UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

On the relationship between the cognitive and the communal : A complex systems perspective

Vetchinnikova, Svetlana

Filppula, Markku; Klemola, Juhani; Mauranen, Anna; Vetchinnikova, Svetlana

2017

http://hdl.handle.net/10138/313288

Vetchinnikova, S 2017, On the relationship between the cognitive and the communal : A complex systems perspective. in M Filppula, J Klemola, A Mauranen & S Vetchinnikova (eds), Changing English : Global and local perspectives. Topics in English Linguistics [TiEL], vol. 92, De Gruyter Mouton, Berlin, Boston, pp. 277–310. https://doi.org/10.1515/9783110429657-015

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

Svetlana Vetchinnikova

On the relationship between the cognitive and the communal: a complex systems perspective

Abstract: This paper presents a specific take on the relationship between the global and the local in language. In particular, it draws a distinction between the cognitive and the communal plane of language representation and attempts to model the relationship between the two using complexity theory. To operationalise this relationship and examine it with corpus linguistic methods, it proposes a concept of a cognitive corpus, setting it against the more usual idea of a corpus as representing the language of a certain community of speakers. As a case study, the paper compares the properties of chunking at the cognitive and communal planes. The study shows that (1) chunks at the cognitive plane seem to be more fixed than at the communal, (2) their patterning at the communal plane can be seen as emergent from the patterning observable in individual languages, but that (3) there is also similarity in the shape of the patterning across the two planes. These findings suggest that although the processes leading to multi-word unit patterning are different at each of the planes, the similarity in the shape the patterning takes might be regarded as an indication of the fractal structure of language which is a common property of complex adaptive systems. For example, Zipf's law, which is able to model the patterning at each of the planes, can be seen as one of the symptoms of such structure. Since the cognitive and the communal planes of language are in constant interaction with each other, such conceptualisation suggests intriguing implications for ongoing change in English and the role second language users might play in it.

Keywords: cognitive corpus, individual language, lexical bundles, chunking, complex adaptive system, fractal, Zipf's law, emergence, fixing, unit of meaning, approximation, *it is* ADJ *that*, second language use, ELF

> [A fractal is] a rough or fragmented geometric shape that can be split into parts, each of which is (at least approximately) a reduced-size copy of the whole.
> Mandelbrot 1982: 34

Svetlana Vetchinnikova, University of Helsinki

DOI 10.1515/9783110429657-015

Copyright 2017. De Gruyter Mouton.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

EBSCO Publishing : eBook Collection (EBSCOhost) - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI
AN: 1615781 ; Filppula, Markku, Klemola, Juhani, Mauranen, Anna, Vetchinnikova, Svetlana.; Changing English : Global and Local Perspectives
Account: s5205063.main.ehost

1 Introduction

There is a number of ways *language* can be split into parts. If language behaves as a complex adaptive system (CAS), then in each case (1) a part will be a reduced-size copy of the whole, a fractal and (2) the relationship between the parts and the whole will be characterised by emergence. This should also apply to the distinction between the global and the local, focused on in this volume. I will now introduce this hypothesis in more detail.

The CAS approach to language modelling (see notably Larsen-Freeman 1997; Ellis and Larsen-Freeman 2006; Ellis 2011; Larsen-Freeman and Cameron 2008; de Bot et al. 2007) has developed from a usage-based, emergentist theoretical orientation (Hopper 1987; Bybee and Hopper 2001; Tomasello 2003). In this paradigm, the understanding of how language works is in stark contrast to the traditional view where grammar rules are seen to work top-down rather than merely be a description of what happens bottom-up. Conversely, language structure is seen to emerge dynamically in usage and to be shaped by "interrelated patterns of experience, social interaction, and cognitive processes" (Beckner et al. 2009: 2). This view entails an intricate relationship between every individual and every language event and language change at the global level. On the one hand, as Larsen-Freeman (1997) points out:

> [There is] no distinction between current use and change/growth, they are isomorphic processes. Every time language is used, it changes. As I write this sentence, and as you read it, we are changing English. [. . .] as the user's grammar is changed, this sets in motion a process, which may lead to change at the global level. (Larsen-Freeman 1997: 148)

At the same time, every instance of language use, necessarily performed by an individual, is itself a product of interaction of different forces, such as the user's cognitive processes, previous experience of language and social motivations. As a result, we have a multiply embedded system with a complex interrelationship between the communal and the individual. As Beckner et al. (2009: 15) observe "an idiolect is emergent from an individual's language use through social interactions with other individuals in the communal language, whereas a communal language is emergent as the result of the interaction of the idiolects".

