Glottometrics 40, 2018, 1-45

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

Alexander Mehler, 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., prosummers 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 (Kittur et al. 2007b; Laniado 2011).
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 Wikipedia (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 Wikipedia – 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.

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 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. 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).

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).
Alexander Mehler, Rüdiger Gleim, Andy Lücking, Tolga Uslu and Christian Stegbauer

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).

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

---

See Arazy et al. (2011) for a related model of structure formation on two levels.
On the Self-similarity of Wikipedia Talks:
A Combined Discourse-analytical and Quantitative Approach

– called Level II –, which is established by (a mixture of) user(group)-related processes of 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 or-
ganization 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).
On the Self-similarity of Wikipedia Talks: a Combined Discourse-analytical and Quantitative Approach

\[ T_d = (V_d, A_d, r_d, \text{author}_d, \text{content}_d, \text{ord}_d, \text{signature}_d) \]  

- \( r_d \) is the root of \( T_d \) so that both can be used interchangeably to denote \( d \).
- \( V \) is partitioned into three non-overlapping subsets \( \{r_d\}, V^p_d = \{v_{i_1}, \ldots, v_{i_a}\} \) of vertices denoting posts and \( V^s_d = \{v_{i_{a+1}}, \ldots, v_{i_n}\} \) of vertices denoting sections. That is, we assume a bijection between posts in \( d \) and elements of \( V^p_d \) as well as between sections in \( d \) and elements of \( V^s_d \).
- 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 \( V^s_d \) which are processed as follows: (1) for each top-level section \( v \in V^s_d \), we generate arcs of the sort \((r_d, v) \in A\); (2) for all sections \( w \in V^s_d \) directly dominated by some \( v \in V^s_d \), we generate arcs of the sort \((v, w) \in A\); (3) for all posts \( w \in V^p_d \) directly dominated by \( v \in V^s_d \), we generate \((v, w) \in A\); (4) for all top-level posts \( w \in V^p_d \) 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 \( \text{outdegree}(v) = 0 \). \( \text{depth}(v) \) 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 \( i \)-th neighborhood of \( v \) in \( T_d \). \( N^p_i(v) = N_i(v) \cap V^p_d \) restricts this neighborhood to posts. For any \( i \in \mathbb{N} \), for which \( N_i(v) \neq \emptyset \), we call \( i \) the \( i \)-th 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 \( P = (v_{i_1}, \ldots, v_{i_j}, \ldots, v_{i_k}) \in \mathbb{P}(T_d) \), \( \text{in}(P) = v_{i_1} \) is called the start and \( \text{out}(P) = v_{i_k} \) 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, \( (v_{i_{j-1}}, v_{i_j}) \in P : (v_{i_{j-1}}, v_{i_j}) \in A \). \( \text{length}(P) \) 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 \( \text{out}(P) \in L(T_d) \). The set of all threads of \( T_d \) is denoted by \( \mathbb{T}(T_d) \subseteq \mathbb{P}(T_d) \). Finally, by \( \text{lcp}(v, w) \) 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 content \(_d\).\(^{11}\)
- \( \text{ord}_d \subseteq V^2_d \) 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

\[ \text{author}_d : V^p_d \to \mathcal{A}(D) = \mathcal{A} \cup \mathcal{I} \cup \{?\} \]  

\(^{10}\) 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.
mapping posts onto the names of their authors as part of their signature:

$$\text{signature}_d: V^p_d \rightarrow (\mathcal{A}(D), \mathbb{N})$$

Since this paper does not consider the temporal structure of Wikicussions we do not consider the time values of signatures.\(^{\text{12}}\) In the case of posts \(v \in V^p_d\) tagged by the proper name \(x\) and the time string \(y\), we set \(\text{signature}_d(v) = (x, y)\) and \(\text{author}_d(v) = x \in \mathcal{A}\) where \(y\) codes the time at which \(v\) was posted or last changed. If \(v\) is an unsigned post tagged by \(\text{signature}_d(v) = (x, y)\) 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 \(\text{author}_d(v) = x \in \mathcal{I}\). In the case of anonymous posts \(v\), for which we assume signatures of the form \(\text{signature}_d(v) = (?, ?)\), we set \(\text{author}_d(v) = ?\).

