Glottometrics 40
2018

RAM-Verlag

ISSN 2625-8226

Glottometrics

Indexed in ESCI by Thomson Reuters and SCOPUS by Elsevier

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.

Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.

Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).

Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.

Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies.**

Herausgeber – Editors

External academic peers for Glottometrics

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

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 40 (2018), Lüdenscheid: RAM-Verlag, 2017. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar (Open Access).
Bibliographische Deskription nach 40 (2018) **ISSN 2625-8226**

Contents

Glottometrics 40, 2018, 1-45

On the Self-similarity of Wikipedia Talks: a Combined Discourse-analytical and Quantitative Approach[1]

Alexander Mehler,[2] Rüdiger Gleim, Andy Lücking, Tolga Uslu and Christian Stegbauer

Abstract: Do the talk pages in Wikipedia, referred to as *Wikicussions*, exhibit effects of mass communication? In order to provide an answer to this question, we assess Wikicussions from the point of view of dialog theory and identify characteristics specific to this webgenre. We then show that webgenres of this sort evolve into a state of multidimensional scale invariance that is simultaneously reflected on several syntactic and pragmatic dimensions – irrespective of the underlying topic being discussed and the composition of the underlying community of discussants. We also show that a system exhibiting multidimensional scale invariance interferes with thematic classification. The resulting *confusability* of the gestalt of Wikicussions in terms of their thematic provenance and their underlying participation structure is not just caused by the predominance of small units. Rather it also concerns larger or even largest Wikicussions. According to these findings, we distinguish two sorts of self-similarity of Wikipedia's discussion space: horizontally, regarding thematically demarcated subparts of this space, and vertically regarding the gestalt of top-level sections in relation to Wikicussions. Our analysis is exemplified by means of the discussion space of the German Wikipedia. The results suggest that a quantitative *discourse* analysis of *big dialogical data* as provided by Wikicussions is a promising way to explain the peculiarities of this medium: it can be a starting point for a corresponding theory formation.

Keywords: webgenre, Wikicussion, dialog theory, quantitative discourse analysis, multidimensional scale invariance, self-similarity

1. Introduction

Wikipedia is a genuine webgenre (Santini et al. 2010) that integrates several subgenres such as articles, portals, and so-called talk pages. Talk pages are the subject of this article. They manifest multiparty multi-threaded online conversations to which multitudes of discussants (i.e., prosumers in the sense of Tapscott and Williams (2008)) may participate. Talk pages serve as forums for debating the content of collaboratively written articles in order to improve, for example, their quality – as in the case of task-oriented *article talk pages* (Gómez et al. 2011) – or to communicate self-expression – as in the case of *user talk pages*[3] (Kittur et al. 2007b; Laniado

[1] This article is dedicated to Reinhard Köhler on the occasion of his 65th anniversary.

[2] Text Technology Lab, Goethe University Frankfurt, Robert-Mayer-Straße 10, D-60325 Frankfurt am Main, Germany. Mail: mehler@em.uni-frankfurt.de

[3] Note that self-expression, for example, by means of authority claims also concerns article talk pages (Bender et al. 2011; Oxley et al. 2010; Marin et al. 2011).

et al. 2011; Laniado et al. 2012; Iosub et al. 2014)).[4] Article talk pages serve a wide range of functions in support of coordinating work on Wikipedia (Backstrom et al. 2013). This includes, for example, planning of editing activities, conflict resolution, communicating or negotiating Wikipedia's goals, norms and policies, or extending Wikipedia as a knowledge base or even as a software system (Bryant et al. 2005; Arazy et al. 2011; Viégas et al. 2004; Viégas et al. 2007; Schneider et al. 2011; Schneider et al. 2012). In this way, talk pages transport *social influence* in social communities of online collaborating users in a way that never existed before the advent of this medium: Wikipedia's prosumers build "online communities of practice" (Bryant et al. 2005; Hara et al. 2010) for *knowledge sharing* as well as for *sharing practices of knowledge sharing*. Whereas the shareability (Freyd 1983) of the former kind of knowledge is addressed by article talk pages, the shareability of the latter meta-knowledge is the topic of so-called *Wikipedia* talk pages (Hara et al. 2010).

The status of Wikipedia as a novel webgenre and of talk pages as one of its subgenres is justified in several ways. Researchers claim, for example, that wiki media have fundamentally changed the way people communicate since they affect fundamental processes such as opinion formation and collective problem solving (Wang et al. 2012). Others claim that Wikipedia has changed the status of collective work (Welser et al. 2011) outweighing the work on ancestor genres (e.g., offline encyclopedia). This qualitative innovation is said to be accompanied by a quantitative one regarding the "exponential growth of asynchronous online conversations" (Hoque and Carenini 2015) manifested by media such as Twitter, blogs and talk pages. Unlike face-to-face dialogs or multilogs, online discussions are open in terms of space, time (Kaltenbrunner and Laniado 2012), participation structure and subtopics under discussion though being restricted by the framing topic of the corresponding article.

Wikipedia establishes the largest encyclopedia that ever existed (Iosub et al. 2014) by means of the collaboration of a multitude of editors in a self-organized manner subject to a loose governance (Arazy et al. 2011). As a by-product of writing encyclopedias, this cooperation is also seen as a source for the formation of collective memories (Ferron and Massa 2014). In order to approach these and related goals, Wikipedia has to balance out (i) the needs of a wide range of users regarding (ii) a variety of subgoals subject to (iii) a diversity of boundary conditions thereby entering into fluent equilibria of all included variables (as exemplified by Kittur et al. (2007a) regarding Wikipedia's participation structure):

1. The first range of variables comprises (readers in the role of) so-called *free-riders* (Antin and Cheshire 2010), *lurkers* (Preece et al. 2004), *serendipitous editors* (Antin and Cheshire 2010), *legitimate peripheral participators* (Bryant et al. 2005), *low-edit users* (Kittur et al. 2007a) and *novices* (Bryant et al. 2005; Kittur et al. 2007a; Schneider et al. 2012) as well as high-edit *Wikipedians* (Welser et al. 2011), *specialized experts* (Bryant et al. 2005; Kriplean et al. 2008; Iosub et al. 2014; Arazy et al. 2011) reflecting labor division (Kriplean et al. 2008) and *elite users* (Kittur et al. 2007a). These types of users – henceforth subsumed under the notion of *Wikipedians* – span a social hierarchy (Iosub et al. 2014) undergoing Wikipedia's socialization process (Arazy et al. 2011).
2. The range of variables relating to subgoals includes but is not limited to securing *validity*, *veridicality* or *objectivity* (Arazy et al. 2011), *reliability* or *coherence*, *coverage* or *completeness* (Arazy et al. 2011), *readability* (Kaltenbrunner and Laniado 2012), *transparency* or *manageability*, *flexibility* and *extensibility* or *openness*.

[4] Beyond these two namespaces of talk pages, Hara et al. (2010) distinguish seven additional such spaces in Wikipedia (cf. Viégas et al. 2007).

3. Finally, boundary conditions are exemplified by *infrastructure* (Welser et al. 2011), *development status of the wiki software* (Viégas et al. 2004), *social participation structure* and the *change of the world* as a source of ever new topics, which are described in articles or discussed in corresponding talk pages.

The process of balancing out these factors is necessarily accompanied by conflicts among users affected by conflicting needs, degrees of expertise, divergent social roles[5] and statuses within and outside of Wikipedia (Kittur et al. 2007b; Arazy et al. 2011; Marin et al. 2011; Welser et al. 2011). The literature lists various processes of cooperation and competition which aim to solve or stimulate conflicts during the build-up of Wikipedia (Kittur et al. 2007b; Kaltenbrunner and Laniado 2012). These processes range from *direct user communication* (e.g., by means of suggestions or complaints within talk pages), *maintenance work* (e.g., by combating vandalism or by making reverts (Kittur et al. 2007b)) and *implicit coordination* to the *conventionalization* (Lewis 1969) of rules and procedures of content production by *explicit coordination* between privileged editors.[6] According to Kittur and Kraut (2008), implicit coordination is illustrated by situations where articles are written by small groups of editors who are not explicitly coordinated, while all other authors act as uncoordinated supporters. On the other hand, conventions resulting from explicit coordination relate to requirements of the sort that contributions to talk pages should be signed by their authors.

In general, Wikipedia's guidelines specify that talk pages should contribute to improving the corresponding articles. To this end, the German Wikipedia requires, for example, that discussions should be structured in a way that secures their comprehensibility. Technical instructions are given to enable participants to better follow such conventions, while the interface for writing talk pages does not impose too many restrictions to support these conventions (cf. Backstrom et al. 2013). In any case, the official purpose of article talk pages is not to serve as platforms for personal conversations or self-portrayal. However, one may observe such rule-breaking behavior on the side of individual interlocutors. In this sense, conflicts can be expected between the conventions negotiated at the community level and compliance with these rules at the level of individual users.

A conventional system as manifested by Wikipedia and its implementation by the underlying community can be seen as a "negotiated culture". (Stegbauer 2016). Such a system is affected by a range of social (e.g., acceptability of rules), situational (e.g., thematic salience) and cognitive restrictions (e.g., time pressure) regarding the contributions of ever new interlocutors joining the talks. Concerning the formation of Wikipedia's subgenres, one can detect several macroscopic effects of collaboration and competition among Wikipedians under such restrictions. This concerns, for example, the reduction of work on article content in favor of work on administrative or talk pages (Kittur et al. 2007b; Kaltenbrunner and Laniado 2012). In addition, a kind of functional diversification is observed in relation to the development of subgenres. This applies in particular to talk pages, which are not only intended to ensure the quality of articles, but also serve as a platform for discussing or disseminating policies in support of community building (cf. Schneider et al. 2011). A third example is given by Schneider et al. (2012) who distinguish specialized discussions about *Articles for Deletion* as a sort of a second-order subgenre. *Microscopic* counterparts of these macroscopic effects concern the syntactic, semantic, pragmatic and temporal structure and dynamics of single talk pages (Kumar et al. 2010).

The duality of macro- and microscopic diversification is mirrored by processes of social

[5] For example, administrators, vandal fighters or social networkers (Welser et al. 2011).

[6] Note that detecting conflicts may help identifying related controversial articles because of a sort of assortative mixing (Newman 2002) among controversial pages (Dori-Hacohen et al. 2016).

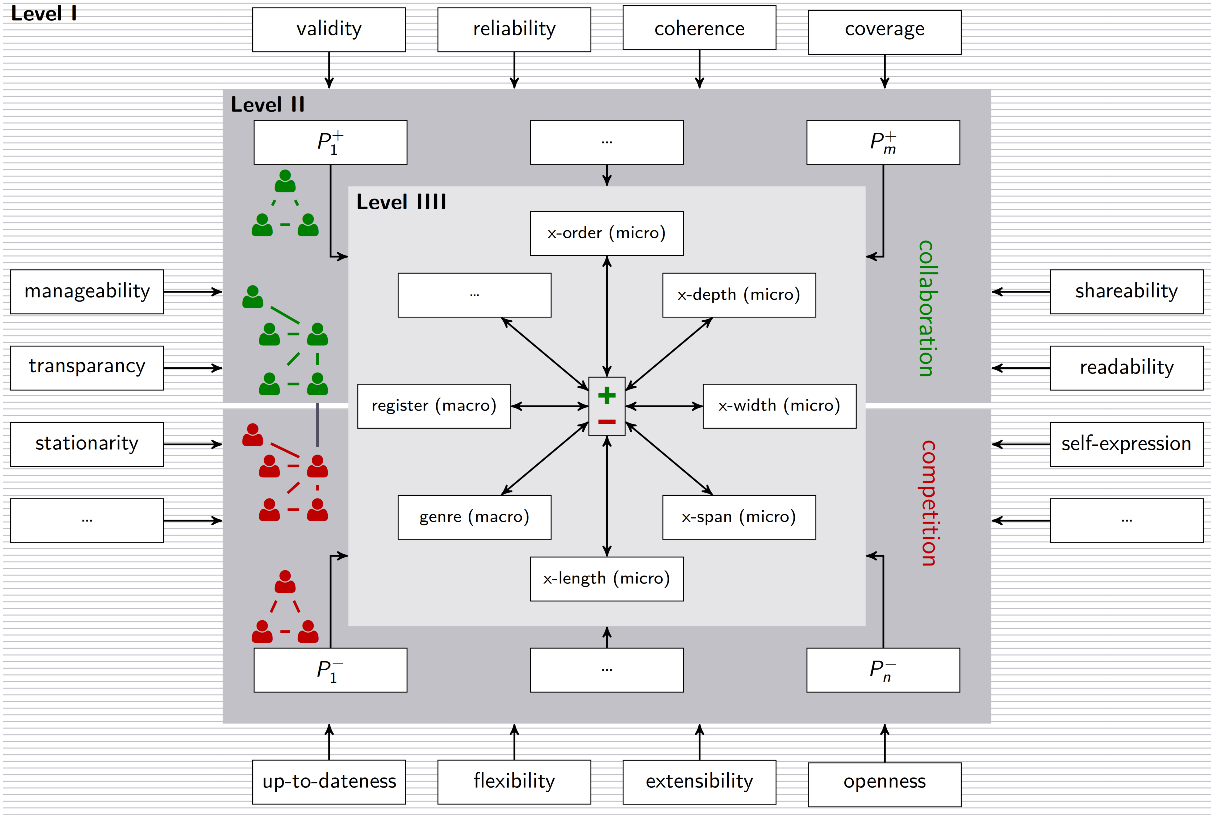

Figure 1. A three-level synergetic model of structure formation manifested by talk pages. *Order*, *depth*, *width*, *span* and *length* denote characteristics of tree-like structures of talk pages (cf. Section 4). Macro-level effects (regarding, for example, the diversification of subgenres) are distinguished from their micro-level counterparts (regarding the gestalt of single pages). P_i^+ and P_j^- denote collaboration and competition processes, respectively, whose supporters participate to interaction networks (see Section 2).

differentiation regarding the roles and statuses of Wikipedians (Arazy et al. 2017). On the highest level of resolution, this concerns the distinction between content- and administration-related users who compose heterogeneous teams of editors and posters. Arazy et al. (2011), for example, show that role membership correlates with the degree of activity in writing pages of certain type(s) (i.e., subgenres in the sense described above). It can also be shown that the composition of the group has an influence on conflict resolution, which ultimately determines the quality of cooperation.

In accordance with our brief introduction to talk pages one can distinguish three levels of independent variables (i.e., enslaving order parameters in the sense of synergetics (Haken 1998; Köhler 1993)) on the one hand and (enslaved or) dependent variables of structure formation in Wikipedia on the other. This scenario is depicted in Figure 1:[7] on the external level of order parameters – called *Level I* – we distinguish system requirements (or needs) that Wikipedia is supposed to meet for its user community. This concerns the subgoals enumerated on page 2 above. On the inner *Level III* we localize the structure and dynamics of Wikipedia's subgenres. More specifically, we distinguish the formation of subgenres as a whole from laws of structure formation within single instances of these genres. Both layers are mediated by a mid-layer – called *Level II* –, which is established by (a mixture of) user(group)-related processes of

[7] See Arazy et al. (2011) for a related model of structure formation on two levels.

conflict prevention-oriented collaboration and conflict stimulation-oriented competition. We assume that Level II is structured by the affiliation of agents to social roles and the formation of groups and networks of users who collaborate or compete in writing pages.

Under this regime, the question arises whether the structure of talk pages manifests a fluent equilibrium of partly competing requirements operating as order parameters on Level I (see Figure 1) in such a way that the pages' shape is distributed in a law-like manner, which neither depends (i) on the subject under discussion nor (ii) on the concrete composition of the underlying group of participants. The shape of talk pages could then be influenced by a distribution of the degree to which Wikipedia's rule system is followed by Wikipedians, which guarantees sufficient participation opportunities for new users and at the same time keeps the entire system clear and manageable. As a result, interlocutors could expect to take part in discussions on an individual basis, while actually behaving in such a way that their contributions become similar (in a way to be defined in this article) beyond the boundaries of both the (i) thematic structure and the (ii) participation structure of talk pages. If this is true, we expect Wikipedia to behave in a self-similar way at the level of talk pages (Level III) as a result of a compromise between divergent requirements on Level I (see Figure 1) and their mediation through processes of collaboration and competition on Level II.

The analysis of the law-like shape of talk pages and its relation to the latter hypothetical equilibrium is at the core of the present article. This equilibrium state is likely affected by at least two boundary conditions: firstly, at any point in time, a minority of topics is highly salient while the majority of them is peripheral. Consequently, we expect a power law-like distribution of thematic salience. This assumption relates to the power law of semiological preference (cf. Tuldava 1998). Secondly, at any point in time, only a small fraction of discussants is highly active while the majority of them tends to submit very few posts – this assumption relates to the power law scaling of human behavior (Wang et al. 2012). Both restrictions may finally result in a skewed distribution of the gestalt of talk pages that below will be modeled in terms of multidimensional power laws. In summary, we provide results regarding the following areas:

1. *Semiotic modeling:* We provide a bimodal model of structure formation in article talk pages based on *syntactic* and *pragmatic* features. By analogy to Kittur et al. (2007b) and many others (see Section 2), this is done by reusing or inventing easy-to-compute quantities that scale well on datasets as large as those derived from Wikipedia.
2. *Multidimensional scale invariance:* In line with many approaches to structure formation in webgenres, we provide evidence of scale invariance. However, unlike these approaches we systematically draw on our bimodal model by showing that syntactic and pragmatic structures coincide in terms of their scale-freeness. That is, talk pages exhibit a kind of multidimensional scale invariance that is accompanied by a statistical dependence of the random variables being involved. In this sense, one knows a lot about the rank of a discussion according to one feature if knowing its rank according to some other features of the same model. In other words, in webgenres such as talk pages, random variables of syntactic structure and pragmatic participation structure are mutually informative.
3. *Vertical self-similarity:* Since talk pages are subdivided into threads discussing different aspects of the same article, one may ask whether the latter finding also holds for structures spanned by the top-level sections of talk pages. By showing that multidimensional scale invariance is also prevalent on this constituent level, we demonstrate that talk pages exhibit a kind of *vertical* self-similarity: *top-level sections are similar to the talk pages to which they contribute.*
4. *Horizontal self-similarity:* Next, we show that multidimensional scale invariance makes

the subject area of talk pages undetectable. That is, we observe self-similarity with respect to thematic partitions of Wikipedia's discussion space: by knowing the structure of a discussion in terms of our bimodal feature model, one does not know the corresponding thematic area. In other words, discussions in Wikipedia take place on an equal footing *irrespective of (i) the subject area under discussion and (ii) the composition of the underlying group of posters*. Though being structurally indistinguishable, discussions are well separable in terms of the underlying subject area. To demonstrate this, we build a neural network language model of text vocabularies and show that it allows for classifying articles and discussions according to their topics. This experiment shows that the structure of talks in Wikipedia is self-similar, though the corresponding vocabularies used to manifest the topics under discussion are not.

5. *Bridging webgenre analysis and dialog theory:* Talk pages are a rather new genre of dialogical communication. Thus, differences of face-to-face communication and online discussions, which make time and space dispensable variables of communication, have not been systematically described so far. To address this gap, we additionally provide an assessment of talk pages, henceforth alternatively called *Wikicussions*, from the point of view of dialog theory. This will finally allow us to explain multidimensional scale invariance as a consequence of the fact that Wikicussions externalize common ground in a way that allows for joining the same talk irrespective of time, space, topic and participation structure.

The article is organized as follows: Section 2 reviews related work. In Section 3, we characterize talk pages in terms of theories of dialogical communication. In Section 4, we present our model of discussions as manifested by talk pages. In this context, we introduce a template model of syntactic, semantic, pragmatic and temporal structure formation. In Section 5, we instantiate this model on the level of syntax and pragmatics, present our corpus of more than 600,000 discussions and describe its preprocessing in terms of *Natural Language Processing* (NLP). Section 6 describes our findings which are discussed in Section 7. Finally, Section 8 gives a conclusion and looks at future work.

2. Related Work

Talks in online media such as Twitter, blogs or Wikipedia have been object of a range of approaches to *Quantitative Discussion Analysis* (QDA). This concerns the syntactic, semantic, functional and temporal dynamics of online talks. A fifth research perspective relates to generative models of random trees approximating the structure of real talks. Further, talk pages have been used to derive social interaction networks of discussants to assess their collaboration. Alternatively, one analyzes the editing or posting behavior of users to predict their social roles.

Syntactic content, for example, is the object of the study of Laniado et al. (2011) who compute statistics of talk pages of the English Wikipedia. By applying a range of measures of tree-like structures, they observe power law-like distributions of order and depth (cf. Section 5.4) thereby hinting at a sort of syntactic self-similarity. Gómez et al. (2011) observe order distributions that fail to be adapted by power laws. See also Wang et al. (2012) who alternatively adapt the log-normal and the negative binomial model to the size distribution of posts. Another kind of content analysis is provided by Laniado et al. (2012) and by Iosub et al. (2014) who perform sentiment analyses of the lexical content of talk pages using prior polarities of words to analyze social roles and statuses of Wikipedians and gender-related characteristics. Iosub et al. (2014) measure, for example, assortative mixing (Newman 2002) of editors and posters according to their sentiment profiles.

A central topic of QDA regards the question whether the course of a discussion depends on the underlying subject area. Kaltenbrunner et al. (2009) focus on this question by example of discussions in an online forum. Using a confusion matrix of the depth and width (cf. Section 5.4) of tree-like structures, they observe a contingency between topic and a four-part classification of trees. They also observe temporal habits regarding the participation of interlocutors depending on the key subject (e.g., politics). Laniado et al. (2011) describe a related association between topic and structure. They give evidence that the gestalt of a talk page (in terms chaining, depth etc.) partly depends on its topic (classified according to 21 macro-categories of Wikipedia's category system). Another kind of content dependence is considered by Gonzalez-Bailon et al. (2010) who observe that political discussions tend to be wider and deeper (in terms of the trees spanned by their posts). The temporal dynamics of talk pages and its relation to the dynamics of edits in the underlying articles is addressed by Kaltenbrunner and Laniado (2012). Characteristics of waiting times between consecutive posts are computed by Wang et al. (2012) though not by example of talk pages.

Another key topic of QDA concerns the functional dynamics of online discussions. Viégas et al. (2007) provide a functional analysis of talk pages by reference to eleven functions, where *request for coordinating editing activities* is the most frequent one. Schneider et al. (2011) contrast this list with five categories of talk pages. While Viégas et al. (2007) analyze talk pages in terms of macro functions, a rather micro-functional approach is presented by Bender et al. (2011) who distinguish authority claims (cf. Oxley et al. 2010) from alignment moves as two dialog acts concerning self-presentation and expression of alignment among interactants. Marin et al. (2011) focus on a single type of authority claim described by Viégas et al. (2007) to present a model for automatically detecting its instances on the level of sentences and posts. This approach gives rise to automatically segmenting talk pages by means of machine learning. In line with this approach, Ferschke et al. (2012) tag dialog acts within talks of the Simple English Wikipedia.

Unlike the approaches reviewed so far, so-called generative models aim at providing a *tertium comparationis* for testing hypotheses about structural peculiarities of online discussions. Gómez et al. (2011) introduce such a model in terms of a variant of the preferential attachment model for generating random trees. An alternative model of the growth dynamics of conversation threads is proposed by Wang et al. (2012). It describes the probability by which a new post replies to a given one as a function of the overall number of all replies. See also Kumar et al. (2010) for a growth model of discussion threads.

Talk pages allow for deriving so-called (*user*) *interaction networks* as 2nd order observation units. These are graphs in which social entities (e.g., prosumers) are represented by nodes whose edges denote social relations among the entities under consideration (Qin et al. 2015; Yasseri et al. 2012; Iosub et al. 2014; Stegbauer and Bauer 2010). A relation may be established, for example, if one poster replies to the post of another one. Further information can be explored to weight or attribute nodes and edges. Iosub et al. (2014), for example, explore sentiment data to color the nodes of reply-based user networks. In this context, one distinguishes two approaches: *local models* concern network representations of the interaction structure among posters of the same discussion while *global models* aim at network representations of the interaction structure of several discussions or of the discussion space as a whole (cf. Cogan et al. 2012).

By exploring the editing/posting behavior of users one can predict their social roles as done, for example, by Welser et al. (2011). Analogously, Kittur et al. (2007b) and Kittur and Kraut (2008) study the impact of implicit/explicit coordination on article quality and on the organization of subgenres. In this context, they attribute a major role to talk pages: discussions within such pages range from single expressions to complex processes of finding a consensus on the scope of the corresponding article. Their findings indicate that explicit coordination within talk pages is manifested by rather small groups of specialized agents thereby hinting at the need to elaborate fine-grained classifiers for automatically attributing social roles to Wikipedians. This research bridges between QDA and computational sociology for which media such as Wikipedia are still the first choice.

Note that while the overwhelming majority of approaches reviewed so far focuses on a single language, namely English, Hara et al. (2010) provide a cross-cultural analysis of article and user talk pages by example of four languages, thereby extending the beaten tracks. In our case, this happens by example of the German Wikipedia.

3. Online Discussions from a Dialogical Point of View

Though functional analysis (e.g., by means of dialog act tagging) is a central task of QDA, theoretical assessments of the status of online discussions in relation to face-to-face communication are hardly found in the literature. In this section, we provide such a comparative analysis by classifying written discussions that are technically mediated by talk pages in the context of linguistic theories of dialog. Henceforth, we call this kind of discussions *Wikicussions*. The most striking difference between Wikicussions and face-to-face dialogs is that the former are not regimented by the immediate language action and perception cycle being constitutive for the latter (Pickering and Garrod 2013). The separation of production and comprehension within postings implies that phenomena which are intricately bound up with interactivity and forward modeling in language use differ between Wikicussions and dialog. We briefly highlight five issues in this regard, namely *incremental multi-speaker utterances*, the *build-up of common ground*, *task-orientedness*, *multilogs*, and *strategic conversations*, thereby pointing out differences and commonalities between Wikicussions and dialog.

To begin with, dialog is characterized by phenomena such as (i) multi-speaker utterances (Poncin and Rieser 2006) or collaborative turns (Lerner 2004), (ii) predictions concerning upcoming speech behavior (Kutas et al. 2011) and end of turns (Ruiter et al. 2006), (iii) fragments and non-sentential speech (Fernández and Ginzburg 2002). Except for elliptical speech, Wikicussions are devoid of these examples which all point to units of dialog smaller than turns or utterances, known as *micro conversational events* (Poesio 1995; Poesio and Traum 1997). Dialog theories aim at spelling out a "grammar of interaction" in order to account for the tight dialogical coupling at the micro conversational level (Ginzburg and Poesio 2016; Kempson et al. 2016), and draw on psycholinguistic research on dialog processing (e.g., Tanenhaus et al. 1995; Pickering and Garrod 2013). Wikicussions, on the contrary, do not exhibit incremental interaction at the micro conversational level, but operate on the level of postings or turns (Sacks et al. 1974).

In contrast to the ephemeral nature of spoken dialog, however, postings within Wikicussions remain visible and are even archived. This has repercussions on the structure and content of *Common Ground* (CG) (Stalnaker 2002; Clark 1996). During conversations, interlocutors build up a body of agreed knowledge and enter conversations with shared backgrounds. This CG has public and private shares. Public CG is not only established incrementally, but each conversational participant develops her own take on the publicly available information (Poesio and Rieser 2010). Technically speaking, interlocutors are assigned their own dialog game boards each (Ginzburg 1994). Being accessible permanently, Wikicussion postings give rise to a different model of public CG: CG is abstracted from the conversational participants; familiarity with CG contents is steadily licensed by perceptual access (Clark et al. 1983). In other words, instead of a perspectival dialog game board for each interlocutor, a Wikicussion can be associated with a single dialog game board. A consequence being that CG is externalized and available to an in principle unlimited audience (although each discussant may have a private CG as well). This is in sharp contrast to dialog, where accessing (memorized aspects of) CG is restricted to the (memory of) authoring interlocutors. The dialog history, therefore, is accessible to the public and for that reason every stage of that history provides a possible entry point for third-parties. In this sense, Wikicussions are not bound to the inevitable progressing nature of dialog, as reflected in information-state update models (Traum and Larsson 2003), which can refer to, but not go back to previous states in dialog history.

The publicity of dialog history leads to issues of multi-speaker dialog, or *multilog*. As Dignum and Vreeswijk (2004) argue, multilog is not just a number of dialogs running in parallel. Already for combinatorial reasons, multiperson communication gives rise to a richer variety of participant roles, including, for instance, group addresses, overhearers and bystanders (Goffman 1981). Most of these roles have a direct correspondence in Wikicussions (e.g., overhearers become "overreaders"), but their technical underpinning also gives rise to special roles such as *administrator*. In any case, however, the interaction protocol of multilog has to acknowledge multi-party addressing and provides more roles than *speaker* and *hearer*. Accordingly, scaling up from dialog to multilog involves taking multi-party addressing and different roles into account. This mainly affects grounding: a Wikicussion initiating posting brings up a *Question Under Discussion* (QUD) (Ginzburg 2012; Roberts 2012) with respect to the wording, statement, or evaluation of a section of the associated article. QUD downdating has to be correctly distributed over the Wikipedians. A main complication in this respect is that participants in different roles as well as participants conjoining in various coherent collectivities or coalitions build up different shared contexts (Schober and Clark 1989; Lerner 1993). In such groups of participants, grounding may distribute over each group member and a single group member may act as a grounding "proxy" for its coalition (Eshghi and Healey 2016). While in face-to-face dialog the dynamics of such coalitions is interleaved with the progredient nature of interaction in time, in Wikicussions also "retroactive" context sharing is apparent, e.g., when a user conjoins a previous coalition representing a certain position at a (much) later time. As a consequence, multilogs give rise to different CG structures compared to dialog (Ginzburg and Fernández 2005). Speaking in terms of dialog protocols, multi-party addressing gives rise to multilog histories that exhibit the structure of a directed acyclic graph rather than of a tree (cf. Fernández and Endriss 2007). Despite lending themselves for multilog conversations due to their publicity, Wikicussions are not able to capture such graph structures *simply for technical reasons*. The wiki software does not allow for multi-party addressing by single posts. If we do not find Wikicussion multilogs in this sense, this reflects technical circumstances rather than pointing at a dialog *A Priori*.

By and large then, resolving a QUD is the driving force of a Wikicussion in the first place: Wikicussions follow a specific task and being task-oriented further distinguishes them from dialogs that not only are usually initiated by greeting and counter-greeting moves, but also are thematically open (think of chatting and small talk). Above all, however, every participant of a Wikicussion is assumed to sincerely intend to contribute to resolving an underlying QUD as reflected in Wikipedia's Wikiquette. This prerequisite can be questioned, of course. For drawing implicatures from postings or dialog utterances, credibility and coordination of intentions about the conversational goals of the interlocutors have to be warranted (Grice 1975). Worries in this regard are bound up with so-called *strategic conversations* (Asher and Lascarides 2013). These are conversations where participants entertain different conversational intentions, although they may cooperate rhetorically. Common examples of strategic conversation are court hearings or political debates. Since according to the Gricean view of communication, interlocutors infer indirect meanings partly by recognizing the other's intention, misdirecting intentions of strategic speech violate for instance sincerity conditions underlying implicatures. Since we do not know which discussants in Wikicussions or in dialogs have strategic aims or intend proper QUD resolving, there is the risk of drawing implicatures that are not *safe* (Asher and Lascarides 2013) in both cases.

In sum, Wikicussions differ from dialogs in not being subject to the incremental interactivity at the micro conversational level, exclusively being QUD-driven and task-oriented, and in giving rise to a public and externalized common ground abstracted away from the authoring participants. Both Wikicussion and dialog may be involved in strategic conversations which can be detected only with reference to speakers' intentions. Further, both Wikicussion and dialog scale up to multilog in principle while talk pages fail to manifest acyclic graph structures due to rather contingent technical reasons.

4. On the Logical Document Structure of Wikicussions

In this section, we account for the tree-like structure of Wikicussions on article talk pages in terms of graph theory. This graph model will be used later on to quantify the structure of Wikicussions and to classify them accordingly.

Building blocks of talk pages are *sections* and *posts* (cf. Backstrom et al. 2013) (also called *turns* (Marin et al. 2011), *comments* or *replies* to precedent posts) partly entering into adjacency pairs and finally spanning tree-like structures.[8] While posts are normally signed by their author and the date and time of creation, for sections such assignments are only indirectly accessible via the revision history. In order to measure statistical characteristics of discussions, we bijectively map each article onto a single tree-like representation comprising all its talk pages. Generally speaking, the same *article page* can be related to several *article talk pages*. Among the latter pages, a single page contains the article's latest discussions while all other pages are "archived" as sub-pages (cf. Laniado et al. 2011). Archived talk pages collect older, so to speak, "outsourced" threads as part of an article's overall discussion. Each of these talk pages – whether archived or not – that belongs to the same article will be mapped onto a single representation of the underlying Wikicussion. The reason for this approach is to get an overall picture of all threads debating the same article. Note that we characterize Wikicussions by additionally drawing on their top-level sections as dominating separate conversations (for the notion of conversation in this context cf. Marin et al. 2011).

We use graph theory for a formal treatment of the structure of Wikicussions. More specifically, we utilize the notion of a *rooted ordered directed tree* to model their document structure.[9] Let $D = \{d_1, \ldots, d_m\}$ be a corpus of Wikicussions and $d \in D$. Each discussion $d \in D$ is represented as a tree such that $T_D = \{T_{d_1}, \ldots, T_{d_m}\}$ is the set of the resulting tree representations. For any $d \in D$, T_d is defined as follows (below we apply this definition to top-level sections of discussions):

[8] Our terminology departs from related work which distinguishes the initial post or *conversation root* (Cogan et al. 2012) from subsequent comments. We subsume both under the notion of a post.

[9] See Laniado et al. (2011) and Kaltenbrunner and Laniado (2012) for a reference model of this approach. For an alternative model based on forests see Krishnan et al. (2016).

$$T_d = (V_d, A_d, r_d, \mathrm{author}_d, \mathrm{content}_d, \mathrm{ord}_d, \mathrm{signature}_d) \tag{1}$$

- r_d is the root of T_d so that both can be used interchangeably to denote d.
- V is partitioned into three non-overlapping subsets $\{r_d\}$, $V_d^p = \{v_{i_1}, \ldots, v_{i_j}\}$ of vertices denoting posts and $V_d^s = \{v_{i_{j+1}}, \ldots, v_{i_n}\}$ of vertices denoting sections. That is, we assume a bijection between posts in d and elements of V_d^p as well as between sections in d and elements of V_d^s.
- For any pair of adjacent posts $v, w \in V$ for which w replies directly to v we generate an arc $(v, w) \in A$. Exceptions to this rule are induced by the elements of V_d^s which are processed as follows: (1) for each top-level section $v \in V_d^s$, we generate arcs of the sort $(r_d, v) \in A$; (2) for all sections $w \in V_d^s$ directly dominated by some $v \in V_d^s$, we generate arcs of the sort $(v, w) \in A$; (3) for all posts $w \in V_d^p$ directly dominated by $v \in V_d^s$, we generate $(v, w) \in A$; (4) for all top-level posts $w \in V_d^p$ directly dominated by d (i.e., by the html-h1 element of the respective (archival) site), we generate arcs of the sort $(r_d, w) \in A$. That is, r_d dominates all top-level posts and sections (all of depth 1).
- Based on these preliminaries, we introduce several auxiliaries: $L(T_d) \subset V$ is the set of all vertices v of $\mathrm{outdegree}(v) = 0$. $\mathrm{depth}(v)$ is the length of the shortest path from r_d to $v \in V$. $N_i(v) = \{w \in V \mid \delta(v, w) = i\}$ is the set of all vertices of equal shortest distance $\delta(v, w) = i$ from v, called the ith *neighborhood* of v in T_d. $N_i^p(v) = N_i(v) \cap V_d^p$ restricts this neighborhood to posts. For any $i \in \mathbb{N}$, for which $N_i(v) \neq \emptyset$, we call i the ith *level* of the subtree of T_d rooted by v. Note that $N_0(r_d) = \{r_d\}$.
- $\mathbb{P}(T_d)$ is the set of all *paths* in T_d. For any path $P = (v_{i_1}, \ldots, v_{i_j}, \ldots, v_{i_k}) \in \mathbb{P}(T_d)$, $\mathrm{in}(P) = v_{i_1}$ is called the *start* and $\mathrm{out}(P) = v_{i_k}$ the *end vertex* of P. Further, $v_{i_1}, \ldots, v_{i_{j-1}}$ are *predecessors* and $v_{i_{j+1}}, \ldots, v_{i_k}$ are *successors* of v_{i_j}. We say that v_{i_j} *is dominated by* any of its predecessors. We write $(v_{i_{j-1}}, v_{i_j}) \in P$ or $v_{i_j} \in P$ to denote arc- or node-related constituents of P; that is, $\forall (v_{i_{j-1}}, v_{i_j}) \in P : (v_{i_{j-1}}, v_{i_j}) \in A$. $\mathrm{length}(P)$ is the number of arcs $(v_{i_{j-1}}, v_{i_j}) \in P$. For any pair of vertices v, w, for which w is dominated by v, $P(v, w)$ denotes the unique path in T_d from v to w. A *thread* $P \in \mathbb{P}(T_d)$ is a path starting in r_d and ending in some leaf $\mathrm{out}(P) \in L(T_d)$.[10] The set of all threads of T_d is denoted by $\mathbb{T}(T_d) \subseteq \mathbb{P}(T_d)$. Finally, by $\mathrm{lcp}(v, w)$ we denote the *lowest common predecessor* of $v, w \in V$, that is, the highest-level predecessor dominating v and w (Mir et al. 2013).
- Since this paper concentrates on the syntactic and pragmatic structure of Wikicussions, we do not give a formal definition of $\mathrm{content}_d$.[11]
- $\mathrm{ord}_d \subseteq V_d^2$ defines a total order among the children $N_1(v)$ of all vertices $v \in V_d$ reflecting the vertical ordering of sections and posts in talk pages and the temporal ordering of archived pages.
- A post is not necessarily signed by a signature informing about its author and the date of its creation or of its last change (see Figure 2 for a visual depiction of such a scenario). In order to distinguish between signed and unsigned posts, we partition the range of the authorship function

$$\mathrm{author}_d \colon V_d^p \to \mathcal{A}(D) = \mathcal{A} \cup \mathcal{I} \cup \{?\} \tag{2}$$

[10] This notion departs, for example, from Backstrom et al. (2013) who define a *full thread* to include the initial post together with all dominated comments. See also Kumar et al. (2010) who define threads to be trees.

[11] This will be done in a follow-up paper.