Here, the concept of *emergence* takes on a new meaning. Hopper (2011) distinguishes between *emerging* and *emergent* grammar, pointing out that what is commonly foregrounded by the former is inquiry into the historical origins of present-day grammar and, as a consequence, its conceptualisation as a stable system which *has emerged*, while the latter sees grammar as always temporary, ephemeral and provisional. While forming a fundamental background, neither

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

of the two interpretations captures the specifics of emergence conceptualised as a property of a CAS (see e.g. Ellis and Larsen-Freeman 2006). A CAS is a product of interaction of different types of elements where each element is itself a product of interaction of even smaller elements. In other words, we see multiply embedded complex systems at different levels of abstraction. It is the relationship between the levels that can be characterised by emergence. At every level, a CAS *emerges* from the interaction of the elements at a lower level, that is, it does not equal the sum of the elements but arises from their interaction, and thus can have properties which are not present at the lower level. In this way, a traffic jam is not the property of an automobile but emerges in the interaction of many automobiles, or, to be more precise, their trajectories (de Bot and Larsen-Freeman 2013: 17).

At the same time, another common property of complex adaptive systems, namely, fractal structure, also called scale-free self similarity, predicts that an element or component part participating in the interaction at the lower level will have the same shape as the system as a whole. In other words, multiply embedded complex systems located at different levels or scales of a certain dimension will be self-similar. For example, Mandelbrot (1963) examined variation in cotton prices over short and long time periods and found that the pattern of change was similar regardless of the scale. Language appears to be an entity which is particularly suitable for fractal analysis since we routinely talk about langauges at different levels of abstraction, such as English language, British language, academic language, newspaper language, Early Modern English language, child language, Hip Hop nation language, individual language.

In principle, there seems to be three major dimensions along which one can split language into parts: (1) across different groups of speakers, from individuals and discourse communities to nations and global networks, (2) across different levels of language organisation, such as phonological, morphological, lexico-grammatical, discoursal and (3) across different time-scales . We can tentatively refer to them as 'social', 'structural' and 'temporal' dimensions (cf. e.g. Ellis 2006; de Bot et al. 2013). The first two can also be thought of as breadth and depth, to borrow the terms from vocabulary studies (Anderson and Freebody 1981; Read 2004). If the hypothesis about the fractal structure of language is correct, we should be able to see inherent similarity in shape whether we zoom in or zoom out along each of the dimensions. Thus, if we take the temporal dimension, for example, short-term and long-term changes in language can exhibit similarity in their patterning. At the same time, the temporal dimension can be kept separate as there is no need to impose it on either the social or the structural dimension: the behaviour of complex systems on these two dimensions, including the property of emergence which might not be so intuitively obvious, can be studied from a synchronic point of view.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

The contrast of the global and the local of this volume is then situated on the social dimension, that of breadth or spread. Thus, we can think of English as comprised of different varieties, dialects and registers (e.g. see Mair this volume for a suggestion of a currently relevant taxonomy for English). In this paper, I am drawing a distinction between the individual and the communal plane of language representation which is another way of breaking up language on the social dimension. In principle, the properties of the relationship between these two planes should be applicable to other divisions into planes on the social dimension. Here, I equate the individual with the cognitive for reasons I discuss in Section 3. I will refer to the levels of the social dimension as planes to set them apart from the time-scales of the temporal dimension and the levels of the structural dimension.

In what follows, I first further explain in which way the communal plane of language representation can be seen as emergent from the interaction of idiolects by giving two examples of observed language patterning. It must be mentioned that while complex system properties must apply to language representation across different levels of its organisation, this paper contextualises the model by focusing on the type of patterning which is often referred to as 'phraseological' (see Section 4 for an outline of this approach). Thus, the examples examined in Section 2 directly relate to the case study described in Sections 4, 5 and 6. But before that, in Section 3, I argue for the feasibility of making a connection between the individual and the cognitive and suggest a concept of a cognitive corpus. Then, in Section 4, I move on to exploring the relationship between the individual and the communal by looking at the properties of chunking at each of them and describe the data I am going to use for this purpose. In Section 5, I focus on the variation in the construction *it is* ADJ *that* at each of the planes to see whether the relationship between the two can be described as emergent and fractal, in line with CAS predictions. In Section 6, I take a more quantitative approach to inspect whether there is further support for the chunking differences observed in Section 5. In the final section, I summarise the findings and relate them to research on grammaticalization and ongoing change in English as possible avenues for further research.

2 Emergence of the communal from the cognitive

In this section, I will show in which way language representation at the communal plane can be seen as emergent from the cognitive and what implications this might have for understanding the mechanisms underlying phraseological patterning observable at each of the planes. To do this, I will take two linguistic

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

units, a more lexical and a more grammatical one, as my examples and see how they can be represented at each of the planes.