- Next, we apply this apparatus to Top-Level Sections (TLS) \(s \in S = \{ s \mid \exists T_d \in T_D : T_d = (V_d, A_d, \ldots) \land s \in V^s_d \land (r_d, s) \in A^s_d \}\). That is, by

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

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 (\(\sigma\)), semantic (\(\mu\)), pragmatic (\(\pi\)) and temporal (\(\tau\)) structure formation.\(^{\text{13}}\) To this end, we distinguish the syntactic, semantic, pragmatic and temporal order (as a function of \(V\)), depth (as a function

\(^{\text{12}}\) This will be the object of a follow-up paper.

\(^{\text{13}}\) 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:

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

(5)

For a given $x$, $\text{order}[x] : T_D \cup S_D \rightarrow \mathbb{R}, \ldots, \text{length}[x] : T_D \cup S_D \rightarrow \mathbb{R}$ 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) $\pi$-order of a discussion, its (syntactic) $\sigma$-depth or its (temporal) $\tau$-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

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

(6)
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. 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 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
On the Self-similarity of Wikipedia Talks: a Combined Discourse-analytical and Quantitative Approach

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).

<table>
<thead>
<tr>
<th></th>
<th>e-all</th>
<th>e-reg</th>
<th>e-ano</th>
<th>e-bots</th>
<th>p-all</th>
<th>p-reg</th>
<th>p-ano</th>
<th>p-bots</th>
</tr>
</thead>
<tbody>
<tr>
<td>e-all</td>
<td>1.00</td>
<td>0.98</td>
<td>0.89</td>
<td>0.84</td>
<td>0.95</td>
<td>0.97</td>
<td>0.78</td>
<td>0.12</td>
</tr>
<tr>
<td>e-reg</td>
<td>0.98</td>
<td>1.00</td>
<td>0.83</td>
<td>0.79</td>
<td>0.94</td>
<td>0.99</td>
<td>0.73</td>
<td>0.11</td>
</tr>
<tr>
<td>e-ano</td>
<td>0.89</td>
<td>0.83</td>
<td>1.00</td>
<td>0.69</td>
<td>0.85</td>
<td>0.84</td>
<td>0.83</td>
<td>0.14</td>
</tr>
<tr>
<td>e-bots</td>
<td>0.84</td>
<td>0.79</td>
<td>0.69</td>
<td>1.00</td>
<td>0.78</td>
<td>0.79</td>
<td>0.68</td>
<td>0.11</td>
</tr>
<tr>
<td>p-all</td>
<td>0.95</td>
<td>0.94</td>
<td>0.85</td>
<td>0.78</td>
<td>1.00</td>
<td>0.95</td>
<td>0.82</td>
<td>0.43</td>
</tr>
<tr>
<td>p-reg</td>
<td>0.97</td>
<td>0.99</td>
<td>0.84</td>
<td>0.79</td>
<td>0.95</td>
<td>1.00</td>
<td>0.73</td>
<td>0.12</td>
</tr>
<tr>
<td>p-ano</td>
<td>0.78</td>
<td>0.73</td>
<td>0.83</td>
<td>0.68</td>
<td>0.82</td>
<td>0.73</td>
<td>1.00</td>
<td>0.17</td>
</tr>
<tr>
<td>p-bots</td>
<td>0.12</td>
<td>0.11</td>
<td>0.14</td>
<td>0.11</td>
<td>0.43</td>
<td>0.12</td>
<td>0.17</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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 edits16, does not adequately consider archived pages17 and requires extensive processing on edit histories.

Unlike Ferschke et al., who analyze a small sample, we focus on all talk pages of the

---

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).
German Wikipedia including archived pages. That is, we doubt that the latter approach scales 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 \( _T(d) \) 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 \( _T(d) \) as follows:

\[
s(T_E(d), _T(d)) = 1 - \frac{\epsilon(T(d), _T(d))}{\max(|T(d)|, |T(d)|)}
\]  

\( \epsilon \) 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).
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.