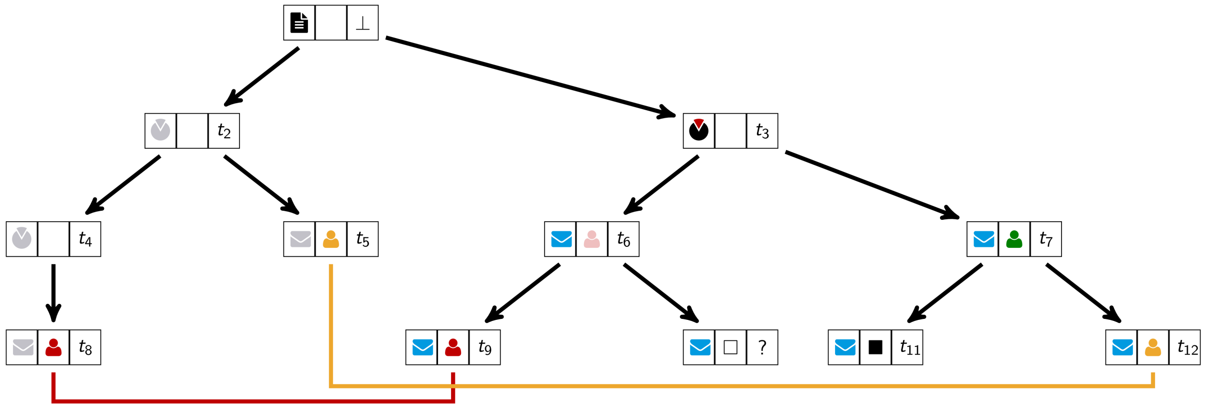

Figure 2. Schematic depiction of a Wikicussion including threads from archived talk pages (indicated by grayed circular segments ◕ dominating archived posts ✉). Each node is represented by a tripartite vector: the first dimension denotes the type of content unit (i.e., a section ◕ or a post ✉), the second the signing author 👤 (where posts by the same author are denoted by same-color icons such that the corresponding vertices are connected by same-color edges), the third dimension codes the time of creation of the section or post (where ? characterizes unsigned (leaf) nodes of unknown posting time). Unregistered users are denoted by unfilled boxes □, while bots are denoted by filled boxes ■. ⊥ denotes the time of creation of the corresponding Wikicussion.

mapping posts onto the names of their authors as part of their signature:

$$\mathrm{signature}_d \colon V_d^p \to (\mathcal{A}(D), \mathbb{N}) \tag{3}$$

Since this paper does not consider the temporal structure of Wikicussions we do not consider the time values of signatures.[12] In the case of posts $v \in V_d^p$ tagged by the proper name x and the time string y, we set $\mathrm{signature}_d(v) = (x, y)$ and $\mathrm{author}_d(v) = x \in \mathcal{A}$ where y codes the time at which v was posted or last changed. If v is an unsigned post tagged by $\mathrm{signature}_d(v) = (x, y)$ such that x is an IP address of the set $\mathcal{I}$ of all IP addresses used to tag posts in corpus D, we set $\mathrm{author}_d(v) = x \in \mathcal{I}$. In the case of anonymous posts v, for which we assume signatures of the form $\mathrm{signature}_d(v) = (?, ?)$, we set $\mathrm{author}_d(v) = ?$.

- Next, we apply this apparatus to *Top-Level Sections* (TLS) $s \in S = \{s \mid \exists T_d \in T_D \colon T_d = (V_d, A_d, \ldots) \wedge s \in V_d^s \wedge (r_d, s) \in A_d\}$. That is, by

$$T_s = (V_s, A_s, r_s, \mathrm{author}_s, \mathrm{content}_s, \mathrm{ord}_s, \mathrm{signature}_s) \tag{4}$$

we denote the tree-like representation of the TLS s which we compute by analogy to T_d. Finally, $S_D = \{T_s \mid s \in S\}$ is the set of all tree-like representations of TLS in corpus D. See Figure 2 for a depiction of the model introduced so far.

As we model Wikicussions and their top-level sections as trees, we can now introduce a tree-related *Feature Template* (FT) regarding different meta-dimensions of *syntactic* (σ), *semantic* (μ), *pragmatic* (π) and *temporal* (τ) structure formation.[13] To this end, we distinguish the syntactic, semantic, pragmatic and temporal *order* (as a function of V), *depth* (as a function

[12] This will be the object of a follow-up paper.

[13] The first three of these dimensions coincide with Morris' (Morris 1938) triadic sign model: while syntactics considers relations of signs among each other, semantics is concerned with signs in relation to their meanings. Finally, pragmatics focuses on signs in relation to their users.

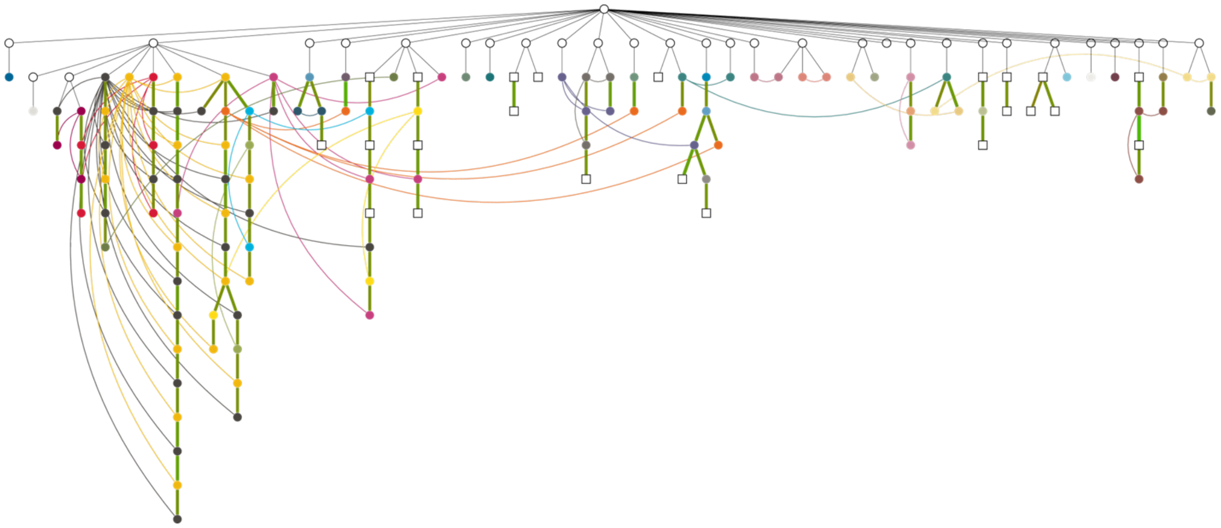

Figure 3. Visual depiction of the Wikicussion of the German Wikipedia article about *Hillary Clinton* represented as a tree extending the method of (Pascual-Cid and Kaltenbrunner 2009; Laniado et al. 2011) by including three additional types of information: (1) vertex color distinguishes users as well as types of nodes: users are depicted by colored circles, sections by white circles, unregistered users by unfilled boxes and bots by filled black boxes. (2) Saturation of edge color signals semantic similarity of adjacent posts: the more similar the reply, the "greener" the line; the less similar, the redder the line. Semantic similarity is measured by means of word embeddings (Mikolov et al. 2013) computed over the complete space of articles and discussions in the German Wikipedia. (3) Lateral edges interlink posts of the same user. (See also (Weninger et al. 2013) for using color to code information within tree-like representations of threaded discussions.)

of T_d's threads), *width* (as a function of the neighborhoods N_i), *level* (as a function of the levels spanned by the N_i's), *span* (as a function of the set $N_1(r_d)$ of child nodes), and *length* (as a function of the set $L(T_d)$ of leaf nodes). Thus, for the operator variable $x \in \{\sigma, \mu, \pi, \tau\}$, template FT takes the following form:

$$FT[x] = (\text{order}[x], \text{depth}[x], \text{width}[x], \text{level}[x], \text{span}[x], \text{length}[x]) \tag{5}$$

For a given x, $\text{order}[x]\colon T_D \cup S_D \to \mathbb{R}, \ldots, \text{length}[x]\colon T_D \cup S_D \to \mathbb{R}$ are functions operating on the order, depth, …, and length of their tree-like operands to calculate x-related quantities for them. For reasons of simplicity, we denote $\text{order}[x]$, $\text{depth}[x]$ etc. by x-order, x-depth etc. In this way, we can speak, for example, of the (pragmatic) π-order of a discussion, its (syntactic) σ-depth or its (temporal) τ-span. Since in this paper we focus on syntactic and pragmatic features, we get a 12-dimensional vector for quantifying the gestalt of Wikicussions and top-level sections starting from the six-dimensional feature template FT. The corresponding feature set, which will be extended in Section 5.4 and 5.5 to map additional quantities of the gestalt of discussions, is denoted by

$$F = \{X_1, \ldots, X_n\},\ n \geq 12 \tag{6}$$

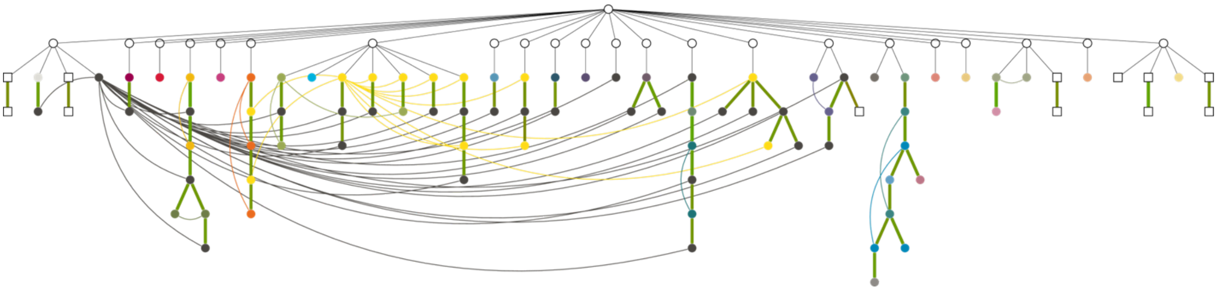

Figure 4. Depiction of the Wikicussion of the German Wikipedia article about *Donald Trump*.

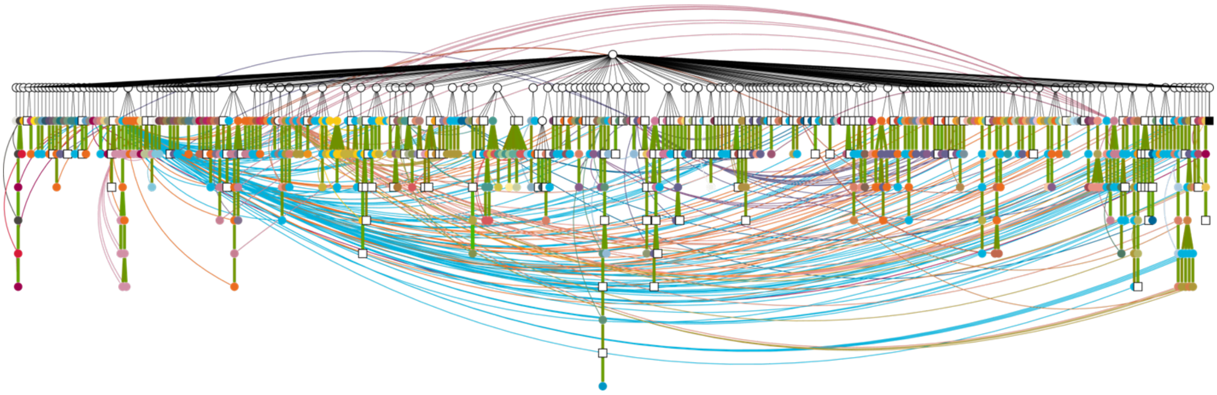

Figure 5. Depiction of the Wikicussion of the German Wikipedia article about *Europa*.

5. Data and Method

5.1. Data

We instantiate the Wikicussion model of Section 4 by means of all article talk pages of the German Wikipedia. To this end, we explore the XML dump from 2016-02-03.[14] Our corpus consists of 710,995 talk pages corresponding to 687,888 articles. In the case of 12,676 articles, multiple (partly archived) talk pages exist all of which are integrated into the corresponding Wikicussion. We call this corpus of 687,888 data units *Corpus of geRman Wikipedia Discussions* (CRoWD). See Figure 6 for statistics of editing and posting activities regarding CRoWD. Figure 7 and 8 show the distribution of this data over time. Obviously, there is a high correlation between same-named user groups when comparing their editing and posting activities (see the distance correlations in Table 1): edits of registered users, for example, correlate with postings of registered users by a value of 0.99. At the same time, activities of bots on talk pages (p-bots) do not correlate with activities of any other user group (including e-bots operating on articles). The following subsections describe and evaluate our procedure of preprocessing CRoWD in order to achieve tree-like representations of Wikicussions according to Section 4.

5.1.1. Text-technological Preprocessing

We use the Wikipedia offline-reader XOWA[15] for parsing CRoWD. The corresponding articles are parsed by means of Sweble (Dohrn and Riehle 2013). Generally speaking, there are two approaches to segment posts in talk pages, to interrelate them within the tree-like structure of a

[14] `https://dumps.wikimedia.org`

[15] `http://xowa.org`

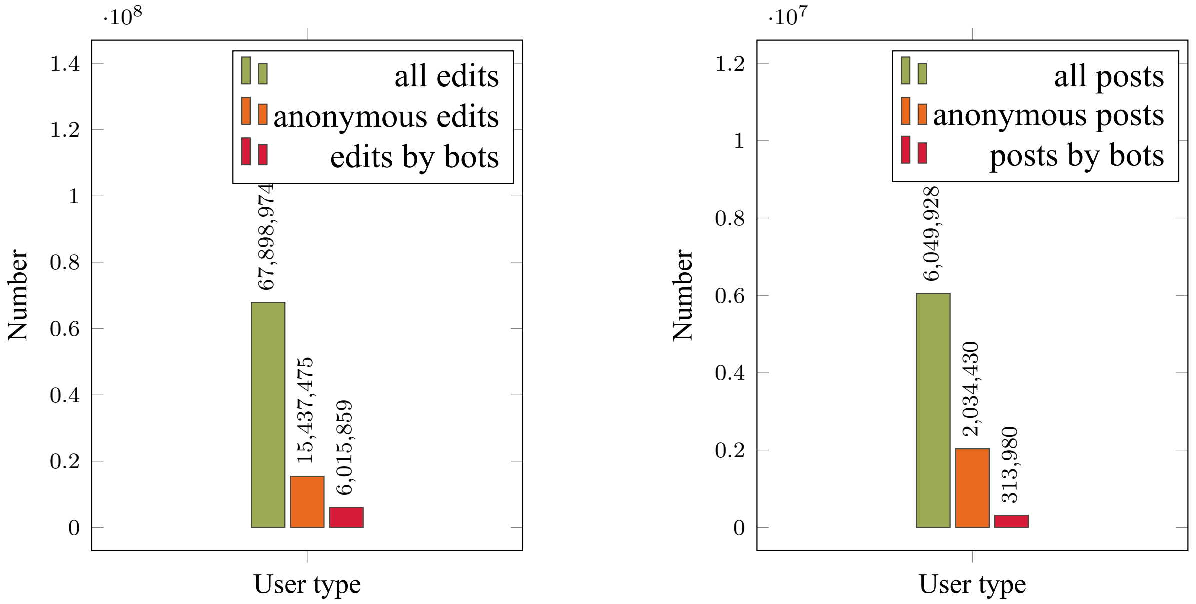

Figure 6. Left: edits on *article* pages; right: posts on corresponding *talk* pages.

Table 1
Distance correlations of edit and posting activities of registered users, anonymous users and bots (*e*: edits, *p*: posts, *reg*: registered, *ano*: anonymous).

	e-all	*e*-reg	*e*-ano	*e*-bots	*p*-all	*p*-reg	*p*-ano	*p*-bots
e-all	1.00	0.98	0.89	0.84	0.95	0.97	0.78	0.12
e-reg	0.98	1.00	0.83	0.79	0.94	0.99	0.73	0.11
e-ano	0.89	0.83	1.00	0.69	0.85	0.84	0.83	0.14
e-bots	0.84	0.79	0.69	1.00	0.78	0.79	0.68	0.11
p-all	0.95	0.94	0.85	0.78	1.00	0.95	0.82	0.43
p-reg	0.97	0.99	0.84	0.79	0.95	1.00	0.73	0.12
p-ano	0.78	0.73	0.83	0.68	0.82	0.73	1.00	0.17
p-bots	0.12	0.11	0.14	0.11	0.43	0.12	0.17	1.00

Wikicussion and to date them:

1. Laniado et al. (2011) and many others rely on extractions based on text indentions, signatures and supplementary heuristics. Though this method is suitable for large datasets including archived pages its problems relate to the heuristics used to extract boundaries of unsigned posts which do not allow for detecting their authorship and timestamps.
2. Ferschke et al. (2012) develop an approach for extracting talk pages by means of computing edit differences of revisions of their edit history. They accurately detect boundaries of inserted posts and also link them to users and timestamps. However, this method is error-prone on editorial edits[16], does not adequately consider archived pages[17] and requires extensive processing on edit histories.

Unlike Ferschke et al., who analyze a small sample, we focus on all talk pages of the German Wikipedia including archived pages. That is, we doubt that the latter approach scales

[16] Editorial edits are contributions which relate to spelling corrections, rearrangements of posts, or editorial remarks at the beginning of a talk page.

[17] Topics of large discussions are periodically moved to archive pages. In such cases the bot or user performing the edition would be detected as the author of the post(s).

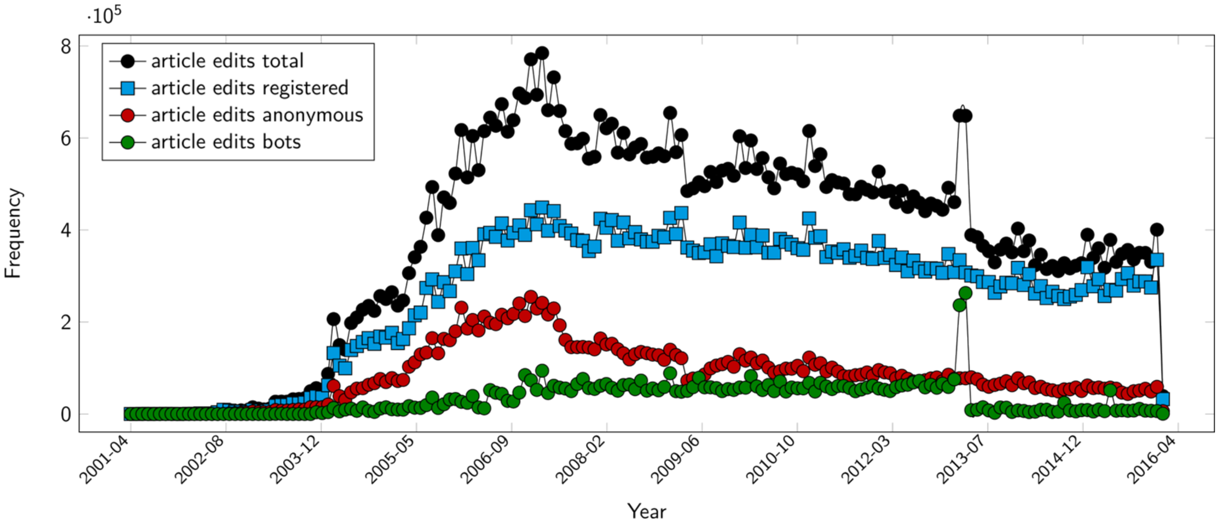

Figure 7. Frequency distribution of edits over time within the German Wikipedia (blue: registered users, red: anonymous users, green: bots, black: all users).

well with a corpus of the size considered here so that we opted for the former approach. A hybrid solution bringing together the best of both approaches is left for future work.

5.1.2. Evaluation

To perform an evaluation of extracting Wikicussions, we created a gold standard based on a corpus D' of 100 randomly selected talk pages. Further, we considered Wikicussions of at least 4 nodes (including posts and sections) in order to reflect the Zipfian nature of this data (see Section 6). We utilize the tree edit distance (Zhang and Shasha 1989) to compute the similarity between our gold standard and its automatically extracted counterpart by setting the penalty of node inserts, deletions and replacements to 1. That is, we measure how many edit operations are minimally needed to transform an extracted tree $T(d)$ into the corresponding gold standard-tree $\dot{T}(d)$ and relate the overall edit cost to the trees' order. Two nodes are seen to be equal if both are equally entitled sections or if both are posts signed by the same name and time. We compute the similarity s of $T(d)$ and $\dot{T}(d)$ as follows:

$$s(T_E(d), \dot{T}(d)) = 1 - \frac{\epsilon(\dot{T}(d), T(d))}{\max(|\dot{T}(d)|, |T(d)|)} \tag{7}$$

ϵ is the tree edit distance. For the 50 largest talk pages in D', we achieve an average similarity s of 0.82. For all 100 discussions, we achieve a score of 0.89. This outcome is competitive regarding a related approach to extracting talk pages from the German Wikipedia (Margaretha and Lüngen 2014).

5.2. Natural Language Processing (NLP)

In order to get access to content-related features of discussions, we lemmatize all tokens of CRoWD while tagging their parts of speech (POS) and grammatical categories. To this end, we utilize a variant of MarMoT (Müller et al. 2013), that is, a POS tagger based on non-linear conditional random fields especially trained for tagging German texts (Eger et al. 2016). This instance of MarMoT, henceforth called *gMarMoT*, shows competitive results particularly in

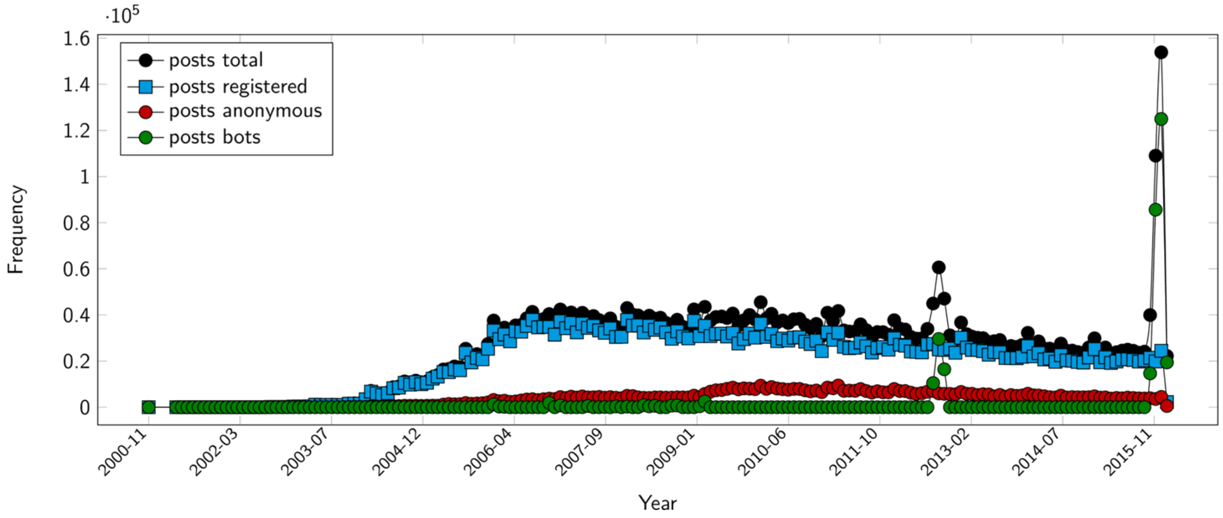

Figure 8. Frequency distribution of posts over time within the German Wikipedia (blue: registered users, red: anonymous users, green: bots, black: all users). Unsigned posts (without timestamps) are excluded. Posts dated by posters outside of the valid time-frame (before the date of creation of the discussion or after the date of its download) are also excluded.

Table 2
Characteristics of CRoWD processed by means of gMarMoT.

Attribute	Articles	Talk pages (users)	Talk pages (bots)
Tokens	520,873,655	355,297,065	7,744,486
Wordforms	10,998,351	6,522,102	545,785
Syntactic words	20,605,609	11,401,559	562,966
Lemmas	12,201,422	7,063,276	555,011
Nouns	4,545,234	2,980,206	14,262
Named entities	2,942,400	1,078,195	34,242

the case of out-domain scenarios. Since gMarMoT has been trained on a corpus that does not contain any sample of written orality (Koch and Oesterreicher 1985), we face a scenario of out-domain tagging when processing Wikicussions. However, since we trained gMarMoT by example of a manually tagged Wikipedia article (on *genetics*) (Lücking et al. 2016), the second half of CRoWD (comprising articles) is better addressed. Note, however, that we do not only risk a loss in accuracy when tagging Wikicussions. Rather, because of the lexical diversity of Wikipedia (comprising many special languages), out-domain tagging is also at stake when switching from the thematic domains of the training corpus (e.g., *biology*) to other domains in Wikipedia (e.g., *music*). Nevertheless, to the best of our knowledge, gMarMoT currently belongs to the best performing taggers for German. By using it, we leave the narrow range of wordforms as typically explored in text mining and get access to lemmas. See Table 2 for statistics of CRoWD as a result of being tagged by means of gMarMoT.

5.3. A Blueprint for Measuring Multidimensional Scale Invariance

Power-laws are the only scale-free probability distributions (Newman 2005): they measure a kind of skewness that exists for the respective distribution at whatever scale one looks at it. Thus, when studying scale invariant structure formation in Wikicussions as motivated in Section 1, power-laws are the first choice. To this end, we explore a range of syntactic (§5.4) and pragmatic (§5.5) features of structure formation. For each of them we compute for each Wikicussion in CRoWD the corresponding feature value. In this way, we learn about the skewness of structure formation in Wikicussions according to the given feature. To keep our model simple, we use the following blueprint for this measurement: let $X_i \in F = \{X_1, \ldots, X_n\}$ (as defined in Section 5.4 and 5.5), $T_D = \{T_{d_1}, \ldots, T_{d_m}\}$ (see Section 4) and $X_i(T_D) = \{X_i(T_{d_1}), \ldots, X_i(T_{d_m})\}$ be the set of X_i-values of tree-like representations T_{d_j} of Wikicussions $d_j \in D$. Then,

$$P(X_i \geq x_k) = \frac{|\{T_{d_j} \in T_D \,|\, X_i(T_{d_j}) \geq x_k\}|}{m} \tag{8}$$

is the complementary-cumulative distribution of X_i over T_D. Based on these preliminaries, our first hypothesis is:

Hypothesis 1 For all features $X_i \in F$:

$$P(X_i \geq x) \sim \beta_{X_i} x^{-\alpha_{X_i}},\ \alpha_{X_i}, \beta_{X_i} \in \mathbb{R}^+ \tag{9}$$

We additionally consider probability distributions of feature values over S_D of tree-like representations T_s of top-level sections $s \in S$ (see Section 4). Each such section ideally addresses a single subtopic of the corresponding Wikicussion. In this way, we get access to thematically more homogeneous subtrees of T_d. As will be shown in Section 7, this approach allows for relating our analysis of scale invariance to the notion of self-similarity in web-based genres (cf. Dill et al. 2002).

So far, our blueprint concerns scale invariance along single dimensions. Our idea is that structure formation in Wikicussions is scale invariant along multiple dimensions such that skewness along one dimension tends to be correlated with skewness along other dimensions. To show this, we define rank numbers

$$r_{ij} = k \text{ iff } X_i(d_j) \text{ is the } k\text{th smallest value of } X_i(D) \tag{10}$$

for each discussion $d_j \in D$ and each feature $X_i \in F$. Rank numbers allow for comparing features w.r.t. the rankings they induce over D. This is done by means of Spearman's rank correlation[18] so that our second hypothesis is:

Hypothesis 2 The rankings of Wikicussions induced by different features in F tend to correlate.

Hypothesis 2 entails that knowing the gestalt of a discussion according to a feature $X_i \in F$ (say, *syntactic depth*) informs about its gestalt according to other features $X_j \in F$ (say, *pragmatic order*). The more features in F are correlated in this way, the denser the feature network whose edges $\{X_i, X_j\}$ are weighted by correlation values $\rho(X_i, X_j)$.[19] In order to measure this density, we compute the following statistics by analogy to the connection coefficient of Egghe and Rousseau (2003):

$$\text{density}_{P_\alpha}(F) = \frac{2}{|F|(|F|-1)} \sum_{X_i, X_j \in F, i \neq j, P_\alpha(\rho(X_i, X_j))} |\rho(X_i, X_j)| \tag{11}$$

[18] Note that in our context, distance correlation (Székely and Rizzo 2009) is no alternative to Spearman's rank correlation because of inducing a prohibitively large time effort facing a corpus of Wikicussions as large as D.

[19] Note that Gonzalez-Bailon et al. (2010) also perform a correlation analysis but only of 6 features (partly syntactic and partly pragmatic in the sense considered here – this includes, amongst others, an h-index (Kaltenbrunner and Laniado 2012) of the number of posts of users per discussion).

The operator $P_\alpha(\cdot)$ checks for the significance of Spearman's rank correlation given the significance threshold α. The higher the absolute values $|\rho(X_i, X_j)|$ among the more features, the higher $\mathrm{density}_{P_\alpha}$. Conversely, if all pairs of features are uncorrelated or if their correlation is insignificant at level α (i.e., $P_\alpha(\cdot) = \mathrm{false}$), then $\mathrm{density}_{P_\alpha} = 0$. Based on these considerations, our third hypothesis is:

Hypothsis 3 For feature set F, $\mathrm{density}_{P_\alpha}(F) \gg 0$.

We now define the features that we used in our study to instantiate this blueprint.

5.4. Syntactic Features

Syntactic measures reflect the complexity of tree-like representations of discussions (Kaltenbrunner and Laniado 2012). Let T_x denote a discussion tree T_d or a section tree T_s rooted by r_x. Then, we instantiate the feature template FT of Section 4 as follows:

1. The *syntactic order* of T_x, denoted by σ-order(T_x), is the number of its vertices. Its *syntactic depth*, denoted by σ-depth(T_x), is the length of the longest thread in T_x (cf. Gonzalez-Bailon et al. 2010). This measure has already been taken for threads in talk pages by Laniado et al. (2011) who additionally compute the following h-index (cf. Gómez et al. 2008) also considered here:[20]

   $$h\text{-index}(T_x) = \max\{i \in \{0, \ldots, \mathrm{depth}(T_x)\} \mid \forall 0 \leq j \leq i \colon |N_j(r_x)| - 1 \geq j\} \qquad (12)$$

   The *syntactic width* (already considered by Kaltenbrunner et al. (2009) for quantifying tree-like structures of online discussions – cf. Gonzalez-Bailon et al. (2010)) of T_x is defined as:

   $$\sigma\text{-width}(T_x) = \max\{|N_i(r_x)| \mid i = 0..\,\mathrm{depth}(T_x)\} \qquad (13)$$

   The *syntactic level* of minimal depth in T_x, denoted by σ-level(T_x), is the smallest number i, for which Expression 13 takes its maximum (Mehler 2011). Note that the levels $j < i$ are "narrower" than σ-width(T_x) and therefore branch out afterwards, while levels larger than i either terminate or do not branch beyond σ-width(T_x). At first glance, this feature seems to be pointless since it is unlikely that the threads of a discussion (generated by different interlocutors) are coordinated in a way to shape σ-level(T_x) in a lawful manner. However, a correlation of σ-level(T_x) with other tree-related features of discussions (e.g., σ-depth) may hint at a law-like organization of discussions in terms of the formation of levels. The *syntactic span* of T_x, denoted by σ-span(T_x), is the number of its child nodes. Finally, the *syntactic length* σ-length(T_x) is the number of its leafs (for these two quantities cf. Köhler 1999).
2. Note that the span of a section may contain nodes of different types, that is, sections and posts. Regarding the meta-dimensions of semantics and pragmatics, this distinction is relevant: posts are the smallest communication units in Wikicussions serving certain communicative functions (as part of dialog acts) and manifesting truth-functional sentences. Therefore, we consider alternative definitions of σ-order, …, σ-length by restricting counting units to posts: $\hat{\sigma}$-order(T_x) is the number of posts in T_x, $\hat{\sigma}$-depth(T_x) the maximum number of posts in threads, $\hat{\sigma}$-width(T_x) the size of the largest neighborhood of r_x consisting only of posts, $\hat{\sigma}$-level(T_x) the smallest i maximizing the latter quantity, $\hat{\sigma}$-span(T_x) the size of the lowest-level non-empty neighborhood $N_i(r_x)$ containing at least one post and $\hat{\sigma}$-length(T_x) the number of all terminating posts.

[20] For a measure of the temporal dynamics of this index see Kaltenbrunner and Laniado (2012).

3. As a measure of tree-like structures interrelating two characteristics (i.e., order and depth), we compute the dependency value of trees, denoted by $\text{depend}(T_x)$, as introduced by Altmann and Lehfeldt (1973):

$$\text{depend}(T_x) = \frac{2\sum_{i=1}^{\text{depth}(T_x)+1} i|\{v \mid \delta(r_x, v) = i-1\}|}{|V_x|(|V_x|+1)} \in (0, 1] \tag{14}$$

   The higher the order of T_x, the more vertices are subordinated in T_x, the higher the value of $\text{depend}(T_x)$. Analogously, the deeper T_x in terms of syntactic depth, the larger the value of $\text{depend}(T_x)$. As a second measure of imbalance, we compute the relative h-index of tree-like structures as:

$$h\text{-balance}(T_x) = \frac{h\text{-index}(T_x)}{\sigma\text{-depth}(T_x)} \in (0, 1] \tag{15}$$

   Obviously, line graphs of order $n \to \infty$ are of lowest h-balance, while star graphs of order 3 are of highest h-balance.
4. Next, we utilize three measures of hypertext theory (Botafogo et al. 1992) that have been used to classify websites (Mehler et al. 2007): the *Absolute Child Imbalance* (ACI) (measuring the imbalance of a node as the variance of the orders of all trees dominated by its child nodes), the *Absolute Depth Imbalance* (ADI) (computing the variance of depths instead of orders), and the *stratum* measuring the deviation of a graph from a same-order line graph (the higher $\text{stratum}(T_x)$, the more hierarchically structured T_x). We complement this subset of features by the *Absolute Width Imbalance* (AWI), which computes widths instead of depths to calculate the imbalance of a tree, and by $\text{resolution}(T_x)$ (Thorley and Wilkinson 2007), which computes the ratio of all branches in T_x minus 1 to $|V_x| - 2$: the more branches, the more likely the discussion is thematically diversified.
5. We also compute the so-called *total cophenetic index* $\text{coph}(T_x)$ of Mir et al. (2013) which calculates the balance of a tree as the sum of depths of all lowest common predecessors of all pairs of leafs in T_x:

$$\begin{aligned} \text{coph}(T_x) &= \begin{cases} \sum\limits_{v_i, w_j \in L(T_x), 1 \leq i < j \leq l = |L(T_x)|} \text{depth}(\text{lcp}(v_i, w_j)) & : l' > 3 \\ 0 & : \text{else} \end{cases} \\ &\in [0, \binom{l'}{3}] \end{aligned} \tag{16}$$

   Since we do not focus on phylogenetic trees, but consider more general trees, $\text{coph}(T_x)$ is in the range of $[0, \binom{l'}{3}]$, such that l is the number of leafs in T_x and l' is the expected number of leafs in a caterpillar graph of order $|T_x| + k$, $k = |\{v \in V(T_x) \mid \text{outdegree}(v) = 1\}|$, and

$$l' = l + k \tag{17}$$

   Rather than $\text{coph}(T_x)$, we calculate its normalized variant

$$\text{imbal}(T_x) = \begin{cases} \frac{\text{coph}(T_x)}{\binom{l'}{3}} : l' > 3 \\ 0 : \text{else} \end{cases} \in [0, 1] \tag{18}$$

   as a measure of imbalance: the higher its value, the more imbalanced T_x where so-called caterpillar trees of the same order $|V_x|$ as T_x are maximally imbalanced (Mir et al. 2013).

One reason to compute imbal is that while the measures taken from Botafogo et al. (1992) are based on the notion of dispersion, imbal is more easily interpreted regarding the range of isomorphism classes of same-order trees (see Mir et al. (2013) for more details). Obviously, the depth of a Wikicussion that is maximally imbalanced in terms of imbal is dominated by a single pair of threads (forming a caterpillar tree), while a maximally balanced discussion (forming a star graph) branches only onto the first level thereby maximizing the number of different threads given its order. Note that imbal does not necessarily distinguish between trees of different orders displaying the same pattern of structuring (e.g., as a star graph or a caterpillar tree).

6. Finally, we compute the Wiener index as proposed by Goel et al. (2016) and Krishnan et al. (2016) in order to round up our syntactic model of Wikicussions:

$$\text{Wiener}(T_x) = \frac{2}{|V_x|(|V_x|-1)} \sum_{v_i,w_j \in V_x, i<j} \delta(v,w) \tag{19}$$

   Wiener(T_x) is known to distinguish between star graphs (for which it is minimized) and large trees with many small branchings (Goel et al. 2016).

While ACI, ADI and AWI are basically measures of imbalance focusing on a single reference quantity (order, depth and width), depend, h-balance, stratum, imbal and Wiener are more holistic measures relating to the overall gestalt of a tree. Finally, σ-order, σ-depth, σ-level, σ-width, σ-span and σ-length are simple statistics of tree-like structures. By computing these features, tree-like structures are mapped onto vectors that can be made input to distribution analysis and classification. Indeed, some of these measures have already been applied for classifying webgenres (Botafogo et al. 1992; Mehler et al. 2007; Mehler 2011) and syntactic structures (Altmann and Lehfeldt 1973; Köhler 1999; Abramov and Mehler 2011).

5.5. Pragmatic Features

Pragmatic features relate to the participation structure of Wikicussions. Regarding feature template FT of Section 4, they can be introduced as follows:

1. π-order$(T_x) = |\{\text{author}_x(v) \mid v \in V_x^p\}|$ is the number of different authors of posts belonging to T_x. This measure has also been used by Gonzalez-Bailon et al. (2010) to characterize discussion trees.
2. π-depth(T_x) is the largest number of different authors contributing to the same thread:

$$\pi\text{-depth}(T_x) = \max_{P \in \mathbb{T}(T_x)} \{|\{\text{author}_x(v) \mid \exists v \in V_x^p(P)\}|\} \tag{20}$$

   where $V_x^p(P)$ is the set of all posts of thread P.
3. π-width(T_x) is the largest number of different participants posting on the same level:

$$\pi\text{-width}(T_x) = \max_{i=1..\,\text{depth}(T_x)} \{|\{\text{author}_x(v) \mid \exists v \in V_x^p\colon \text{depth}(v) = i\}|\} \tag{21}$$

4. π-level(T_x) is the smallest number i maximizing the latter quantity.
5. By analogy to σ-span, π-span(T_x) is the number of different participants posting on

$$N^p_{\min}(r_x) = \arg\min\nolimits_{N_i^p(r_x) \neq \emptyset, i=1..\,\text{depth}(T_x)} \{i\} \tag{22}$$

   comprising at least one post, that is:

$$\pi\text{-span}(T_x) = |\{\text{author}_x(v) \mid \exists v \in N^p_{\min}(r_x)\}| \tag{23}$$

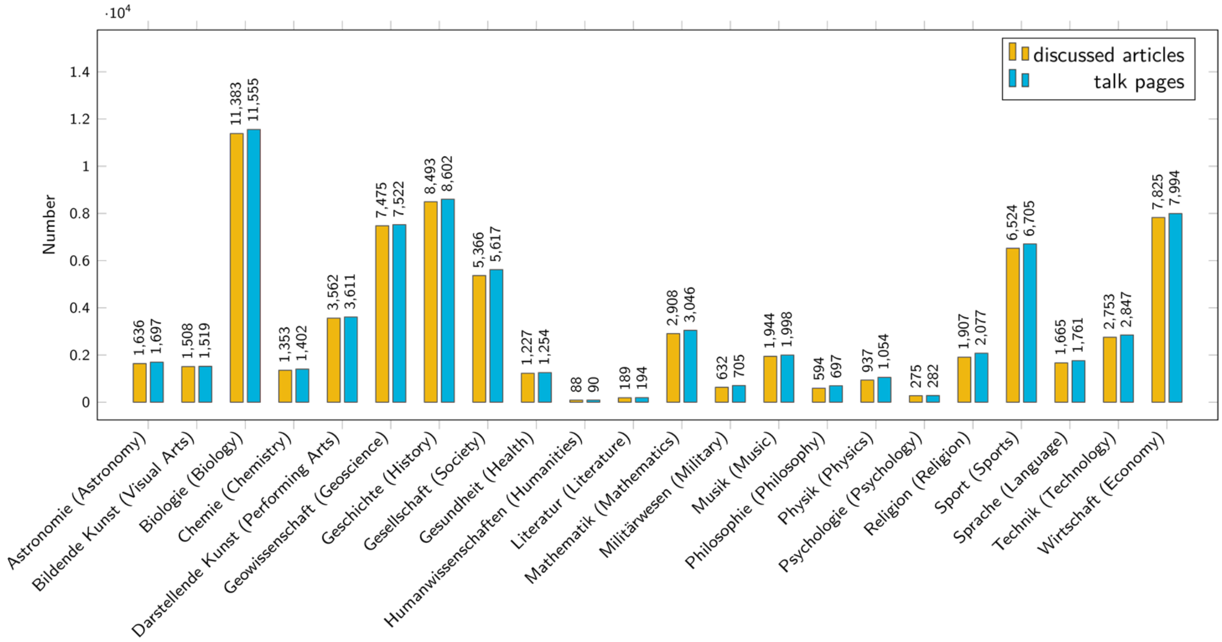

Figure 9. Number of discussed articles and associated talk pages per OCS category.