I will start with Sinclair's unit of meaning (Sinclair 1996, 2004), a form-meaning pairing which, in contrast to many other conceptualisations of meaningful units, allows for fixed as well as variable components. Its fixed components are obligatory: the core, the most invariable formal element, and the semantic prosody, which is defined here as the communicative purpose of a unit (Vetchinnikova 2014; see also Hunston 2007 who defines semantic prosody as a discourse function of a unit of meaning). The variable components are optional: they are collocation, colligation and semantic preference. Collocation is a verbatim association between two or more words. In contrast, colligation and semantic preference are abstracted associations: association with a grammatical feature and association with a semantic set respectively. A well-known example of a unit of meaning is the case of *naked eye* (Sinclair 1996). *Naked eye* itself is a collocation because it is a verbatim co-occurrence of two words. It also has a semantic preference for words from the semantic set of 'visibility', like *seen*, *discernible*, *visible*, rather than collocates with just one of them, and colligates with the class of prepositions, including *by*, *with* or *via*, again rather than collocates with just one of them. The whole patterning associates with the communicative purpose of saying that something is difficult to see, which is the semantic prosody of this unit.

The patterning of a unit of meaning was first revealed through corpus observations of language (Sinclair 1996), i.e. at the communal plane. But there is evidence for its psycholinguistic reality too (Vetchinnikova 2014). Yet, the fact that the unit of meaning seems to be represented at both planes, cognitive and communal, does not yet mean that the processes leading to its emergence at each of the planes are the same.

At the cognitive plane, the existence of colligation and semantic preference can be explained by effects of frequency on entrenchment and the properties of human memory, which is stronger for meaning than for linguistic form (Bock and Brewer 1974; Gurevich et al. 2010). As such, if a sequence or a certain component of it has not been frequent enough (or the type/token ratio is tipped towards the higher type frequency) in the language experience of a user, its representation in memory is abstracted in grammatical or semantic terms rather than is verbatim. Thus, a language user might associate the verb *undergo* with a semantic set of words meaning some kind of 'change' (*surgery*, *transformation*, *change*, *treatment*, *operation*) rather than with a specific word from this set. This abstracted representation would be explained by the cognitive reality of semantic preference. At the same time it is possible for a language user to associate *undergo* with e.g. *change* in particular due to his/her specific experience of

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

language, i.e. have the unit represented as a collocation in memory. Over time, due to continuous change in the experience of language, these associations can also change: become more fixed, from abstracted to verbatim or collocational, and loosen, abstract from a collocation to a semantic preference or a colligation. These reverse processes have been called *fixing* (Vetchinnikova 2014) and *approximation* (Mauranen 2012).

At the communal plane, the picture is a bit different. When we examine the patterning of the verb *undergo* and find that it co-occurs with words belonging to the same semantic set, such as *surgery*, *change*, *transformation*, *operation*, *endoscopy* (BNC) and therefore can postulate the category of semantic preference for it, it does not yet mean that this semantic preference is valid for all language speakers, or for all native speakers or for all speakers of British English. It is just as well possible that the category of semantic preference we observe in a corpus is a result of averaging[1] across speakers: that is, each of the speakers represented in the BNC might have his/her own collocational preference which together look like a semantic preference. Therefore we can say that semantic preference exists at both planes but is driven by different mechanisms.

Let us imagine for a moment that due to certain socio-economic developments *undergo change* becomes a fixed expression at the communal plane, say, to refer to restructuring of an organisation due to severe cuts in budget.[2] Then the processes of fixing at the individual plane and conventionalisation at the communal plane can also be seen as similar, even though again the mechanisms underlying them are different. It seems that this similarity across different planes can be regarded as an example of scale-free self-similarity or fractal scaling (Mandelbrot 1982; Gleick 1987), mentioned at the outset of this paper. Fractal scaling is also another common property of complex systems. I will come back to this idea in the analysis of data in Section 5.

But let me give another example of how emergence might be conceptualised. Mollin (2009a) studies the distribution of maximiser adverbs, such as *absolutely*, *completely*, *entirely* and *totally* in a three-million corpus of Tony Blair's public speeches and finds a clear preference for specific combinations, for example,

1 See Larsen-Freeman 2013 for the discussion of the problem of averaging and thus abstracting away from variability in research on language development, including second language research.
2 It is important to mention at this point that in Sinclair's conceptualisation of lexis and meaning, when a combination of words starts to be treated as a unit, i.e. on the idiom principle, it always means a change in meaning, a meaning shift, however small (e.g. Sinclair 2004; Cheng et al. 2009). This view is not dissimilar to views expressed in the study of grammaticalization in relation to the mechanisms underlying language change which suggests a possibility for cross-fertilisation between these theoretical frameworks, but which would not be further discussed here for reasons of space.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

completely unacceptable, *entirely understand* or *absolutely blunt*. Such preferences sometimes align with the BNC collocational patterns, sometimes do not and sometimes are clearly "Blairisms". To me, this might mean that the grammatical category of maximiser adverbs as an abstraction of regularities in language patterning emerges only at the communal plane, when we aggregate different speakers and different discourses and average across them. To put it in other words, many aspects of linguistic structure might be an emergent property of language at the communal plane and might not be present at the individual level.