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

instance of MarMoT, henceforth called gMarMoT, shows competitive results particularly in 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,

$$P(X_i \geq x_k) = \frac{|\{T_{d_j} \in T_D \mid X_i(T_{d_j}) \geq x_k\}|}{m}$$  \hspace{1cm} (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 $X_i \in F$:

$$P(X_i \geq x) \sim \beta x, x^{-\alpha x_i}, \alpha_{X_i}, \beta_{X_i} \in \mathbb{R}^+$$  \hspace{1cm} (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

$$r_{i,j} = k \text{ iff } X_i(d_j) \text{ is the } k\text{th smallest value of } X_i(D)$$  \hspace{1cm} (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

\(^{18}\) 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).
density, we compute the following statistics by analogy to the connection coefficient of Egghe and Rousseau (2003):

\[
density_{\alpha}(F) = \frac{2}{|F|(|F| - 1)} \sum_{X_i, X_j \in F, i \neq j, P_\alpha(\rho(X_i, X_j))} |\rho(X_i, X_j)|
\]  

(11)

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

**Hypothesis 3** For feature set \( F \), \( \density_{\alpha}(F) \gg 0 \).

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 \( \sigma\text{-order}(T_x) \), is the number of its vertices. Its syntactic depth, denoted by \( \sigma\text{-depth}(T_x) \), 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:

\[
h\text{-index}(T_x) = \max\{i \in \{0, \ldots, \text{depth}(T_x)\} \mid \forall 0 \leq j \leq i: |N_j(r_x)| - 1 \geq j\} \]

(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:

\[
\sigma\text{-width}(T_x) = \max\{|N_i(r_x)| \mid i = 0..\text{depth}(T_x)\}
\]

(13)

The syntactic level of minimal depth in \( T_x \), denoted by \( \sigma\text{-level}(T_x) \), is the smallest number \( i \), for which Expression 13 takes its maximum (Mehler 2011). Note that the levels \( j < i \) are “narrower” than \( \sigma\text{-width}(T_x) \) and therefore branch out afterwards, while levels larger than \( i \) either terminate or do not branch beyond \( \sigma\text{-width}(T_x) \). 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 \( \sigma\text{-level}(T_x) \) in a lawful manner. However, a correlation of \( \sigma\text{-level}(T_x) \) with other tree-related features of discussions (e.g., \( \sigma\text{-depth} \)) may hint at a law-like organization of discussions in terms of the formation of levels. The syntactic span of \( T_x \), denoted by \( \sigma\text{-span}(T_x) \), is the number of its child nodes. Finally, the syntactic length \( \sigma\text{-length}(T_x) \) 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 \( \sigma \)-order, \ldots, \( \sigma \)-length by restricting counting units to posts: \( \delta \)-order\( (T_x) \) is the number of posts in \( T_x \), \( \delta \)-depth\( (T_x) \) the maximum number of posts in threads, \( \delta \)-width\( (T_x) \) the size of the largest neighborhood of \( r_x \) consisting only of posts, \( \delta \)-level\( (T_x) \) the smallest \( i \) maximizing the latter quantity, \( \delta \)-span\( (T_x) \) the size of the lowest-level non-empty neighborhood \( N_i(r_x) \) containing at least one post and \( \delta \)-length\( (T_x) \) the number of all terminating posts.

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):

\[
\text{depend}(T_x) = 2\sum_{i=1}^{\text{depth}(T_x)+1} i \left\{ v \mid \delta(r_x, v) = i - 1 \right\} / |V_x|(|V_x| + 1) \in (0, 1) \quad (14)
\]

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:

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

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 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 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 \textit{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 \( T_x \):

\[
\text{coph}(T_x) = \left\{ \begin{array}{ll}
\sum_{v_i, w_j \in L(T_x), 1 \leq i < j \leq |L(T_x)|} \text{depth}(\text{lcp}(v_i, w_j)) & : l' > 3 \\
0 & : \text{else}
\end{array} \right. \quad \in [0, \binom{l'}{3}] \quad (16)
\]