6. π-length(T_x) is the number of authors of those posts that are leafs in T_x:

$$\pi\text{-length}(T_x) = |\{\mathrm{author}_x(v) \mid \exists v \in L(T_x) \cap V_x^p\}| \tag{24}$$

7. Finally, we compute the overlap of the set of authors contributing to T_x and those contributing to the article discussed by T_x:

$$\pi\text{-overlap}(T_x) = \frac{|\,\mathrm{author}_x(T_x) \cap \mathrm{author}_x(\mathrm{article}(T_x))|}{|\,\mathrm{author}_x(T_x) \cup \mathrm{author}_x(\mathrm{article}(T_x))|} \tag{25}$$

   where

$$\mathrm{author}_x(T_x) = \{\mathrm{author}_x(v) \mid v \in V_x^p\} \tag{26}$$

   is the set of authors of T_x and $\mathrm{author}_x(\mathrm{article}(T_x))$ is the set of authors of the corresponding article. We calculate π-overlap to find out if articles are usually discussed by their own authors: in the event of a high degree of overlap between the two groups of authors, we get evidence of the dominance of a small group of authors, provided that the authorship of articles is distributed according to a power law. In such a case, a small set of interlocutors dominates both writing and discussing articles.

Note that Backstrom et al. (2013) also characterize threads by the number of their different commenters. They observe a bimodal scenario in which focused threads generated by a small number of commenters are contrasted by expansionary threads generated by a large number of commentators commenting only once. Further note that authorship of articles and discussions can be anonymous, so pragmatic quantities are likely to be noisy.

5.6. Thematic Classification

Modeling multidimensional scale invariance aims at a macroscopic picture of Wikicussions irrespective of the underlying topic. However, on a *mesoscopic* level of analysis, one may assume

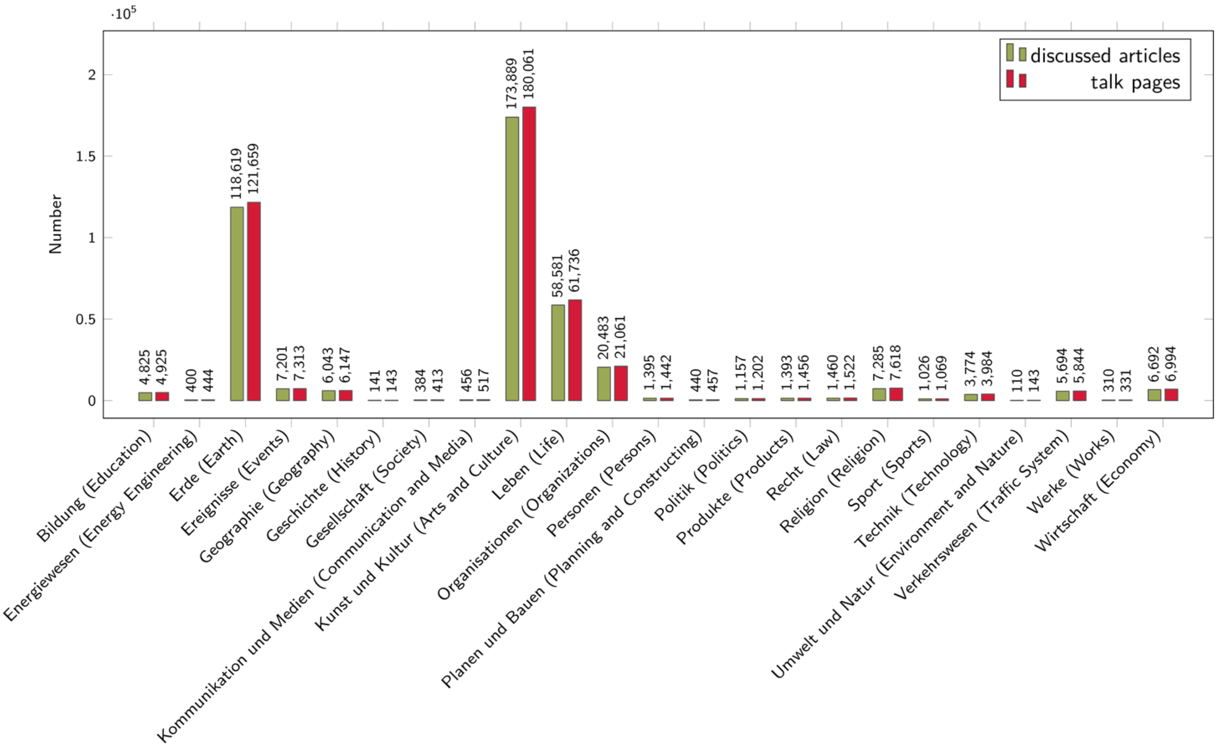

Figure 10. Number of discussed articles and associated talk pages per OCP category.

an influence of the subject area of an article on the gestalt of the corresponding discussion. We may expect, for example, that Wikicussions of political topics are more controversial than those of mathematical or computational ones and, therefore, result in different gestalts. In order to shed light on this mesoscopic dependency on subject area, we finally perform a classification using the feature model F. That is, each discussion $d \in D$ is represented by a feature vector $\vec{d}$ whose dimensions are defined by the elements of F. The aim is to perform an experiment in which the classifier learns to predict the subject area of a discussion based on these features. In order to arrive at a sufficiently abstract classification of as many discussions as possible, we utilize a subset of the *main topic classification* of the German Wikipedia[21] extended by categories reflecting the OECD classification of the fields of science and technology (OECD 2007). Henceforth, this selection is called *OECD-oriented Category Selection* (OCS). See Figure 9 for the corresponding distribution of articles and discussions per topic of this classification. Note that all topics collected by OCS denote categories within Wikipedia's category system (Mehler 2011): they are either directly dominated by the so-called *!Hauptkategorie* (*main category*) or in a short distance to the category *Kategorie:Sachsystematik* (*Category:Main topic classifications*). Table 3 lists all categories selected in this way.

As a *tertium comparationis* we consider a thematic partitioning of the discussion (and of the corresponding article) space called *Optimized Category-based Partition* (OCP). This partitioning, which covers 687,888 Wikicussions of 23 different thematic classes (see Figure 10 for the corresponding distribution of articles and discussions per target class), has been computed by means of a bottom-up algorithm climbing up Wikipedia's category graph till a partitioning of the discussion space is reached. This means that the intersection of any pair of OCP categories

[21] Cf. `https://de.wikipedia.org/wiki/Kategorie:Sachsystematik`. See Mehler and Stegbauer (2012) for a related approach to classifying article networks regarding a subset of main fields of knowledge.

Table 3

Macro-topics used as target classes for classifying Wikicussions based on their syntactic and pragmatic features. P0 denotes the path *!Hauptkategorie:Sachsystematik* in the category graph of Wikipedia, P1 = *Wissen:Wissenschaft:Wissenschaft_nach_Fachgebiet*, P2 = *P1:Naturwissenschaft*, P3 = *Kunst und Kultur* and P4 = *Kunst nach Gattung*.

No.	Subject Area	Translation
1.	P0:P2:Archäologie	*archeology*
2.	P0:P2:Astronomie	*astronomy*
3.	P0:P3:P4:Bildende Kunst	*fine arts*
4.	P0:P2:Biologie	*biology*
5.	P0:P2:Chemie	*chemistry*
6.	P0:P3:P4:Darstellende Kunst	*performing arts*
7.	P0:P2:Geowissenschaft	*geoscience*
8.	P0:Geschichte	*history*
9.	P0:Gesellschaft	*society*
10.	P0:Gesundheit	*health* (including *medicine*)
11.	P0:P1:Humanwissenschaften	*human sciences*
12.	P0:P3:P4:Literatur	*literature*
13.	P0:P1:Mathematik	*mathematics*
14.	P0:Militärwesen	*military*
15.	P0:P3:P4:Musik	*music*
16.	P0:P1:Philosophie	*philosophy*
17.	P0:P2:Physik	*physics*
18.	P0:P1:Psychologie	*psychology*
19.	P0:Religion	*religion*
20.	P0:P1:Sozialwissenschaft	*social sciences*
21.	P0:Sport	*sports*
22.	P0:P3:Sprache	*language*
23.	P0:Technik	*engineering*
24.	P0:Wirtschaft	*economics*

is empty: not a single discussion refers to an article that is directly or indirectly assigned to more than one of these categories. Note that we first transformed Wikipedia's category graph into a tree-like structure using a breadth-first algorithm to compute this partitioning. In order to secure that OCS induces a partitioning over the set of discussions by analogy to OCP, we excluded two categories: *Archäologie* (*archeology*) and *Sozialwissenschaft* (*social science*). This results in a list of 22 OCS-categories (see Figure 10). Based on these considerations our fourth hypothesis is:

Hypothesis 4 The syntactic and pragmatic gestalt of Wikicussions is neither affected by the underlying subject area nor by the participation structure of discussants in this area.

Supposed that our OCP- and OCS-related classifications are successful (in terms of high F-scores[22]), this hypothesis would be falsified. On the other hand, a failing classification would give evidence that the gestalt of Wikicussions is confusable *across the borders of both (i) thematic structure and (ii) participation structure*. To show this, we will additionally show that discussants hardly contribute across the borders of classes as comprised by OCS or OCP. This classification will be contrasted with two baselines: the first baseline concerns the thematic separation of the corresponding article space in terms of the articles' lexical content. The second baseline applies the same procedure to the lexical content of Wikicussions. In line with Hypothesis 4, an interesting scenario is then, for example, that while Wikicussions are structurally

[22] The F-score of a classification is the harmonic mean of its precision and recall.

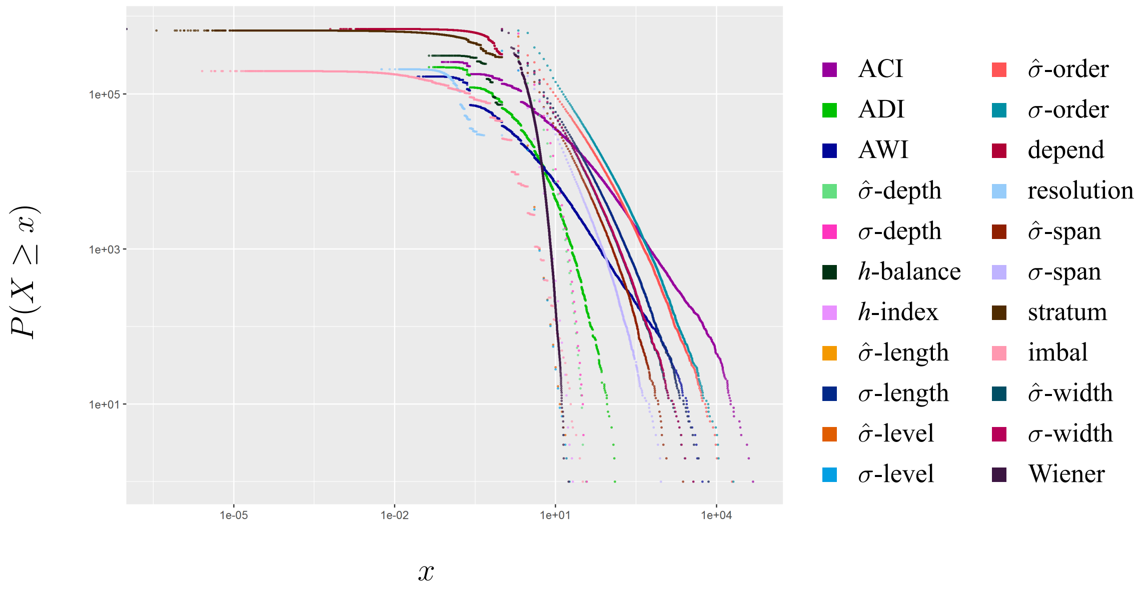

Figure 11. Log-log plots of the distributions of syntactic features of Wikicussions. The abscissa refers to the corresponding features values; for the ordinate (complementary cumulative distribution) see Expression 9 (of Hypothesis 1).

confusable across the borders of OCS-classes, they are nevertheless separable in terms of their lexical content.

6. Results

We present results regarding the macroscopic analysis of multidimensional scale invariance (Hypotheses 1-3 of Section 6.1) and the mesoscopic classification of Wikicussions by means of syntactic, pragmatic and lexical features (Hypothesis 4 of Section 6.2).

6.1. Multidimensional Scale Invariance

We start with Hypothesis 1 (of page 18) concerning the scale invariance of syntactic and pragmatic features. Tables 4 and 5 show that with a single exception, feature values of Wikicussions *and* of *Top-Level Sections* (TLS) are distributed in a scale free manner. The only exception is the h-balance of Wikicussions, for which the *Adjusted Coefficient of Determination* (ADI) is below 0.9. In all other cases, ADI is higher than 90%. In the majority of cases fitting is nearly perfect allowing for statements of the following sort: *While the majority of units (Wikicussions or TLS) is rather unstructured (because of comprising only a single or few posts), there is a very small group of units exhibiting a rich structure in terms of the focal feature.* For example, only very few Wikicussions are highly imbalanced (imbal), exhibit longer threads (σ-depth), span broader discussions (σ-width), written by many different participants (π-order), initially (π-span) and finally (π-length). The majority of discussions exhibits the exact opposite of this scenario. At the same time, we observe that many features exhibit a power law-exponent smaller than 2 – especially in the case of Wikicussions. According to Newman (2005), this indicates that the underlying theoretical power law neither has a finite variance nor a finite expected value making the question for typical observations along these dimensions pointless. In line with this observation, we cannot speak about the typical pragmatic depth (π-depth), authorship overlap

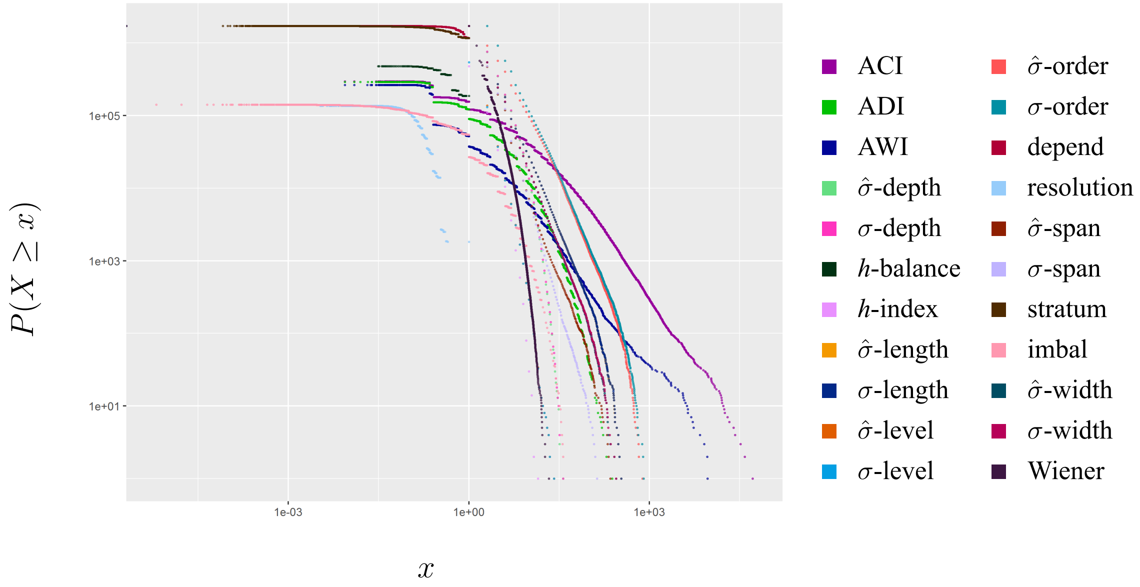

Figure 12. Log-log plots of the distributions of syntactic features of top-level sections (for the axes see Figure 11).

(π-overlap) or imbalance (imbal) of a Wikicussion.

Altogether, these observations indicate that with a single exception (i.e., h-balance in the case of Wikicussions), power laws fit the value distributions of features in F (see Figures 11–14 for the corresponding log-log-plots). At the same time, the log-normal model fails so that it does not provide an alternative.[23] Under this regime, we consider Hypothesis 1 as *not* being falsified: *Wikicussions and TLS are scale-free along a multitude of syntactic and pragmatic dimensions as comprised by the feature set F.*

We now turn to Hypothesis 2 (of page 18), that is, correlation analysis. To this end, we study the degree, by which the rank (Expression 18) of a Wikicussion/TLS according to one feature is correlated with its rank according to other features of F (see Table 4 and 5). We use these correlations to induce two feature networks denoted by WRN (*Wikicussion-Related Network*) and TRN (*TLS-Related Network*), respectively: for a given threshold of minimal correlation $\min_\rho$ any two vertices (features) of the corresponding Wikicussion- or TLS-related network are linked if their correlation is at least $\min_\rho$.

We start with looking at strongly correlated feature pairings for which $|\rho| \geq \min_\rho = 80\%$. In this case, WRN exhibits a cohesion of 0.17 (see Figure 15). Cohesion is the ratio of the number of existing edges (70) in relation to the overall number of possible edges (i.e., 406). In the case of TRN, cohesion is 0.15. Regarding very strong correlations of at least 90%, the cohesion of WRN is still 0.12 (0.07 in the case of TRN). 18 features contribute to this value (more than 62% of all features). The fraction of vertices belonging to the *Largest Connected Component* (LCC) of WRN is 0.34. Among the pairs of features of highest correlation is $\{\pi\text{-order}, \hat{\sigma}\text{-order}\}$ (for which $\rho = 0.96$): that is, by knowing the rank of a discussion in terms of the number of its different posters, one knows almost perfectly its rank in terms of its number of posts. That is, pragmatic order informs about syntactic order. The same holds, for example, for *pragmatic* versus *syntactic span* and for *pragmatic* versus *syntactic length*. While this is not very surprising (*the more posters, the more posts and vice versa*), one also detects more interesting pairings such as π-level (see page 21) and $\hat{\sigma}$-level (see page 19): if a discussion d is, for example, broadest

[23] Power laws were fitted by means of MATLAB, log-normal distributions by means of R.

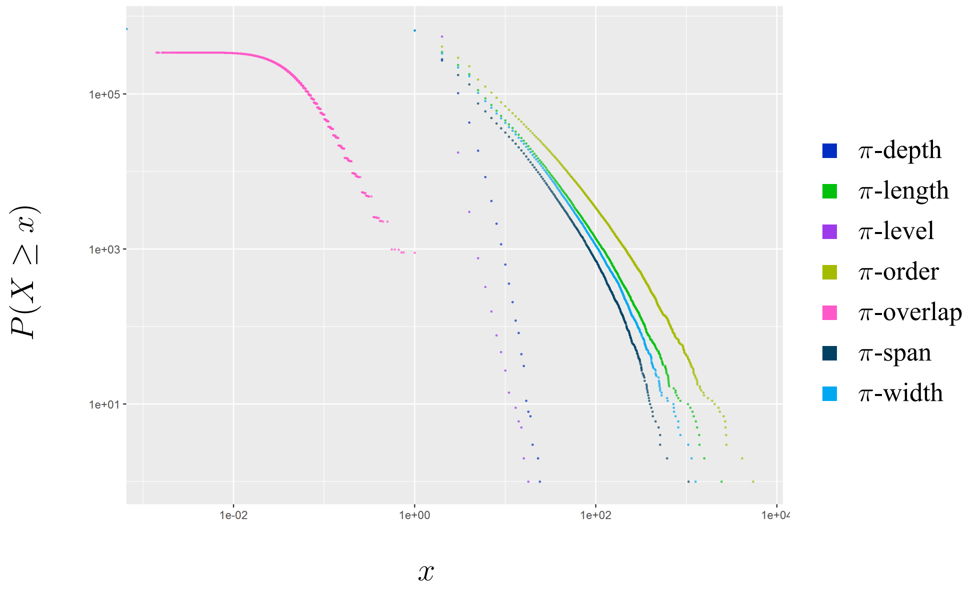

Figure 13. Log-log plots of pragmatic features of Wikicussions (for the axes see Figure 11).

at the beginning of a Wikicussion tree, it tends to have the highest number of different authors on a level at least nearby or identical with its syntactic counterpart. That is, by knowing the level of maximum syntactic width of a discussion d, one is indirectly informed about its level of broadest participation: the deeper the syntactic level of d, the lower its rank, the lower the rank of d in terms of π-level, the deeper its pragmatic level. In other words: *thematic broadness requires a correspondingly broad participation of different authors*. Note that this interpretation presupposes that the more threads in a discussion, the broader its thematic spectrum.

Another interesting pairing (of strong correlation $|\rho| > 80\%$) concerns π-depth and $\hat{\sigma}$-depth ($\rho = 0.89$). It indicates that the longer the longest thread of a Wikicussion, the higher its corresponding rank number in terms of both $\hat{\sigma}$-depth and π-depth, the higher the number of different authors contributing to that thread. In other words, thematic specification (in terms of long threads about the same topic) tends to require a broader participation structure – and not contributions by only a few or just a pair of authors. Note that pairings of syntactic length, order, span and width are also among the strongest correlations being measured. This observation indicates that (except for syntactic level) syntactic structuredness correlates along several dimensions: what is big (σ-order) is also broad (σ-width) and long (σ-length) and has a wide span (σ-span).

By considering medium level correlations of at least 50%, we arrive at a feature network (WRN) exhibiting a cohesion of 0.48 and an LCC comprising a fraction of 0.90 of all 29 features. Thus, regarding Hypothesis 2, we state that though many feature pairings neither exhibit a strong, nor a medium correlation, the amount of those which do is remarkably high: Figure 15 shows that for a lower bound of $\min_{|\rho|} = 0.5$, the average correlation of the remainder feature pairs is 0.73 – in the case of both WRN and TRN. This lower bound results in a network (WRN) exhibiting a cohesion of 0.48 in which only π-overlap is isolated. In order to raise cohesion, one has to reduce $\min_{|\rho|}$. An interesting case concerns the bound $\min_{|\rho|} = 0.34$, for which *cohesion* (0.640) equals more or less *average correlation* (0.649). That is, when considering feature pairs of a correlation of at least 34% (weak correlation), cohesion and average correlation are nearly the same. Under this regime, the fraction of features belonging to the LCC is 1: At this level, there is no feature that is not correlated with another one. In the case of TRN this "break-even point" is induced by the bound $\min_{|\rho|} = 0.39$ (for which the size of the LCC is 96.5%.). In line

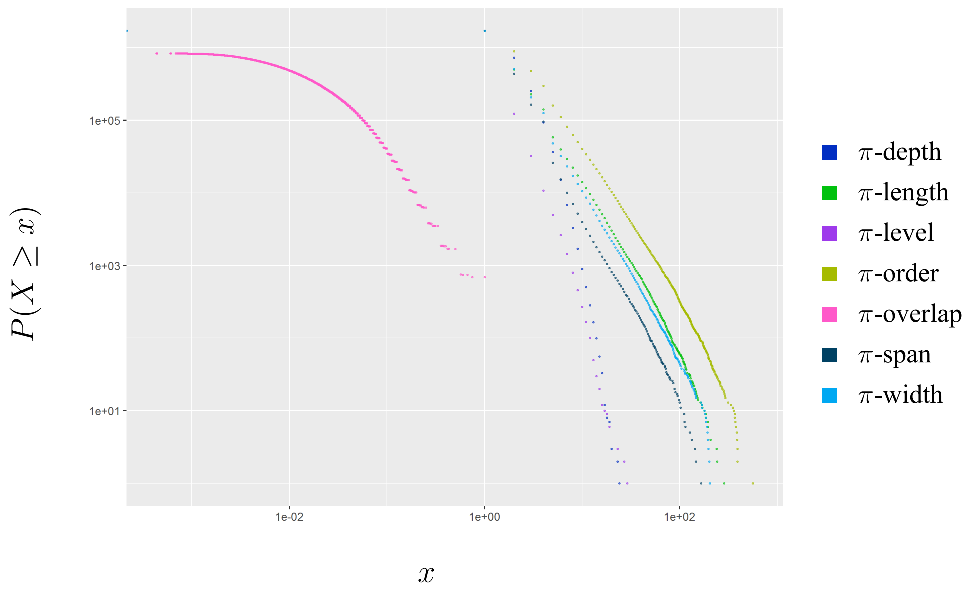

Figure 14. Log-log plots of pragmatic features of top-level sections (for the axes see Figure 11).

with these observations, we get strong hints at a correlative association among a larger group of features: in these cases, one is informed about the rank of a Wikicussion (TLS) along one feature when knowing its rank along other features of the same group – at least in terms of weak correlations (and on average in terms of medium correlations). At the same time, we observe that Wikicussions and TLS exhibit a very similar dynamics as a function of $\min_{|\rho|}$ – below, this observation will be related to the notion of self-similarity. Note that we only consider significant rank correlations ($p = 0.05$). Note also that outliers are robust in being uncorrelated regarding the majority of features. π-overlap (i.e., of authorship of articles and corresponding Wikicussions), for example, gets isolated for $\min_{|\rho|} = 0.37$. This feature provides rather independent information not being covered by $F \backslash \{\pi\text{-overlap}\}$. That is, by knowing the degree of overlap between the authorship of an article and its corresponding Wikicussion, one is not informed about any other aspect of the gestalt of the latter: higher overlap values do not indicate larger, broader or deeper discussions. Another example is stratum (getting isolated for $\min_{|\rho|} = 0.57$). That is, for a lower bound of $\min_{|\rho|} = 0.57$, only two features are isolated: π-overlap and stratum. In light of these results we do not get enough evidence for falsifying Hypothesis 2 with respect to both Wikicussions and top-level sections.

To get an overall picture regarding Hypothesis 3 (see page 19), consider Table 6. The average correlation in WRN is 0.398. Note that since the LCC covers all features already for a near zero correlation of 0.02, $\text{density}_{P_\alpha}(F)$ equals average correlation (unlike in Figure 15 we consider all significant correlations). In the case of TRN, $\text{density}_{P_\alpha} = 0.384$. Things look different in the case of mono- (of either syntactic or pragmatic features) and bimodal networks (of syntactic features in relation to pragmatic ones). Pragmatic features are more correlated among each other than syntactic ones. In the bimodal case, syntactic and pragmatic features correlate more or less on an equal footing. Putting these observations together, we observe a trend in terms of weak correlations somehow approaching the level of medium correlations. Thus, for feature set F as a whole we do not get enough evidence in support of Hypothesis 3: multidimensional scale invariance frequently occurs in the system but is not omnipresent. However, in the case of top-level sections, average correlation is of medium level indicating a tendency towards higher-level correlations among all features.

Table 4
Syntactic features of Wikicussions (left) and their top-level sections (right), the exponent α of the power law fitted to the corresponding distribution, the *adjusted coefficient of determination* of power law fitting (pl), the p-value of the Kolmogorow-Smirnow test (values smaller than 0.1 indicate failure) followed by the p-value of the Shapiro-Wilk test (values smaller than 0.05 indicate failure) of fitting the log-normal distribution (l-n). The last column in the table on the right indicates the page on which the corresponding index is defined.

Quantity	α	pl	l-n
ACI	-0.72	1.00	0.07 (0.00)
ADI	-1.47	1.00	0.11 (0.00)
AWI	-1.15	1.00	0.08 (0.00)
depend	-0.32	1.00	0.32 (0.00)
h-balance	-1.97	0.80	0.28 (0.00)
h-index	-3.45	0.99	0.09 (0.99)
imbal	-7.01	0.99	0.34 (0.00)
resolution	-7.89	0.99	0.20 (0.00)
σ-depth	-3.03	0.98	0.06 (1.00)
$\hat{\sigma}$-depth	-2.52	0.98	0.06 (1.00)
σ-length	-1.34	1.00	0.05 (0.08)
$\hat{\sigma}$-length	-1.37	1.00	0.05 (0.05)
σ-level	-8.22	0.99	0.13 (0.88)
$\hat{\sigma}$-level	-8.11	0.95	0.14 (0.85)
σ-order	-1.48	1.00	0.05 (0.01)
$\hat{\sigma}$-order	-1.46	1.00	0.05 (0.01)
σ-span	-1.72	1.00	0.06 (0.26)
$\hat{\sigma}$-span	-1.54	1.00	0.06 (0.14)
σ-width	-1.47	1.00	0.05 (0.10)
$\hat{\sigma}$-width	-1.47	1.00	0.06 (0.08)
stratum	-1.43	0.99	0.28 (0.00)
Wiener	-4.54	1.00	0.08 (0.00)

Quantity	α	pl	l-n	p.
ACI	-0.79	1.00	0.06 (0.00)	20
ADI	-1.24	1.00	0.11 (0.00)	20
AWI	-2.56	0.98	0.07 (0.00)	20
depend	-0.08	1.00	0.34 (0.00)	20
h-balance	-1.49	0.91	0.25 (0.00)	20
h-index	-4.74	1.00	0.08 (1.00)	19
imbal	-2.23	1.00	0.33 (0.00)	20
resolution	-6.74	0.95	0.17 (0.00)	20
σ-depth	-2.75	0.98	0.06 (1.00)	19
$\hat{\sigma}$-depth	-2.77	0.98	0.06 (1.00)	19
σ-length	-2.08	1.00	0.07 (0.20)	19
$\hat{\sigma}$-length	-2.20	1.00	0.08 (0.13)	19
σ-level	-4.44	1.00	0.13 (0.78)	19
$\hat{\sigma}$-level	-4.48	0.96	0.13 (0.79)	19
σ-order	-2.02	1.00	0.08 (0.02)	19
$\hat{\sigma}$-order	-1.90	1.00	0.08 (0.02)	19
σ-span	-3.26	0.99	0.13 (0.10)	19
$\hat{\sigma}$-span	-2.72	0.99	0.10 (0.14)	19
σ-width	-2.24	1.00	0.07 (0.33)	19
$\hat{\sigma}$-width	-2.45	1.00	0.09 (0.18)	19
stratum	-0.29	1.00	0.29 (0.00)	20
Wiener	-4.19	1.00	0.06 (0.00)	21

6.2. Classification Experiments

Our analysis of multidimensional scale invariance shows that syntactic and pragmatic features are well fitted by power laws (Hypothesis 1), while many pairs of them are highly correlated (Hypothesis 2). However, we also observe that the overall feature system exhibits a medium average correlation (objecting Hypothesis 3). One may think that these findings simply result from the fact that the majority of discussions are rather structureless by comprising a few posts. Seemingly, Wikicussions of this sort make a Zipfian organization and corresponding correlations likely. Though we may object this argument by hinting at how we fit power laws (according to the method presented in Newman (2005)), we now undertake a classification experiment which shows that structural separability does not exist for Wikicussions and that this finding is independent of the degree to which they are structured. To this end, we first show that articles discussed by Wikicussions are thematically separable using a state-of-the-art classifier called `fastText` (Joulin et al. 2016) developed as an efficient alternative to time-consuming deep learners (see Table 7). This is followed by a second classification showing that the same data space is not separable when relying on feature model F (see Section 5). Since this feature model is purely numeric, we use a deep learner instead of `fastText` as a classifier (see Table 8). Both classification scenarios are carried out by example of the OCS- and the OCP-based partitioning of the article and the discussion space (see Section 5.6). Note that according to Hypothesis 4 (on page 24), we expect that the discussion space is *not* separable by means of the feature model F.

We start with OCP (see Figure 16): by classifying the longest 100 articles, one observes an increase of F-score up to 0.461. However, this causes an increase of training effort by raising

Table 5
Pragmatic features of Wikicussions (left) and their top-level sections (right). For the legend of the columns see Table 4.

Quantity	α	pl	l-n
π-depth	-4.39	0.97	0.09 (0.98)
π-length	-1.80	1.00	0.06 (0.16)
π-level	-7.87	0.94	0.15 (0.81)
π-order	-1.38	1.00	0.05 (0.07)
π-overlap	-23.72	1.00	0.14 (0.00)
π-span	-1.83	1.00	0.05 (0.38)
π-width	-1.63	1.00	0.06 (0.21)

Quantity	α	pl	l-n	p.
π-depth	-4.86	0.97	0.09 (0.98)	21
π-length	-2.54	1.00	0.08 (0.32)	22
π-level	-4.67	0.96	0.11 (0.91)	21
π-order	-2.18	1.00	0.08 (0.13)	21
π-overlap	-52.81	0.99	0.21 (0.00)	22
π-span	-2.98	0.99	0.10 (0.29)	21
π-width	-2.68	1.00	0.10 (0.15)	21

Table 6
Average correlation values $\mathrm{density}_{P_\alpha}$, $\alpha = 0.05$, of $|F|(|F|-1)/2 = 406$ pairings of $|F| = 29$ features computed for Wikicussions and their *Top-Level Sections* (TLS).

Scope	#Features	#Pairings	$\mathrm{density}_{P_\alpha}(F)$
Features of Wikicussions	29	406	0.398
syntactic	22	231	0.337
pragmatic	7	21	0.456
syntactic ↔ pragmatic	29	154	0.404
Features of top-level sections	29	406	0.384
syntactic	22	231	0.331
pragmatic	7	21	0.435
syntactic ↔ pragmatic	29	154	0.388

the number of training cycles (epochs) up to 5,000 (see Figure 16). Obviously, longer articles are not so well separable when considering OCP as the target classification. Note that in all these experiments, we randomly split the set of observations into 70% of items used for training and 30% used for testing. A different scenario is given when classifying the complete article space (Figure 16): in this case, the F-score finally reaches 0.796 by requiring "only" $1{,}000$ training epochs. At the same time, by applying the same classification model to discussions (by operating on their vocabulary), one gets an F-score of 0.301 regarding the 100 largest Wikicussions and of 0.536 when regarding all discussions (Figure 16). In this scenario, Wikicussions are less separable in terms of their vocabulary when using a state-of-the-art classifier like `fastText` while the corresponding articles are well separable. Further, by considering a small subset of large items (articles or discussions), the F-score drops significantly.

More or less the same scenario is induced by the OCS-based partitioning (see Figure 17). However, all F-scores are now higher than in the case of OCP. Further, while the largest 100

Table 7
Parameter setting of the `fastText`-based classifier.

Parameter Name	Parameter Value
Classifier	`fastText` (Joulin et al. 2016)
Learning rate	0.05
Size of hidden layer	100
Size of context window	5
Lower bound of word occurrences	1
Max length of word n-gram	3

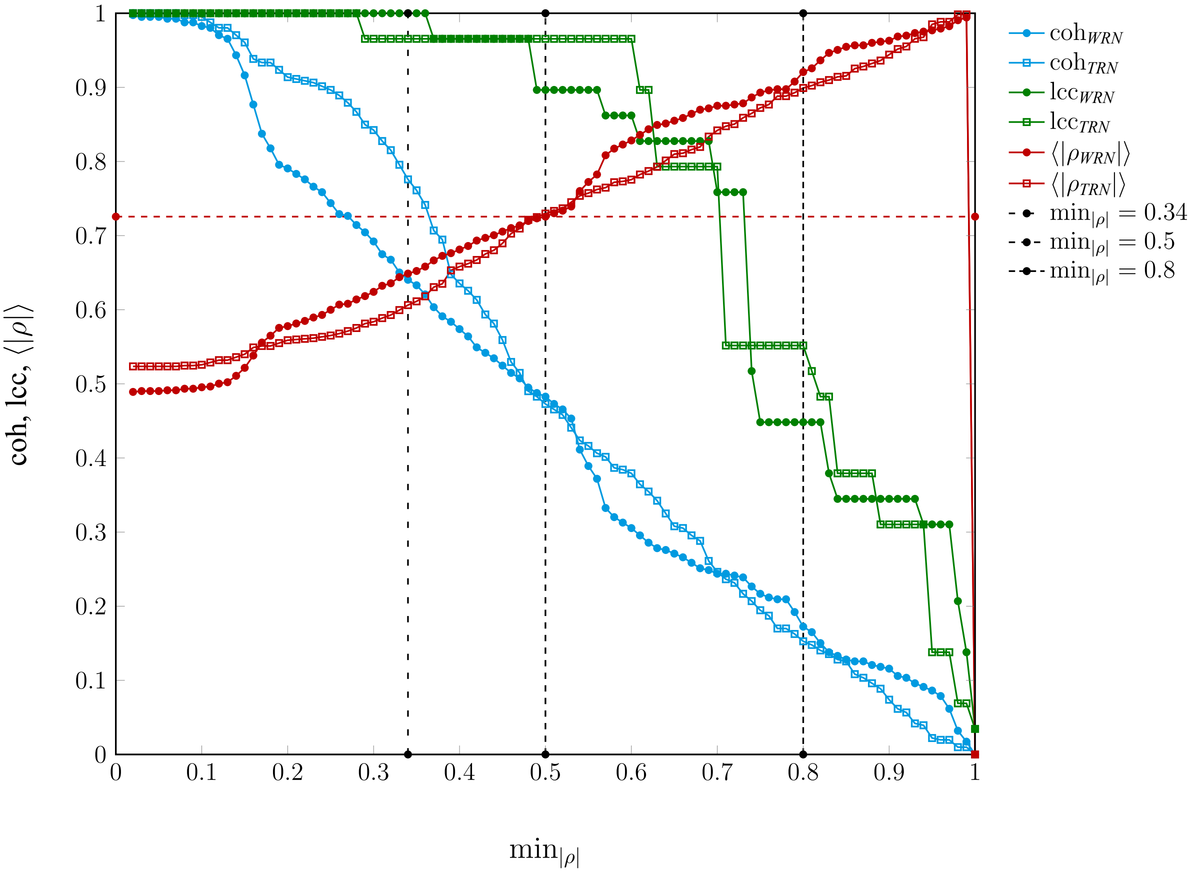

Figure 15. Cohesion (coh), fraction of vertices belonging to the largest connected component (lcc) and average correlation ($\langle|\rho|\rangle$) as a function of minimal allowable correlation $0.2 < \min_\rho \leq 1$ (i.e., for a given value of $\min_\rho$, only those pairings (edges) are considered within the resulting feature network, whose correlation is at least $\min_\rho$) distinguished for the Wikicussion- (WRN) and TLS-related feature network (TRN).

articles and discussions are separable according to an F-score of 0.724 and 0.547, respectively, the complete article space is separable according to an F-score of 0.922. These higher scores may reflect the fact that the number of documents collected by OCS is much smaller than the one comprised by OCP. In any event, these findings also indicate that classifying articles and discussions by means of their lexical content is possible when looking for the underlying subject area being described (in the case of articles) or being discussed (in the case of discussions). In any event, Wikicussions are less separable than articles. This may be explained by their prevalent task-orientedness (see Section 3) making them confusable across thematic borders.

Now, we turn to the feature model of Wikicussions elaborated in Section 5 and try to classify the same items according to OCP and OCS, respectively, in order to shed light on Hypothesis 4 (on page 24). The corresponding F-scores are depicted in Figure 18 for increasing numbers of largest discussions, starting with the 10 largest ones and ending with the 1,000 largest ones. In this scenario, we generally observe very low F-scores and a drop down near to zero for increasing sample sizes. Obviously, classifying discussions based on the feature model of Section 5 fails: Wikicussions are not thematically distinguishable according to their syntactic and pragmatic structure as considered here. Figure 18 also shows that the same diagnosis holds when considering all Wikicussions comprised by OCS and OCP, respectively. Thus, separability is not a function of the degree of structuredness – neither smaller nor longer Wikicussions

Table 8
Parameter setting of the neural network-based classifier.