If this hypothesis is correct, it might help to explain often conflicting findings of the studies examining the psycholinguistic reality of co-occurrence patterns observed in corpora, an issue which has recently drawn attention of many scholars (e.g. Hoey 2005; Mollin 2009b; Ellis and Frey 2009; Ellis et al. 2009; Durrant and Doherty 2010). The solution it suggests is that corpus-linguistically attested patterns do not necessarily have to be psycholinguistically real in order to be valid observations of language at the communal level. It is possible, and plausible, that the patterning at the two levels is different, in a qualitative way.

3 A *cognitive* corpus

In many ways, the distinction between the individual and communal levels of language is not new. For example, in the study of language contact we find the distinction between transfer at the level of the society and at the level of individual already in Weinreich (1953). Mauranen (2012 and this volume) adds another, microsocial level, and builds a framework based on three interrelated levels: the societal or macrosocial, the individual or cognitive and the level of social interaction between speakers or the microsocial. Both Weinreich (1953) and Mauranen (2012) make an explicit connection between the individual and the cognitive, after all cognition is the property of an individual. But can we make the same connection in corpus linguistic research and treat a corpus of an individual's language use as his/her cognitive corpus, that is, a corpus enabling observation of individual's cognitive processes and representations? Let us have a look at what research on idiolects can tell us.

The study of idiolects is not a particularly popular topic in almost any field of linguistics, mostly for reasons of limited generalisability as it seems, but not a non-existent one. Idiolectal preferences have been studied in forensic linguistics for the purposes of authorship identification (Coulthard 2004; Wright 2015) and in a few other corpus linguistic studies, such as Mollin (2009a) and Barlow (2013). All these studies come to the conclusion that idiolectal preferences are clearly identifiable and are able to distinguish an individual from other language

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

speakers or from the "communal average".[3] In a way these studies, and especially Barlow (2013), continue in the tradition of studies on language variation and suggest that there is no reason to stop at the already well acknowledged fact of register variation: idiolectal variation seems to be just as palpable.

The case of idiolects has also been brought to attention in historical sociolinguistic studies of grammaticalization, but in a bit different way (see Raumolin-Brunberg and Nurmi 2011 for a review). These studies are interested in the role of the individual in language change. Taking a historical perspective and using the benefit of hindsight, they examine in which way language changes attested at the communal plane reflect in language use of specific individuals across their life span. What they find is "a great deal of variation between individuals concerning their participation in ongoing linguistic changes" (Raumolin-Brunberg and Nurmi 2011: 262). In fact, Raumolin-Brunberg and Nurmi observe that "[t]he patterns that arise from studies of large groups of people do not necessarily surface in the language of individuals" (262). This conclusion makes one think whether what we see is mere variation or whether it is possible that the relationship between the individual and communal is not straightforward but rather can indeed be described by emergence.

One thing this brief glance at research on idiolects shows is that interestingly while collective corpora have been used in research on psycholinguistic reality of corpus-attested patterns, as mentioned in the previous section, individual corpora have not (though, see Vetchinnikova 2014). I will try to give several theoretical arguments in support of such uses, i.e. in support of cognitive corpora. First, in the tradition of usage-based linguistics, there is no propensity to make a distinction between competence and performance. Therefore, language produced by an individual is the language available to him/her. Second, as discussed in the previous section, the distinction is also not made between language use and language change. Language acquisition can thus be seen as a language change at the individual or cognitive plane. Therefore, again, language produced by an individual can be taken to represent his/her stage of language acquisition/development/change at this particular moment in time. And lastly, language produced by an individual is a product of his/her cognition. Therefore, it reflects the properties of cognition just as individual's answers to physiological and psychological tests used in cognitive science do.

One reason why cognitive corpora have not been compiled might lie in the obvious practical problems of collecting all the language an individual produces

3 In relation to this, see also studies in psychology (Molenaar 2004; Molenaar 2008; Molenaar and Campbell 2009; van Geert 2011) showing that "we cannot argue from group to individuals" (Schumann 2015: xv), that is, individual trajectories cannot be inferred from aggregated data.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

even for a short period of time (e.g. Mollin 2009a refers to this problem as a limitation of her study). Yet, given the evidence from studies on register variation (most notably Biber 1988 and Biber et al. 1999), it is reasonable to hypothesise that individual language use will exhibit the same kind of variation across different domains of language use. Therefore, it might not be absolutely necessary to compile a 24/7 cognitive corpus, but rather enough to focus on a specific domain, for example someone's academic writing, spoken communication at work or online interaction in social media as a representative sample of his/her language use in this context. At least it would be clear what such a corpus is representative of and what it can be compared to.