Since we do not focus on phylogenetic trees, but consider more general trees, \( \text{coph}(T_x) \) is in the range of \( [0, \frac{l'}{3}] \), 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

\[
l' = l + k \quad (17)
\]
Rather than $\text{coph}(T_x)$, we calculate its normalized variant
\[
\text{imbal}(T_x) = \begin{cases} 
\frac{\text{coph}(T_x)}{|V_x|} : l' > 3 & \in [0, 1] \\
0 & \text{else}
\end{cases}
\]  
(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:
\[
\text{Wiener}(T_x) = 2 |V_x|(|V_x| - 1) \sum_{v_i,v_j \in V_x; i < j} \delta(v, w)
\]  
(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, $\sigma$-order, $\sigma$-depth, $\sigma$-level, $\sigma$-width, $\sigma$-span and $\sigma$-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. $\pi$-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. $\pi$-depth($T_x$) is the largest number of different authors contributing to the same thread:
\[
\pi\text{-depth}(T_x) = \max_{P \in \mathbb{V}(T_x)} \{ |\{\text{author}_x(v) \mid \exists v \in V_x^P(P)\}| \}
\]  
(20)

where $V_x^P(P)$ is the set of all posts of thread $P$.

3. $\pi$-width($T_x$) is the largest number of different participants posting on the same level:
\[
\pi\text{-width}(T_x) = \max_{i=1..\text{depth}(T_x)} \{ |\{\text{author}_x(v) \mid \exists v \in V_x^P : \text{depth}(v) = i\}| \}
\]  
(21)
4. \(\pi\)-level\((T_x)\) is the smallest number \(i\) maximizing the latter quantity.
5. By analogy to \(\sigma\)-span, \(\pi\)-span\((T_x)\) is the number of different participants posting on

\[
N_{\min}^p(r_x) = \arg \min_{|l| \neq \emptyset, i=1..\text{depth}(T_x)} \{l\}
\]

comprising at least one post, that is:

\[
\pi\text{-span}(T_x) = |\{\text{author}(v) | \exists v \in N_{\min}^p(r_x)\}|
\]

6. \(\pi\)-length\((T_x)\) is the number of authors of those posts that are leaves in \(T_x\):

\[
\pi\text{-length}(T_x) = |\{\text{author}(v) | \exists v \in L(T_x) \cap V_x^p\}|
\]

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

\[
\pi\text{-overlap}(T_x) = \frac{|\text{author}_x(T_x) \cap \text{author}_x(\text{article}(T_x))|}{|\text{author}_x(T_x) \cup \text{author}_x(\text{article}(T_x))|}
\]

where

\[
\text{author}_x(T_x) = \{\text{author}(v) | v \in V_x^p\}
\]

is the set of authors of \(T_x\) and \(\text{author}_x(\text{article}(T_x))\) is the set of authors of the corresponding article. We calculate \(\pi\)-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 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).

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

Henceforth, this selection is called \textit{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 \textit{Hauptkategorie} (main category) or in a short distance to the category \textit{Kategorie:Sachsystematik} (Category:Main topic classifications). Table 3 lists all categories selected in this way.

As a \textit{tertium comparationis} we consider a thematic partitioning of the discussion (and of the corresponding article) space called \textit{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 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: \texttt{Archäologie} (archeology) and \texttt{Sozialwissenschaft} (social science). This results in a list of 22 OCS-categories (see Figure 10). Based on these considerations our fourth hypothesis is:

\textbf{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\footnote{The $F$-score of a classification is the harmonic mean of its precision and recall.}), 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 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 ($\sigma$-depth), span broader discussions ($\sigma$-width), written by many different participants ($\pi$-order), initially ($\pi$-span) and finally ($\pi$-length). The majority of discussions exhibits the exact opposite of this
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.

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

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 (\( \pi \)-depth), authorship overlap (\( \pi \)-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

\(^{23}\) Power laws were fitted by means of MATLAB, log-normal distributions by means of R.
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).