Parameter Name	Parameter Value
Type	feedforward neural network
Learning rate	0.05
Number of Epochs	500
Number of hidden layers	1
Size of hidden layer	100
Lower bound of word frequency	1

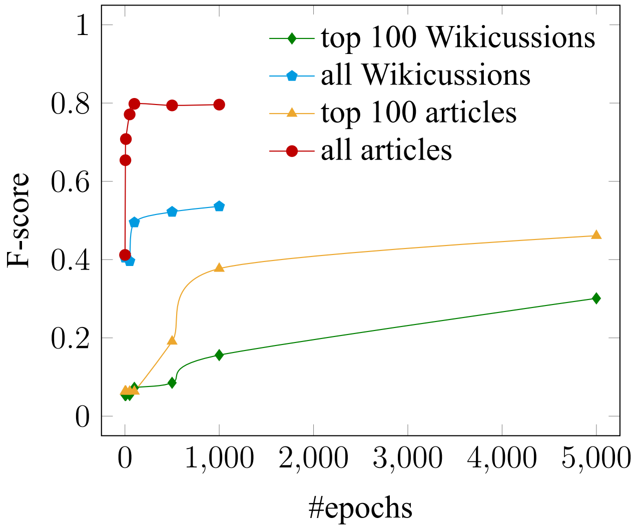

Figure 16. F-scores of classifying Wikicussions regarding the OCP scenario.

are separable by means of the feature model F. It remains to show that this finding is not biased by the fact that articles of different topics tend to be discussed by the same community. This is depicted in Figures 19 and 20 w.r.t. OCS: on average, the Fuzzy Jaccard coefficient of posters who contribute to Wikicussions of different OCS-categories is 5.4%. By disregarding bots, this average is reduced to 4.4%. By additionally disregarding sysops, we get 4.1% fuzzy overlap on average. The Fuzzy Jaccard is computed as follows (cf. Ramli and Mohamad 2009) ($\mathcal{A}(D)$ is the set of author IDs of posters contributing to Wikicussions in corpus D – see Expression 2):

$$\forall A, B \in \text{OCS}\colon\ J_\mu(A, B) \quad = \quad \frac{\sum_{x \in \mathcal{A}(D)} \mu_{A \cap B}(x)}{\sum_{x \in \mathcal{A}(D)} \mu_{A \cup B}(x)} \tag{27}$$

where

$$\mu_A(x) = \frac{\text{number of posts of poster } x \text{ to Wikicussions of category } A}{\text{number of all posts to Wikicussions of category } A} \tag{28}$$

In the case of the classical Jaccard coefficient we get 4.26% (all posters), 4.2% (without bots) and 3.87% (without bots and sysops). Hence, posters tend to concentrate their posts to Wikicussions of a single category: the thematic participation induced by OCS is almost parallelized by a partitioning of the underlying space of posters. We additionally computed random distributions of posts over categories – on average, this results in a fuzzy overlap of categories sharing posters of 29%. Comparing the vectors of random overlaps with those being observed using a t-test,

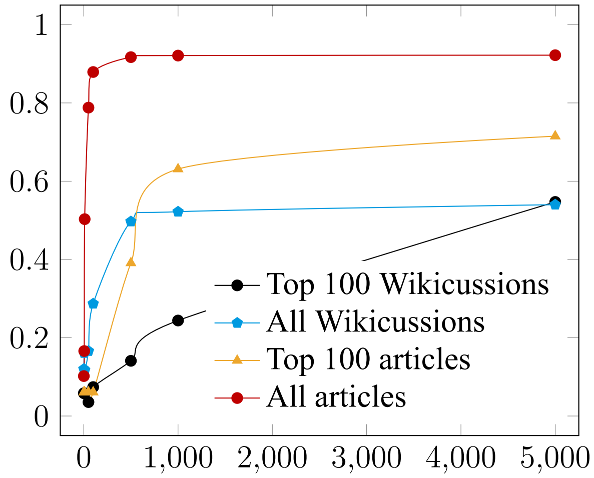

Figure 17. F-scores of classifying Wikicussions according to the OCS scenario.

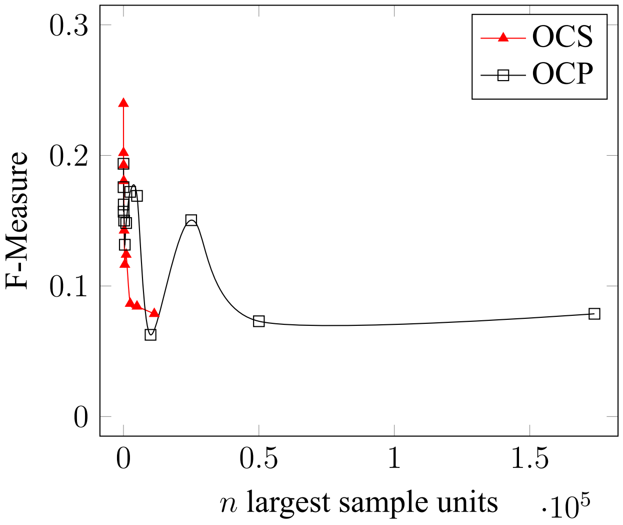

Figure 18. F-scores of classifying Wikicussions based on syntactic and pragmatic features.

we see that the observed overlaps are significantly smaller than their random counterparts (p-value $<$ 2.2e-16). Thus, Wikicussions of different OCS-categories are almost formed by non-overlapping communities. In sum, we do not get enough evidence for falsifying Hypothesis 4.

7. Discussion

7.1. Multidimensional Scale Invariance

As shown in Section 6, online communication as exemplified by Wikicussions evolves into a state of scale invariance that is simultaneously reflected on several syntactic and pragmatic dimensions – irrespective of the underlying topic being discussed and irrespective of the composition of the underlying community of posters. Ideally, we expect a discussion forum as provided by talk pages to be both (i) thematically diversified in the sense of unfolding a wide range of subtopics and (ii) participatory in the sense of attracting a wide range of discussants on an equal footing. Seemingly, such a discussion is rarely found in Wikipedia while the likelihood

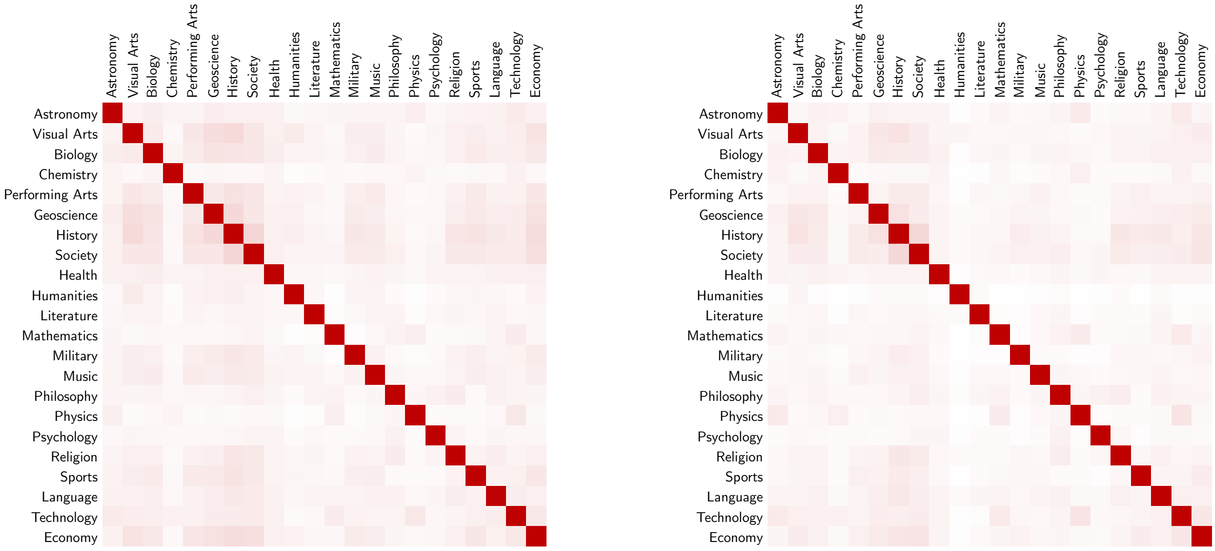

Figure 19. Heatmaps showing the rather tiny overlap of communities of posters contributing to Wikicussions of different topics according to OCS: with bots/sysops (left) and without bots/sysops (right).

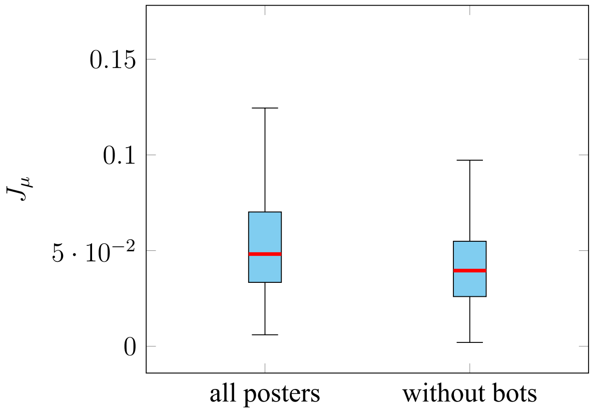

Figure 20. Boxplots of the Fuzzy Jaccard overlap of posters of pairs of OCS categories.

to observe examples increases according to a power-law when syntactic and pragmatic variety decrease simultaneously. In other words, “poverty” of syntactical structuredness coincides more or less with “poverty” of pragmatic structuredness, and vice versa. That is, for the range of characteristics studied here the interplay of their scale-free behavior is such that if a manifestation of a (top-level section of a) Wikicussion is rare according to one dimension, it also tends to be rare according to a greater subset of other dimensions of the same feature set F.

Suppose that discussions in Wikipedia exhibit this kind of multidimensional scale invariance. *How does then a typical discussion look like?* Obviously, it would rather be very small with respect to the number of participants, topics (sections), turns (or posts) and subtopics (addressed within the same thread). Note that if discussions are distributed in terms of a family of power-laws, depending on the exact values of their exponents, one may even question the existence of a “typical discussion structure”. This is exactly, what we found in Section 6. Thus, when trying to analyze Wikicussions by relying on samples of large discussions, one runs the risk of overestimating one’s findings by considering far too rare cases. In other words, “typical” Wikicussions do not exist in terms of the variables considered here so that sampling Wikicussions, say, for the task of linguistic modeling is problematic when trying, for example, to build a *theory* of webgenres based thereon rather than just giving a picture of far too seldom phenomena. This is not to say that Wikicussions do not exhibit, for example, patterns of rhetorical structure or of argumentation. Rather, what one will not find is a typical size of such structures or a typical participation structure underlying them so that selecting and analyzing "longer" or even longest samples bears the risk of overestimating the kind of structure formation under consideration.

Suppose now a corpus of discussions that unfold in a deep (measured by the length of their threads) as well as in a branching manner (calculated as a function of the diversity of topics being addressed). If in such a case the participation shrinks such that only a couple of discussant or even a single interlocutor – who may finally coincide with the main author of the corresponding article – dominates the discussion, we finally arrive at an example of a kind of *mass communication*: few discussants write for many rather inactive recipients – *free-riders* in the sense of Antin and Cheshire (2010). Seemingly, the discussion space of the German Wikipedia tends to exhibit effects like this. In other words, posts tend to be posted by a smaller group of interlocutors while the majority of posters rather act as hardly active posters better not called "prosumers". In line with the linguistic model of Section 3, this scenario is coincident with a situation in which only a few or a single group member acts as a grounding "proxy" (Eshghi and Healey 2016) for the corresponding community. Under this regime, Wikicussions depart from dialogical communication in which common ground results from cooperating interlocutors.

Supposed that this diagnosis is not contradicted by a far more elaborated feature model, the question is raised how to arrive at higher degrees of participation securing more open, more active Wikicussions possibly allowing for higher article quality. At least one can interpret our findings as hinting at such a requirement. From this point of view it is *not only* a problem that the number of editors decreases over time. The same would also hold for the number of discussants and their rather hierarchical participation structure. However, we also stated that many correlations observed in Section 6 are not high enough to speak about a complete scale-free system (Hypothesis 4). Moreover, we did not yet consider a semantic model of Wikicussions. Such an elaboration would ideally be based on a dialog theoretical model of online communication as sketched in Section 3. To this end, one requires a model of common ground elaborated enough to capture the gist of single posts. However, one also needs a model of dialog acts and moves by example of online discussions that allows for mapping their functional structure (cf. Ferschke et al. 2012) while being computable based on corpora as large as those considered here. As a matter of fact, such a model is still future work.

How can we explain the Zipfian scale invariance of the gestalt of Wikicussions detected here? In Section 3 we argued that unlike face-to-face dialogs or multilogs, Wikicussions are rather open in terms of space and time (Kaltenbrunner and Laniado 2012) as well as in terms of participation structure and the sub-topics under discussion though being restricted by the framing topic of the corresponding article. That is, at any point in time, agents newly entering the conversation or re-entering the discussion may decide to link their posts to whatever turns being manifested in the past thereby evolving the tree-like structure of Wikicussions. Apparently, it is this type of the extensibility of Wikicussions that characterizes their scale invariance – across thematic and community-induced boundaries:

1. Firstly, we have to distinguish processes of thematic *innovation* according to which Wikipedia grows continually by articles about ever-new topics which, at the beginning of their life cycle, are not discussed. Secondly, we observe that already active agents continually re-enter already established Wikicussions to add new posts. By analogy with Simon (1955), we may speak about a mixture-process of thematic-participatory *association*

where agents coherently resume already established threads. Evidently, in cases where these agents contribute to novel articles, a mixture of thematic innovation and participatory association is given. Thirdly, the appearance of new agents entering the conversation space manifests a process of *participatory* innovation. In the case that they contribute to novel articles, a mixture of thematic and participatory innovation appears, while when they contribute to already existing talks, a mixture of thematic association (to already established topics) and participatory innovation is given.

2. Starting from this confusion matrix of thematic and participatory innovation and association, respectively, one can speculate about the emergence of scale invariance: both processes continually shift Wikicussions to higher “frequency classes” in terms of the (syntactic and pragmatic) statistics considered so far. Thereby, processes of thematic innovation ensure that “zero-class” Wikicussions enter the scene again and again. Apparently, the latter processes occur more frequently than the former ones so that one finally observes the characteristic dominance of *structural hapax* as described in Section 6 by means of the notion of multidimensional scale invariance. From a semantic point of view one can speak of a kind of *thematic hapax* as a result of the creation of ever new articles whose topics hardly get salient, so that these articles are unlikely to be discussed shortly after being created.

An indispensable prerequisite of this dynamics relates to the sort of extensibility of Wikicussions assessed in Section 3. In this sense, multidimensional scale invariance as detected here may be seen as a simple consequence of both the peculiarities of the webgenre *Wikicussion* (which are neither manifested by dialogs nor by multilogs) and the hypothetical scale-invariant distribution of thematic salience. According to this interpretation, one can speak of a dissolution of the boundaries of space and time: by having the possibility to reply to any turn at any time, scale invariant structures emerge that are characterized by infinite expected values and variances. These structures will now be related to the notion of self-similarity.

7.2. Self-similarity

Section 6.2 demonstrates that a system exhibiting multidimensional scale invariance in terms of (more or less) parallelized power laws along a whole regiment of features interferes with thematic classification. The resulting *confusability* of the gestalt of Wikicussions in terms of thematic provenance and the underlying participation structure is not just caused by the predominance of small units manifesting *structural hapax*. Rather it also concerns larger or even largest units (see Figure 18). Thus, we may speak of the self-similarity of Wikipedia’s discussion space in cases where its subsystems are demarcated thematically.

Generally speaking, scale invariance has been related to (approximate or statistical) self-similarity of fractal structures (Feder 1988; Harris and Stöcker 1998). Self-similar structures are related to a certain power-law, but not necessarily vice versa. However, this relation gives rise to speculate about the self-similarity of Wikicussions characterized by a multitude of homological power-laws.

Self-similarity can be analyzed on horizontal and vertical scale. Starting from a reference system (e.g., the Web) whose self-similarity is predicated, we speak of *horizontal* self-similarity on a given level of observational resolution (e.g., the level of websites), if the corresponding observational units (e.g., single sites) tend to be similar according to the operative similarity function (e.g., of structural similarity). In this sense, Dill et al. (2002), for example, describe the Web as consisting of interconnected, thematically unified clusters each exhibiting a bowtie structure.

In contrast to this, we speak of *vertical* self-similarity, if the latter similarity is observed for observational units in a recursive manner so that the structure of wholes resembles the one of their parts. This sort of self-similarity is exemplified by nested bowtie structures also observed by Dill et al. (2002) by example of the Web. While measuring horizontal self-similarity means comparing parts of the same whole, type or species, units are compared with their components in the case of vertical self-similarity.

According to the experiments of Section 6.2, we observe horizontal self-similarity regarding the distribution of tree-like gestalts of Wikicussions across the boundaries of subject areas and communities of posters: one does not know the underlying topic when knowing the gestalt of a Wikicussion in terms of our feature model. Articles of different subject areas are discussed in a way that results in similarly structured discussions. This observation is reflected by the fact that posters rarely cross the borders of communities as partitioned by OCS: people who tend to discuss, e.g., articles about biology hardly also discuss articles, say, about astronomy. Apparently, the self-similarity of Wikicussions does not result from larger intersections of commonly active posters, but from the self-organization of distributed communities of discussants in the sense sketched above. Beyond that we also observed vertical self-similarity by showing that *Top-Level Sections* (TLS) mirror the structure of Wikicussions. This relates to scale invariance and our correlation analysis. Figure 15 shows that correlation-based feature spaces derived from Wikicussions and from TLS are very similar in terms of the dynamics of cohesion, the sizes of largest connected components and average edge weights. Thus, Wikicussions simultaneously manifest two kinds of self-similarity: on horizontal and vertical scale.

Note that our findings coincide somehow with Laniado et al. (2011), but with the difference that we considered a larger set of statistics. Note also that while we observed self-similarity across the borders of subject areas, Laniado et al. indicate a contingency of membership to such classes and location parameters of syntactic features. However, since we performed an experiment of the size of OCP and OCS, we assume that the gestalt of a Wikicussion does not depend on the underlying topic and that multidimensional scale invariance is the main reason for this failure. Thus, we assume horizontal and vertical self-similarity for Wikicussions as long as there is no study falsifying this observation.

Based on these observations, we again ask for an analogy between self-similarity on the one hand and fractality on the other. Fractal linguistic structures have been studied with respect to Menzerath-Altmann's law (Altmann and Schwibbe 1989) and, thus, regarding interrelations of different levels of linguistic resolutions in natural language texts (Hřebíček 1992; Hřebíček 1995; Leopold 2001; Andres and Rypka 2012). More recently, Najafi and Darooneh (2015) apply the notion of fractal structures in automatic text analysis in order to develop a method for keyword extraction. A more critical view on using the concept of fractal structures can be attributed to Köhler (1997) and partly also to Leopold (2001) – but see Köhler (2014) for a more recent study of the notion of linguistic motifs from the point of view of fractality. Though we note the connection of self-similarity and scale invariance, we also have to detect certain differences: in the present article, we developed our apparatus by example of a rather non-mainstream linguistic construct which we call Wikicussion. Moreover, a direct translation of power-law exponents to fractal dimensions is problematic – see Leopold (2001) for a seminal account of what it means to ensure the interpretability of linguistic quantities in relation to fractality. At least we need more research to attribute fractality to the self-similar structure of Wikicussions and their top-level sections. A second analogy of our findings with respect to multidimensional scale-invariance relates to the omnipresence of scale invariance in many complex networks regarding, for example, degree distributions (Barabási and Albert 1999; Dorogovtsev and Mendes 2001). However, we have to state that Wikicussions are *tree-like* structures and that we stratified our model according to syntactic and pragmatic dimensions each of which had been further differentiated according to several sub-dimensions in order to finally account for *multidimensional* scale-invariance. In this way, we find out that Wikicussions are mainly scale-invariant tree-like structures, in terms of their syntactic and pragmatic structure.

8. Summary and Outlook

We developed and experimented with a model of multidimensional scale invariance by example of talk pages. We simultaneously studied syntactic and pragmatic features derived from a multimodal feature template. To this end, we computed 29 features by studying the power law-like scaling of the corresponding value distributions. We showed that with a single exception, all features are scale-free while a larger subset of them exhibits at least medium or even strong correlations. Both findings indicate a tendency towards multidimensional scale invariance: Wikicussions evolve into a state of scale-freeness that is simultaneously reflected by a whole regiment of dimensions. At the same time, we showed that while articles (and partly also discussions) are well separable by exploring their vocabularies using a state-of-the-art classifier based on neural networks (Joulin et al. 2016), this classification fails when considering our two-modal feature model. This finding points to a kind of horizontal self-similarity, which makes the shape of Wikicussions confusing beyond the boundaries of subject area and participation structure. At the same time, we detected a sort of vertical self-similarity according to which top-level sections mirror the structure of Wikicussions. In this way, we can begin to speculate on the fractality of this medium. Finally, by contrasting Wikicussions with dialogs and multilogs, we identified the extensibility of the former across space, time, subject area and participation structure as a probable candidate for explaining their scale invariance and self-similarity.

In our future work we want to extend the bridge between classical dialog theory on the one hand and computational webgenre analysis on the other, which we have developed in this article. This can be done by means of a model of common ground that captures the gist of posts and the functional structure of Wikicussions in terms of dialog acts while being computable by example of corpora as large as Wikipedia. Regarding our statistics, we plan to include semantic and temporal features into our comparative study of scale invariance. Another extension concerns *multimodal measures* operating on at least two modes (e.g., syntax and semantics). An example of such a multimodal quantity is given by the temporal dynamics of the h-index and by the m-index of Kaltenbrunner and Laniado (2012) relating the h-index to time. Finally, we plan a comparative analysis of Wikicussions of different languages.

Acknowledgment

Financial support by the Federal Ministry of Education and Research (BMBF) via the *Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences* CEDIFOR) is gratefully acknowledged. We would also like to thank our reviewers for their helpful hints, which helped us to improve this article.

References

Abramov, O. & Mehler, A. (2011). “Automatic Language Classification by Means of Syntactic Dependency Networks”. In: *Journal of Quantitative Linguistics* 18.4, pp. 291–336.

Altmann, G. & Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. München: Fink.

Altmann, G. & Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Olms, Hildesheim, Zürich, New York: Olms.

Andres, J. & Rypka, M. (2012). "Self-similar fractals with a given dimension and the application to quantitative linguistics". In: *Nonlinear Analysis: Real World Applications* 13.1, pp. 42–53. DOI: 10.1016/j.nonrwa.2011.07.009.

Antin, J. & Cheshire, C. (2010). "Readers Are Not Free-riders: Reading As a Form of Participation on Wikipedia". In: *Proc. of CSCW '10*. Savannah, Georgia, USA, pp. 127–130. DOI: 10.1145/1718918.1718942.

Arazy, O., Liifshitz-Assaf, H., Nov, O., Daxenberger, J., Balestra, M. & Cheshire, C. (2017). "On the "How" and "Why" of Emergent Role Behaviors in Wikipedia". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Portland, Oregon, USA: ACM, pp. 2039–2051. DOI: 10.1145/2998181.2998317.

Arazy, O., Nov, O., Patterson, R. & Yeo, L. (2011). "Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict". In: *Journal of Management Information Systems* 27.4, pp. 71–98. DOI: 10.2753/MIS0742-1222270403.

Asher, N. & Lascarides, A. (2013). "Strategic Conversation". In: *Semantics and Pragmatics* 6.2, pp. 1–62. DOI: 10.3765/sp.6.2.

Backstrom, L., Kleinberg, J., Lee, L. & Danescu-Niculescu-Mizil, C. (2013). "Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry". In: *Proc. of WSDM '13*. Rome, Italy, pp. 13–22. DOI: 10.1145/2433396.2433401.

Barabási, A.-L. & Albert, R. (1999). "Emergence of Scaling in Random Networks". In: *Science* 286, pp. 509–512.

Bender, E. M., Morgan, J. T., Oxley, M., Zachry, M., Hutchinson, B., Marin, A., Zhang, B. & Ostendorf, M. (2011). "Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages". In: *Proc. of LSM '11*. Portland, Oregon, pp. 48–57.

Botafogo, R. A., Rivlin, E. & Shneiderman, B. (1992). "Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics". In: *ACM Transactions on Information Systems* 10.2, pp. 142–180.

Bryant, S. L., Forte, A. & Bruckman, A. (2005). "Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia". In: *Proc. of GROUP '05*. Sanibel Island, Florida, USA, pp. 1–10. DOI: 10.1145/1099203.1099205.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., Schreuder, R. & Buttrick, S. (1983). "Common Ground and the Understanding of Demonstrative Reference". In: *Journal of Verbal Learning and Verbal Behavior* 22.2, pp. 245–258. DOI: 10.1016/S0022-5371(83)90189-5.

Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A. & Tucci, G. (2012). "Reconstruction and Analysis of Twitter Conversation Graphs". In: *Proc. of HotSocial '12*. Beijing, China, pp. 25–31. DOI: 10.1145/2392622.2392626.

Dignum, F. P. M. & Vreeswijk, G. A. W. (2004). "Towards a Testbed for Multi-Party Dialogues". In: *Advances in Agent Communication*. Ed. by F. Dignum. Berlin/Heidelberg: Springer, pp. 212–230. DOI: 10.1007/978-3-540-24608-4_13.

Dill, S., Kumar, R., Mccurley, K. S., Rajagopalan, S., Sivakumar, D. & Tomkins, A. (2002). "Self-similarity in the web". In: *ACM Trans. Internet Technol.* 2.3, pp. 205–223.

Dohrn, H. & Riehle, D. (2013). "Design and implementation of wiki content transformations and refactorings". In: *Proc. of OpenSym '13*, 2:1–2:10.

Dori-Hacohen, S., Jensen, D. & Allan, J. (2016). "Controversy Detection in Wikipedia Using Collective Classification". In: *Proc. of SIGIR '16*. Pisa, Italy, pp. 797–800. DOI: `10.1145/2911451.2914745`.

Dorogovtsev, S. N. & Mendes, J. F. F. (2001). "Language as an evolving word web". In: *Proceedings of The Royal Society of London. Series B, Biological Sciences* 268.1485, pp. 2603–2606.

Eger, S., Gleim, R. & Mehler, A. (2016). "Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art". In: *Proc. of LREC'16*. Portorož (Slovenia).

Egghe, L. & Rousseau, R. (2003). "BRS-compactness in networks". In: *Mathematical and Computer Modelling* 37.7-8, pp. 879–899.

Eshghi, A. & Healey, P. G. T. (2016). "Collective Contexts in Conversation: Grounding by Proxy". In: *Cognitive Science* 40.2, pp. 299–324. DOI: `10.1111/cogs.12225`.

Feder, J. (1988). *Fractals*. New York: Plenum Press.

Fernández, R. & Endriss, U. (2007). "Abstract Models for Dialogue Protocols". In: *Journal of Logic, Language and Information* 16.2, pp. 121–140. DOI: `10.1007/s10849-006-9032-z`.

Fernández, R. & Ginzburg, J. (2002). "Non-Sentential Utterances: A Corpus Study". In: *Traîtement Automatique de Languages* 43.2, pp. 13–42.

Ferron, M. & Massa, P. (2014). "Beyond the encyclopedia: Collective memories in Wikipedia". In: *Memory Studies* 14.1, pp. 22–45.

Ferschke, O., Gurevych, I. & Chebotar, Y. (2012). "Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages". In: *Proc. of EACL '12*. Avignon, France, pp. 777–786.

Freyd, J. J. (1983). "Shareability: The Social Psychology of Epistemology". In: *Cognitive Science* 7, pp. 191–210.

Ginzburg, J. (1994). "An Update Semantics for Dialogue". In: *Proc. of the first International Workshop on Computational Semantics*. Tilburg, The Netherlands.

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford, UK: Oxford University Press.

Ginzburg, J. & Fernández, R. (2005). "Action at a Distance: the Difference between Dialogue and Multilogue". In: *Proc. of DIALOR'05*. Nancy, France.

Ginzburg, J. & Poesio, M. (2016). "Grammar Is a System That Characterizes Talk in Interaction". In: *Frontiers in Psychology* 7, p. 1938. DOI: `10.3389/fpsyg.2016.01938`.

Goel, S., Anderson, A., Hofman, J. & Watts, D. J. (2016). "The Structural Virality of Online Diffusion". In: *Management Science* 62.1, pp. 180–196. DOI: `10.1287/mnsc.2015.2158`.

Goffman, E. (1981). *Forms of Talk*. University of Pennsylvania Publications in Conduct and Communication. Philadelphia, Pennsylvania: University of Pennsylvania Press.

Gómez, V., Kaltenbrunner, A. & López, V. (2008). "Statistical Analysis of the Social Network and Discussion Threads in Slashdot". In: *Proc. of WWW '08*. Beijing, China, pp. 645–654. DOI: `10.1145/1367497.1367585`.

Gómez, V., Kappen, H. J. & Kaltenbrunner, A. (2011). "Modeling the structure and evolution of discussion cascades". In: *Proc. of HT '11*. Eindhoven, The Netherlands, pp. 181–190.

Gonzalez-Bailon, S., Kaltenbrunner, A. & Banchs, R. E. (2010). "The structure of political discussion networks: a model for the analysis of online deliberation". In: *Journal of Information Technology* 25, pp. 230–243.

Grice, H. P. (1975). "Logic and Conversation". In: *Syntax and Semantics*. Ed. by P. Cole **&** J. Morgan. Vol. 3, Speech Acts. New York: Academic Press, pp. 41–58.

Haken, H. (1998). "Can we Apply Synergetics to the Human Sciences?" In: *Systems. New Paradigms for Human Sciences*. Ed. by G. Altmann & W. A. Koch. Berlin/New York: De Gruyter.

Hara, N., Shachaf, P. & Hew, K. F. (2010). "Cross-cultural Analysis of the Wikipedia Community". In: *J. Am. Soc. Inf. Sci. Technol.* 61.10, pp. 2097–2108. DOI: `10.1002/asi.v61:10`.

Harris, J. W. & Stöcker, H. (1998). *Handbook of Mathematics and Computational Science*. New York: Springer.

Hoque, E. & Carenini, G. (2015). "ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations". In: *Proc. of IUI '15*. Atlanta, Georgia, USA, pp. 169–180. DOI: `10.1145/2678025.2701370`.

Hřebíček, L. (1992). *Text in Communication: Supra-Sentence Structures*. Bochum: Brockmeyer.

Hřebíček, L. (1995). *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*. Trier: Wissenschaftlicher Verlag.

Iosub, D., Laniado, D., Castillo, C., Fuster Morell, M. & Kaltenbrunner, A. (2014). "Emotions under Discussion: Gender, Status and Communication in Online Collaboration". In: *PlOS ONE* 9.8, pp. 1–23. DOI: `10.1371/journal.pone.0104880`.

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016). "Bag of Tricks for Efficient Text Classification". In: *CoRR* abs/1607.01759.

Kaltenbrunner, A., Gonzalez-Bailon, S. & Banchs, R. E. (2009). "Communities on the Web: Mechanisms Underlying the Emergence of Online Discussion Networks". In: *Proc. of WebSci'09*. Athens, Greece.

Kaltenbrunner, A. & Laniado, D. (2012). "There is No Deadline – Time Evolution of Wikipedia Discussions". In: *Proc. of WikiSym'12*. Linz, Austria.

Kempson, R., Cann, R., Gregoromichelaki, E. & Chatzikyriakidis, S. (2016). "Language as Mechanisms for Interaction". In: *Theoretical Linguistics* 42.3-4, pp. 203–276. DOI: `10.1515/tl-2016-0011`.

Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B. & Mytkowicz, T. (2007a). "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie". In: *Proc. of CHI '07*.

Kittur, A. & Kraut, R. E. (2008). "Harnessing the wisdom of crowds in Wikipedia: quality through coordination". In: *Proc. of CSCW '08*. San Diego, CA, USA, pp. 37–46.

Kittur, A., Suh, B., Pendleton, B. A. & Chi, E. H. (2007b). "He says, she says: conflict and coordination in Wikipedia". In: *Proc. of CHI '07*. San Jose, California, USA, pp. 453–462. DOI: `10.1145/1240624.1240698`.

Koch, P. & Oesterreicher, W. (1985). "Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte". In: *Romanistisches Jahrbuch* 36, pp. 15–43.

Köhler, R. (1993). "Synergetic Linguistics". In: *Contributions to Quantitative Linguistics*. Ed. by R. Köhler & B. B. Rieger. Dordrecht: Kluwer, pp. 41–51.

Köhler, R. (1997). "Are there Fractal Structures in Language? Units of Measurement and Dimensions in Linguistics". In: *Journal of Quantitative Linguistics* 4.1-3, pp. 122–125.

Köhler, R. (1999). "Syntactic Structures. Properties and Interrelations". In: *Journal of Quantitative Linguistics* 6, pp. 46–57.

Köhler, R. (2014). "The fractal structure of linguistic motifs". In: *Empirical approaches to text and language analysis. Dedicated to Luděk Hřebíček on the occasion of his 80th birthday*. Ed. by G. ltmann, R. Cech, J. Macutek & L. Ulírová. Vol. 17. Studies in Quantitative Linguistics. Lüdenscheid: RAM-Verlag, pp. 94–104.

Kriplean, T., Beschastnikh, I. & McDonald, D. W. (2008). "Articulations of Wikiwork: Uncovering Valued Work in Wikipedia Through Barnstars". In: *Proc. of CSCW '08*. San Diego, CA, USA, pp. 47–56. DOI: 10.1145/1460563.1460573.

Krishnan, S., Butler, P., Tandon, R., Leskovec, J. & Ramakrishnan, N. (2016). "Seeing the Forest for the Trees: New Approaches to Forecasting Cascades". In: *Proc. of WebSci '16*. Hannover, Germany, pp. 249–258. DOI: 10.1145/2908131.2908155.

Kumar, R., Mahdian, M. & McGlohon, M. (2010). "Dynamics of Conversations". In: *Proc. of KDD '10*. Washington, DC, USA, pp. 553–562. DOI: 10.1145/1835804.1835875.

Kutas, M., DeLong, K. A. & Smith, N. J. (2011). "A Look around at What Lies Ahead: Prediction and Predictability in Language Processing". In: *Predictions in the Brain: Using our Past to Generate a Future*. Ed. by M. Bar. Oxford and New York: Oxford University Press. Chap. 15, pp. 190–207.

Laniado, D., Castillo, C., Kaltenbrunner, A. & Fuster-Morell, M. (2012). "Emotions and dialogue in a peer-production community: the case of Wikipedia". In: *Proc. of WikiSym'12*. Linz, Austria.

Laniado, D., Tasso, R., Volkovich, Y. & Kaltenbrunner, A. (2011). "When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages". In: *Proc. of ICWSM'11*.

Leopold, E. (2001). "Fractal structure in language. The question of embedding space". In: *Text as Linguistic Paradigm: Levels, Constituents, Construents: Festschrift in honour of Luděk Hřebíček*. Ed. by R. Köhler, L. Uhlířová **&** G. Wimmer, pp. 163–176.

Lerner, G. H. (1993). "Collectivities in Action: Establishing the Relevance of Conjoined Participation in Conversation". In: *Text – Interdisciplinary Journal for the Study of Discourse* 13.2, pp. 213–246. DOI: 10.1515/text.1.1993.13.2.213.

Lerner, G. H. (2004). "Collaborative Turn Sequences". In: *Conversation Analysis. Studies from the First Generation*. Ed. by G. H. Lerner. Amsterdam and Phildelphia, PA: John Benjamins Publishing Company, pp. 225–256.

Lewis, D. (1969). *Conventions. A Philosophical Study*. Cambridge: Harvard U.P.

Lücking, A., Hoenen, A. & Mehler, A. (2016). "TGermaCorp – A (Digital) Humanities Resource for (Computational) Linguistics". In: *Proc. of LREC'16*. Portorož (Slovenia).

Margaretha, E. & Lüngen, H. (2014). "Building Linguistic Corpora from Wikipedia Articles and Discussions". In: *JLCL* 29.2, pp. 59–82.

Marin, A., Zhang, B. & Ostendorf, M. (2011). "Detecting Forum Authority Claims in Online Discussions". In: *Proc. of LSM '11*. Portland, Oregon, pp. 39–47.

Mehler, A. (2011). "Social Ontologies as Generalized Nearly Acyclic Directed Graphs". In: *Towards an Information Theory of Complex Networks*. Boston/Basel: Birkhäuser, pp. 259–319.

Mehler, A., Gleim, R. & Wegner, A. (2007). "Structural Uncertainty of Hypertext Types. An Empirical Study". In: *Proc. of the RANLP Workshop* Towards Genre-Enabled Search Engines: The Impact of NLP, pp. 13–19.

Mehler, A. & Stegbauer, C. (2012). "On the self-similarity of intertextual structures in Wikipedia". In: *Proc. of HotSocial '12*. Beijing, China, pp. 65–68. DOI: 10.1145/2392622.2392633.

Mikolov, T., Yih, W. & Zweig, G. (2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proc. of NAACL 2013*, pp. 746–751.

Mir, A., Rosselló, F. & Rotger, L. (2013). "A new balance index for phylogenetic trees". In: *Mathematical Biosciences* 241.1, pp. 125–136. DOI: 10.1016/j.mbs.2012.10.005.

Morris, C. W. (1938). *Foundations of the Theory of Signs (International encyclopedia of unified science)*. Chicago: Chicago University Press.

Müller, T., Schmid, H. & Schütze, H. (2013). “Efficient Higher-Order CRFs for Morphological Tagging”. In: *Proc. of EMNLP ’13*, pp. 322–332.

Najafi, E. & Darooneh, A. H. (June 2015). “The Fractal Patterns of Words in a Text: A Method for Automatic Keyword Extraction”. In: *PLOS ONE* 10.6, pp. 1–18. DOI: 10.1371/journal.pone.0130617. URL: https://doi.org/10.1371/journal.pone.0130617.

Newman, M. E. J. (2002). “Assortative Mixing in Networks”. In: *Physical Review Letters* 89.20, p. 208701.

Newman, M. E. J. (2005). “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary Physics* 46, pp. 323–351.

OECD (2007). *Revised Field of Science and Technology (FOS)*. www.oecd.org/science/inno/38235147.pdf.

Oxley, M., Morgan, J. T., Zachry, M. & Hutchinson, B. (2010). ““What I Know is...”: Establishing Credibility on Wikipedia Talk Pages”. In: *Proc. of WikiSym ’10*. Gdansk, Poland, 26:1–26:2. DOI: 10.1145/1832772.1832805.

Pascual-Cid, V. & Kaltenbrunner, A. (2009). “Exploring Asynchronous Online Discussions through Hierarchical Visualisation”. In: *Proc. of 13th Int. Conf. IV*, pp. 191–196. DOI: 10.1109/IV.2009.14.

Pickering, M. J. & Garrod, S. (2013). “An Integrated Theory of Language Production and Comprehension”. In: *Behavioral and Brain Sciences* 36.4, pp. 329–347. DOI: 10.1017/S0140525X12001495.

Poesio, M. (1995). “A model of Conversation Processing Based on Micro Conversational Events”. In: *Proc. of the Annual Meeting of the Cognitive Science Society*. Pittsburgh, pp. 698–703.

Poesio, M. & Rieser, H. (2010). “Completions, Coordination, and Alignment in Dialogue”. In: *Dialogue & Discourse* 1.1, pp. 1–89. DOI: 10.5087/dad.2010.001.

Poesio, M. & Traum, D. (1997). “Conversational Actions and Discourse Situations”. In: *Computational Intelligence* 13.3, pp. 309–347.

Poncin, K. & Rieser, H. (2006). “Multi-Speaker Utterances and Co-Ordination in Task-Oriented Dialogue”. In: *Journal of Pragmatics* 38.5, pp. 718–744. DOI: 10.1016/j.pragma.2005.06.013.

Preece, J., Nonnecke, B. & Andrews, D. (2004). “The top five reasons for lurking: improving community experiences for everyone”. In: *Computers in Human Behavior* 20.2, pp. 201–223. DOI: 10.1016/j.chb.2003.10.015.

Qin, X., Cunningham, P. & Salter-Townshend, M. (2015). “The influence of network structures of Wikipedia discussion pages on the efficiency of WikiProjects”. In: *Social Networks* 43, pp. 1–15.

Ramli, N. & Mohamad, D. (2009). “On the Jaccard index similarity measure in ranking fuzzy numbers”. In: *Matematika* 25, pp. 157–165.

Roberts, C. (2012). “Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics”. In: *Semantics and Pragmatics* 5.6, pp. 1–69. DOI: 10.3765/sp.5.6.

Ruiter, J. P. de, Mitterer, H. & Enfield, N. J. (2006). “Projecting the End of a Speaker’s Turn: A Cognitive Cornerstone of Conversation”. In: *Language* 82.3, pp. 515–535. DOI: 10.1353/lan.2006.0130.