So, in this paper a *cognitive corpus* will be defined as a corpus of an individual's language use which is compiled in a way that enables observation of cognitive aspects of language production. As such, in contrast to most other corpora, which can be called communal, it does not aggregate data from different individuals, but rather focuses on a specific individual. And at the same time, it differs from a corpus of an idiolect compiled for the purposes of sociolinguistic research as it does not sample one's language use across genres, domains or the life span. In contrast, it deliberately tries to get rid of sociolinguistic variables as far as possible to pave the way for examining cognitively important factors of recency and frequency as they work within an individual. As Ellis (2006: 104) writes "[f]requency, recency, and context are [...] the three most fundamental influences on human cognition, linguistic and non-linguistic alike". In a cognitive corpus by keeping the context constant as it were, one should be in a good position to examine the effects of frequency and recency. For example, frequency in such a corpus gains a new significance: rather than serving as a predictor of how likely a pattern is to occur in general, it has a direct relationship to the strength of entrenchment of this pattern in the cognition of an individual. Compiling what I call a cognitive corpus is certainly not the only way to do "cognitive corpus linguistics" (Arppe et al. 2010). And indeed there is a growing number of corpus linguistic studies which aim at examining different aspects of human cognition and Cognitive Linguistics studies which use corpora (see Arppe et al. 2010; Grondelaers et al. 2007; Gilquin and Gries 2009; Gries and Stefanowitsch 2006).[4] Yet, to my knowledge there have not been any studies

4 As I see it, it is important to distinguish between Cognitive Linguistics and cognitive linguistics, such as cognitive corpus linguistics. Cognitive Linguistics is an established field of research with its own traditions and theoretical framework which grew out of the work of notably Lakoff, Langacker and Talmy. Cognitive linguistics, in contrast, is a type of research undertaken by cognitively oriented linguists, i.e. linguists of any theoretical background who take an interest in human cognition and believe that human cognitive processes can at least in part explain phenomena we see in language.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

which use a corpus of an individual's language use for this purpose. In the next section, I will rely on the definition of a cognitive corpus and describe a case study based on the comparison between cognitive and communal corpora.

4 Cognitive vs. communal: focus on chunking

In this case study, I will compare chunking at the cognitive and communal planes. By chunking researchers usually mean a cognitive predisposition of language speakers towards holistic, rather than compositional processing (e.g. Wray 2002), also called sequential processing (e.g. Ellis 1996 or Bybee 2012), or the idiom principle (Sinclair 1987, 2004) which, being one of the domain-general cognitive processes underlying language use, leads to the emergence of complex patterning at all levels of language organisation. This is one of the central tenets of usage-based theory and emergentism (see e.g. Ellis 2003; Bybee 2012). At the same time, such complex patterning is usually explored at the communal rather than cognitive level of language use, for example in collective corpora. Corpus studies have yielded observations of numerous patterns in language use, from various methodologically defined n-grams, skipgrams, phrase-frames (Fletcher 2002), concgrams (Greaves 2009), PoS-grams (Stubbs 2007) to formulaic sequences (Wray 2002), units of meaning (Sinclair 1996), collocational frameworks (Renouf and Sinclair 1991), grammar patterns (Hunston and Francis 2000) and lexical bundles (Biber et al. 1999). Whether we can draw a direct line between the patterning observed in corpora and cognitive processes is an unresolved question. Here I would like to probe the hypothesis that cognitive processes lead to patterning at the individual level, which I call cognitive for precisely this reason. While such patterning certainly feeds into the patterning at the communal level, they are not in direct correspondence, as the property of emergence predicts. So chunking might be co-existent at two levels: language speakers chunk, but language can also have its "chunking processes". Such chunking at the communal plane is probably better described as a *phraseological tendency* of language (Sinclair 1996; Cheng et al. 2009), "syntagmatic organisation in language in use" (Stubbs 2009: 115) or simply a phraseological phenomenon, as corpus linguists usually do. Yet, for the ease of presentation I will also talk about "chunking at the communal level".