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 $\{\pi\text{-order}, \sigma\text{-order}\}$ (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 $\pi$-level (see page 22) and $\sigma$-level (see page 20): if a discussion $d$ is, for example, broadest 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 $\pi$-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 $\pi$-depth and $\sigma$-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 $\sigma$-depth and $\pi$-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 ($\sigma$-order) is also broad ($\sigma$-width) and long ($\sigma$-length) and has a wide span ($\sigma$-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 $\min_{j} = 0.5$, 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 $\pi$-overlap is isolated. In order to raise cohesion, one has to reduce $\min_{j}$. 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 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 ($\rho = 0.05$). Note also that outliers are robust in being uncorrelated regarding the majority of features. $\pi$-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 $F\backslash \{\pi$-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

Figure 12. Log-log plots of the distributions of syntactic features of top-level sections (for the axes see Figure 11).
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: $\pi$-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, $\text{density}_{\rho_c}(F)$ equals average correlation (unlike in Figure 15 we consider all significant correlations). In the case of TRN, $\text{density}_{\rho_c} = 0.384$. 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.

### 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 Wikiscussions 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 23), 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 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 Wikiscussions and of 0.536 when regarding all discussions (Figure 16). In this scenario, Wikiscussions 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 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

---

Figure 14. Log-log plots of pragmatic features of top-level sections (for the axes see Figure 11).
Table 4