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). “A Simplest Systematics for the Organization of Turn-Taking for Conversation”. In: *Language* 50.4, pp. 696–735.

Santini, M., Mehler, A. & Sharoff, S. (2010). “Riding the Rough Waves of Genre on the Web: Concepts and Research Questions”. In: *Genres on the Web: Computational Models and Empirical Studies*. Ed. by A. Mehler, S. Sharoff & M. Santini. Dordrecht: Springer, pp. 3–32.

Schneider, J., Passant, A. & Breslin, J. G. (2011). “Understanding and Improving Wikipedia Article Discussion Spaces”. In: *Proc. of SAC ’11*. TaiChung, Taiwan, pp. 808–813. DOI: 10.1145/1982185.1982358.

Schneider, J., Passant, A. & Decker, S. (2012). “Deletion Discussions in Wikipedia: Decision Factors and Outcomes”. In: *Proc. of WikiSym ’12*. Linz, Austria, 17:1–17:10. DOI: 10.1145/2462932.2462955.

Schober, M. F. & Clark, H. H. (1989). “Understanding by Addressees and Overhearers”. In: *Cognitive Psychology* 21.2, pp. 211–232. DOI: 10.1016/0010-0285(89)90008-X.

Simon, H. A. (1955). “On a Class of Skew Distribution Functions”. In: *Biometrika* 42, pp. 425–440.

Stalnaker, R. (2002). “Common Ground”. In: *Linguistics and Philosophy* 25, pp. 701–721.

Stegbauer, C. (2016). *Grundlagen der Netzwerkforschung: Situation, Mikronetzwerke und Kultur*. Wiesbaden: Springer-VS.

Stegbauer, C. & Bauer, E. (2010). “Die Entstehung einer positionalen Struktur durch Konflikt und Kooperation bei Wikipedia: Eine Netzwerkanalyse”. In: *Medienwandel als Wandel von Interaktionsformen: von frühen Medienkulturen zum Web 2.0*. Ed. by A. Mehler **&** T. Sutter. Wiesbaden: Verlag für Sozialwissenschaften, pp. 231–255.

Székely, G. J. & Rizzo, M. L. (2009). “Brownian distance covariance”. In: *The Annals of Applied Statistics* 3.4, pp. 1236–1265.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995). “Integration of Visual and Linguistic Information in Spoken Language Comprehension”. In: *Science* 268.5217, pp. 1632–1634. DOI: 10.1126/science.7777863.

Tapscott, D. & Williams, A. D. (2008). *Wikinomics: How Mass Collaboration Changes Everything*. New Jork: Portfolio.

Thorley, J. L. & Wilkinson, M. (2007). *The RadCon Manual v. 1.1.3*. http://www.bmnh.org/systematics/RadConManual/.

Traum, D. R. & Larsson, S. (2003). “The Information State Approach to Dialogue Management”. In: *Current and New Directions in Discourse and Dialogue*. Ed. by J. Kuppevelt, R. W. Smith **&** N. Ide. Vol. 22. Amsterdam, The Netherlands: Springer, pp. 325–353.

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Trier: Wissenschaftlicher Verlag.

Viégas, F. B., Wattenberg, M. & Dave, K. (2004). “Studying Cooperation and Conflict Between Authors with History Flow Visualizations”. In: *Proc. of CHI ’04*. Vienna, Austria, pp. 575–582. DOI: 10.1145/985692.985765.

Viégas, F. B., Wattenberg, M., Kriss, J. & Ham, F. V. (2007). “Talk before you type: Coordination in Wikipedia”. In: *Proc. of HICSS 40*. Society Press.

Wang, C., Ye, M. & Huberman, B. A. (2012). “From User Comments to On-line Conversations”. In: *Proc. of KDD ’12*. Beijing, China, pp. 244–252. DOI: 10.1145/2339530.2339573.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G. & Smith, M. (2011). “Finding Social Roles in Wikipedia”. In: *Proc. of iConference ’11*. Seattle, Washington, USA, pp. 122–129. DOI: 10.1145/1940761.1940778.

Weninger, T., Zhu, X. A. & Han, J. (2013). “An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community”. In: *Proc. of ASONAM ’13*. Niagara, Ontario, Canada, pp. 579–583. DOI: 10.1145/2492517.2492646.

Yasseri, T., Sumi, R., Rung, A., Kornai, A. & Kertész, J. (2012). “Dynamics of conflicts in Wikipedia”. In: *PloS ONE* 7.6, e38869.

Zhang, K. & Shasha, D. (1989). "Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems". In: *SIAM Journal on Computing* 18.6, pp. 1245–1262. DOI: 10.1137/0218082.

Glottometrics 40, 2017, 46-50

Linking Elements of German Compounds in the Texts of Technical Science

Anastasia Gnatciuc[1]
Hanna Gnatchuk[2]

Abstract: In this investigation we deal with the analysis of German compounds in the technical texts. Our attention is drawn to the study of linking elements in Book “Wirtschaftsinformatik” by H. R. Hansen et al (2015). The corpus of our study includes 206 German compounds taken from 20 pages of the book under consideration. The data have undergone statistical processing. The outcomes can be of great use to typological research.

Keywords: *German, compounds, technical language*

1. Introduction: some remarks on German compounds

It is a well-known fact that any language is constantly undergoing certain linguistic changes: i.e. the appearance of new words or the disappearance of old words. E. Donalies (2005), W. Fleischer (2012), M. D. Stepanova (1953) admit the leading tendency of compounds in the German language. Stepanova (1953) takes the view that German compounds consist of two or several roots or the roots with word-building affixes. She classifies German compounds according to two principles:

1. According to structural and genetic types of compounds;
2. According to syntactic and semantic connections between the components of compounds.

It must be remarked that various other definitions of classes are possible, and the kind of joining may be scaled. Here, we shall adhere to Stepanova’s view.

Considering structural and genetic types of compounds, Stepanova distinguishes a) *complete unions;* b) *incomplete unions;* c) *shifts*. As far as the *complete unions* are concerned, there is no linking element in their structure: *der Schreibtisch, der Seemann, das Landhaus*. The in*complete unions* are the compounds where the first component is attached to another component by means of a certain linking element. On the whole, it is possible to find the following linking elements in the German language: *-(e)s, -(e)n, -(e)r, -e, -i, -o, -ens*: *Neoroinformatik, Männerstimme, Hundefell, Weizenbau, Geisteswissenschaft, Nachtigall, Pferdestahl, der Kindergarten.* Focusing on the origin of the following German linking elements of the compounds *–(e)n, -(e)r, -e,* they refer to the inflections of genitive case, plural form; *-(e)s* refers to the inflection of genitive case, singular form. These functions were lost at a certain period of time. Nowadays, they are used in the following cases of German compounds:

[1] Anastasia Gnatciuc. Alpen-Adria Universität, Institut für Slawistik, Universitätsstraße 65-67.
[2] Hanna Gnatchuk. Alpen-Adria Universität, Institut für Slawistik, Universitätsstraße 65-67. Email: agnatchuk@gmail.com

- Linking element -*s*- is used after a specifying word with the so-called “heavy” suffix (= the suffix which includes 1 or two consonants): *-heit, -keit, -schaft,* etc: *Frei**heits**liebe*;
- Linking element *–e, -er, -en* can be used in the plural forms of certain nouns: *Hundefell, Datenmodelle, Fächerkatalog, Kindergarten*. M.D. Stepanova (1953) emphasizes the fact that the usage of a linking element depends upon the traditions found in the language.

The main function of a linking element is to make the pronunciation of a compound easier. Sometimes one can find two or more consonants between two roots of a compound. In order to overcome the difficulties in the pronunciation of these words, one uses a linking element: i.e. *Stell**en**markt, Zustand**s**diagram, Präsentation**s**technik*.

The linking elements are not to be found in the group of *shifts*. The shifts are the compounds whose components are combined as a word combination (phrase) or a sentence: i.e. *die Blindekuh = eine blinde Kuh*. In such a way, the classification according to structural and genetic types is based upon the presence or absence of the linking element.

Stepanova (1953) divides German compounds according to syntactic and semantic connections between the components. On this basis she distinguishes the following types:

a) Determinative compound-nouns;
b) Coordinative compound-nouns;
c) Words-sentences or imperative names.

Dealing with the *determinative compound-nouns*, the first component of this type determines or specifies the second element: *Schwarzbrot, Weißbrot*. In most cases the first component is represented by the basic word of any part of speech: *der Französischunterricht, das Elternhaus, der Klassenleiter*. In addition to it, the last element (the basic word) of the compounds of this type determines gender, type of declension as well as plural forms. As far as the lexical aspect of the compound is concerned, the first and the second elements are of great importance;

Coordinative compounds consist of the compounds connected by coordinating conjunctions (i.e. *and, but*). None of these elements can specify the meaning of another component: *das Katz-Maus-Spiel, althochdeutsch*. In this case, the components of a compound belong to the same grammatical and lexical system. Unlike the determinative compounds, the coordinative ones are less productive.

The imperative noun-compounds are represented by sentences with a verb in imperative form: *das Vergiss-mein-nicht, das Ruhr-mich-nicht-an*. Imperative compounds in German are less productive as a word-building pattern.

2. The analysis of German compounds in terms of the linking elements

The task of the empirical research consists in classifying German compounds in the technical texts according to their linking elements. We aim here to reveal the frequencies of German compounds according to the linking elements and draw a comparison with English compounds.

The data of the research are represented by the book “Wirtschaftsinformatik” (2015) where 20 pages have been studied. In such a way, we have selected German compounds on each page. As a result, we’ve received 7 types of linking elements of the German compounds. The results are illustrated below:

1) **Compounds without joining elememts:** *Schriftsteller, Grenzverschiebung, Spielkarten, Konsumartikel, Teilaspekt, Deutschland, Weltkrieg, Frankreich,*

Fachkraft, Fachrichtung, Kulturhoheit, Finanznote, Fachbereit, fachbereit, Wechselweise, Wechselwirkung, Ingenieurinformatik, Ingenieurseite, Regeltechnik, Sprachraum, Titeländerung, Grundlagen, Fachredaktion, Sprachschöpfung, Informatikstudium, Bibliothekwissenschaftler, Fragestellung, Handbedingung, Grundbegriff, Systemtheorie, Spieltheorie, Prozessautomatisierung, Abschlussarbeit, Informatikausbildung, Programmieraufwand, Medizininformatik, Mustererkennung, Informatikfrage, Naturwissenschaft, Informatikmethode, Theorieverständnis, Strukturwissenschaft, Ingenieurwissenschaft, Technikwissenschaft, Theoriebildung, Systemprogrammierung, Schutzwall, Notlösung, Computerindustrie, Autoindustrie, Kopfarbeit, Netzwerk, Stausauger, Handwerk, Kunstfertigkeit, Restbestand, Informatikfirme, Modellbildung, Computergrafik, Lehrbereich, Lehrplan, Lehrbuch, deutschsprachig, mittelfristig, erfolgreich, informatikrelevant, kampfbereit, wertfrei, Informatikfachbereich, Zeitschriftartikel, Fachzeitschrift, Informatiklehrbuch, Stichwortgebe, Hochschule, Neugründung, Neubestimmung, Bereitstellung, vollziehen, hochqualifizieren, wahrnehmen, gleichberechtigen, bereitstellen, Langfristziel, Nebenfachregelung, Übersichtband, gleichzeitig, formallogisch, teilnehmen, Kopfzerbrechen, allgemein, alltäglich, Allmachtphantasie, Nebenfach, außeruniversitär, kontraproduktiv, Grenzüberschreitung, wenngleich, Selbstverständnis, drittmittelvorhaben, vielfältig, auseinandersetzung, Selbstüberschätzung;

2) **-s-**: *Geltungsanspruch, Informationsgesellschaft, Antrittsvorlesung, Forschungsansatz, Gehaltsskalen, Bundesrepublik, Ausbildungsprogramm, Eröffnungsrede, Wirtschaftswissenschaftler, Forschungsprogramm, Arbeitsgruppe, Wirtschaftsinformatik, Beschreibungsverfahren, Kommunikationsaspekt, Ausbildungsgang, Informationstätigkeit, Informationswissenschaft, Dokumentationswissenschaft, Informationsnutzer, Forschungsinstitut, Forschungsfrage, Forschungsbereich, Anwendungsbereich, Gründungsphase, Forschungsvorhaben, Inhaltsverzeichnis, Schaltungsentwurf, Informationsbegriff, Geisteswissenschaft, Wissenschaftsklassifikation, Formierungsansatz, Automatisierungstechnik, Kodierungstheorie, Entfaltungsmöglichkeit, Kooperationsmöglichkeit, Anwendungslücke, Anpassungsdruck, Forschungsprogramm, Abgrenzungsentscheidung, Korrekturheitsproblem, Anwendungsproblem, Verwaltungsinformatik, Betriebswirtschaft, Informationsbegriff, Klassifikationsproblem, Forschungsführer, Prüfungsarbeit, Verfahrenstechnik, Volkswirtschaft, Gründungsphase, Rationalisierungstechniken, Produktionsbereich, Unternehmensberater, Jahrestagung, Diskussionsbeitrag, Lösungsansatz, Geschmacksfrage, Kooperationspartnerin, Berufsentscheidung, Unterstützungssystem, Engelsgeduld, Arbeitsprozess, Komplexitätstheorie, Modellierungsmöglichkeit, Zustandsdiagramm, Vermittlungstechnik, Präsentationstechnik, forschungspolitisch, wissenschaftspolitisch, verwaltungsrechtlich, Bundesforschungsminister, Überlebensstrategie, Alltagsgegenstand, Berufsalltag, überarbeitungsbedürftig;*

3) **– en-:** *Datenverarbeitung, Studiengang, Studienfach, Datentechnik, Stellenwert, Datenschutz, Methodenlücke, Studienführer, Stellenmarkt, Rahmenrichtung, Weizenbau, Datenmodelle, maschinennahe, Datenbanktechniken;*

4) **– er-:** *Fächerkatalog, Rechnernetz, Rechnerstruktur, Rechnerunterstützung, Speichertechnik, Halbleitertechnik;*

5) **Hyphenized compounds:** *Informatik-Werk, Akademie-Definition, Mensch-Maschine-Kommunikation, künstliche-Intelligenz-Forschung, geistig-philosophisch;*

6) **– e -:** *Gerätetechnik, Hundefell, Gerätefixiert;*

7) – **o** -: *Neuroinformatik*

The results are also represented in Table 1.

Table 1
The frequencies of linking elements for German compounds in technical texts

Rank	Pattern	Frequencies	Computed values (Zipf-Alekseev)
1	**Compounds without joining elements**	102	102.04
2	**-s-**	75	74.72
3	**-en-**	14	15.81
4	**-er-**	6	2.85
5	**Hyphenized compounds**	5	0.53
6	**-e-**	3	0.11
7	**-o-**	1	0.02

		206	a = 1.6829,
			b = -3.0766,
			c = 102.0406
			R^2 = 0.9958

In such a way, we have revealed 7 types of linking elements for German compounds in the technical texts. Similar to English compounds (Gnatchuk, 2016) in the scientific texts, blank compounds proved to be the most productive (= have the highest frequencies).

As is usual in any linguistic data, if there are several classes, they are not created with the same intensity. It has been shown many times that classes with different frequencies can always be at least ranked. The simplest approach has been initiated by G.K. Zipf who used inductively the power function which is adequate in many cases. It must be noted that considering some regularity a probability distribution or a simple function does not change the theoretical background. The only difference is the normalizing defined for our model, not given in the reality. We shall adhere to the functions which can easily be derived from the unified theory (cf. Wimmer, Altmann 2005). Since in some cases the power function does not yields satisfactory results, caused perhaps by the effects originating with the speaker or hearer or with the complexity of definition of classes, Mandelbrot (1959) derived a slightly more complex function. But in many cases human perception and impact are not straightforward (linear) but intuitively transformed in a logarithmic operation, Alekseev (1987) proposed another function given as

(1) $y = c^*x^{a + b \log x.}$

which can be obtained from the differential equation

(2) $$\frac{dy}{y} = \frac{A + B\log x}{Cx}dx$$

after reparametrization, in which the effect of speaker/writer is rather logarithmic. Applying (1) to our data we obtain the results presented in the last column of Table 1. The simple Zipf function yields R^2 = 0.86, the Zipf-Mandelbrot function yields R^2 = 0.92 and the Zipf-Alekseev function yields R^2 = 0.9958.

Changing the manner of classification or quantification one would obtain other sequences but all of them would follow the Zipf-Alekseev function. To obtain a still more realistic image, one could insert in the differential equation y-1 (instead of y).

References

Alekseev, Pavel M. (1987). Quantitative Typology of Texts. In: *Glottometrika 8* (Hrsg. Ingeborg Fickermann). Bochum: Brockmeyer, 202.

Donalies, E. (2005). *Die Wortbildung der deutschen Gegenwartssprache. 4. Auflage*. Berlin: Walter de Gruyter.

Gnatchuk, H. (2016). A Quantitative Analysis of English Compounds in Scientific Texts. *Glottometrics 33, 1-7.*

Fleischer, W., Barz, I. (2012). *Wortbildung der deutschen Gegenwartssprache.* 4. Auflage. Walter de Gruyter.

Hansen, H.R., Mendling, J., Neumann, G. (2015). *Wirtschaftsinformatik*. De Gruyter Studium

Mandelbrot, B. (1959). A Note on a Class of Skew Distribution Functions. Analysis and Critique of a Paper by H. Simon. *Information and Control 2, 90-99.*

Stepanova, M.D. (1953). *Slovoobrazovanije sovremennogo nemetskogo jaz'ka* [Word formation of contemporary German]. M.: Izdatelstvo literatur‘ na inostrann'h jaz'kah (in Russian)

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

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge, Mass.

Glottometrics 40, 2018, 51-62

On the Development of Old Czech (En)clitics

Kosek Pavel, Čech Radek, Navrátilová Olga, Mačutek Ján

Abstract. The presented study deals with the historical development of Czech (en)clitics (AuxP). Based on the data from the previous research (Kosek, 2015a,b, 2017), it focuses on the development of one group the Czech (en)clitics – on the preterite auxiliary forms. In the article, three hypotheses are formulated and then tested on the data gained from selected parts of historical Czech Bible translations. The suggest that there were two significant word order positions of historical Czech (en)clitics: 1. the post-initial position, i.e. after first word / phrase, 2. the contact position, i.e. an (en)clitic is located immediately before (pre-verbal position) or after (post-verbal position) its syntactically or morphologically superordinate item (the post-verbal position is the more frequent variant of the both variants of the contact positions). Since the time when the oldest analyzed text was translated, the post-initial position has had the status of the basic word order position of the Czech (en)clitic, while the contact position has had the status of a stylistically, pragmatically or textually motivated position. It seems that the contact position begins to retreat only in 19th century and hence the definitive historical change of Czech auxiliary (en)clitics in the sole second position clitics was realized not before 19th or 20th century.

Keywords: Czech, enclitics

1. Introduction

The development of language characteristics is determined by various kinds of mechanisms influencing human linguistic behaviour. Over time, some mechanisms grow stronger and finally acquire the status of a grammatical rule, while other mechanisms gradually weaken and can even disappear. In this study, we attempt to model the development of the word order of Czech preterite auxiliary (en)clitics (AuxP) of the type *nesl jsem* 'I carried' in Old Czech. We hypothesize that this development is systematically determined by a hierarchy of mechanisms and also that the "strength" of these mechanisms is not constant but varies in different periods. As for the hierarchy of mechanisms, there seem to be three strong mechanisms which can be expressed in the form of stochastic rules for the word order of (en)clitics:

> R1: if an (en)clitic appears in a clause, use it after the initial phrase of the clause (i.e. in post-initial position)[1]:
> (1) a. [*v zahradě*]$_1$ se$_2$ [*starý strom*]$_3$ [*rázem*]$_4$ [*skácel*]$_2$
> ′ In the garden, an old tree suddenly fell.′
> *in garden*$_{\text{LOC.M.SG}}$ REFL$_{\text{ACC}}$ *old*$_{\text{NOM.M.SG}}$ *tree*$_{\text{NOM.M.SG}}$ *suddenly fall* $_{\text{PART.PRET.ACT.M.SG}}$
>
> R2: if rule (R1) is not applied, use the (en)clitic in a postposition of a verb;
> (1) b. [*v zahradě*]$_1$ [*skácel* se]$_2$ [*starý strom*]$_3$ [*rázem*]$_4$

[1] For an illustration of this phenomena, we use Czech examples with the reflexive pronominal (en)clitic "se" that were quoted by Ertl (1924), or by their known variations in historical Czech; enclitics in these examples are underlined.

ʹ In the garden, an old tree suddenly fell.ʹ
in garden$_{\text{LOC.M.SG}}$ *fall* $_{\text{PART.PRET.ACT.M.SG}}$ REFL$_{\text{ACC}}$ *old*$_{\text{NOM.M.SG}}$ *tree*$_{\text{NOM.M.SG}}$ *suddenly*

R3: if rule (R2) is not applied, use the (en)clitic in a preposition of a verb.
(1) c. [*v zahradě*]$_1$ [*rázem*]$_2$ [*se skácel*]$_3$ [*starý strom*]$_4$
ʹ In the garden, an old tree suddenly fell.ʹ
in garden$_{\text{LOC.M.SG}}$ *suddenly* REFL$_{\text{ACC}}$ *fall* $_{\text{PART.PRET.ACT.M.SG}}$ *old*$_{\text{NOM.M.SG}}$ *tree*$_{\text{NOM.M.SG}}$

All other positions of (en)clitics are a result of miscellaneous factors, such as functional sentence perspective or rhythmical factors in poetry-like texts; however, these miscellaneous factors are not strong and systematic enough to be detected in our model. Needless to say, some word order positions can represent mere fluctuations which occur in any dynamic system.

The idea of the hierarchy of the above-mentioned rules is inspired by Kosek's (2015a,b, 2017) description of the development of auxiliary (en)clitics. Kosek's studies (2015a,b, 2017) also show that the distribution of (en)clitics in different word order positions has changed during the historical development of the language and it is influenced by text type and style[2]. However, Kosek did not use any statistical tests to observe whether the differences can be interpreted either as chance or as a consequence of pragmatic factors (i.e., time and style). To gain a deeper insight into the issue of (en)clitic properties and their development, we postulated the following hypotheses:

H1: there are significant differences among the distributions of word order positions in different historical periods;

H2: there is an increasing proportion of the post-initial position during the course of the historical development;

H3: there are significant differences among distributions of word order position in different styles.

The reasons for these hypotheses are presented in detail in Section 2. Further, the results of statistical testing are used to interpret the relations between the hierarchy of stochastic rules R1 – R3 and pragmatic factors.

The article is organized as follows. Section 2 presents the main characteristics of (en)clitics in historical Czech. Section 3 describes the language material, and the methodology is introduced in Section 4. Section 5 presents and discusses the results of the study, and Section 6 presents the conclusions of the article.

2. *(*En)clitics in historical Czech

(En)clitics are traditionally defined as expressions which never appear independently and which are attached to a preceding clause element (word or phrase). The Old Czech enclitics can be divided into two groups:

1. The archaic (en)clitic particles *ž*(*e*)*, li/le/l, ti/tě/ť, s*(*i*), which are characterized by different grammatical functions, for example the question marker (*li/le/l*), the focus marker (*ti/tě/ť*) or the indefinite determiner (-*s*(*i*)).

2. The pronominal and verbal (en)clitics: a) pronominal forms *mi* 'to me'*, si* 'to oneself' (REFL)*, ti* 'to you'*; ho* 'him'*, mu* 'to him'*, sě* (> *se*) 'to oneself'(REFL)*, tě* 'you'

[2] In the Czech language of the Baroque period, there are obvious differences in the distributions of (en)clitic positions in different text types, in, e.g. in Bible translations, chronicles, sermons, chap-books, etc.

(ACC); b) preterite auxiliary forms (AuxP) Sg 1Ps (*nesl*) *jsem* ‘I carried’, 2Ps (*nesl*) *jsi/s…*; c) conditional auxiliary forms (AuxC) Sg 1Ps (*nesl*) *bych* ‘I would carry’, 2Ps (*nesl*) *by…* .

The two groups display different properties as far as word order is concerned (for more detail see the encyclopedia by Karlík et al., 2017, specifically the entry *Vývoj českých klitik*, or Kosek, 2017); in this paper we therefore focus solely on (en)clitics from the second group. This group of (en)clitics is distinctly heterogeneous because it contains forms that differ in their origin and grammatical status. The pronominal and verbal (en)clitics occur in historical Czech in several word order positions. These positions are demonstrated by above quoted Ertl's (1924) examples or their variants with the reflexive pronominal (en)clitic *se* that show all in historical Czech known clausal positions of enclitic:

1. in the post-initial position shown in examples (1a.)–(1e.),
2. in the contact or verbal adjacent position shown in examples (1b.), (1f.)–(1i.),
3. in the position in the middle of a clause without contact with its syntactically superordinate item shown in example (1j.).

(1) d. [*starý strom*]$_1$ se [*rázem*]$_2$ [***skácel***]$_3$ [*v zahradě*]$_4$
'An old tree suddenly fell in the garden.'
old$_{\text{NOM.M.SG}}$ *tree*$_{\text{NOM.M.SG}}$ REFL$_{\text{ACC}}$ *suddenly fall* $_{\text{PART.PRET.ACT.M.SG}}$ *in garden*$_{\text{LOC.M.SG}}$
e. [*starý*]$_1$ *se*$_3$ [*strom*]$_1$ [*rázem*]$_2$ [***skácel***]$_3$ [*v zahradě*]$_4$
f. [*starý strom*]$_1$ [***skácel** se*]$_2$ [*v zahradě*]$_3$ [*rázem*]$_4$
g. [*starý strom*]$_1$ [*v zahradě*]$_2$ [*rázem*]$_3$ [***skácel** se*]$_4$
h. [*starý strom*]$_1$ [*rázem*]$_2$ [*se **skácel***]$_3$ [*v zahradě*]$_4$
i. [*starý strom*]$_1$ [*v zahradě*]$_2$ [*rázem*]$_3$ [*se **skácel***]$_4$
j. [*starý strom*]$_1$ [*rázem*]$_2$ *se* [*v zahradě*]$_3$ [***skácel***]$_4$

There are two different competing theoretical concepts of the post-initial position in Old Czech (similar to other Slavic languages, e.g. Serbian and Croatian):

1) position after the first modified phrase (so-called 2D position – Halpern 1995) – demonstrated by example (1d.);
2) position after the first stressed word in a sentence (so-called 2W position – Halpern 1995) – demonstrated by example (1e.).

A similar distinction may be found within the contact position:

1) postposition[3] of an (en)clitic after the syntactically or morphologically superordinate item[4] (the syntactically superordinate phrase is placed in the second position in the sentence after a modified initial phrase) – shown in examples (1b.), (1f.), (1g.);
2) anteposition[5] of an (en)clitic before the syntactically superordinate item (the syntactically superordinate item is positioned at the end of a phrase) – demonstrated by examples (1c.) (1h.), (1i.).

There is a third word order pattern of Older Czech clitics: a clitic separated from its superordinate item could appear deep in the middle of a clause (in our analysis this is termed *clause-medial isolated / non-contact position).* This position is also demonstrated by example (1j.).

As our research of the historical development of the Czech pronominal (en)clitics is only in its early stages, we have not yet collected a sufficient volume of data. For this reason, we tested the hypotheses from Section 1 on data obtained by Kosek’s previous research into

[3] This position is also called the *post-verbal position.*

[4] For the sake of simplicity, we classify words which are grammatically/syntactically superior to pronominal and verbal (en)clitics as clitic regents (Toman, 2004). A similar approach was taken by Lešnerová (2002, p. 325); she considers these relations to be cases of morphological and syntactic dependency.

[5] This position is also called the *pre-verbal position.*

the development of the preterite auxiliary (AuxP) in historical Czech Bible translations (Kosek 2015 a,b, 2017). As we mentioned above, AuxP is also a stable (en)clitic and as such it can be tested.

To test the hypotheses, we consider only the three positions which represent the strongest tendency in the word order of (en)clitics.[6] Specifically, we distinguish the post-initial position, i.e. the position after the first phrase of the clause, as shown in the following example (2) from the oldest Czech Bible translation *Bible olomoucká* (the enclitic is underlined; the list of abbreviations of the analyzed Czech Bible translations can be found in Table 1):

(2) [*Onť*] *jest našě nemoci přijal* BiblDrážď Mt 8,17–18[7]
′He took our diseases’
he$_{\text{NOM.M.SG}}$+FOC *be*$_{\text{AUX.PRET.3}}$$^{\text{ps}}$$_{\text{.SG}}$ our$_{\text{ACC.F.PL}}$ diseases$_{\text{ACC.F.PL}}$ take$_{\text{PART.PRET.ACT.M.SG.,}}$

the non-postinitial postverbal position demonstrated by example (3) (the verb is bolded):

(3) [*Tehdy*][*veliká búřě*] ***učini*** *sě na moři,...* BiblDrážd Mt 8,24
′Then great storm developed on the sea... ′
then great$_{\text{NOM.F.SG}}$ storm$_{\text{NOM.F.SG}}$ develope$_{\text{AORIST.ACT.3}}$$^{\text{ps}}$$_{\text{.SG}}$ in sea$_{\text{LOC.F.SG,}}$

and the non-post-initial pre-verbal position, illustrated by example (4):

(4) [*Proč*][*my a duchovníci*][*často*] *sě* ***postíme***…? BiblDrážd Mt 9,14
′Why do we and the Pharisees fast often…?′
why we$_{\text{NOM.PL}}$ and Pharisees$_{\text{NOM.M.PL}}$ often REFL$_{\text{ACC}}$ fast$_{\text{PRES.3}}$$^{\text{ps}}$$_{\text{.SG.}}$

Table 1
List of the abbreviations of the analyzed historical and new Czech Bible translations

BiblDrážď		*Bible drážďanská* (Kyas, ed., 1981; 1985; 1988; Kyas et al., eds, 1996; Pečírková et al., eds., 2009).
BiblOl		*Bible olomoucká* (ibid.)
BiblMlyn		*Bible mlynářčina* (the third quarter of the 15th c.). NK v Praze (Sg. XVII.A.10), Available at Manuscriptorium http://www.manuscriptorium.com.
BiblBen		*Biblí česká v Benátkách tištěná*. 1506. Venice.
BiblMel		*Biblí česká*. 4th edition, Prague 1570.
BiblKral		*Biblí české díl šestý*. Kralice 1593.
BiblSvat		*Druhý díl Biblí totižto Nový zákon*. Prague 1677
ČSPB		*Český studijní překlad bible.* Available at http://www.obohu.cz/bible/.

All other positions occur only very rarely, so we pooled them together in the category “Others”, see Tables 2 and 3 and Figures 1 and 2.

[6] A very detailed analysis of these phenomena is presented in Kosek (2015a,b, 2017).

[7] A complete translation of the Old Czech examples would lengthen this paper to an unacceptable extent; for this reason, we generally cite one example of a particular phenomenon, with a simple gloss of the relevant parts of examples (the glossed parts of the examples are indicated by a vertical line ⎜). The English Bible-translations have been taken from the New English Translation (NET Bible) (http://www.bible.org/netbible/index.htm) or from the Clementine Vulgate (http://vulsearch.sourceforge.net).

Table 2
The distribution of positions of (en)clitics in different translations of the Gospel of Matthew. The translations are ordered chronologically.

	date	post-initial position	pre-verbal position	post-verbal position	others
BiblDrážď	14th c.	292	8	34	3
BiblOl	1417	197	9	22	1
BiblMlyn	15th c.	322	8	108	0
BiblBen	1506	462	7	97	1
BiblMel	15704	148	1	25	1
BiblKral	1594	133	1	27	0
BiblSvat	1677	115	0	19	0
ČSPB	1994/2007	142	0	0	0

Table 3
The distribution of positions of (en)clitics in selected books of *Bible olomoucká*.

	post-initial position	pre-verbal position	post-verbal position	others
BiblOl Mt[8]	197	9	22	1
BiblOl Lk	396	8	36	3
BiblOl Rev	129	2	28	0
BiblOl Gn	434	7	45	0
BiblOl Is	293	12	71	1
BiblOl Sir	99	12	31	0

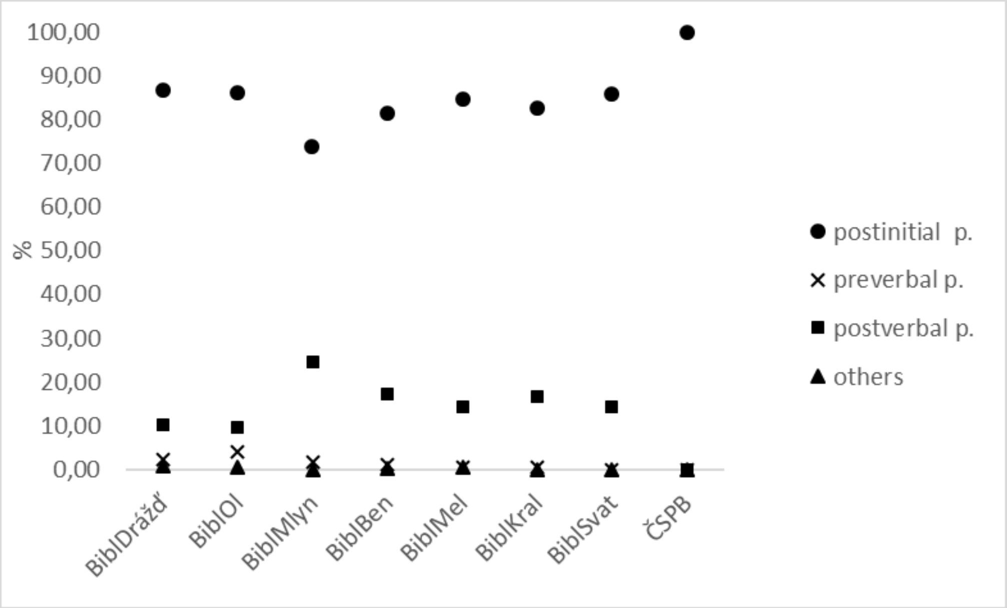

Figure 1. Proportions of positions of (en)clitics in different translations of the Gospel of Matthew. The translations are ordered chronologically.

[8] See Section 3 for the abbreviations of the Bible books analyzed.

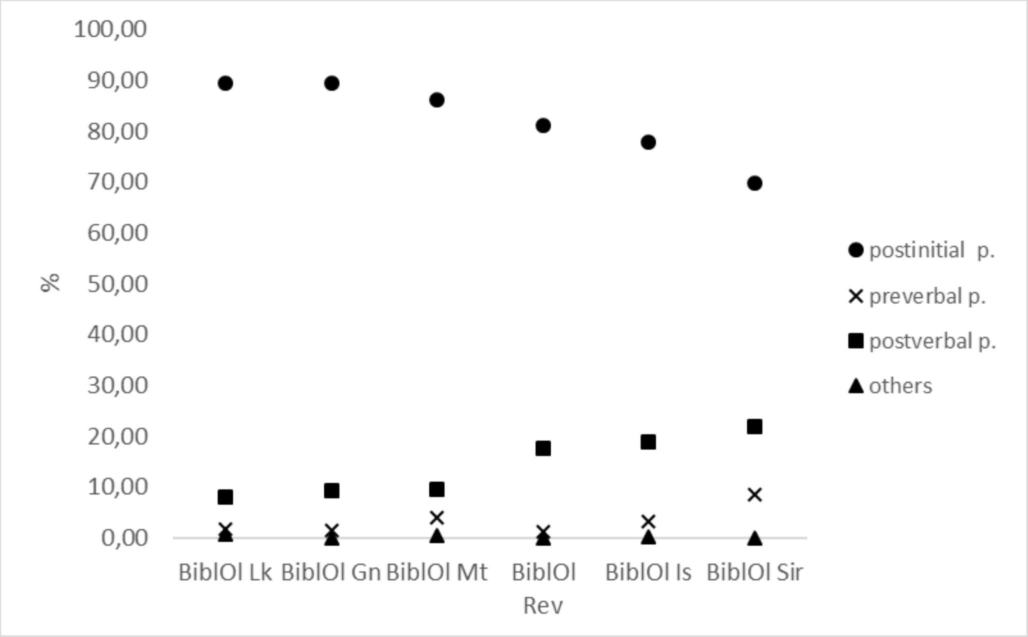

Figure 2. Proportions of positions of (en)clitics in selected books of *Bible olomoucká*. The books are ranked in descending order of post-initial position.

3. Language material

As has been mentioned above, the development of (en)clitic word order is investigated using data which served as the material for research of the preterite auxiliary (AuxP) in historical Czech Bibles (Kosek, 2015a,b, 2017). As the Czech language possesses a long literary tradition and hence a large number of textual sources, the number of texts to be investigated needs to be reduced. It appears very convenient to choose a text type that was adapted into Czech in the very early stages of Czech literary history and one which (being a relatively stable textual formation) has remained present throughout the history of Czech literature until today, i.e. Bible translations. The complete text of the Bible was first translated into Czech during the second half of the 14th century, and it was repeatedly re-translated/adapted (Kyas, 1997; Vintr, 2008) during the following centuries. Since the number of text variants of Old Czech Bibles is large, the Bible text is extensive, and the data are annotated manually, the corpus of texts to be investigated has to be reduced to samples consisting of:

1. parts of the New Testament and the Old Testament *(Gospel of Matthew* Mt, *Gospel of Luke* Lk, *Book of Revelation* Rev, *Book of Genesis* Gn, *Book of Isaiah* Is, *Book of Sirach* Sir[9]) from one of the oldest complete Czech Bible translations (*Bible olomoucká*, from 1417[10]),

2. the Gospel of Matthew from Bible translations originating in different historical periods (*Bible drážďanská* from the end of the 14th century, *Bible olomoucká*, *Bible mlynářčina* from the last quarter of the 15th century, *Bible benátská* from the beginning of the 16th

[9] As the books of the Old Testament differ from the books of the New Testament in their extent, we reduced our analysis to the following chapters: Gn 1–28, Is 14–40, Sir 1–29.

[10] *Bible olomoucká* is a younger copy of a pre-text that had been written in the middle of 14th century (Kyas, 1997; Vintr, 2008).

century, *Bible Melantrichova* from the second half of the 16th century, *Bible kralická* from the turn of the 17th century, *Bible svatováclavská* from the turn of the 18th century), see Table 3. The chosen texts present all documented developmental stages of the Czech language from the oldest Bible translation (e.g. from the 14th century) to the present day. In order to obtain a complete picture of the development of (en)clitics, we also included among the texts for analysis the *Český studijní překlad bible* (ČSPB, Czech Study Bible Translation), which represents a translation into the modern Czech language (it does not contain apparent archaisms).

3. **Methodology**

The data presented in Tables 1 and 2 were used for testing the hypotheses. They are a typical example of categorical data (Agresti, 2013). The hypothesis on the homogeneity (in this context, the homogeneity corresponds to equal proportions of different (en)clitic positions in all translations) of such data is most often tested by the χ^2-test (Snedecor – Cochran, 1989). As only the asymptotic – as opposed to exact – distribution of the test statistic is known, the expected frequencies cannot be too small (otherwise results would not be reliable). Our observed frequencies are (very) low for some (en)clitic positions, and in addition there is no general consensus on the minimum acceptable values of the expected frequencies. Therefore, we used simulated p-values (Ross 2006). Consequently, we do not present degrees of freedom (this notion, relevant if the computation of p-values is based on the asymptotic distribution of the test statistic, has no sense for simulated p-values). All computations were performed in the statistical software environment R.[11]

4. Results

According to hypothesis H1, there should be significant differences among the distributions of word order position in different historical periods. To test this hypothesis, we used the data presented in Table 1, and we found significant differences among the distributions (χ^2 = 95.092, p-value < 0.001). To obtain a deeper insight into the development of word order, we tested differences between distributions in pairs of chronologically subsequent texts[12]. Specifically, the difference between BiblDrážď and BiblOl was tested first, followed by the difference between BiblOl and BiblMlyn and so on. The results are presented in Table 4.