As my data, I will use a corpus of interaction within one blog over a period of seven years. In order not to mention the actual name of the blog, let us call it the Diachronic Blog Community Corpus (the DBCC). The corpus comprises comments posted to an exceptionally active blog by over 4,000 unique commenters over 7 years, amounting to 7.3 million words in more than 73,000 comments.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Note that the corpus contains comments to the blog entries only and excludes the blogs themselves. This includes 1.77 million words of comments written by a single person, the author of the blog. In addition to this exceptionally large individual contribution, there are five more commenters who produced from 160k to 250k words and ten more who produced from 50k to 100k words. Thus, the corpus allows sampling of language representation at three different levels: 1) the individual or the cognitive level – language use of the author of the blog and its most active commenters taken individually; 2) the communal level – all comments, excluding the heavily represented commenters (24 commenters who contributed over 400 comments to the blog); 3) and the inter-individual micro-social level (Mauranen 2012 and this volume) – the interaction between the most active commenters, including the author of the blog. In this way, the corpus operationalises a complex systems perspective on language. It is important to mention that there are native as well as non-native speakers of English among the active and regular commenters of the blog. The author of the blog is a non-native speaker. In this sense, the blog exhibits typical ELF interaction. Also, blog comments are relatively spontaneous and unedited in contrast to, for example, books or articles which can be more easily put together in an individual corpus of language production. Thus, blog comments are more likely to reflect the common usage patterns of an individual rather than adherence to established norms of standard language introduced through many stages of revision and editing common in more formal writing.

For the purpose of the present article, I will use the following subsets of data extracted from the DBCC:

1) Josef_1750k, which contains ca. 1,750,000 words written by the main author of the blog;
2) Non24_1750k, which contains ca. 1,750,000 words extracted from the DBCC excluding 24 of its most frequent commenters as well as Josef himself;
3) five additional cognitive sub-corpora (C1 to C5 where C1 stands for Commenter 1, C2 for Commenter 2 etc.), each equalling ca. 150,000 words produced by a single commenter.

It is thought feasible to compare these sub-corpora because all of them contain comments posted in response to blog entries or other comments. That is, the corpora are matched in terms of genre. The difference between them might reside in the types of comments made since their contributors in the Non24 corpus are by definition less active in blog discussions that the author of the blog and most active commenters. Figure 1 shows the mean length of comments calculated for five 150k samples extracted chronologically from Non24_1750k and Josef_1750k as well as five additional 150k sub-corpora (Commenters 1 to 5).

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

It shows that: 1) Josef's comments tend to be on average somewhat longer than comments of other contributors; 2) there is variation in the mean length of comments Josef contributes across time; 3) there is some variation in the mean length of comments different people contribute (C1 to C5 bars); 4) Non24 corpus, comprised of comments by over 4, 000 commenters, is relatively homogeneous with regard to the mean length of comments.

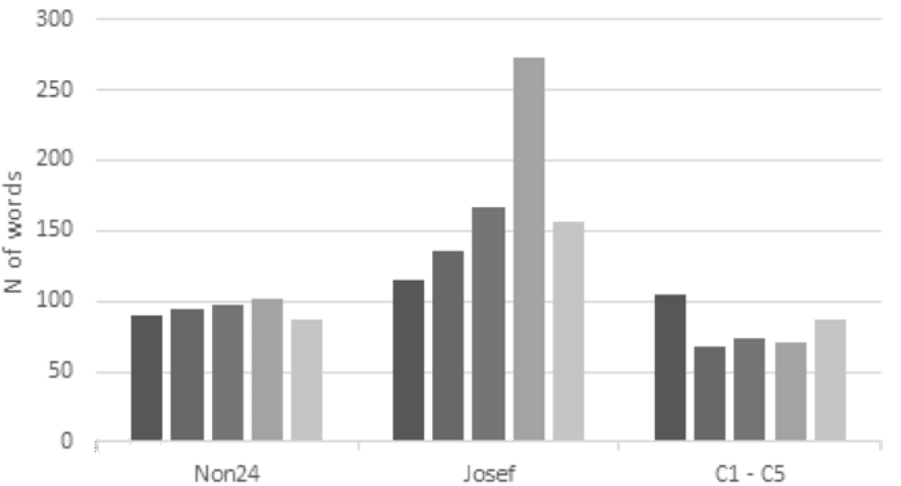


Figure 1: Mean length of comments across sub-corpora (150k samples)

These features of the sub-corpora should not have a large confounding effect in the retrieval of chunks or multi-word units. What might interfere are extremely short or extremely long comments: the former because they might not give possibility for a multi-word unit to occur, the latter because they might become substantially different in terms of text structure. As an extremely short comment, I will treat any comment which is less than five words[5], and as an extremely long, any comment longer than 500 words since 500–570 words is a very common maximum length of a comment across the corpora. Yet, the numbers of such comments are relatively small and equal across all corpora: the proportion of

5 Since in part of the analysis I will use lexical bundles or immediate co-occurrences of 4 words, I have taken the impossibility for a lexical bundle to occur as a criterion for an extremely short comment. In principle a 4-word comment can be a lexical bundle. However, in practice, when comments of up to 4 words were searched for lexical bundles occurring at least twice, none were found. Comments of up to 5 words yielded one lexical bundle in Josef_1750k: *Thanks for the link*. This is thus used as a yardstick.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

short comments per 150k corpus varies between 0 and 2.3 per cent, the proportion of long comments between 0 and 1.2 per cent, with the exception of Josef3_150k (5.3%) and Josef4_150k (16.3%): he clearly had a period when he wrote longer comments, this is visible in the mean length of comments presented in Figure 1 too. This dynamics of Josef's writing will be taken into account in the analysis to follow (see Appendix 1 for more details on the statistics of comments).