Syntactic features of Wikicussions (left) and their top-level sections (right), the exponent $\alpha$ 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.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>$\alpha$</th>
<th>$pl$</th>
<th>$l-n$</th>
<th></th>
<th>Quantity</th>
<th>$\alpha$</th>
<th>$pl$</th>
<th>$l-n$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACI</td>
<td>0.72</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>ACI</td>
<td>0.79</td>
<td>0.06</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>ADI</td>
<td>1.47</td>
<td>0.11</td>
<td>0.00</td>
<td>0.00</td>
<td>ADI</td>
<td>1.24</td>
<td>0.11</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>AWI</td>
<td>1.15</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>AWI</td>
<td>2.56</td>
<td>0.07</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>depend</td>
<td>0.32</td>
<td>0.32</td>
<td>0.00</td>
<td>0.00</td>
<td>depend</td>
<td>0.08</td>
<td>0.34</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>$h$-balance</td>
<td>1.97</td>
<td>0.28</td>
<td>0.00</td>
<td>0.00</td>
<td>$h$-balance</td>
<td>1.49</td>
<td>0.25</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>$h$-index</td>
<td>3.45</td>
<td>0.09</td>
<td>0.99</td>
<td>0.99</td>
<td>$h$-index</td>
<td>4.74</td>
<td>0.08</td>
<td>1.00</td>
<td>19</td>
</tr>
<tr>
<td>imbal</td>
<td>7.01</td>
<td>0.34</td>
<td>0.00</td>
<td>0.00</td>
<td>imbal</td>
<td>2.23</td>
<td>0.33</td>
<td>0.00</td>
<td>21</td>
</tr>
<tr>
<td>resolution</td>
<td>7.89</td>
<td>0.20</td>
<td>0.00</td>
<td>0.00</td>
<td>resolution</td>
<td>6.74</td>
<td>0.17</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>$\sigma$-depth</td>
<td>3.03</td>
<td>0.06</td>
<td>1.00</td>
<td>1.00</td>
<td>$\sigma$-depth</td>
<td>2.75</td>
<td>0.06</td>
<td>1.00</td>
<td>19</td>
</tr>
<tr>
<td>$\sigma$-depth</td>
<td>2.52</td>
<td>0.06</td>
<td>1.00</td>
<td>1.00</td>
<td>$\sigma$-depth</td>
<td>2.77</td>
<td>0.06</td>
<td>1.00</td>
<td>20</td>
</tr>
<tr>
<td>$\sigma$-length</td>
<td>1.34</td>
<td>0.05</td>
<td>0.08</td>
<td>0.00</td>
<td>$\sigma$-length</td>
<td>2.08</td>
<td>0.07</td>
<td>0.20</td>
<td>19</td>
</tr>
<tr>
<td>$\sigma$-length</td>
<td>1.37</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>$\sigma$-length</td>
<td>2.20</td>
<td>0.08</td>
<td>0.13</td>
<td>20</td>
</tr>
<tr>
<td>$\sigma$-level</td>
<td>8.22</td>
<td>0.13</td>
<td>0.88</td>
<td>0.88</td>
<td>$\sigma$-level</td>
<td>4.44</td>
<td>0.13</td>
<td>0.78</td>
<td>19</td>
</tr>
<tr>
<td>$\sigma$-level</td>
<td>8.11</td>
<td>0.14</td>
<td>0.85</td>
<td>0.85</td>
<td>$\sigma$-level</td>
<td>4.48</td>
<td>0.13</td>
<td>0.79</td>
<td>20</td>
</tr>
<tr>
<td>$\sigma$-order</td>
<td>1.48</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>$\sigma$-order</td>
<td>2.02</td>
<td>0.08</td>
<td>0.02</td>
<td>19</td>
</tr>
<tr>
<td>$\sigma$-order</td>
<td>1.46</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>$\sigma$-order</td>
<td>1.90</td>
<td>0.08</td>
<td>0.02</td>
<td>20</td>
</tr>
<tr>
<td>$\sigma$-span</td>
<td>1.72</td>
<td>0.06</td>
<td>0.26</td>
<td>0.26</td>
<td>$\sigma$-span</td>
<td>3.26</td>
<td>0.13</td>
<td>0.10</td>
<td>19</td>
</tr>
<tr>
<td>$\sigma$-span</td>
<td>1.54</td>
<td>0.06</td>
<td>0.14</td>
<td>0.14</td>
<td>$\sigma$-span</td>
<td>2.72</td>
<td>0.10</td>
<td>0.14</td>
<td>20</td>
</tr>
<tr>
<td>$\sigma$-width</td>
<td>1.47</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
<td>$\sigma$-width</td>
<td>2.24</td>
<td>0.07</td>
<td>0.33</td>
<td>19</td>
</tr>
<tr>
<td>$\sigma$-width</td>
<td>1.47</td>
<td>0.06</td>
<td>0.08</td>
<td>0.08</td>
<td>$\sigma$-width</td>
<td>2.45</td>
<td>0.09</td>
<td>0.18</td>
<td>20</td>
</tr>
<tr>
<td>stratum</td>
<td>1.43</td>
<td>0.28</td>
<td>0.00</td>
<td>0.00</td>
<td>stratum</td>
<td>0.29</td>
<td>0.29</td>
<td>0.00</td>
<td>20</td>
</tr>
<tr>
<td>Wiener</td>
<td>4.54</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>Wiener</td>
<td>4.19</td>
<td>0.06</td>
<td>0.00</td>
<td>21</td>
</tr>
</tbody>
</table>

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 23). 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 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) ($A(D)$ is
On the Self-similarity of Wikipedia Talks:  
a Combined Discourse-analytical and Quantitative Approach