Table 4
Results of testing of differences between distributions of pairs of chronologically subsequent texts. Values represent adjusted p-values of the chi-squared test (adjusted by the Benjamini-Hochberg-Yekutieli procedure). Bolded values denote a significant difference (α < 0.05), and N means that the test cannot be applied (because of the lack of the data for some positions).

	BiblOl	BiblMlyn	BiblBen	BiblMel	BiblKral	BiblSvat	ČSPB
BiblDrážď	>0.999						
BiblOl		**0.0057**					
BiblMlyn			**0.0371**				
BiblBen				>0.999			

[11] www.r-project.org

[12] As we test several hypotheses simultaneously, p-values must be adjusted (see Hochberg – Tamhane, 1987, in general, and Benjamini – Yekutieli. 2001, for the procedure selected).

BiblMel					>0.999		
BiblKral						N	
BiblSvat							N

The results show that significant differences are mainly caused by the specificity of BiblMlyn – it differs significantly from both the preceding BiblOl and the following BiblBen. There is an obvious increase in the proportion of the post-verbal position (see Figure 1) in the analyzed text from this Bible. This increase can be seen as a consequence of the loss of simple past forms (aorist and imperfect). These past forms disappeared in the 15th century, and the translators of BiblBen (and also BiblMlyn) may have used the non-typical post-verbal positions of AuxP to preserve the ceremonial character of the Biblical language instead of using the extinct forms of the aorist and imperfect (Kyas 1997, p. 132–133). Further, another specific position among the Bible translations is occupied by the newest translation (ČSPB), where the post-initial position is realized exclusively. Differences between the other chronologically subsequent translations are not significant; thus, no evident developmental tendency can be claimed. This finding, however, falsifies our second hypothesis. In other words, we cannot see the development of the distribution of the (en)clitics as a gradual development culminating in the contemporary situation. Instead, the results seem to reveal that the complete predominance of the post-initial position is not an outcome of a long-term development but rather a consequence of a relatively abrupt shift. The most likely explanation is that the contemporary situation is a result of a change that happened in Modern Czech (i.e. during the 20th century – Ertl, 1924, p. 266–267; Avgustinova – Oliva, 1997, p. 26; Toman, 2004, p. 74; Kosek, 2011, p. 320).

Finally, we hypothesize that there should be significant differences among distributions of word order position in different styles. The data presented in Table 2 were used for the hypothesis testing, and we found significant differences among the distributions (χ^2 = 74.895, p-value < 0.001). To obtain a deeper insight into the impact of the style on the distribution of (en)clitics, we tested differences between distributions in all pairs of texts; the results of the tests are presented in Table 5 and Figure 3.

Table 5
Results of testing of differences between distributions of all pairs of chosen texts of BiblOl. Values represent adjusted p-values of the chi-squared test (adjusted by the Benjamini-Hochberg-Yekutieli procedure). Bolded values denote a significant difference (α < 0.05), and N means that the test cannot be applied (because of the lack of the data for some positions).

	BiblOl Mt	BiblOl Lk	BiblOl Rev	BiblOl Gn	BiblOl Is
BiblOl Lk	>0.999				
BiblOl Rev	0.1512	0.0521			
BiblOl Gn	0.2357	>0.999	N		
BiblOl Is	0.0992	**0.0062**	>0.999	**0.0062**	
BiblOl Sir	**0.0093**	**0.0062**	N	N	0.1512

The results show that the style has some impact on the distribution of (en)clitics.

Roughly half of the pairs differ significantly in this respect. A closer look at Table 5 reveals the extraordinary position of the text Sir. This result is not surprising because Sir is a wisdom book, which mainly contains advices and instructions. As such, the text of Sir approaches the form of maximas or aphorisms.

5. Conclusion

The presented study focuses on the development of Czech preterite auxiliary (en)clitics (AuxP). Based on the data from the previous research (Kosek, 2015a,b, 2017), three hypotheses were formulated in Section 1. These hypotheses were statistically tested. The tests corroborated hypotheses H1 ("there are significant differences among the distributions of word order position in different historical periods") and H3 ("there are significant differences among the distributions of word order position in different styles"). On the contrary, hypothesis H2 ("there is an increasing proportion of the post-initial position during the historical development") was rejected.

Nevertheless, these results need to be taken critically and subjected to a further investigation on language material from other historical Czech Bible translations or other types of texts:

1. The differences among the distributions of word order positions of AuxP (en)clitics are evident especially between the historical Czech Bible translations on the one hand and the modern biblical translation (*Český studijní překlad bible*) on the other. This fact implies that the definitive historical change of Czech auxiliary (en)clitics in the sole second position clitics was not realized until the modern Czech period (i.e. in the 19th and the 20th century), which was observed by several scholars in the past (see Section 4 for more details). However, these findings have not been explored sufficiently thoroughly yet, and they deserve a special attention in the future.
2. Differences among distributions of word order positions of AuxP (en)clitics in the examined biblical translations were partially influenced by the historical change of simple past forms of aorist and imperfect. However, the development of the Czech pronominal (en)clitics is not influenced by such historical change. Hence, the forthcoming research will examine how much these differences in the word order of AuxP have been influenced by factors other than the developmental trends of word order of Czech (en)clitics.
3. The analysis of the selected biblical books both from the New and Old Testament suggested differences among distributions of word order positions of AuxP in (en)clitics caused by style. The forthcoming research should look for further manifestations of stylistic differences in word order of historical Czech (en)clitics.

One of challenges for a future research is to analyze the pronominal (en)clitics using the same methodological procedure, with the aim to gain knowledge of a more general behaviour of (en)clitics in Czech.

Acknowledgement
This study was supported by the project *Development of the Czech pronominal (en)clitics* (GAČR GA17-02545S).

Pavel Kosek
Masaryk University, Department of Czech Language (Brno), kosek@phil.muni.cz

Radek Čech
University of Ostrava, Department of Czech Language (Ostrava), cechradek@gmail.com

Olga Navrátilová
Masaryk University, Department of Czech Language (Brno), olganav@mail.muni.cz

Ján Mačutek
Masaryk University, Department of Czech Language (Brno) / Comenius University, Department of Applied Mathematics and Statistics (Bratislava), jmacutek@yahoo.com

References

Agresti, A. (2013). *Categorical Data Analysis.* Hoboken (NJ): Wiley.

Avgustinova, T. – Oliva, K. (1997). On the nature of the Wackernagel position in Czech. In: Junghanns, U. – Zybatow, G. (eds.), *Formale Slavistik: 25-47.*, Frankfurt a. M: Vervuert Verlag.

Benjamini, Y. – Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29, 1165–1188.*

Ertl, V. (1924). Příspěvek k pravidlu o postavení příklonek. *Naše řeč 8(9), 257–268; 8(10), 293–309.*

Franks, S. – King, T.H. (2000). *A Handbook of Slavic Clitics*. Oxford: Oxford University Press.

Gebauer, J. (1929). *Historická mluvnice jazyka českého IV. Skladba* (Trávníček, F., ed.). Praha: Nákladem České akademie věd a umění.

Gebauer, J. (1958) [1906]. *Historická mluvnice jazyka českého III/1. Tvarosloví — časování*. 2nd ed., Praha: ČSAV.

Grepl, M. – Karlík, P. (1998). *Skladba češtiny.* Praha: Votobia.

Halpern, A. (1995). *On the Placement and Morphology of Clitics*. Stanford: CSLI.

Havránek, B. (ed.) (1968). *Staročeský slovník. Úvodní stati, soupis pramenů a zkratek.* Praha: Academia.

Hochberg, Y. – Tamhane, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Jakobson, R. (1971). Les enclitiques slaves. In: *Selected Writings. Volume II: Word and Language: 16-22.*. The Hague: Mouton &. Co.

Junghanns, U. (2002). Klitische Elemente im Tschechischen: eine kritische Bestand-aufnahme. In: Daiber, T. (ed.), *Linguistische Beiträge zur Slavistik IX: 117-150.* München: Sagner.

Karlík, P. – Nekula, M. – Pleskalová, J. (eds.). (2017). *Nový encyklopedický slovník češtiny*. Praha: Nakladatelství LN.

Kosek, P. (2011). *Enklitika v češtině barokní doby*. Brno: Host.

Kosek, P. (2015a). Development of Word Order of Preterite Auxiliary Clitics in Old Czech Bibles. In: Ziková, M. – Caha, P. – Dočekal, M. (eds.), *Slavic Languages in the Perspective of Formal Grammar. Proceedings of FDSL 10.5, Brno 2014. Frankfurt a. M., Berlin, Bern, Bruxell: 178-198.* P. Lang.

Kosek, P. (2015b). Slovosled auxiliáru préterita v českých raně novověkých biblích. *Linguistica Brunensia, 63(2), 109–124.*

Kosek, P. (2017). Wortstellung des Präteritum-Auxiliars in der alttschechischen Olmützer Bibel. *Die Welt der Slaven, 62(1), 22–41.*

Kosta, P. – Zimmerling, A. (2013). Slavic Clitic Systems in a Typological Perspective. In: Schürcks, L. – Giannakidou, A. – Etxeberria, U. (eds.), *Nominal Constructions in Slavic and Beyond: 439-486*. Boston: de Gruyter.

Kyas, V. (1997). *Česká Bible v dějinách národního písemnictví*. Praha: Vyšehrad.

Kyas, V. (ed.) (1981). *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století. I. Evangelia.* Praha: Academia.

Kyas, V. (ed.) (1985). *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století s částmi Bible litoměřicko-třeboňské. II. Epištoly. Skutky apoštolů. Apokalypsa.* Praha: Academia.

Kyas, V. (ed.). (1988). *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století. III. Genesis–Esdráš*. Praha: Academia.

Kyas, V. – Kyasová, V. – Pečírková, J. (eds.). (1996). *Staročeská bible drážďanská a olomoucká: kritické vydáni nejstaršího českého překladu bible ze 14. století. IV. Tobiáš–Sirachovec.* Padeborn: F. Schöningh.

Lamprecht, A. – Šlosar, D. – Bauer, J. (1986). *Historická mluvnice češtiny*. Praha: SPN.

Lešnerová, Š. (2002). Postavení příklonky „se" v textu Kryštofa Haranta „Cesta z Království českého... do Země svaté...". In: Hladká, Z. – Karlík, P. (eds.), *Čeština – univerzália a specifika*, 4: 325-327. Brno: Nakladatelství LN.

Lenertová, D. (2004). Czech pronominal clitics. *Journal of Slavic Linguistics*, 12(1–2), p. 135–171.

Merhaut, L. et al. (eds.) (2008). *Lexikon české literatury*, Part 4/II U–Ž, Dodatky A–Ř. 1882–1887. Praha: Academia.

Migdalski, K. (2009). On two types of Wackernagel cliticization in Slavic. In: Reich, J. et al. (eds.), *Formal Approaches to Slavic Linguistics: 147-162.* Ann Arbor: Michigan Slavic Publishers.

Pancheva, R. (2005). The rise and fall of second-position clitics. *Natural Language and Linguistic Theory, 23(1), 103–167.*

Pečírková, J. et al. (eds.) (2009). *Staročeská Bible drážďanská a olomoucká s částmi Proroků rožmberských a Bible litoměřicko-třeboňské. V/1 Izaiáš–Daniel, V/2 Ozeáš–2. kniha Makabejská.* Praha: Academia.

Radanović-Kocić, V. (1996). The placement of Serbo-Croatian clitics: A prosodic Approach. In: Halpern, A. – Zwicky, A. (eds.), *Approaching second: second position clitics and related phenomena:329-445.* Stanford: CSLI.

Ross, S.M. (2006). **Simulation. Burlingt**on (MA): Academic Press.

Snedecor, G.W. – Cochran, W.G. (1989). *Statistical Methods*. Ames (IA): Iowa State University Press.

Toman, J. (2004). Ertlova diskuse českých klitik. In: Hladká, Z. – Karlík, P. (eds.), *Čeština — univerzália a specifika 5: 73-79.* Brno: MU.

Trávníček, F. (1935). *Historická mluvnice československá*. Praha: Melantrich.

Trávníček, F. (1956). *Historická mluvnice česká 3. Skladba*. Praha: SPN.

Vintr, J. (2008). Bible (staroslověnský překlad, české překlady). In Merhaut, L. et al. (eds.) *Lexikon české literatury*, Part 4/II U–Ž, Dodatky A–Ř. Praha: 1882–1887. Academia.

Zikánová, Š. (2009). *Postavení slovesného přísudku ve starší češtině* (1500–1620). Praha: Karolinum.

Internet sources

Kučera, K. – Řehořková, A. – Stluka, M. DIAKORP: Diachronní korpus, version 6. Ústav Českého národního korpusu FF UK, Praha 2015. Available at http://www.korpus.cz (accessed 30 June 2017).

Manuskriptorium. Available at http://www.manuscriptorium.com/cs (accessed 30 June 2017).

New English Translation (NET Bible). Available at http://www.bible.org/netbible/index.htm) (accessed 30 June 2017).

Vokabulář webový [on-line]. Version 0.4.2. Oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i. Available at http://vokabular.ujc.cas.cz (accessed 30 June 2017).

Glottometrics 40, 2018, 63-76

Adnominal Aggregation

Sergej Andreev[1]*, Fengxiang Fan*[2]*, Gabriel Altmann*

Abstract. In the present study we concentrate to one property of adnominals as they appear in Russian texts: their aggregation which can be expressed e.g. by measuring the distances between equal adnominals and as a similarity of adnominals occurring in sentences in distance x = 1,2,...

Keywords: Aggregation, adnominals, Skinner hypothesis, Russian, similarity

Distance

Aggregation is an accumulation of identical elements in near distance to one another. According to a hypothesis of B. Skinner (1939, 1941, 1957) the neurons in our head obtain a stimulus but slowly cease firing if the stimulus dies out or is replaced by another. Thus there are many short distances between identical elements and the distances increase. It means that whatever subconscious entity is scrutinized, the number of small distances between the positions of the same entity is greater than that of large ones. What is more, the distances follow a regularity that can be derived and tested.

Up to now researchers have studied phonemes (Altmann 1968), parts of speech (Tuzzi et al. 2012), words of equal length (Zörnig 2012), hexameter types (Strauss et al. 1984; Altmann, Köhler 2015) and applied Markov chains (Brainerd 1974), Poisson processes (Herdan 1966; Králík 1977), similarity measures (Altmann, Köhler 2015), random distributions (Zörnig 1984a,b, 1986, 2012). Here we shall examine the placing of adnominals in Russian texts, compute their distances and the similarity of sentences and express the results formally in order to test Skinner`s hypothesis in this domain. Adnominals are presented in form of abbreviations and represent classes of expressions.

The Russian adnominals are presented in the Appendix. For the analysis we have chosen 40 modern Russian texts, out of which 20 were written by female authors and 20 by male ones.

The distance between two equal adnominals can be computed in the usual Euclidean way but since we have a text, i.e. a linear sequence, we may express it simply by the number of steps that are necessary to arrive at the same adnominal. Consider e.g. the sequence *ABCDAC*. Starting from the first *A* we need 4 step to meet another *A;* starting from the first *C* we need 3 steps to meet another C behind the first one, etc. One can compute the distance also in terms of the number of other elements laying between two equal ones, but this is merely a displacing the resulting numbers one place to the left. In the above case, there would be a distance 3 between the *A'*s, distance 2 between the *C'*s, etc.

Let us first consider text T 1 written by Demidova. We obtain the sequence of adnominals as follows: (the list of adnominals can be found in the Appendix)

[1] Sergej Andreev : smol.an@mail.ru;
[2] Fengxiang Fan: fanfengxiang@yahoo.com

[A,G,PTY,PR,G,RC,G,PTY,DETQ,DETF,PR,A,PT,G,A,G,A,A,A,A,DETS,A,G,A,DETS,PR
,G,DETS,PR,DETS,AP,DETQ,I,A,G,DETS,PT,A,A,DETS,A,I,A,DETS,RC,A,RC,A,PT,PR,
PT,A,DETQ,AP,DETQ,DETF,DETH,A,DETF,DETF,A,DETV,DETQ,A,DETS,DETS,
DETS,A,G,DETQ,DETF,A,A,DETS,A,PR,PR,PTY,DETS, A,A,DETQ, DETQ,G,DETS,
AO,PTY,A,A,A,A,A,A,A,DETQ,PR,DETS,PR,G, DETQ,DETS,A,G,G,G,A,DETF,A,PR,
AP,PR,A,AP,DETQ,DETS,DETQ, DETS,A, DETQ,DETS,A,DETS,A,A,PR,PR,DETS,A,
A,G,DETF,RC,RC,A,A,PR,DETS,AP,A, PTY,A,RC,PT,DETQ,A,A,PR,A,A,DETQ,
DETQ,G,PT,A,PR,A,DETQ,DETH,G,A,DETF,A,PT,A,G,PR,DETF,DETQ,G,PR,A,PR,
A,A,G,DETS,DETQ,G,PTY,DETQ,A,A,G,DETF,PT,DETF,PTY,A,A,DETF,A,DETS,
A,G,PR,A,A,DETS,A,G,DETS, DETN,A,G,DETQ,PR,G,A,APAJ,RC,A,PTY,ADV,A,
DETS,AP,G,G,A,A,APAJ, PTY,A,RC,A,A,A,A,PTY,A,A,A,I,A,PR,A,PR,PR,PR,G, DETS,
G,A,A,PR,APX, PTY,A,RC,A,A,PT,A,A,G,DETS,A,A,G,A,G,AY,PTY,A,PT,G,PT,A,A,A,
G,PT, A,G,AY,RC,A,PR,PT,PR,G,A,A,DETQ,G,A,G,DETN,A,A,A,A,ADV,A,A,A,A,G,
A,A,PR,PTY,DETQ,A,A,RC,APAJ,RC,A,A,G,PR,DETQ,CN,A,G,A,A,PR,DETQ,DETS,
A,DETF,A,DETS,PT,G,A,A,A,G,APAJ,A,PR,G,PTY,A,RC,A,A,RC,DETQ,PTY,A,G,
DETQ,G,A,PTY,A,RC,G ,G,DETF, G,RC,PT ,PT,PT,RC,A,A,A,A,PR,DETF,A,PR,
DETF,A,DETS,PT,G,A,DETQ,A,PR,DETF,DETS,G,DETS,A,PR,A,A,G,APAJ,ADV,
DETF,A,DETF,RC,A,A,G,PR,G,DETQ,A,DETS,G,A,A,PR,G,PR,A,DETS,DETS,A,PT,
A,APAJ,DETQ,A,DETQ,A,DETS,G,A,PR,A,DETS,A,RC,DETS,A,A,PR,A,CN,G,A,
APAJ,A,PR,DETS,G,A,G,DETF,PR,PR, DETS,AY,A,A,DETF,DETS,A,A,DETF,A,A,
A,PTY,A,PTY,A,A,A,G,A,A,A,A,PT,G,A,G,PTY,A,PR,AY,DETS,DETQ,A,PTY,PTY,
G,A,G,AY,PT,A,PTY,I,PTY,A,A,PR,A PR,A,DETF,A, PTY, RC,A,G,PTY,AY,PR,A,CN,
PR ,PR,DETS,PT,G,A,A,A,G,DETQ,DETS,A,PR,DETS,A,A,G,DETS,A,PTY,A,A,PTY,G,
DETF, DETS,A,DETS,PR,DETS,PR,A,DETF,DETS,PR,A,A,G,RC,A,DETF,DETF,A,
DETS,DETS,G,DETF,A,CN,APAJ,A,PR,DETF,A,DETQ,PTY,G,A,A,G,A,A,DETQ,DETF
,A,G,DETS,A,DETS,DETS,A,DETH,A,DETF,DETH,A,G,A,PT,CN,A,G,A,A,PR,PT,
PR,DETF,A,DETS,G,A,A,DETN,A,A,DETS,DETS,PR,A,PTY,DETQ,PR,A,G,G,G,G,A]

There are 10 different adnominals between the first and second A; 2 different between the first and second G, etc. For the first 10 symbols we obtain the sequence:

[10,2,4,6,1,38,6,68,22,45,…

Computing all distances in this text we obtain the results presented in Table 1. The distances in Table 1 are placed in the order of appearing of individual symbols which are separated by a comma. Computing the number of distances 1,2,3,… we obtain, as a matter of fact, a discrete distribution and our aim is to find it empirically and substantiate it theoretically. Up to now, the authors have used the power function (Wimmer et al. 2003), the Zipf-Alekseev function (Altmann, Köhler 2015), Markov chains (Brainerd 1974), the Poisson process and the exponential function (Herdan 1966, Králík 1977), the geometric distribution (Epstein 1953; Spang-Hanssen 1956; Yngve 1956), the urn model, the negative binomial distribution (Strauss et al. 1984; Zörnig 2012) and the Euclidean distance (Ferrer-i-Cancho 2004) on some empirical linguistic data. Here our aim is to test the validity of some of the models for adnominals in text.

Here we present only the full results of Text 1 because the tables are very long. We take into account only distances 0 to 100 (number of elements lying between repetitions). For all the other data we show only the parameters and the determination coefficient. Note that some distances are absent, e.g. 26,27,31 etc.

Table 1
Sequential distances between equal adnominals and their frequency in Text 1.
Fitting by the exponential function

Dist.	Frequ	Comp	Dist.	Frequ	Comp	Dist.	Frequ	Comp
0	95	98.04	19	1	1.65	48	1	1.00
1	80	75.54	20	5	1.50	49	1	1.00
2	57	58.25	21	5	1.38	52	1	1.00
3	55	44.98	22	4	1.29	54	1	1.00
4	28	34.78	23	1	1.22	55	2	1.00
5	20	26.95	24	5	1.17	59	1	1.00
6	21	20.93	25	4	1.13	64	1	1.00
7	15	16.31	28	3	1.06	66	1	1.00
8	8	12.76	29	2	1.05	69	1	1.00
9	10	10.03	30	3	1.04	73	1	1.00
10	17	7.94	32	3	1.02	74	1	1.00
11	5	6.33	33	1	1.02	77	1	1.00
12	5	5.09	34	1	1.01	78	1	1.00
13	8	4.14	35	1	1.01	79	1	1.00
14	6	3.42	38	3	1.00	84	1	1.00
15	2	2.86	42	1	1.00	85	2	1.00
16	1	2.43	44	1	1.00	91	1	1.00
17	3	2.09	45	2	1.00	94	1	1.00
18	4	1.84	46	1	1.00	100	1	1.00
a = 97.0373, b = 0.26379, R^2 = 0.9779								

As can be seen, the determination coefficient is very high showing that in this case the repetitions decrease exponentially.

Table 2
Fitting the exponential function to all female texts

Text	**a**	**b**	**R^2**	**Text**	**a**	**b**	**R^2**
1	97.0373	0.2638	0.9779	**11**	107.7699	0.4016	0.9652
2	92.7013	0.2237	0.9717	**12**	132.0156	0.3420	0.9773
3	105.6463	0.2906	0.9698	**13**	133.1568	0.3528	0.9734
4	102.2859	0.2749	0.9753	**14**	91.5264	0.2388	0.9676
5	112.5936	0.2946	0.9710	**15**	131.4905	0.2996	0.9746
6	59.0131	0.2599	0.9387	**16**	181.5941	0.3443	0.9756
7	113.5505	0.2876	0.9767	**17**	192.8984	0.3218	0.9690
8	118.6625	0.3796	0.9829	**18**	103.6521	0.3173	0.9421
9	149.4904	0.3789	0.9877	**19**	85.8716	0.3154	0.9536
10	200.1767	0.4357	0.9700	**20**	156.0166	0.4863	0.9673

Table 3
Fitting the exponential function to all male texts

Text	a	b	R^2	Text	a	b	R^2
21	128.9019	0.3442	0.9761	**31**	106.5279	0.2880	0.9435
22	176.6420	0.5350	0.9758	**32**	89.7675	0.3502	0.9719
23	210.8997	0.5127	0.9789	**33**	110.7374	0.3460	0.9849
24	73.8525	0.3289	0.9675	**34**	78.1615	0.3253	0.9662
25	149.3626	0.4111	0.9802	**35**	112.7567	0.2540	0.9694
26	117.7593	0.2357	0.9719	**36**	72.6885	0.2175	0.9808
27	74.4225	0.3284	0.9747	**37**	111.4385	0.2650	0.9853
28	285.4658	0.3570	0.9840	**38**	118.3882	0.3328	0.9582
29	183.6775	0.3771	0.9818	**39**	131.0613	0.3246	0.9692
30	119.8481	0.4799	0.9720	**40**	99.2373	0.3853	0.9560

Having a number of parameters, we may compare the female and male texts in a simple way. Parameter *a* gives merely the beginning of the distribution; it depends also on the text length. Parameter *b* can be considered a characteristic property, namely the strength of the Skinner tendency. Here we may consider it a quite simple variable. For female texts we obtain values in the interval <0.22; 0.49>, for the male texts in the interval <0.22; 0.54>. The mean of female values is 0.3255, that of male values is 0.3499, i.e. slightly greater because some male texts have some higher means. But one cannot speak about a general difference. In order to find a significant difference which would be valuable also for psycholinguists and socio-linguists one must investigate a number of languages. For Russian cf. Andreev (2017).

We can preliminarily state that in the case of adnominals the distances follow the exponential function, at least in Russian. It is to be remarked that one tries to find a well fitting function or distribution which is as simple as possible because the parameters must be interpreted. As to the exponential function, we start from the assumption that the relative rate of change of frequencies is constant, i.e. *dy/y = bdx*, where *b* is the characteristic constant of the given text, text type, age, gender, etc. Needless to say, sometimes more parameters must be applied but one should avoid polynomials.

Similarity

The other way of examining aggregation is the comparison of the adnominals within the individual sentences. The text is segmented according to sentences; the adnominals are the only remaining symbols (everything else is omitted) and their similarity is computed. Let the first sentence be called A and the second B, then using one of the many similarity measures we compute the number of adnominals in sentence A as n_A and those in sentence B as n_B. Further we compute the number of identical adnominals whereby an adnominal can be counted only once, i.e. the adnominals may be placed alphabetically, and then we compute

$$S = 100\frac{|A \cap B|^2}{|A| * |B|}$$

that is, the intersection of A and B divided by the product of the cardinal numbers of the two sets (sentences). Many times one attaches also a second element taking into account the pairs of elements but here we can omit it. For the sake of illustration let us take the first 2 sentences (S1 and S2) in text T 10 (// shows the end of sentence):

A,G//
A,G,A,AP,AP,A,G,AP,A,A,PR,A,A,PT,A,PT,A,G,A,A//

The cardinal numbers of S1 und S2 are n_{S1} = 2, n_{S2} = 20. The intersection of these two sentences is 2, namely A and G, whereby we count A and G only once, hence S(1/2) = $100*2^2/(2*20) = 400/40 = 10$. Computing the similarity of all sentence pairs in distance 1, we can get the mean of the similarities. This procedure must be performed for all sentence pairs in distance 1,2,3,… and the means show the course of similarity. The results are presented in Table 2. Here we compute maximally to the distance x = 10.

Looking at the results in Table 2 we may state that Skinner's hypothesis holds. The similarity decreases with increasing distance. Our aim is to find a function which would express this fact for all texts. We conjecture that the placing of adnominals is a special quality of syntax and in its classified form cannot be constructed consciously. Hence starting from the unified theory we may consider the simple hypothesis that the relative rate of change of similarity is constant, i.e. we use again the exponential function. If we find an exception, then the text must be scrutinized separately and if possible, the author must be asked how he wrote the text. This is a very complex problem; it is much simpler to investigate texts in other languages.

Considering the results in the table one may state that none of them corroborates Skinner's hypothesis, i.e. in no text one can find a monotonous decrease of similarities. In several cases one finds the last differences are even larger than the first ones. Since adnominality is a very abstract property, one may conjecture that Skinner's hypothesis holds for rather lower levels of language. On higher levels, there is rather an irregular oscillation. This may be caused by pauses in writing, a posteriori changes, intervention of editors, etc.

Table 2a
Means of adnominal similarities between sentences in distance x = 1,2,3,…
Female texts

Distance	T1	T 2	T 3	T 4	T 5
1	16.1956	15.4934	18.1199	15.1334	16.1553
2	14.9602	14.7306	16.2085	18.1413	15.0703
3	16.3718	17.5327	17.2433	13.6852	13.5046
4	18.2473	14.2497	17.3862	16.5451	13.5642
5	16.6893	13.3640	16.8648	15.0986	12.5284
6	14.5004	15.0814	16.8200	15.6380	11.1530
7	15.9376	14.8332	17.0525	15.2502	11.9351
8	15.3788	12.5987	16.7478	15.9574	10.2084
9	15.7070	15.8950	19.0341	16.9676	12.4794
10	16.9668	13.5975	15.9067	14.1621	12.2468

Table 2a
Means of adnominal similarities between sentences in distance x = 1,2,3,…

Distance	T 6	T 7	T 8	T 9	T 10
1	12.7795	13.7416	20.5144	19.8945	16.5387
2	13.4745	12.8790	15.8058	19.8304	16.4535
3	14.9731	11.2152	15.1827	17.8238	16.3063
4	11.3719	13.9475	14.0343	15.7162	17.1411
5	15.7877	13.1772	16.3520	19.6372	14.5359
6	15.1170	13.5606	16.2102	17.6671	16.6178
7	11.1894	13.7069	15.2744	18.5938	14.7794
8	11.0822	14.1742	16.0637	19.2404	17.6067
9	14.6973	12.1640	15.9503	19.0271	18.0022
10	12.7994	12.5785	15.0990	21.1556	16.6563

Table 2a
Means of adnominal similarities between sentences in distance x = 1,2,3,…

Distance	T 11	T 12	T 13	T 14	T 15
1	18.6107	14.7948	14.8190	20.5701	14.9528
2	17.5661	13.0235	14.8182	20.0939	14.2934
3	20.7910	14.4367	15.7392	14.7751	14.4104
4	17.4882	13.0634	14.8112	17.3153	16.6940
5	17.1937	13.1735	15.6213	16.5602	14.8567
6	16.3349	12.9243	15.4227	16.7353	12.6086
7	15.4470	12.9982	13.9964	14.8126	14.5538
8	17.9473	11.7734	16.4739	16.4410	14.9512
9	14.0062	15.0732	13.0218	15.7291	14.6862
10	14.8707	13.3246	14.1624	16.2405	13.6196

Table 2a
Means of adnominal similarities between sentences in distance x = 1,2,3,…

Distance	T 16	T 17	T 18	T 19	T 20
1	18.3193	15.4941	12.0435	17.7265	21.3514
2	17.7138	17.1145	10.6049	16.3342	20.3786
3	15.4400	16.6168	11.5630	13.6617	21.6887
4	16.1811	14.4796	12.7373	16.7717	20.8065
5	14.8152	17.1125	11.4093	18.9133	20.0692
6	14.6645	17.6473	10.0999	13.6093	17.7088
7	16.3396	15.5554	10.6903	14.1830	17.4763
8	15.9540	16.2500	14.8975	16.3631	18.5873
9	17.4582	17.1064	10.5587	14.8689	16.4492
10	17.6060	15.0210	12.4207	14.9958	15.0663

Table 2b
Means of adnominal similarities between sentences in distance x = 1,2,3,…
Male texts

Distance	T 21	T 22	T 23	T 24	T 25
1	18.7990	23.3695	21.1191	21.8687	18.8090
2	17.6886	22.8482	22.5520	19.0455	19.6479
3	21.5620	21.2619	20.0732	17.6541	20.2252
4	19.5722	20.8100	19.7426	21.3534	18.4480
5	16.7169	18.9438	17.9678	21.2411	19.5591
6	20.1602	19.4431	19.3518	18.0857	16.9700
7	19.0647	21.5493	20.9475	19.7467	18.6896
8	18.4260	20.3500	20.7789	18.3893	16.4080
9	19.8753	19.1853	20.3172	18.7211	17.3968
10	19.6958	20.6542	18.3993	19.2157	16.7248

Table 2b
Means of adnominal similarities between sentences in distance x = 1,2,3,…

Distance	T 26	T 27	T 28	T 29	T 30
1	16.4513	15.2842	18.9349	22.1506	25.1507
2	15.2284	16.9644	18.4892	20.5168	24.3280
3	15.0192	13.2706	17.4807	18.2208	26.5047
4	14.4913	15.3102	16.2591	17.1718	26.5355
5	14.9477	12.4349	16.2652	17.1861	26.7036
6	14.2991	14.1804	16.6005	19.6297	25.3807
7	13.4451	13.2122	18.0897	20.5028	24.7995
8	12.2394	14.5083	18.6447	18.9805	24.1738
9	11.5202	13.9400	15.7310	17.6981	25.6291
10	14.3313	14.6715	17.1817	18.2597	22.7601

Table 2b
Means of adnominal similarities between sentences in distance x = 1,2,3,…

Distance	T 31	T 32	T 33	T 34	T 35
1	19.5502	17.5241	16.7339	15.6451	16.1736
2	18.7324	15.3280	16.0493	15.3684	18.2190
3	16.1009	19.1602	17.0375	18.7415	16.6852
4	18.0216	16.9499	16.7730	15.6331	17.1805
5	14.3648	14.4765	16.8268	17.4344	18.8979
6	11.9264	15.8863	15.7939	16.8633	16.8605
7	13.6889	16.3304	15.1468	14.1206	16.9391
8	15.4918	16.1697	16.1490	14.0474	17.1904
9	14.5973	14.7497	13.8555	17.6503	17.4432
10	15.6504	22.0192	13.8394	14.6619	17.7721

Table 2b
Means of adnominal similarities between sentences in distance x = 1,2,3,…

Distance	T 36	T 37	T 38	T 39	T 40
1	14.5166	15.7598	15.2778	15.0126	11.7550
2	13.9163	16.2083	14.7470	14.4320	14.2477
3	15.0566	19.5740	13.4280	12.8450	15.4354
4	15.3524	22.0257	12.3532	10.4894	15.0900
5	13.2392	18.6835	15.2297	13.9366	12.9129
6	11.9091	16.9267	13.8857	15.5184	15.3057
7	13.6443	17.2278	12.8922	10.5455	10.1299
8	13.5848	16.4573	15.1109	11.9410	12.5730
9	12.7995	18.7690	13.2590	11.9316	11.7825
10	13.0612	14.2342	14.9270	12.9129	14.0018

However, if we consider a whole group of texts, we may state that (1) Skinner's hypothesis holds as a whole, even if not for individual texts, (2) there is a difference between female and male texts. Male texts have a higher adnominal similarity; that means, male writers are more stereotype, the firing of neurons holds longer, while female writers are more flexible (or make more corrections, or write more slowly, etc.). This fact can be shown both vertically (according to distance) as well as horizontally (taking all distances and computing an average). If we compute the averages of distance means separately for female and male writers, we obtain the numbers presented in Table 3.

Table 3
Averages of distance means in female and male texts

Distance	Female averages of means	Male averages of means
1	16.66	17.99
2	15.97	17.73
3	15.65	17.77
4	15.63	17.48
5	15.69	16.90
6	15.01	16.75
7	14.73	16.54
8	15.43	16.58
9	15.44	16.34
10	14.92	16.75

The Figure below shows that male averages end where female ones begin. Nevertheless, the number of data is too small to search for a mathematical model. Inductive search did not give good results, and further texts must be analyzed

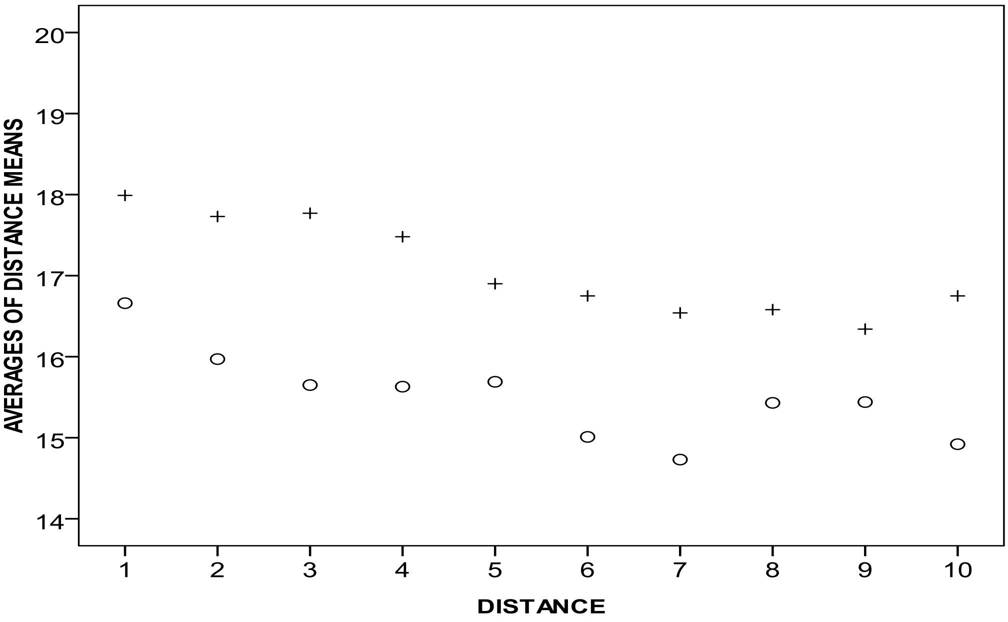

Figure 1. Averages of distance means in female and male texts.
Circles: female; pluses: male

For further comparison, the female and male texts were ordered according to decreasing average of means of all texts. The results are presented in Table 4.

Table 4
Decreasing ordering of female and male texts

Female texts				Male texts			
Text	Mean	Text	Mean	Text	Mean	Text	Mean
T 20	18.96	T 19	15.74	T 30	25.20	T 32	16.86
T 9	18.86	T 4	15.66	T 22	20.84	T 34	16.02
T 3	17.14	T 13	14.89	T 23	20.12	T 33	15.92
T 11	17.03	T 2	14.74	T 24	19.53	T 31	15.81
T 14	16.92	T 15	14.56	T 21	19.16	T 27	14.37
T 10	16.46	T 12	13.46	T 29	19.03	T 26	14.19
T 16	16.44	T 6	13.33	T 25	18.29	T 38	14.11
T 17	16.24	T 7	13.11	T 37	17.59	T 36	13.71
T 1	16.09	T 5	12.88	T 28	17.37	T 40	13.23
T 8	16.05	T 18	11.70	T 35	17.34	T 39	12.96

Though we do not propose any function, one may see in Figure 2 that the male texts lie higher than the female ones. Again, a sign of longer associations with male writers.

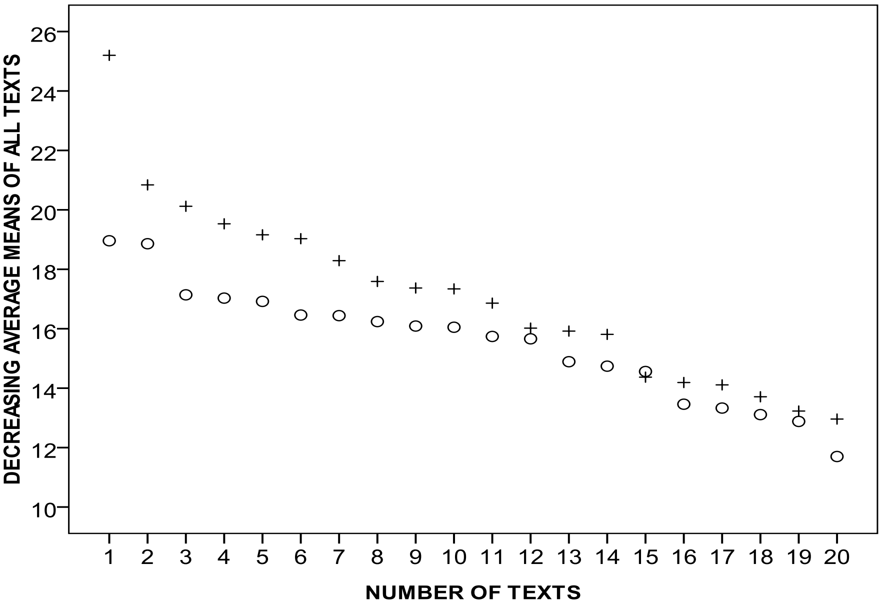

Figure 2. Decreasing ordering of female and male texts. Circles: female; pluses: male. On the x-axis 1—20 respectively represent T 20, T 9, T 3, T 11, T 14, T 10, T 16, T 17, T 1 , T 8, T 19, T 4, T 13, T 2, T 15, T 12, T 6, T 7, T 5, T 18 for female texts and T 30, T 22, T 23, T 24, T 21, T 29, T 25, T 37, T 28, T 35, T 32, T 34, T 33, T 31, T 27, T 26, T 38, T 36, T 40, T 39 for male texts

From the linguistic point of view, this is merely a simple statement. For psychologists, it could present a problem which could be solved by relating the distance forming to firing of neurons. Of course, this is a task for the future when we shall have a free access to the human mind.