5 The case of *it is* ADJ *that*: is there evidence for emergence and fractality?

To probe the relationship between the cognitive and the communal and examine the predictions of complexity theory as applied to language, I will first focus on one construction – *it is* ADJ *that* – and see how it is represented at the two planes.

It is ADJ *that* is a well-known construction, very common for academic language in particular. It belongs to a wider grammar pattern *it* v-link ADJ *that* as documented in *Collins COBUILD Grammar Patterns* (Francis et al. 1998: 480) which itself, making a grammatical generalisation, can be subsumed under 'introductory' or 'anticipatory' *it* structures which "make forward reference to produce an end-focus" (Carter and McCarthy 2006: 891), so characteristic of English in general (see e.g. Quirk et al. 1985). It is also possible to classify it as an extraposed *that*-clause (Biber et al. 1999: 671–675). Further, it is often noted that this pattern carries an evaluative meaning (e.g. Francis 1993, Biber et al. 1999), while Hunston and Sinclair (2000) actually use the pattern[6] (or, "a collection of several patterns" as they say, [84]) to identify evaluative adjectives showing that in fact evaluation is the primary purpose of the pattern. Groom (2005) further points out that this evaluative function combined with the impersonality of *it* used as a grammatical subject makes the pattern very useful for writers of academic texts in particular. Charles (2004) has analysed the use of the pattern in academic discourse and noted further regularities of its more specific realisation, *it is clear/apparent/obvious/evident that*, i.e. the one where the adjectives filling the slot in the *it* v-link ADJ *that* grammar pattern fall into the 'obvious' group according to *Collins COBUILD Grammar Patterns* (Francis et al. 1998: 481). Hunston has later used Charles's findings to reinterpret the observed regularities as semantic sequences, i.e. "sequences of meaning elements rather than as formal sequences" (Hunston 2010 [2008]: 7) since they are much more abstract

6 "*it* + link verb + adjective group + clause" (Hunston and Sinclair 2000: 84)

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

than simple co-occurrences of words as her representation of them clearly suggests:

- 'Logical basis + *it is clear that* + claim'
- 'Consensual information + *it is clear that* + claim'
- '*It is clear that* + claim + exception or caveat' (Hunston 2010 [2008]: 21)

What this substantial body of previous research on the pattern suggests is that from the point of view of language organisation *it is* ADJ *that* is an instance of a grammatical structure which, when it comes to actual language use, is at the same time lexically restricted, for example, to only a few semantic sets of adjectives, with a very clear communicative purpose of evaluation. Taking a step down from this rather high level of generality and looking at the use of the pattern with its adjectives from the specific 'obvious' group and focusing only on academic discourse, we detect further regularities in the way meaning elements are arranged. Still, even at this level of specificity, when we have restricted our observations to the use of a specific group of instances in a particular discourse, the pattern displays a lot of variability. What happens if we look at the patterning of the construction at the individual level?

In my analysis, I am deliberately simplifying the grammar pattern *it* v-link ADJ *that* to *it* BE ADJ *that* so that there is only one variable slot left since there are reasons to expect specific lexicalised preferences for the verb-slot across individual languages.[7] Thus, the most direct way to examine such predictions is to leave just one possible point of variability. It is also reasonable to assume that *it* BE ADJ *that* is probably the prototypical variant of constructions at progressively higher levels of abstractness.

As presented in Table 1, *it* BE ADJ *that* occurs 585 times in Josef_1750k with 69 different adjectives[8]. In comparison, it occurs twice less often in Non24_1750k but with almost the same number of different adjectives which results in a larger type-token ratio. Larger type-token ratio usually means more diversity in lexical choice. Yet, lexical diversity does not seem to be a plausible explanation here since the number of different adjectives used in the construction is almost exactly the same in the two corpora.

Table 1: Type-token distribution of adjectives in *it* BE ADJ *that* in Josef_1750k and Non24_1750k

Corpus	Adj.: Types	Adj.: Tokens	TTR
Josef_1750k	69	585	0.12
Non24_1750k	61	288	0.21

7 Variation in the tenses of the verb BE will probably be determined by the context only.

8 The DBCC was tagged with Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al. 2003).

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Interestingly, if we look at the type-token distribution of adjectives within the construction, it turns out to be Zipfian in both corpora as approximately linear log-log plots show (Figures 2 and 3), where the slope (y) is close to –1 and coefficient of determination (r^2) is close to +1.

