The mathematics of meaning***

Katharina Prochazka & Gero Vogl, University of Vienna, Austria: ***A diffusion model for language shift in Austria***

Li Haotian, Zhejiang University, China: ***Quantitative properties of the dialogue in Brothers Karamazov***

Emmerich Kelih, University of Vienna, Austria: ***How plausible is the hypothesis that the population size is related to the phoneme inventory size? Comments from a quantitative linguistics point of view***

Hanna Gnatchuk, Alpen-Adria University, Austria: ***Testing hypotheses on English compounds***

Yingxian Zhang & Yue Jiang, Xi'an Jiaotong University, China: ***A VSM-based algorithm for comparing online translation and human translation of English passives into Chinese***

Yuichiro Kobayashi, Misaki Amagasa & Takafumi Suzuki, Toyo University: ***Investigating the chronological variation of lyrics of popular songs through lexical indices***

Jiří Milička & Karolína Vyskočilová, Charles University, Prague: ***Models of noisy channels that speech gets over***
Relja Vulanović & Oliver Ruff, Kent State University at Stark, Ohio, USA: ***Measuring the degree of violation of the One-Meaning–One-Form Principle***
Martina Benešová, Dan Faltýnek & Lukáš Zámečník, Palacký University Olomouc, Czech Republic: ***Functional explanation in quantitative linguistics***
Aiyun Wei, Haitao Liu & Jingqi Yan: ***Quantitative features of vocabulary in Zhuang Language***
Gertraud Fenk-Oczlon, Alpen-Adria-Universität Klagenfurt, Austria: ***Are some languages spoken more quickly than others? A quantitative typological study***
Xinying Chen[1] & Heng Chen[2], 1Xi'an Jiaotong University; 2Zhejiang University, China: ***A diachronic study of Chinese words based on Google N-gram data***
Wei Huang, Beijing Languange and Culture University, China: ***Script complexity distribution based on construction theory of Chinese character form***
Lu Wang, Dalian Maritime University, China: ***Polyfunctionality studies in German, Dutch, English and Chinese***
Olga Nevzorova, Alfiia Galieva & Dzhavet Suleymanov, The Tatarstan Academy of Sciences, Kazan Federal University, Russian Federation: ***Assessment of linguistic complexity of the Tatar language using corpus data (an attempt to study)***
Edward Klyshinsky[1] & Varvara Logacheva[2], 1Higher School of Economics, Moscow, Russia; 2University of Sheffield, UK: ***Quantitative evaluation of morphological ambiguity of European languages***
Haruko Sanada, Rissho University, Japan: ***Quantitative interrelations of properties of the complement and the adjunct***
Adam Pawłowski, Krzysztof Topolski, Piotr Malak, Jan Kocón & Michał Marcińczuk, University of Wrocław, Poland: ***Statistical distributions of lexeme frequencies in a text corpus. A comparative analysis***
Petra Steiner, Institut für deutsche Sprache, Mannheim: ***Quantitative properties of case syncretism and valency of Indo-European languages***
Rafał L. Górski & Maciej Eder, Polish Academy of Sciences, Poland: ***Four cases of diachronic change in Polish, and Piotrowski law***
Radek Čech[1], Ján Mačutek[2], Michaela Koščová[2] & Markéta Lopatková[3], 1University of Ostrava, Czech Republic; 2Comenius University, Slovak Republic; 3Charles University Prague, Czech Republic: ***Quantitative analysis of syntactic dependency in Czech***
Jacques Savoy, University of Neuchatel, Switzerland: ***Stylistic evolution of US political speeches***
Ramon Ferrer I Cancho, Universitat Politecnica de Catalunya, Spain: ***Why verb initial languages are less optimized with respect to dependency lengths than other languages?***
Hongxin Zhang, Haitao Liu & Bingli Liu, Zhejiang University, China: ***Distribution of combined structures of parts of speech and syntactic functions***
Andrei Beliankou & Sven Naumann, University of Trier: ***Syntactic complexity of dependency structures***
M. F. Ashurov & V. V. Poddubny, National Research Tomsk State University, Tomsk, Russian Federation: ***Stream-based RF-measure in text classification***

Lukáš Zámečník, Vladimir Matlach & Dan Faltýnek, Palacký University Olomouc, Czech Republic: ***Statistical trends manifested in protein analysis (Case study, QUITA)***

Juhong Zhan & Yue Jiang, Xi'an Jiaotong University, China: ***Application of quantitative linguistic properties to translatorship recognition***

Peter Grzybek & Ernst Stadlober, University of Graz, Austria: ***Why the lognormal distribution seems to be a good model in quantitative film analysis***

Michele A. Cortelazzo[1], Paolo Nadalutti[2], Stefano Ondelli[2] & Arjuna Tuzzi[1], 1Università di Padova; 2Università di Trieste, Italy: ***Authorship attribution and text clustering for contemporary Italian novels***

Elena Yagunova & Tatiana Nikulina, St.-Petersburg State University, Russian Federation: ***Context predictability methods in Twitter social network analysis***

Georgios Mikros, Vassiliki Pouli & Ephtychia Triantafyllou, National and Kapodistrian University of Athens, Greece: ***Personality prediction in Facebook status updates using multilevel Ngram profiles (MNP) and word features***

Katharina Prochazka & Gero Vogl, University of Vienna, Austria: ***Using physics to model language diffusion: benefits and challenges***

Hao Sun & Mingzhe Jin, Doshisha University, Japan: ***Authorship attribution of Yasunari Kawabata's Novels: Who Actually Wrote Otome no minato and Hana nikki***

Patrick Juola[1], Sean Vinsick[1] & Michael V. Ryan[2], 1Duquesne University; 2EthosIO: ***Are most people on Twitter introverts? A distributional analysis of personality types on Twitter***

Martina Benešová & Petra Vaculíková, Palacký University Olomouc, Czech Republic: ***Segmentation for MAL on content-semantic levels***

Sheila Embleton, Dorin Uritescu & Eric S. Wheeler, York University, Toronto, Canada: ***An expanded quantitative study of linguistic vs geographic distance using Romanian dialect data***

Xiaxing Pan[1], Bingli Liu[2] & Xinying Chen[1], 1National Huaqiao University; 2Xi'an Jiaotong University, China: ***Harmony in diversity: language data codes in English-Chinese poetrytranslation - Quantitative study on Shakespearean sonnets translation***

The social program included a guided tour through the historical city center and a wine-tasting evening in a historical subterranean vault.