Conclusions

In the case of modern Russian literary texts we may state that the distances between equal adnominals follow a decreasing exponential trend is inconformity with the Skinner hypothesis. The fact that we present it as a function – and not as a distribution – does not play any role because mathematical models are merely translations of our hypotheses into a formal language, not the truth. Besides, many distances are missing, hence using a function is more appropriate.

As to the similarity of sentences containing only adnominals we see that texts written by male authors have both greater distances between equal sentences and the ordering of texts according to the averages of means of distances is greater for men than for women. One can hypothetically infer that men are more stereotype, the verbal incitements live longer but this hypothesis must be tested in many texts and many languages. And of course, not only with the adnominals but on all levels.

References

Altmann, G. (1968). Some phonic features of Malay shaer. *Asian and African Studies 4, 9-16.*

Altmann, G., Köhler, R. (2015). *Forms and degrees of repetitions in texts*. Berlin/Munich/ Boston: de Gruyter.

Andreev, S. (2017). Gender differences based on attributive relations (In print).

Brainerd, B. (1976). On the Markov-nature of the text. *Linguistics 176, 6-30.*

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E 70, 056135.*

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Berlin: Springer.

Králík, J. (1977). An application of exponential distribution law in quantitative linguistics. *Prague Studies in Mathematical Linguistics 5.233-235.*

Skinner, B.F. (1939). The alliteration in Shakespeare's sonnets. A study in literary behavior. *Psychological Record 3, 186-192.*

Skinner, B.F. (1941). A quantitative estimate of certain types of sound-patterning in poetry. *The American Journal of Psychology 54, 64-79.*

Skinner, B.F. (1957). *Verbal Behavior*. Acton: Copley Publishing Group.

Spang-Hanssen, H. (1956). The study of gaps between repetitions. In: Halle, M., (ed.), *For Roman Jakobson, 497-502.* The Hague: Mouton.

Strauss, U., Sappok, Ch., Diller, H.J., Altmann, G. (1984). Zur Theorie der Klumpung von Textentitäten. *Glottometrika 7, 73-100.* Bochum: Brockmeyer.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, 42-48.*

Tuzzi, A., Popescu, I-I., Zörnig, P., Altmann, G. (2012). Aspects of the behavior of parts-of-speech in Italian texts. *Glottometrics 24, 41-69.*

Wimmer, G., Altmann, G., Hřebiček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov*. Bratislava: VEDA.

Yngve, V. (1956). Gap analysis and syntax. *IRE Transactions PGIT 2, 106-112.*

Zörnig, P. (1984a) The distribution of the distance between like elements in a sequence I. In: *Glottometrika 6, 1-15.* Bochum: Brockmeyer

Zörnig, P. (1984b). The distribution of the distance between like elements in a sequence II. In: *Glottometrika 7, 1-14.* Bochum: Brockmeyer.

Zörnig, P. (1986). A theory of distance between like elements in a sequence. In: *Glottometrika 8, 1-22.* Bochum: Brockmeyer.

Zörnig, P. (2013). Distances between words of equal length in a texts. In: Köhler, R., Altmann, G. (eds.), *Issues in Quantitative Linguistics 3, 117-129.* Lüdenscheid: RAM-Verlag.

Appendix

A – adjective (*Бледное* лицо – Pale face; Человек *спокойный* – *Man calm).

ADV – adverb (Комната *наверху* – Room upstairs; *Назад* козырьком – *With the backwards peak).

AO – adjective in an elliptical construction (У меня есть один красный карандаш и один *синий*. – *I have one red pencil and one blue).

AP – apposition (Его *костюм, галстук, рубашка* – вся одежда была абсолютно новой – His suit, tie, shirt – all clothes were brand new; Незнакомец, *мужчина среднего возраста*, подошел ко мне – The stranger, a middle- aged man, came up to me).

APAJ - type of apposition based on adjoinment type of connection with the head word, i.e. its syntactic links with the head word are not based on either agreement, or government (Гостиница «*Байкал*»; слово «*привет*» – The hotel Baikal; the word 'hello').

APX – type of apposition expressed by a proper name which agrees in number, case and gender with the appositive (Хирург *Иванов*, капитан *Смоллетт* – Surgeon Ivanov, Captain Smollett).

AY – adjectival phrase (*Бледное от волнения* лицо – *Pale from anxiety* face; Лицо, *бледное от волнения* – Face pale from anxiety).

CN – compound word with attributive relations of two stems, one of which is a modifier) *Страдальцы*-мальки – Sufferers-fries; Спортсмен-*чемпион* – sportsman-champion).

DAT – dative case (Письмо *другу* – Letter to a friend).

DETF – demonstrative pronoun (*Этот* дом – This house; Книга *эта* – моя. – *Book this is mine).

DETH – indefinite pronoun (*Какие-то* книги – Some books; Книги *какие*-то – *Books some).

DETN – negative pronoun (*Никакой* ошибки – No mistake; Знакомств *никаких* не желаю – *Acquaintances any I do not want).

DETQ – *qualifying* pronoun (*Все* книги – All the books; Книги *все* – *Books all).

DETS – possessive pronoun (*Его* друг – His friend; Книги *мои* здесь – *Books mine are here).

DETV – relative pronouns (Я спросил, *какая* книга пропала – I asked which book was missing; Интересно, экономия *какая* будет – It is interesting economy what will happen).

DETW – interrogative pronoun (*Какая* книга пропала? – Which book is missing?; А машина *какая* там была? – *And car which was there?)

G – genitive case (*Отца* брат – *Of the father brother; Книга *брата* – Book of the brother).

I – infinitive (*Поехать* желание было, *собирать* вещи желания не было – *To go there was a wish, to pack things – there was no wish; Желание *узнать* – (Wish to learn).

INSTR – instrumental case (Восхищение *книгой* – Fascination with the book).

PR – prepositional noun (*на плече* чехол – On the shoulder a cover; Книга *для детей*–Book for children).

PT – participle (*Разбитый* стакан – Broken glass; Чудеса *невиданные* – Miracles unseen).

PTY – participial construction (*Разбитый* на куски стакан – Broken to pieces glass; Книга, *потерянная несколько дней назад* – Book lost a few days ago).

RC – subordinate clause (Это тот человек, *который может нам помочь* – This is the man who can help us; Вот план, *что делать дальше* – Here is a plan what to do next; Это – место, *где мы встретились* – This is the place where we met).

Appendix

Female authors

Author	Title	Year	Words in the abstract	Ad-nominals
T 1 S. Demidova	***Rubinovaja vernost'*** (*Ruby fidelity*). Novel.	2007	3723	614
T 2 D. Dontsova	***Kleopatra s parashjutom*** (*Cleopatra with a parachute*). Novel.	2013	4294	612
T 3 D. Dontsova	***In', Jan' i vsjakaja drjan'*** (*Yin-yang and various stuff*). Novel.	2008	4559	556
T 4 D. Dontsova	***Prodjuser koz'ej mordy*** (*Producer of dirty tricks*). Novel.	2008	4082	600
T 5 A. Marinina	***Kazn' bez zlogo umysla*** (*Execution without bad intentions*). Novel.	2015	4053	616
T6 A. Marinina	***Stechenie obstojatel'stv*** (*Coin-cidence of circumstances)* Novel.	1992	2591	370
T7 A. Marinina	***Ukradennyj son*** (*Stolen dream*) Novel.	1994	4605	637
T 8 D. Rubina	***Belaja golubka Kordovy*** (*White dove of Cordova*). Novel.	2009	4352	848
T 9 D. Rubina	***Poslednij kaban iz lesov Pontevedra*** (*The last boar from the woods of Pontevedra*). Novel.	1998	3055	653
T 10 D. Rubina	***Topolev pereulok*** (*Topolev alley*). Long story.	2015	3835	858
T 11 V. Tokareva	***Lavina*** (*Avalanche*). Long story.	1955	4532	508
T 12 V. Tokareva	***Moi muzhchiny*** (*My men*). Long story.	2015	4565	652
T 13 V. Tokareva	***Tihaja muzyka za stenoj*** (*Soft music behind the wall*). Long story.	2012	4537	644
T 14 L. Tret'jakova	***Damy i gospoda*** (*Ladies and gentlemen*). Novel.	2008	3180	574
T 15 L. Tret'jakova	***Krasavitsy ne umirajut*** (*Beatiful women don't die*). Novel.	1998	2982	734
T 16 L. Ulitskaja	***Zelenyj shater*** (*Green marquee*). Novel.	2011	4437	884
T 17 L. Ulitskaja	***Iskrenne vash Shurik*** (*Yours truly Shurik*). Novel.	2006	3796	957
T 18 T. Ustinova	***Oligarh s Bol'shoj Medveditsy*** (*Oligarch from the Big Dipper*). Novel.	2004	4749	587
T 19	***Vselenskij zagovor*** (*Cosmic	2016	4228	467

T. Ustinova	*conspiracy*). Novel.			
T 20 T. Ustinova	***Moj general*** (*My general*). Novel.	2002	4076	567

Male authors

Author	**Title**	**Year**	**Words in the abstract**	**Ad-nominals**
T 21 B. Akunin	***Table-Talk.*** Story.	2006	3966	608
T 22 B. Akunin	***Pikovyj valet*** (*Jack of spades*). Long story.	1999	4043	650
T 23 B. Akunin	***Turetskij gambit*** (*Turkish gambit*). Novel.	1998	5225	777
T 24 A. Bushkov	***Piran'ja. Vojna oligarhov*** (*Piranha. War of oligarchs*). Novel.	2007	2573	354
T 25 A. Bushkov	***Piran'ja protiv vorov*** *(Piranha against thieves)*. Novel.	2001	3834	673
T 26 A. Bushkov	***Tanets Beshenoj*** (*The dance of the rabid*). Novel.	2001	4839	767
T 27 M. Veller	***Laokoon.*** Story.	1993.	2438	375
T 28 M. Veller	***Marina.*** Long story.	1993	7557	1292
T 29 M. Veller	***Pjatiknizhie*** (*The Torah*). Long story.	2009	3461	762
T 30 S. Dovlatov	***Inostranka*** (*A foreign woman*). Long story.	1986	2802	461
T 31 V. Erofeev	***Russkaja krasavitsa*** (*Russian beauty*).	1990	4206	577
T 32 D. Koretskij	***Antikiller.*** Novel.	1995	2948	422
T 33 D. Koretskij	***Antikiller-5.*** Novel.	2014	3937	528
T 34 D. Koretskij	***Antikiller-6.*** Novel.	2016	2815	414
T 35 V. Pelevin	***Operatsija "Burning Bush"*** (*Operation "Burning Bush"*).	2010	3392	676
T 36 V. Pelevin	***Assasin.***	2008	3565	487
T 37 V. Pelevin	***Grecheskij variant*** (*Greek variant*). Novel.	**1977**	2891	616
T 38 Z. Prilepin	***Obitel'*** (*Convent*). Novel.	2014	4523	618
T 39 Z. Prilepin	***Patologii*** (*Pathologies*). Novel.	2005	3968	664
T 40 Z. Prilepin	***Sher amin'*** (*Cher amen*). Story.	2016	3774	457

Glottometrics 18, 2017, 77-90

Probability Distribution of Syntactic Divergences of Determiner *his*-(adjective)-Noun Structure in English-to-Chinese Translation

Biyan Yu & Yue Jiang[*]

Abstract. Studies of translation divergences reveal that lexical divergences in English-to-Chinese translation some regularities which can be modeled by a probability distribution. Besides lexical divergences, we hypothesize that there might be a great number of syntactic divergences since Chinese and English belong to two typologically divergent languages. In this study, we investigated whether the Chinese translation of English determiner *his*-(adjective)-noun structure (DNS (*his*) hereafter) diversifies in syntactic construction and whether the distribution of diversified Chinese translations conforms to the regularity of diversification process introduced by Altmann (2005). It is found that the twenty-two Chinese syntactic constructions corresponding to one single English DNS (*his*) do form a decreasing rank-frequency distribution, which is in comfortable agreement with the usual Zipf-Alekseev function. This diversification process may be ascribed to both linguistic and cultural factors such as contextual factors, translator's subjectivity, functional equivalents, and the tendency for minimal effort in language coding and decoding during translating process.

Key words: syntactic divergence; probability distribution; translation

1. Introduction

Diversification processes (Altmann, 2005) refer to the course when "the attribute space of an entity expands in one or more dimensions" (p. 646). Take the word *word* for example, lexical denotation is one of its attributes, along with its form, articulation, etc. Semantic diversification happens when *word* is interpreted as 1) a single unit of language, 2) a short conversation, 3) pragmatic functions such as warning, advice, or praise, or 4) a promise. This phenomenon of polysemy is a typical diversification process. Besides, allophony, allomorphy, homonomy, variation of grammatical rules, and many other language phenomena all belong to diversification processes. As a fundamental phenomenon in linguistics, diversification is significantly involved in language evolution. In terms of synergetic linguistics, it is an aspect of self-regulation omnipresent in all language phenomena. Language is a special view of reality created by the given folk. If languages of the same family diverge in their evolution, they do not reflect the reality in the same way. Such a diversified evolution leads to divergences also in the description of the reality.

The study of diversification starts from three hypotheses (Altmann, 2005: 646). The primary one is stated as:

[*] Address correspondence to: Yue Jiang, School of Foreign Studies, Xi'an Jiaotong University, 710049, Xi'an, Shaanxi, China. Email address: yuejiang58@163.com.

The classes made up by diversification form a decreasing rank-frequency distribution or another (not necessarily monotonous) discrete distribution according to whether the classes represent a nominal or a numerical variable respectively. This assumption can be explained further as the disparity among the diversified classes of an entity, forming a decreasing distribution according to their frequencies. It has been the ground of many research models and has received considerable verification in quantitative linguistics. Fang & Liu (2015) applied it to describe translation and found that different English translation versions of Chinese character *dao* 道 (literally meaning "said") selected from a Chinese literary masterpiece *Hongloumeng* (*A Dream of Red Mansions*) are distributed according to a diversification process, which is consistent with the modified right-truncated Zipf-Alekseev distribution.

Translation is indeed a diversification process, which has been clarified by translation theories and proved by translation practice. Toury (2004) describes translation as subject to social-cultural constraints of several types and varying degrees: "general, relatively absolute *rules* on the one hand and pure *idiosyncrasies* on the other. Between these two poles lies a vast middle-ground occupied by inter subjective factors commonly designated *norms*." Diversified constraints exert influence on translator's choices and thus generate translation with diversified degrees of adequacy and acceptability. In Tong (1994), the diversification of translation refers to the process in which the repeated meaning in a source text is conveyed by more than one construction in a target text, which helps avoid the consequence that butter to butter is no relish. Song (2012) argues that the activity of translation involves so many parameters such as the translation skopos, reference books, time, targeted readers, publishers, and translators themselves that it always shows non-linear causality in translation. In other words, the transformation from a source text to a target text is not limited to a one-to-one pattern but one-to-many. As a result, various variables give rise to the diversification process in translation which manifests itself in both lexical and syntactic divergences.

Much research has been conducted to investigate the lexical divergences in the translation between English and Chinese. Given the causes of lexical divergences, it is very much likely that there are a great number of syntactic divergences in either English-to-Chinese or Chinese-to-English translation since English and Chinese belong to two typologically divergent languages. Köhler & Altmann (2000) find that the properties of syntactic constructions and categories are lawfully distributed according to a few probability distributions. Thus, we hypothesize that such syntactic divergences can also be found in translation between English and Chinese, which also follow some kind of regularity like the lexical divergences reported by Fang & Liu (2015).

Determiner *his*-(adjective)-noun structure is selected for this study for four reasons. First, literature review reveals that the determiner system cannot be analyzed within the existing classification of translation divergences and thus deserves further study (for cogent description of translation divergences, see, for example, Dorr 1990, 1992, 1993, 1994; Mahesh et al. 2005; and Kulkarni et al. 2013). The second reason is related to the two languages *per se*. Determiner system is a typical grammar gap between English and Chinese. English uses noun phrases to convey both concrete entities and abstract concepts, necessarily with the determiners to determine or modify. Despite its lack of the determiner system as that in English, Chinese has other ways to perform the function of English determiners. This interlingual syntactic and semantic gap may contribute to divergences in the Chinese rendition of English determiners. Third, determiners are seldom used independently in context. As Quirk (1973) says, "determiners themselves have no function independent of the noun they precede" (p. 137). Given this fact, to study determiners in English-to-Chinese translation in a more detailed way, it is indispensable to investigate determiners together with the nouns they determine and sometimes adjectives as well because the latter might also be involved in such a syntactic structure and form a “determiner-(adjective)-noun structure”. Lastly, the possessive determiner *his* is especially representative of English determiners as it is one of the most common high-frequency words according to the word list from BNC (British National Corpus). The study of the English-to-Chinese translation of DNS (*his*) may help investigate syntactic divergences.

The hypothesis we intend to corroborate in this study is that the Chinese translation of English DNS (*his*) diversifies from its English syntactic construction. The following research questions will be addressed in this paper:

1. Do the various Chinese translations of English DNS (*his*) follow a diversification process characterized by a decreasing rank-frequency distribution in terms of syntactic construction?

2. What might be the causes of syntactic divergences in the Chinese translation of English DNS (*his*)?

The issue of translation divergences is of great significance for translation studies. The present study of syntactic divergences concerning the determiner system is to be conducted by referring to some techniques and methods in quantitative linguistics, which is supposed to bring new ideas to the study of translation divergences and promote the formalization of a better-defined framework which may hopefully accommodate all translation variants. It may also help shed some light on the nature of languages when we look for the causes of divergences and facilitate the theoretical development of translation studies.

2. Materials and Method

The data to be used here is retrieved from a corpus consisting of the English version of *Pride and Prejudice* by Jane Austen and its Chinese translation by Zhili Sun. Since it was first published in 1813, the novel has been well received and translated into many languages, including French, German, Danish, Swedish, and Chinese, to name just a few. In China, it has a dozen of translated versions. Among these Chinese translations, a widely endorsed one is accomplished by Zhili Sun and published in 1985. His translation is popular owing to its unique translator style and has had great impact on the introduction of *Pride and Prejudice* to the Chinese readership. This is why we have chosen Sun’s translation as our object of study. For the subsequent statistical analysis, the two texts are saved as electronic form and are aligned sentence to sentence as a parallel Chinese to English corpus.

After digitalizing the texts, we segmented and tagged them. The English source text was part-of-speech (POS) tagged by the Free CLAWS WWW tagger [1] with C7 Tagset. The Chinese translation was segmented, tagged and annotated by using a Chinese segmenting and tagging system called “NLPIR/ICTCLAS 2015[2]” and further checked manually for a more reliable result according to Huang & Liao (2015).

In order to examine the Chinese translation of DNS (*his*) in great detail, we used the RegEx *his_APPGE* as the condition to filter out English sentences that do not have DNS (*his*) in them. We finally obtained 895 English-Chinese sentence pairs where 1,149 possessive determiners *his* occur, after which, SZL’s translation is observed and annotated manually.

[1] It is a free web tagging service for English developed by UCREL at Lancaster that offers access to the latest version of the tagger which is used to POS tag c.100 million words of the British National Corpus (http://ucrel.lancs.ac.uk/claws/trial.html).

[2] It is designed for Chinese and shared on Natural Language Processing & Information Retrieval Sharing Platform (http://ictclas.nlpir.org/).

Every occurrence of DNS (*his*) in source texts is supposed to have its corresponding Chinese translation in whatever way it is presented.

The Chinese translation of *his* should literally be the personal pronoun *ta* 他 followed by the auxiliary word *de* 的. However, the translation usually becomes complicated when this single determiner *his* appears as DNS (*his*) in different contexts, esp. with many adjectives involved between *his* and the noun it modifies. After observation and annotation, we found that noun phrases in M-D (modifier-head) construction, which is one of the attributive structures in Chinese (Huang & Liao, 2015), could be regarded as the most common and frequent translation of DNS (his). Six types of M-D construction have been detected in SZL's translation as listed in Table 1.

Table 1
M-D construction of Chinese translation of English DNS (*his*)

Type	**Structure**
1	*ta* + ude1 + n
2	*ta* + n
3	*ta* + ude1 + a|m|n|b +n
4	*ta* + a|m|n|b + ude1 + n
5	*ta* + ude1 + a|m|n|b + ude1 + n
6	*ta* |this|that + (q) + (a|m|n|b + ude1) + n

ta = Chinese character 他; ude1 = auxiliary word 的; n = noun; a = adjective; m = numeral; b = distinguisher.

This study reveals that nearly half of DNS (*his*) "disappears" after the Chinese rendition. In other words, we cannot find the corresponding noun phrases in the target text. This phenomenon might be categorized as "subtraction" (Nida, 1964:231), and subsequently "implicitation" (Vinay & Darbelnet, 1995:344), which is usually "treated as a stepbrother of explicitation" as a translation universal (Klaudy & Károly, 2005:13) by corpus-based translation researchers (Baker, 1994; Laviosa, 1998, etc.). Another finding is referential substitution, which means that the signification (Saussure, 2001:67) of possessive determiner *his* or DNS (*his*) is performed by the names of the entities they refer to. The prevalent existence of these two syntactic variations in translation prompted us to give a statistical and quantitative probe into their nature.

It is assumed that apart from the M-D construction, omission and referential substitution in translation, there might be other syntactic divergences. Our careful observation and annotation of the Chinese translation reveal that the conveyance of the meaning of DNS (*his*) is no longer confined to the structures of noun phrases but also is realized by the corresponding personal pronouns or related nouns, verbs, or adjectives as various syntactic components (Table 2). Each type of translation is illustrated with one example in Appendix.

As we hypothesize in the beginning of this paper, Chinese translation of English DNS (*his*) diversifies syntactically from its corresponding English syntactic construction, and that the diversification should proceed in a regular manner, presenting a decreasing rank-frequency distribution since the translations are recognized as nominal variables. More precisely, we assume that the investigated distribution obeys the Zipf-Alekseev function. We apply the approach shown in Popescu et al. (2014: 5) given as follows: the relative rate of change in frequencies, i.e. *dy/y* (here *y* = *f(x)*) is given as

$$\frac{dy}{y}=\frac{\mathrm{g}(x)}{\mathrm{h}(x)}dx \qquad （1）$$

where g(x) represents the situation/state in the given language and the striving of the writer/speaker/translator who controls the output given as *g(x) = A + B ln x*.

Table 2
Syntactic components of DNS (*his*) and that of its Chinese translations

Syntactic components of English DNS (*his*)	**Syntactic components of its Chinese translations**
1. Subject 2. Object 3. Complement 4. Predicative	1. Subject 2. Predicate 3. Object 4. Attribute 5. Adverbial 6. Subject & Predicate 7. Adverbial & Predicate 8. Object & Complement 9. Subject & Subject 10. Subject & Complement 11. Subject & Attribute 12. Subject & Object 13. Subject & Adverbial 14. Predicate & Object

The logarithm reminds us of the Weber-Fechner law in psychology (Stout, 1915). The function *h(x)* controls the equilibrating force of the hearer/reader and is given as *h(x)* = *Cx*. Putting these parts together, we obtain

$$\frac{dy}{y}=\frac{A+B\ln x}{Cx}dx. \qquad （2）$$

After solving this differential equation and reparametrizing it, we obtain

$$y=cx^{a+b\ln x}, \quad \mathrm{x}=1,2,3,\dots,\mathrm{n}.$$

This is called Zipf-Alekseev law (*a* and *b* may be then negative or positive), which is adequate for capturing different diversification forms.

3. Results and Discussion

In this section, we will present the results of our study as related to the two research questions in the first section.

3.1. Probability distribution of Chinese translations of English DNS (*his*)

The Chinese translations of English DNS (*his*) in SZL's translation can mainly be categorized into four types, namely, M-D construction, omission, referential substitution, and related personal pronouns, nouns, verbs, or adjectives as various syntactic components. Table 3 shows various Chinese translations of English DNS (*his*) ranked according to their frequencies of occurrence.

Table 3
Diversification of Chinese translation of English DNS (*his*)

Translation	**Rank**	**Frequency**
Omitting	1	421
Subject & Predicate	2	192
<1>	3	164
referential substitution	4	130
<2>	5	97
<6>	6	41
<3>	7	39
<4>	8	19
Predicate	9	14
Adverbial	10	12
Subject & Object	11	12
Object	12	10
Subject	13	9
Subject & Adverbial	14	9
Object & Complement	15	8
Adverbial & Predicate	16	7
Subject & Attribute	17	3
Predicate & Object	18	3
Subject & Subject	19	2
<5>	20	1
Attribute	21	1
Subject & Complement	22	1

<1>~<6> stands for the six M-D constructions listed in Table 1.

We applied the software NLREG to the data shown in Table 3 and obtained the results presented in Table 4 and Figure 1.

Table 4
Fitting the Zipf-Alekseev function to translations of English DNS (*his*) in SZL's translation

X[i]	**F[i]**	**NP[i]**		**X[i]**	**F[i]**	**NP[i]**
1	421	414.51990		12	10	15.39571
2	192	228.63027		13	9	13.12514
3	164	143.73863		14	9	11.28895
4	130	98.17547		15	8	9.78596
5	97	70.90547		16	7	8.54249
6	41	53.31461		17	3	7.50388
7	39	41.33120		18	3	6.62895
8	19	32.82111		19	2	5.88620
9	14	26.57589		20	1	5.25124
10	12	21.86907		21	1	4.70495
11	12	18.24249		22	1	4.23220
a = -0.677846538, b = -0.260519622, c =414.519904, R^2 = 0.9793 (97.93%)						

In this table, X[i] is the observed classes; F[i] is the observed frequency; NP[i] is the calculated frequency according to the usual Zipf-Alekseev function; *a*, *b*, and *c* are the parameters of the function; R^2 represents the Coefficient of Determination.

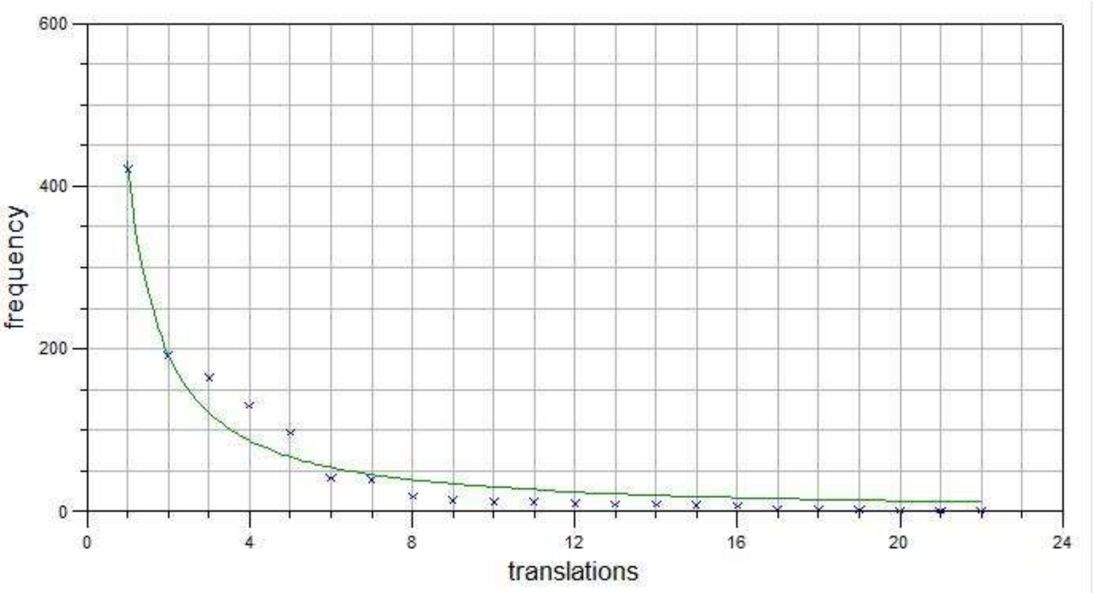

Figure 1. Fitting the usual Zipf-Alekseev function to
different translations of English DNS (*his*) in SZL's translation

The Coefficient of Determination R^2, though defined for linear functions, "may be interesting in many cases and help to enlarge experience with this coefficient in connection with non-linear functions" (Altmann Fitter (3.1), User Guide, p. 10). The determination coefficient R^2 shows that considering the data as a simple function the result is desirable. That is to say, the distribution of DNS (*his*)'s translation fits well with the usual Zipf-Alekseev function. Hence, the various Chinese translations of English DNS (*his*) are distributed according to a diversification process characterized by a decreasing rank-frequency distribution.

3.2. Causes of syntactic divergences in the Chinese translation of English DNS (*his*)

The grounds on which diversification processes start are of much variety. Altmann (2005) puts forward six factors that may contribute to intralingual diversification processes (p. 647-648). We believe that four of them might also have explanatory power over translation divergences. Hence the analysis of the causes of syntactic divergences is carried out under these four factors.

3.2.1. Environmentally conditioned variation

"The form or the meaning of an entity is modified according to the environment in which it is situated" (Altmann, 2005:647). The environment in linguistics is a context, as Fang & Liu (2015) also mentioned as one of the causes of lexical divergence (p. 63). Context is "a highly complex notion" which comprises external (situational and cultural) factors and internal (individual and cognitive) factors (House, 2006:342). Context exerts influence on the overall organization of language, affecting its syntactic, semantic, lexical and phonological structures. The greater the distinction between two contexts or linguistic environments, the more pronounced the translation divergences will be.

The omission of English DNS (*his*) in the Chinese translation is probably attributed to both different cultural and textual contexts. In China, the traditional culture and thoughts such as the doctrine of Confucius and Mencius, Taoist philosophy and many other influential beliefs shape people's modes of thinking into something ethical, holistic, subjective, intuitive, and fuzzy (Lian, 2010). It may possibly explain why the Chinese language is featured by being analytic, paratactic, simplex, dynamic, supple, personal and so on. These features make Chinese depend more on context in conveying certain ideas than English. Reading a paratactic language, one has to infer the relation between entities through semantic rather than syntactic analysis of phrases in context. It is very common in English that one possessive determiner occurs in one sentence for several times owing to its hypotactic nature. In contrast, it is rare in Chinese that the same individual word occurs repeatedly in a succession of sentences unless it is used for some rhetorical purposes such as emphasizing a particular word. In most cases in Chinese, the relation between entities is signified by context. This may comprise to a large extent the cause of omitting DNS (*his*) in the English-Chinese translation.

Referential substitution is another result of translating with the information from context. House (2006) argues that translation is a process of re-contextualization, during which the information from the source language and source context is given a new shape in a new language and new context (p. 343), and "linguistic forms such as personal pronouns, sentence types and modality assume new and contextually determined values" (p. 342). For example, it is not difficult to infer from the context that the noun phrase "his friend" in the third chapter of *Pride and Prejudice* refers to Mr. Darcy. The translator uses the proper name "Mr. Darcy" instead of the noun phrase. Perego (2003) regards this substitution as "specification", "a case of addition of meaning(s), though not necessarily of words" which occurs through "the replacement of a general and wide-ranging word with a more specific and narrow one", and brings about "a clear, more detailed and more transparent meaning" (p. 73). Therefore, this way of translation can facilitate the comprehension of the target texts by the target language readers.

3.2.2. Conscious change

Variation and change in the use of language are more or less a conscious linguistic behavior, esp. in terms of expressing emotional, creative and other intended ideas. "Under certain circumstances it is possible to diversify a feature in language consciously" (Altmann, 2005:647). In translation studies, the role of a translator is usually noticeable and visible in the translating process, which is termed as "translator's subjectivity", which is also regarded as a contributing factor by Fang & Liu (2015:63) to lexical divergences. When trying to convey the content and style of the source text, translators will leave "a kind of thumb-print" (Baker, 2000:245) on their translations, "including open interventions, the translators' choice of what to translate, their consistent use of specific strategies, and especially their characteristic use of language" (Saldanha, 2011:27).

The translator's subjectivity contributes to a translator's conscious variation and change in the choice of words and expressions. It also in a sense contributes to the translation diversification as listed in Table 3. All the diversified types of translations actually are derived from the translator's effort to convey and rewrite the meanings of DNS (*his*), and behind the different frequencies may lie their own conscious or unconscious choice, habit, and preference in their translating process.

3.2.3. Self-regulation

Self-regulation of language triggers the emergence of functional equivalents (Altmann, 2005:647). For example, to distinguish different words, Chinese, as a tone language, has tone diversification, whereas English, as an alphabetic language, has diversification of word length. Diversifications in two languages, different as they are, have the same function and are thus called "functional equivalents".

English has many forms of expression to convey ownership such as content words (*have, own, belong to*), prepositions (*of*), grammatical forms (*-'s*), etc. Among them, possessive determiners occur with a relatively high frequency. Chinese, on the other hand, has no absolutely corresponding word classes but many other ways to convey ownership. For example, the Chinese character *de* 的 can express ownership when it acts as an auxiliary word and collocates with personal pronouns. Using demonstrative elements is also a way to express ownership in Chinese. Theoretically, each form of ownership in English can be translated into any form in Chinese. This is the intrinsic possibility that the functional equivalents provide for translation divergences. In the practice of Chinese translation, nevertheless, the principle of rhythm (Xu, 2003:33) is usually taken into consideration in the use of auxiliary word *de* 的 and demonstrative elements in Chinese sentences. According to Xu (2003), a foot in Chinese basically comprises two syllables, and the presence or the absence of *de* 的 should be conditioned by the rhythmic harmony in the combination of syllables (p. 33-34). The degree of acceptance by the readers to the attributive structures is negatively correlated with the occurrence frequency of *de* 的. Huang & Xu (1997) discovered that the attributive structures can only be well accepted with one and reluctantly accepted with no more than three *de*s 的 (p.

41), which might be decided by the cognitive load from the perspective of dependency grammar (Liu, 2008; Jiang & Liu, 2015). This may account for why six different M-D constructions are involved in translating DNS (*his*).

The way that the signification of DNS (*his*) is realized by relevant personal pronouns, nouns, verbs, or adjectives as various syntactic components in translation can also be explained by functional equivalents. Table 2 shows that as a form of indicating ownership in English, DNS (*his*) usually serves in sentences as either subject or object. However, its Chinese translations serve as various syntactic components in addition to subject and object, and even diversify into two syntactic components. Huang & Liao (2015) observe that the relationships between word classes and syntactic components are simple in English but complicated in Chinese. Figure 2 is a comparison of syntactic functions, or rather, sentence constituents played by English parts of speech with that by their Chinese counterparts (Huang & Liao, 2015: 35). Heavy lines mean major functions, fine lines stand for secondary functions, and dotted lines mean few cases. It can be seen in the figure that most English parts of speech, except nouns, usually can serve only as one sentence constituent whereas Chinese parts of speech, including nouns, verbs, and adjectives, are multifunctional and can serve as five different sentence constituents. The equivalent function of different parts of speech in English and Chinese might account for the multiple divergences in the translation of DNS (*his*).

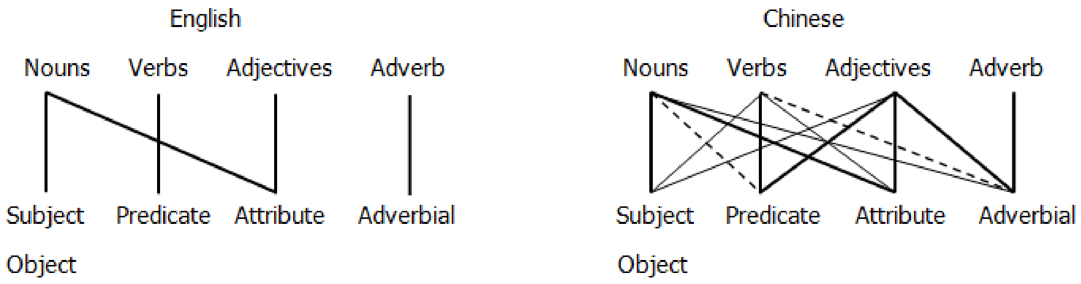

Figure 2. Relationships between parts of speech and syntactic components

3.2.4. The tendency for minimal coding and decoding effort

Verbal interaction, either spoken or written, has been regarded as a process of coding and decoding. It is during this interaction that diversification and unification working in opposite directions occur.

Translation is also a process of coding and decoding information, during which the author of a source text codes his/her information whereas the translator tries to decode the source text to relay its message. This, to some extent, is another effort of coding. It is not as simple as transcoding (automatic word-for-word translation) (Vermeer, 2004), which "may result in a target-language text or utterance that is clumsy, erroneous, or even nonsensical" (Gile, 1995:80). Translators are supposed to decode or capture the language-independent information from a source text and code it in a target language. The case seems to be similar to the translator's subjectivity. However, the two cases should be perceived to be different in that the translator's subjectivity is more conscious in nature whereas the process of coding and decoding is less conscious and even unconscious.

To decode is to comprehend the source text, which "goes beyond the simple recognition of words and linguistic structures" (Gile, 1995:79). Gile (1995) thinks that comprehension should be considered as a continuum going from non-comprehension to full-comprehension, and neither of the two poles is absolute in practice (p. 82). Generally, human comprehension of something is always incomplete. In other words, people tend to stop pursuing thorough understanding of something due to inherent “laziness” (Altmann, 2005:648). This “laziness” may naturally induce or trigger man’s tendency to use minimal effort in decoding. However, the minimal effort can be hardly quantified, which varies with each individual since people’s knowledge reserve might be distinct. Comprehension at different levels of decoding effort may lead to diversification.

The coding phase that a translator is responsible for is when the divergences intensify and take shape. Languages are not isomorphic (Gile, 1995:50). In other words, there are always obvious differences in lexicons and grammars of different languages or subtle differences in their use in context. Even the same message can be conveyed by means of sentences completely different in form (Perego, 2003:67). Translators may unconsciously code the source information in a totally different form in a target language to achieve the idiomaticity of the target language. This might account for why the construction of DNS (*his*) is transformed into relevant personal pronouns, nouns, verbs, or adjectives as various syntactic components to convey its meaning.

4. Conclusion

This study investigated the Chinese translation of English DNS (*his*) at syntactic level in an attempt to validate whether the translation diversifies from English DNS (*his*) in syntactic construction and whether the distribution of various Chinese translations abides by the regularity of diversification processes introduced by Altmann (2005). The results of the study found twenty-two Chinese syntactic constructions corresponding to English DNS (*his*) and they form a decreasing rank-frequency distribution, which agrees with the usual Zipf-Alekseev function.

The causes of the translation divergences in question may be ascribed to a variety of factors, both linguistic and cultural. In this paper, four contributing factors are discussed in detail, including contexts, translator’s subjectivity, functional equivalents, and the minimal effort in the process of coding and decoding. However, it remains to be investigated what is the very factor that initiates the diversification process, as the factors might take effect at the same time (Altmann, 2005).

Surely, there are some limitations in this research. Firstly, the annotation of the Chinese translation of English DNS (*his*) is more or less subjective although it was done by referring to the modern Chinese grammar (Huang & Liao, 2015). Secondly, we chose only one English structure to observe its translation divergences in only one Chinese translator’s work. The findings might be more interesting and more convincing if the scale of the research is enlarged in the future.

Acknowledgments

This work is supported by the Chinese National Social Sciences Foundation (Project No. 17BYY007). A Stylometric research on the linguistic features of translations from different translation directions.

References

Altmann, G. (2005). Diversification processes. In: R. Köhler, G. Altmann & R. G. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook:* 646–658. Berlin: de Gruyter.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target* 12(2): 241-66.

Dorr B. J. (1990). Solving thematic divergences in machine translation. *Meeting of Association for Computational Linguistics*. Association for Computational Linguistics, 127-134.