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

<table>
<thead>
<tr>
<th>Quantity</th>
<th>α</th>
<th>pl</th>
<th>l-n</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>π-depth</td>
<td>-4.39</td>
<td>0.97</td>
<td>0.09 (0.98)</td>
<td></td>
</tr>
<tr>
<td>π-length</td>
<td>-1.80</td>
<td>1.00</td>
<td>0.06 (0.16)</td>
<td></td>
</tr>
<tr>
<td>π-level</td>
<td>-7.87</td>
<td>0.94</td>
<td>0.15 (0.81)</td>
<td></td>
</tr>
<tr>
<td>π-order</td>
<td>-1.38</td>
<td>1.00</td>
<td>0.05 (0.07)</td>
<td></td>
</tr>
<tr>
<td>π-overlap</td>
<td>-23.72</td>
<td>1.00</td>
<td>0.14 (0.00)</td>
<td></td>
</tr>
<tr>
<td>π-span</td>
<td>-1.83</td>
<td>1.00</td>
<td>0.05 (0.38)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantity</th>
<th>α</th>
<th>pl</th>
<th>l-n</th>
<th>p.</th>
</tr>
</thead>
<tbody>
<tr>
<td>π-depth</td>
<td>-4.86</td>
<td>0.97</td>
<td>0.09 (0.98)</td>
<td>21</td>
</tr>
<tr>
<td>π-length</td>
<td>-2.54</td>
<td>1.00</td>
<td>0.08 (0.32)</td>
<td>22</td>
</tr>
<tr>
<td>π-level</td>
<td>-4.67</td>
<td>0.96</td>
<td>0.11 (0.91)</td>
<td>22</td>
</tr>
<tr>
<td>π-order</td>
<td>-2.18</td>
<td>1.00</td>
<td>0.08 (0.13)</td>
<td>21</td>
</tr>
<tr>
<td>π-overlap</td>
<td>-52.81</td>
<td>0.99</td>
<td>0.21 (0.00)</td>
<td>22</td>
</tr>
<tr>
<td>π-span</td>
<td>-2.98</td>
<td>0.99</td>
<td>0.10 (0.29)</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 6  
Average correlation values 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).

<table>
<thead>
<tr>
<th>Scope</th>
<th>#Features</th>
<th>#Pairings</th>
<th>density ( p_\alpha (F) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features of Wikicussions</td>
<td>29</td>
<td>406</td>
<td>0.398</td>
</tr>
<tr>
<td>syntactic</td>
<td>22</td>
<td>231</td>
<td>0.337</td>
</tr>
<tr>
<td>pragmatic</td>
<td>7</td>
<td>21</td>
<td>0.456</td>
</tr>
<tr>
<td>syntactic ↔ pragmatic</td>
<td>29</td>
<td>154</td>
<td>0.404</td>
</tr>
<tr>
<td>Features of top-level sections</td>
<td>29</td>
<td>406</td>
<td>0.384</td>
</tr>
<tr>
<td>syntactic</td>
<td>22</td>
<td>231</td>
<td>0.331</td>
</tr>
<tr>
<td>pragmatic</td>
<td>7</td>
<td>21</td>
<td>0.435</td>
</tr>
<tr>
<td>syntactic ↔ pragmatic</td>
<td>29</td>
<td>154</td>
<td>0.388</td>
</tr>
</tbody>
</table>

the set of author IDs of posters contributing to Wikicussions in corpus \( D \) – see Expression 2):

\[
\forall A, B \in OCS: \quad J_{\alpha}(A, B) = \frac{\sum_{x \in A(D)} \mu_{A \cap B}(x)}{\sum_{x \in A(D)} \mu_{A \cup B}(x)} \tag{27}
\]

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.

Table 7  
Parameter setting of the fastText-based classifier.

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Parameter Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>fastText (Joulin et al. 2016)</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.05</td>
</tr>
<tr>
<td>Size of hidden layer</td>
<td>100</td>
</tr>
<tr>
<td>Size of context window</td>
<td>5</td>
</tr>
<tr>
<td>Lower bound of word occurrences</td>
<td>1</td>
</tr>
<tr>
<td>Max length of word n-gram</td>
<td>3</td>
</tr>
</tbody>
</table>
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).

of 29%. Comparing the vectors of random overlaps with those being observed using a $t$-test, 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 to observe examples increases according to a power-law when syntactic and pragmatic variety
On the Self-similarity of Wikipedia Talks:
a Combined Discourse-analytical and Quantitative Approach

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

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Parameter Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>feedforward neural network</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.05</td>
</tr>
<tr>
<td>Number of Epochs</td>
<td>500</td>
</tr>
<tr>
<td>Number of hidden layers</td>
<td>1</td>
</tr>
<tr>
<td>Size of hidden layer</td>
<td>100</td>
</tr>
<tr>
<td>Lower bound of word frequency</td>
<td>1</td>
</tr>
</tbody>
</table>

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

decreasesimultaneously. 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

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

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.

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 differ-
ences: 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