Dorr B. J. (1992). The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7(3), 135-193.

Dorr, B. J. (1993). Interlingual machine translation: a parameterized approach. *Artificial Intelligence, 63*(1–2), 429-492.

Dorr, B. J. (1994). Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 20(4), 597-633.

Fang, Y., & Liu, H. (2015). Probability distribution of interlingual lexical divergences in Chinese and English: 道 (dao) and said in Hongloumeng. *Glottometrics 32*, 63-87.

Gile, D. (1995). *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam/Philadelphia: John Benjamins.

House, J. (2006). Text and context in translation. *Journal of Pragmatics*, 38(3), 338-358.

Huang, B. & Liao, X. (2015). *Contemporary Chinese language*. Beijing: Higher Education Press.

Huang, Z. & Xu, P. (1997). The optimal amount of *de* in attributive structure in Chinese translation. *Contemporary Rhetoric* 6, 40-41.

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

Klaudy, K., & Károly, K. (2005). Implicitation in translation: empirical evidence for operational asymmetry in translation. *Across Languages & Cultures*, 6(1), 13-28.

Kulkarni, S. B., Deshmukh, P. D., & Kale, K. V. (2013). Syntactic and Structural Divergence in English-to-Marathi Machine Translation. *International Symposium on Computational and Business Intelligence* (pp.191-194). IEEE.

Köhler, R. & Altmann, G. (2000). Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics*, 7(3), 189-200.

Laviosa, S. (1998). Universals in Translation. In: Baker, M. (Eds.) *Encyclopedia of Translation Studies*. London: Routledge.

Lian, S. (2010). *Contrastive Studies of English and Chinese*. Higher Education Press, Beijing.

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

Mahesh, R., Sinha, K., & Thakur, A. (2005). Translation divergence in English-Hindi. Retrieved May 25, 2017, from http://mt-archive.info/EAMT-2005-Sinha-2.pdf.

Nida, E. A. (1964). *Toward a Science of Translating*. Leiden: Brill.

Perego, E. (2003). Evidence of explicitation in subtitling: towards a categorisation. *Across Languages & Cultures*, 4(1), 63-88.

Popescu, I. I., Best, K., & Altmann, G. (2014). *Unified modeling of length in language*. Lüdenscheid, North Rhine-Westphalia: RAM-Verlag.

Quirk, R., Greenbaum, S., Leech G., Svartvik J. (1974). *A grammar of contemporary English*. London: Longman.

Saldanha, G. (2011). Translator Style. *The Translator* 17:1, 25-50.

Saussure, F. (2001). *Course in General Linguistics.* Beijing: Foreign Language Teaching and Research Press. Translated by R. Harris.

Song, Z. (2012). Nonlinear characteristics of the translation process. *Shanghai Journal of Translators* 4, 13-17.

Stout, G. F. (1915). *A manual of psychology* (3rd edition). New York: Hinds, Noble & Eldridge.

Tong, Q. (1994). Diversification of translation. *Shanghai Journal of Translators for Science and Technology* 2, 14-15.

Toury, G. (2004). The Nature and Role of Norms in Translation. In L. Venuti (Eds), *The Translation Studies Reader (Second Edition)*. London & New York: Routledge.

Vermeer, H. J. (2004). Skopos and Commission in Translational Action. In L. Venuti (Eds), *The Translation Studies Reader (Second Edition)*. London & New York: Routledge.

Vinay, J. P. & Darbelnet, J. (1995). *Comparative Stylistics of French and English. A Methodology for Translation*. Amsterdam: John Benjamins. Translated by J. C. Sager and M. J. Hamel.

Xu, C. (2003). The Constraints of *de*'s Presence and Absence. *Contemporary Rhetoric* 2, 33-34.

Appendix

Syntactic components of Chinese translations of DNS (*his*)

(1) Subject

EN: … in_II ***his_APPGE manner_NN1*** of_IO bidding_VVG her_APPGE adieu_NN1 …

CH: … ***他/rr*** 向/p 她/rr 道别/vi …

(2) Predicate

EN: … the_AT consciousness_NN1 of_IO this_DD1 was_VBDZ another_DD1 reason_NN1 for_IF ***his_APPGE resolving_VVG to_TO*** follow_VVI us_PPIO2 …

CH: … 自信/v 有/vyou 这/rzv 点/qt 门路/n ，/wd 才/d ***决定/v*** 紧跟着/d 我们/rr 而/cc 来/vf …

(3) Object

EN: … return_VVI ***his_APPGE affection_NN1*** with_IW sincere_JJ …

CH: … 以/p 同样/b 的/ude1 钟情/vi 报答/v ***他/rr*** …

(4) Attribute

EN: … no_AT man_NN1 in_II ***his_APPGE senses***_NN2 would_VM marry_VVI Lydia_NP1 …

CH: … 一个/mq ***头脑/n 健全/a*** 的/ude1 人/n 是/vshi 不/d 会/v 跟/p 莉迪亚/nrf 结婚/vi 的/ude1 …

(5) Adverbial

EN: … admit_VV0 ***his_APPGE society_NN1*** in_II town_NN1 …

CH: … 在/p 城里/s 也/d 不/d ***和/cc 他/rr*** 来往/vi …

(6) Subject & Predicate

EN: … she_PPHS1 had_VHD been_VBN assured_VVN of_IO ***his_APPGE absence_NN1*** …

CH: … 她/rr 是/vshi 听说/v ***他/rr 不/d 在家/vi*** 才/d …

(7) Adverbial & Predicate

EN: … planning_VVG ***his_APPGE happiness_NN1*** in_II such_DA an_AT1 alliance_NN1 …

CH: … 筹划/v 这/rzv 门/q 亲事/n 会/v ***给/p 他/rr 带来/v 多/m 大/a 幸福/a*** …

(8) Object & Complement

EN: … provoke_VVI ***his_APPGE ridicule_NN1*** …

CH: … 惹/v ***他/rr 嘲笑/v*** …

EN: The_AT perpetual_JJ commendations_NN2 of_IO the_AT lady_NN1 either_RR on_II ***his_APPGE hand-writing_NN1*** …

CH: … 夸奖/v ***他/rr 字/n 写/v 得/ude3 棒/a*** …

(9) Subject & Subject

EN: … ***his_APPGE situation_NN1*** must_VM have_VHI been_VBN benefited_VVN by_II marriage_NN1 …

CH: ***他/rr*** 若是/c 结/v 了/ule 婚/ng，/wd ***境况/n*** 势必/d 会/v 好/a 些/q …

(10)Subject & Complement

EN: ***His_APPGE complexion_NN1*** became_VVD pale_JJ with_IW anger_NN1 …

CH: ***他/rr*** 气/n 得/ude3 ***脸色/n*** 铁青/z …

(11)Subject & Attribute

EN: … ***his_APPGE astonishment_NN1*** was_VBDZ obvious_JJ …

CH: ***达西/nrf*** 又/d 显出/v 非常/d ***惊讶/a*** 的/ude1 样子/n …

(12)Subject & Object

EN: … ***his_APPGE civility_NN1*** was_VBDZ so_RG far_RR awakened_VVN …

CH: ***他/rr*** 终于/d 想起/v 了/ule ***礼貌/a 问题/n*** …

(13)Subject & Adverbial

EN: … give_VVI you_PPY ***his_APPGE reasons_NN2*** for_IF this_DD1 interference_NN1 …

CH: …告诉/v 你/rr ***他/rr 为什么/ryv*** 要/v 干预/v …

(14) Predicate & Object

EN: Her_APPGE answer_NN1 was_VBDZ warmly_RR in_II ***his_APPGE favour_NN1*** ._

CH: 伊丽莎白/nrf 激动/a 地/ude2 回答/v 说/v，/wd 非常/d ***喜欢/vi 他/rr***

Glottometrics 40, 2018, 91-103

A Menzerath-Altmann Model for NP length and Complexity in Maritime English

Yu Yang[1,2], *Se-Eun Jhang*[2]

1. Dalian Maritime University, *yuyang8889@hotmail.com*
2. Correspondence Author, Korea Maritime and Ocean University, *jhang@kmou.ac.kr*

Abstract. This paper investigates the length distribution and the complexity of Noun Phrases (NPs) in maritime English using a Maritime English Corpus (MEC) as the data source. The results reveal that the distributional patterns of the NP length and the relationship between NP length and complexity obey the Menzerath-Altmann law.

Keywords: Maritime English, Corpus, NP, length, complexity, Menzerath-Altmann model

1. Introduction

Maritime English belongs to the domain of English for specific purposes (ESP); it is the lingua franca for people engaged in international maritime transportation, whose throughput accounts for more than 80% of the goods for world trade, and hundreds of thousands of seafarers of different countries speaking different tongues work in this industry communicating in one language - English, among themselves, between labour and management, from ship to ship and between sea and shore. Quite often the captain, other senior officers and crews of one ocean-going ship are from several countries and English is the only language spoken both as a working language and for everyday conversation. Shipping accidents often occur, some being acts of God, and some because of human errors. Both types of sea distress have been thoroughly analyzed, and the nature of several accidents of the latter is determined as linguistic-miscommunication in English. Because of the importance of English, the International Maritime Organization (IMO) under the UN has set English threshold for the international shipping circles and commissioned scholars to research into, and compile course books for, maritime English.

As teachers and researchers of maritime English, we are interested in the syntactic characteristics of maritime English, focusing on its sentence and phrase structures. In an inspiring article carried in *Glottometrics 24*, Wang (2012) examined NP structures and the relationship between length and complexity of NPs of general English with the quantitative approach, and the results proved the effectiveness of such an approach. The conclusion is that the distribution of NPs and their patterns are determined by NP length and that such relationships can be captured with an exponential model and the Nemcová and Serdelová model. The present research follows this path and intends to investigate into the NP structures and the relationships among length, frequency and complexity of NPs in maritime English, searching for mathematical models that can exactly describe such relationships.

2. Data and method

The Maritime English Corpus (MEC) was used as the data source in the research. The corpus contains safety at sea, shipping news, navigational and marine engineering technology, laws, rules and regulations and documents on all the related areas of maritime transportation. The format of the corpus is fixed field plain text, i.e., the first column, which is 14 characters in width, is the code for the origin, data of publication and author of the text chunk, while the rest is the corpus text without any coding. The following is a sample of MEC:

> 0062NE95070036 2.2 ET As for Inward Bound Ships:
> 0062NE95070037 Ships requiring the services of an authorised pilot at the
> 0062NE95070038 N.E. Spit, Sunk Warps or Gravesend Pilot Stations must make
> 0062NE95070039 a provisional notification of arrival to the Thames
> 0062NE95070040 Navigation Service, at boarding/landing station, GRT/GT,
> 0062NE95070041 length overall, draft and destination (name of berth or
> 0062NE95070042 anchorage).

The fixed field coding was first removed, keeping the text per se. The corpus was then syntactically parsed with *Standford Parser v.3.7.0* **(***https://nlp.stanford.edu /software/lex-parser.shtml***),** adding parts of speech tags and phrase and sentence structure trees. Table 1 is the *Standford Parser* syntactic tag sets:

Table 1
Major syntactic tags and the NP components they represent

ADJP	Adjective Phrase	UCP	Unlike Coordinated Phrase
ADVP	Adverb Phrase	VP	Verb Phrase
CONJP	Conjunction Phrase	WHADJP	Wh-adjective Phrase
FRAG	Fragment	WHAVP	Wh-adverb Phrase
INTJ	Interjection	WHNP	Wh-noun Phrase
LST	List marker	WHPP	Wh-prepositional Phrase
NAC	Not a Constituent	X	Unknown or uncertain
NP	Noun Phrase	CC	Coordinating conjunction
NX	Used within certain complex NPs to mark the head of the NP	CD	Cardinal Number
PP	Prepositional Phrase	DT	Determiner
PRN	Parenthetical	EX	Existential there
PRT	Particle	FW	Foreign word
QP	Quantifier Phrase	IN	Preposition or subordinating
RRC	Reduced Relative Clause	JJ	Adjective

JJR	Adjective, comparative	RP	Particle
JJS	Adjective, superlative	SYM	Symbol
LS	List item marker	TO	to
MD	Modal	UH	Interjection
NN	Noun, singular or mass	VB	Verb, base form
NNS	Noun, plural	VBD	Verb, past tense
NNP	Proper noun, singular	VBG	Verb, gerund or present participle
NNPS	Proper noun, plural	VBN	Verb, past participle
PDT	Predeterminer	VBP	Verb, non-3rd person singular present
POS	Possessive ending	VBZ	Verb, 3rd person singular present
PRP	Personal pronoun	WDT	Wh-determiner
PRP$	Possessive pronoun	WP	Wh-pronoun
RB	Adverb	WP$	Possessive wh-pronoun
RBR	Adverb, comparative	WRB	Wh-adverb
RBS	Adverb, superlative		

The following is an extract of the parsed corpus:

```
(ROOT
  (S
    (PP
      (ADVP
        (NP (CD 2.2) (NNS ET))
        (RB As))
      (IN for)
      (NP
        (NP (NNP Inward) (NNP Bound) (NNP Ships))
        (: :)
        (NP
          (NP (NNP Ships))
          (VP (VBG requiring)
            (NP
              (NP (DT the) (NNS services))
              (PP (IN of)
                (NP
                  (NP (DT an) (JJ authorised) (NN pilot))
                  (PP (IN at)
                    (NP (DT the) (NNP N.E.) (NNP Spit))))))))))
    (, ,)
    (NP
      (NP (NNP Sunk) (NNP Warps))
      (CC or)
      (NP (NNP Gravesend) (NNP Pilot) (NNPS Stations)))
    (VP (MD must)
      (VP (VB make)
        (NP
          (NP (DT a) (JJ provisional) (NN notification))
          (PP (IN of)
            (NP (NN arrival))))
        (PP (TO to)
          (NP (DT the) (NNP Thames) (NNP Navigation) (NNP Service)))
        (, ,)
        (PP (IN at)
          (S
            (VP (VBG boarding/landing)
              (NP
                (NP (NN station))
                (, ,)
                (NP (NNP GRT/GT))
                (, ,)
                (NP (NN length) (JJ overall))
                (, ,)
                (NP (NN draft))
                (CC and)
                (NP (NN destination))))))
        (PRN (-LRB- -LRB-)
          (NP
            (NP (NN name))
            (PP (IN of)
              (NP (NN berth)
                (CC or)
                (NN anchorage))))
          (-RRB- -RRB-))))
    (. .)))
```

The non-NPs were removed, keeping only the NPs, as shown below:

```
(NP (CD 2.2) (NNS ET))
      (NP
        (NP (NNP Inward) (NNP Bound) (NNP Ships))
        (: :)
        (NP
          (NP (NNP Ships))
          (VP (VBG requiring)
            (NP
              (NP (DT the) (NNS services))
              (PP (IN of)
                (NP
                  (NP (DT an) (JJ authorised) (NN pilot))
                  (PP (IN at)
                    (NP (DT the) (NNP N.E.) (NNP Spit)))))))))
    (NP
      (NP (NNP Sunk) (NNP Warps))
      (CC or)
        (NP
          (NP (DT a) (JJ provisional) (NN notification))
          (PP (IN of)
            (NP (NN arrival))))
        (PP (TO to)
          (NP (DT the) (NNP Thames) (NNP Navigation) (NNP Service)))
        (, ,)
        (PP (IN at)
          (S
            (VP (VBG boarding/landing)
              (NP
                (NP (NN station))
                (, ,)
                (NP (NNP GRT/GT))
                (, ,)
                (NP (NN length) (JJ overall))
                (, ,)
                (NP (NN draft))
                (CC and)
                (NP (NN destination))))))
        (PRN (-LRB- -LRB-)
          (NP
            (NP (NN name))
            (PP (IN of)
              (NP (NN berth)
                (CC or)
                (NN anchorage))))
```

Quirk et al. (1985, p. 1350) classify NPs into two types, simple NPs and complex NPs. The former refers to nouns without modification, e.g. *John*, *the man*, etc.; the latter are those with modifications. In this research, we also classify NPs into the forgoing two types, but because the ways *Standford Parser* parses a text, the respective structures of the two types of NPs in our research are different from those of Quirk et al. In this article, simple NPs are those with an NP head and non-nested modifiers, each occupying only one line, which contains an NP marker *(NP*, followed by the NP components, e.g.:

```
(NP (DT an) (JJ authorised) (NN pilot))
```

while complex NPs contain nested phrases, including sub-level NPs, and each new line stands for a nested phrase; such an NP is marked by *(NP* which occupies one line itself, followed by nested phrases, each occupying one line as well, with indentation of one or more spaces from the start of the complex NP marker, e.g.:

```
(NP
          (NP (DT the) (NNS services))
          (PP (IN of)
            (NP
              (NP (DT an) (JJ authorised) (NN pilot))
              (PP (IN at)
                (NP (DT the) (NNP N.E.) (NNP Spit))))))))))
```

The last step was extracting all the syntactic tags of the NPs, removing all the rest. Take *(NP (DT the) (NNP N.E.) (NNP Spit))))))))))* as an example, the result of the last step is *NP DT NNP NNP*. All the corpus handling and data processing were done with self-compiled computer programs in R 3.41.

In this study, simple NPs not only include the stand-alone ones but also those contained in complex NPs. Similar to Wang's research, NP length is computed according to the number of syntactic components of the phrase rather than the number of words. For simple NPs, the length is computed according to the number of parts of speech. For example, *CD JJ NNP NN NNS* is a simple NP whose length is 5, while the length of complex NPs is computed according to the number of nested phrases they have. For example, *PP NP NP PP NP NP PP NP NP NP PP NP CC NP NP PP NP* is a complex NP whose length is 17. The complexity is measured in terms of the number of different patterns in NPs of the same length. In the following example:

1. DT CD NN NN NN VBG NN
2. DT JJ JJ NN JJ NN NN
3. DT JJ NN NNP IN NNP NNP
4. DT JJ NNP NNP NNP NNP NNP
5. DT NN CD JJ JJS NN NN
6. DT NNP NNP NNP NNP NN CD
7. DT NNP NNP NNP NNP NN NN
8. DT NNP NNP NNP NNP NNP NNP
9. DT NNP NNPS NNP NNP NNP NN
10. JJ NN NN NN JJ NN NN
11. NN NN NN JJ NN NNP NNP
12. NN NN NN NNP NNP NNP NN
13. NN NN NNS NNS NNS NN NNS
14. NNP NNP JJ NN NNP NNP NNP
15. NNP NNP NNP NNP JJ NNP NNP
16. NNP NNP NNP NNP NNP NNP NNP
17. NNS NNS NNS NN NN NN NNS

These NPs have a length of 7 and consist of 17 different patterns, therefore the complexity of

NPs whose length is 7 have a complexity of 17. All the corpus handling and data processing were done with self-compiled computer programs in R 3.41.

3. Results

3.1. The relationship between length and frequency of simple NPs

The total number of simple NPs is 26,377. As shown in Table 2, the length and its corresponding frequency is in a reverse relationship. Simple NPs with only 2 syntactic components rank the highest, occurring 10,475 times, and those with 15 components occur only once.

Table 2
Simple NP length and its corresponding frequency

Length	Frequency
1	9125
2	10475
3	4761
4	1458
5	409
6	100
7	32
8	4
9	6
10	2
11	3
12	0
13	1
14	0
15	1

Wang (2012) used the following exponential regression model to describe the relationship between NP length and its corresponding frequency with a very good fit:

(1) $$N_{freq} = ae^{bN_{len}}$$

where N_{freq} is NP frequency, N_{len} NP length; *a* and *b* are model parameters. However, the fit of this model noticeably deviates from the simple NP data despite the relatively high R^2, as shown in Figure 1.

In a stimulating article, Altmann (1980) proposed that the longer a linguistic construct the shorter its constituents, which is now known as the Menzerath-Altmann law. He mathematically presented this theory with the following differential equation:

(2) $$\frac{y'}{y} = -c + \frac{b}{x}$$

which can be solved into the following:

(3) $$y = ax^{b}e^{-cx}$$

where *y* is the (mean) size of the immediate constituents, *x* is the size of the construct, and *a*, *b* and *c* are parameters which seem to depend mainly on the level of the units under investigation (Köhler 2012, p. 147). (3) is also known as the Menzerath-Altmann model. We used this model to describe the relationship between simple NP length and its corresponding frequency, with *y* being the frequencies and *x* the length, *a*, *b* and *c* are parameters. The result is excellent, with R^2 = 1, *a* = 7417, *b* = 2.095, *c* = 3.222.

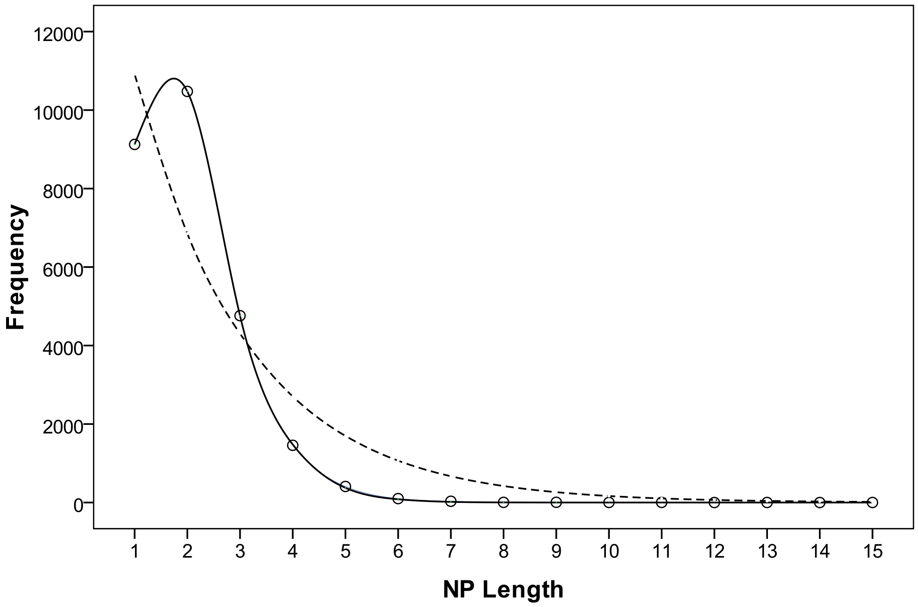

Figure 1. Model fit. The dotted line: the exponential model fit, R^2 = 0.876, *a* = 1737, *b* = −0.465; the solid line, the Menzerath-Altmann model fit, R^2 = 1, *a* = 7,417, *b* = 2.095, *c* = 3.222. The small circles: the observed values.

3.2. The complexity of simple NPs

The 26,377 simple NPs contain 805 different patterns. The complexity of simple NPs of different length is shown in Table 3. Simple NPs with length 4 ranks highest, with 237 different patterns, and the one with length 15 has only 1 pattern.

Table 3
Simple NP length and the corresponding complexity

Length	Complexity
1	27
2	87
3	200
4	237
5	161
6	62
7	17
8	4
9	4
10	2
11	2
13	1
15	1

Wang (2012) used the Nemcová and Serdelová (2005) model, which was intended to describe the relationship between the number of synonyms (*y*) of a word and the length of the word in syllables *x*:

(4) $$y = ax^b e^{cx} + 1 \, .$$

It is a special case of Wimmer & Altmann (2005). We used both (3) and (4) to fit the relationship between the simple NP length and its corresponding complexity. The two models provide the same fit, both R^2 being 0.987. For (3), *a* = 25.929, *b* = 7.853, *c* = 2.175; For (4), *a* = 25.086, *b* = 7.951, *c* = −2.201.

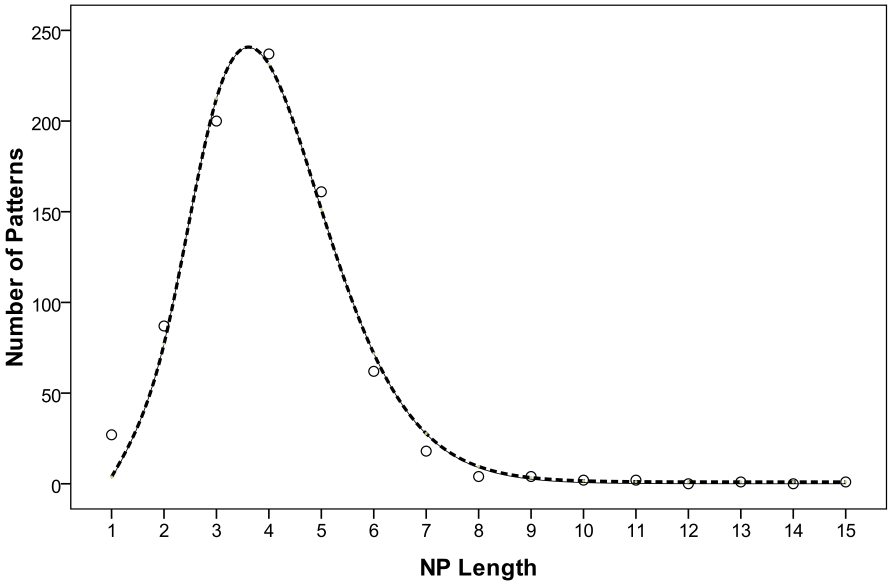

Figure 2 Length and complexity of simple NPs. The dotted line and solid lines are respectively model fit of (3) and (4), the small circles are the observed values.

3.3. The relationship between length and frequency of complex NPs

The total number of complex NPs is 5,865. The relationship between the length of complex NPs and its corresponding frequency is the same as that of simple NPs, i.e., the number of complex NPs decreases along with the increase of length. Complex NPs with only three syntactic components rank the highest, 2,007, and the one with 125 components occurs only once. Table 4 is the relationship between complex NP length and its corresponding frequency.

Table 4
Complex NP length and its corresponding frequency

Length	Freq	Length	Freq	Length	Freq	Length	Freq	Length	Freq
1	51	26	34	51	5	76	0	101	0
2	230	27	21	52	1	77	0	102	1
3	2007	28	16	53	0	78	0	103	0
4	299	29	25	54	2	79	0	104	0
5	408	30	13	55	0	80	0	105	0
6	547	31	14	56	1	81	0	106	0
7	267	32	10	57	0	82	0	107	1
8	239	33	15	58	1	83	0	108	0
9	279	34	10	59	1	84	0	109	0
10	174	35	8	60	0	85	0	110	1
11	138	36	10	61	2	86	0	111	0
12	152	37	5	62	1	87	0	112	0
13	113	38	5	63	1	88	0	113	0
14	101	39	1	64	1	89	0	114	0
15	100	40	4	65	0	90	0	115	1
16	78	41	5	66	2	91	0	116	0
17	73	42	3	67	0	92	0	117	0
18	71	43	3	68	0	93	1	118	0
19	44	44	2	69	1	94	0	119	0
20	66	45	0	70	0	95	0	120	0
21	37	46	4	71	0	96	1	121	0
22	52	47	7	72	1	97	0	122	0
23	34	48	1	73	0	98	0	123	0
24	29	49	4	74	0	99	0	124	0
25	22	50	6	75	0	100	1	125	1

Neither (1) nor (4) can fit the complex NP length and frequency data. (3) provides a reasonably good fit, with $R^2 = 0.828$, $a = 174.52561$, $b = 11.575243$ and $c = 33.830724$.

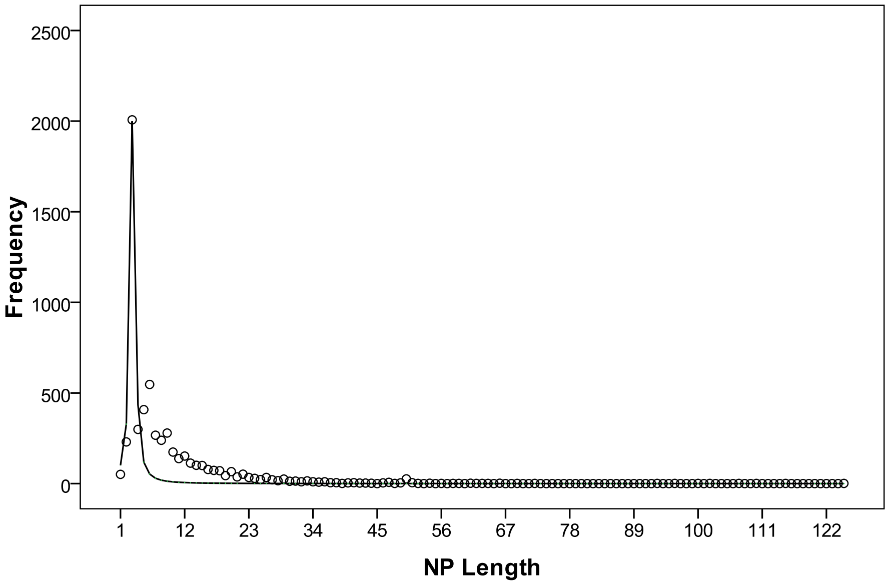

Figure 3. Model fit of (3).

The solid line is the model fit, the small circles the observed values.

3.4. The complexity of complex NPs

The 5,865 complex NPs contain 2,321 different patterns. The complexity of complex NPs of different length is shown in Table 5. Complex NPs with length 5 have 439 different patterns, and those with length 15 have only 2.

Table 5
Complex NP length and the corresponding complexity

Length	Complex	Length	Complex	Length	Complex	Length	Complex	Length	Complex
1	1	26	33	51	5	76	0	101	0
2	27	27	21	52	1	77	0	102	1
3	18	28	16	53	0	78	0	103	0
4	67	29	25	54	2	79	0	104	0
5	116	30	13	55	0	80	0	105	0
6	111	31	14	56	1	81	0	106	0
7	154	32	10	57	0	82	0	107	1
8	171	33	15	58	1	83	0	108	0
9	165	34	10	59	1	84	0	109	0
10	152	35	8	60	0	85	0	110	1
11	132	36	10	61	2	86	0	111	0
12	137	37	5	62	1	87	0	112	0
13	113	38	5	63	1	88	0	113	0
14	101	39	1	64	1	89	0	114	0
15	98	40	4	65	0	90	0	115	1
16	78	41	5	66	2	91	0	116	0

17	73	42	3	67	0	92	0	117	0
18	70	43	3	68	0	93	1	118	0
19	44	44	2	69	1	94	0	119	0
20	66	45	0	70	0	95	0	120	0
21	37	46	4	71	0	96	1	121	0
22	52	47	7	72	1	97	0	122	0
23	34	48	1	73	0	98	0	123	0
24	29	49	4	74	0	99	0	124	0
25	22	50	6	75	0	100	1	125	1

Neither (1) nor (4) can provide any fit for the complex NP length and complexity data. (3) captures such a relationship, with $R^2 = 0.9753$, $a = 6.1631$, $b = 0.2936$ and $c = 2.6618$. Figure 6 is the model fit.

4. Conclusion

This study reveals that, regardless of the types of NPs, the seemingly complex relationships among length, frequency and complexity of NPs in maritime English are highly regular. They all obey the Menzerath-Altmann law, as shown in the results of the good Menzerath-Altmann model fit to the data. This is only a pilot study in maritime English within the framework of quantitative and synergetic approach and the results are encouraging. Further work with this approach in other aspects of maritime English such as sentence structure, semantic and pragmatic characteristics etc. with such an approach needs to be carried out so as to better reveal the quantitative aspect of maritime English.

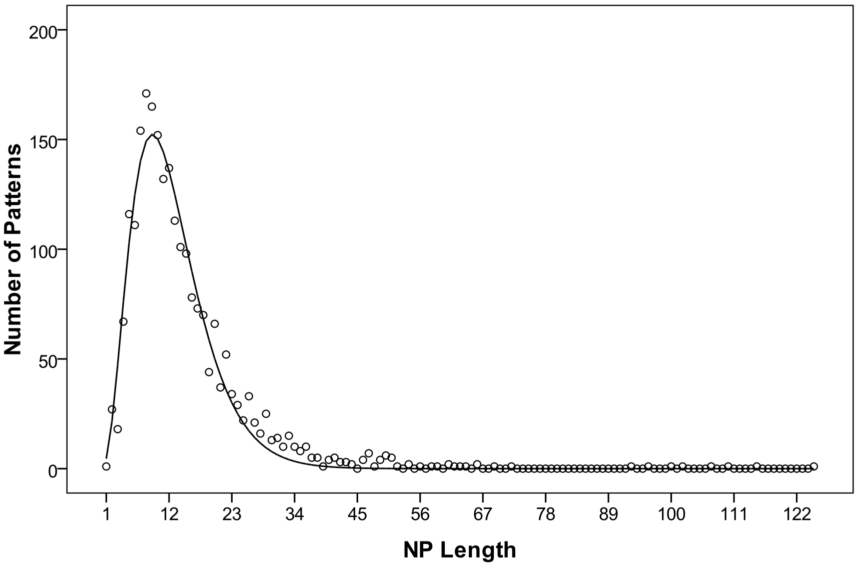

Figure 4. Length and complexity of complex NPs.
Solid line: model fit of (3), small circles: the observed value

References

Altmann, G. (1980). Prolegomena to Menzerath’s Law. In: *Glottometrika 2, 1-10*. Bochum: Brockmeyer.

Köhler, R. (2012). *Quantitative Syntax Analysis*. Mouton de Gruyter, Berlin/Boston.

Nemcová, E., Serdelová, K. (2005). On synonymy of Slovak. In: Altmann, G., Levickij, V. & Perebyinis, V. (eds.), *Problems of Quantitative Linguistics*: *194-209*. Chernivtsi: Ruta.

Quirk R., Sidney G., Geoffrey L., Jan S. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Wang, H. (2012), Length and complexity of NPs in Written English, *Glottometrics 24, 79-87*.

Wimmer, G., Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.). *Contributions to the science of language: Word length and related issues*: *93-117*. Boston: Kluver.

Glottometrics 40, 2018, 104-106

A Dependency Look at the Reality of Constituency

Xinying Chen[1],
Carlos Gómez-Rodríguez[2],
Ramon Ferrer-i-Cancho[3]

Abstract. A comment on "Neurophysiological dynamics of phrase-structure building during sentence processing" by Nelson et al (2017), Proceedings of the National Academy of Sciences USA 114(18), E3669-E3678.

Recently, Nelson et al. (2017) have addressed the fundamental problem of the neurophysiological support for complex syntactic operations of theoretical computational models. They interpret their compelling results as supporting the neural reality of phrase structure. Such a conclusion opens various questions.

First, constituency is not the only possible reality for the syntactic structure of sentences. An alternative is dependency, where the structure of a sentence is defined by word pairwise dependencies (Fig. 1). From that perspective, phrase structure is regarded as an epiphenomenon of word-word dependencies and constituency (in a classical sense as that of X-bar theory) has been argued to not exist (Mel'čuk, 2011). Furthermore, constituency may not be universal and thus its suitability may depend on the language (Evans & Levinson, 2009). Dependency is a stronger alternative for its simplicity, its close relationship with merge (Osborne, Putnam, & Gross, 2011), its compatibility with recent cognitive observations (Gómez-Rodríguez, 2016), its contribution to the cost of individual words even in isolation (Lester & Moscoso del Prado Martin, 2016) and its success over phrase structure in computational linguistics, where it has become predominant (Kübler, McDonald, & Nivre, 2009).

Second, the authors admit that a parser of the sentence might transiently conclude that "ten sad students"... is a phrase consistently with a transient decrease in activity (1st paragraph of p. 4). Unfortunately, their parser does not account for that as shown by the counts in Fig. 2 A of Nelson et al. (2017). In contrast, a standard dependency parser would, because at that point it would close the dependencies opened by "ten" and "sad" (Fig. 1). This raises the question of whether the conclusions depend on the choice X-bar and particular parser as a model of phrase structure. The conclusions by Nelson et al. (2017) may suffer from circularity, namely the positive support for a particular X-bar model could be due to the fact that the source was a toy artificial X-bar grammar. Future analyses would benefit from the use of natural sentences, sentences with realistic probabilities that are also longer and more complex (sentence length does not exceed 10 in Nelson et al. (2017)).

Third, dependency shows the limits of comparing phrase structure models against *n*-gram models with $n = 2$, because only about 50% of adjacent words are linked (Liu, 2008; Ferrer-i-

[1] Foreign Languages Research Center, School of Foreign Studies, Xi'an Jiaotong University, No.28 Xianning West Road, 710049 Xi'an, Shaanxi, P.R. China.

[2] Universidade da Coruña. FASTPARSE Lab, LyS Research Group. Departamento de Computación. Facultade de Informática, Elviña 15071 A Coruña, Spain

[3] Complexity & Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya, Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3, 08034 Barcelona, Catalonia (Spain). Corresponding author, rferrericancho@cs.upc.edu.

Cancho, 2004), thus a bigram model misses 50% of the dependencies. Bigrams are a weak baseline, as the common practice in computational linguistics is using at least smoothed trigram models, and often 5-gram models, to obtain meaningful predictions (Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016). A higher-order lexical *n*-gram model would strengthen the current results. The authors also employ more sophisticated *n*-gram models. One is an unbounded model based on part-of-speech categories, implying a dramatic loss of information with respect to the original words which might explain its poor performance. The other is a syntactic *n*-gram, but not enough information is provided about its definition and implementation. Regardless, since the model is obtained from a corpus derived from a toy grammar and lexicon, its probabilities are likely to be unrealistic and thus it is problematic.

In sum, dependency offers a better approach to the syntactic complexity of languages and merge. *n*-gram models of higher complexity should be the subject of future research involving realistic sentences.

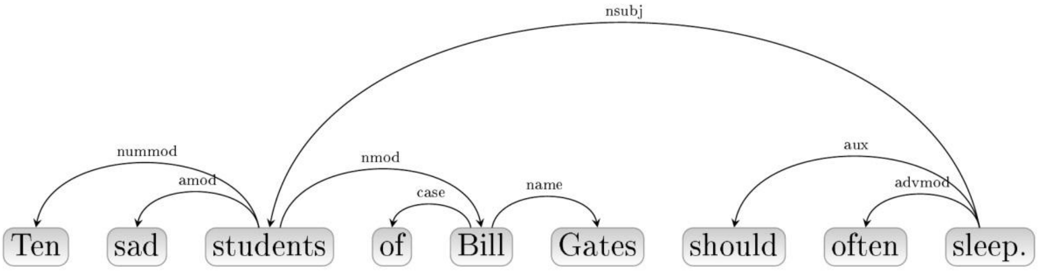

Figure 1: Syntactic dependency structure of the sentence in Fig 2 A of Nelson et al. (2017) according to Universal Dependencies (McDonald et al., 2013).

Acknowledgements

X.C. is supported by the Social Science Fund of Shaanxi State (2015K001). C.G.R is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 714150 FASTPARSE), and by the TELEPARES-UDC project (FFI2014-51978-C2-2-R) from MINECO (Ministerio de Economía y Competitividad). R.F.C is funded by the grants 2014SGR 890 (MACDA) from AGAUR (Generalitat de Catalunya) and the grant TIN2014-57226-P from MINECO.

References

Evans, N., & Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences 32, 429-492.*

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E 70, 056135.*

Gómez-Rodríguez, C. (2016). Natural language processing and the Now-or-Never bottleneck. *Behavioral and Brain Sciences 39, e74.*

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the limits of language modeling.* arXiv preprint arXiv:1602.02410.

Kübler, S., McDonald, R. & Nivre, J. (2009). *Dependency parsing*. Morgan and Claypool Publishers.

Lester, N. A. & Moscoso del Prado Martín, F. (2016). Syntactic flexibility in the noun: Evidence from picture naming. In: Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 2585-2590).* Austin, TX: Cognitive Science Society.

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

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K. Hall, K.B., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu, N. & Lee, J. (2013). Universal dependency annotation for multilingual parsing. *Proceedings of ACL (pp. 92-97).*

Mel'čuk, I. (2011). Dependency in language-2011. In: K. Gerdes, E. Hajicova, & L. Wanner (Eds.)*, Proceedings of the international conference on dependency linguistics, DepLing 2011, Barcelona, September 5-7, 2011.*

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C.P. & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences, 114 (18), E3669-E3678.*

Osborne, T., Putnam, M., & Gross, T. (2011). Bare phrase structure, label-less trees, and specifier-less syntax: Is minimalism becoming a dependency grammar? *The Linguistic Review 28, 315-364.*

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