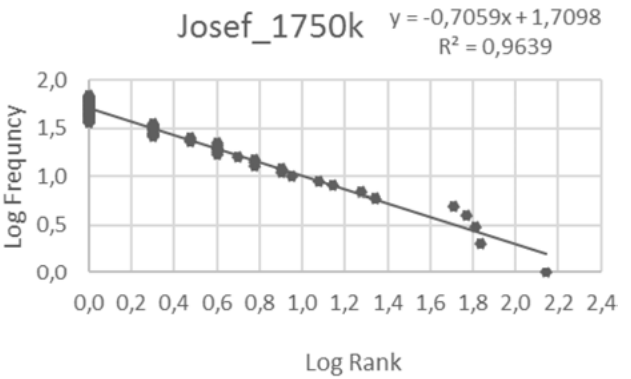


Figure 2: Type-token frequency distribution of adjectives in the *it* BE ADJ *that* in Josef_1750k

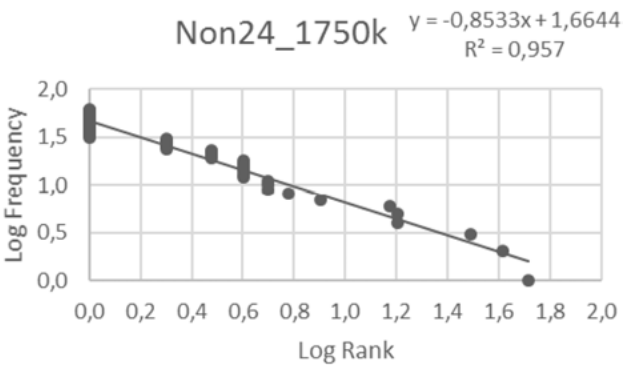


Figure 3: Type-token frequency distribution of adjectives in *it* BE ADJ *that* in Non24_1750k

Looking closer at frequency ranked lists of the 10 most frequent adjectives in each of the two corpora (see Table 2), there are three aspects that become apparent: (1) the order in which different adjectives are preferred in Non 24 and Josef is different; (2) most frequent adjectives on Josef's list are much more frequent

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

than most frequent adjectives on Non24 list[9] and (3) not all adjectives are shared between the two lists (the non-overlapping adjectives are marked in bold)[10].

Table 2: 10 Most frequent adjectives used in *it* BE ADJ *that* in Non24_1750k and Josef_1750k

Non24		Josef	
Adj.	**Freq.**	**Adj.**	**Freq.**
true	52	clear	139
clear	41	**plausible**	69
obvious	31	obvious	65
interesting	16	likely	59
possible	16	true	51
likely	15	unlikely	22
unlikely	8	**conceivable**	19
important	6	important	14
odd	5	good	12
sad	5	possible	9

In fact, both the overlap and the difference between the lists are interesting. In principle, if we assume that the choice of an adjective to fill the open slot in the construction is determined by grammatical rules only, the convergence of four thousand commenters from all parts of the world with Josef in the exhibited preference for certain adjectives is stunning. Even if we take a much more realistic view and acknowledge semantic restrictions, there still remains a large number of adjectives to choose from: for example, Francis et al. 1998 mention 249 different adjectives which commonly occur in this construction in English. So there is clearly a lot of coherence even in a very variable language use such as ELF use. At the same time, there is no evidence for an equilibrium either, such as would be expected in a usage-based scenario driven to its extreme where one's language use would be so adjusted to one's language exposure as to be a mere replica of it. In this case, Josef's preferences would have to reflect the communal average much closer. Yet, he clearly exhibits some individual preferences.

9 It is possible to argue that the main reason for the frequency difference between the most common adjectives on Josef's and Non24 lists is that Josef overall used the construction more often than it was used in Non24 (585 vs 288 occurrences). Yet, it is also possible to reverse the cause-effect explanation: Josef might have used the construction more often because some of the realisations of the construction, such as *it is clear that* and *it is plausible that* have become fixed and therefore come to mind easily adding up to other uses of the construction.

10 Further down the lists the proportion of non-overlapping adjectives increases. In the next 10 most frequent adjectives, it changes from 20% to 80% (i.e. 8 out of 10 adjectives are nor shared). Yet, since the frequencies of occurrence also decrease, this might be due to chance.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

To interpret the distribution of preferences for adjectives in Non24 and Josef, it seems helpful to appreciate that the reasons why some adjectives attain high frequencies on each of the lists are different. *True*, *clear* and *obvious* are frequent in Non24 because they are most popular choices across different speakers. These are the choices different speakers converge on. The fact that all of them are on Josef's list supports this conclusion. In contrast, adjectives on Josef's list are likely to be the result of the process of fixing mentioned in Sections 2 and 3. That is, in frequent use of the construction Josef has started to associate it with specific adjectives. So instead of selecting an adjective from a (semantically) restricted set of variants, he has several adjectives which he commonly uses in the construction and which therefore come to mind first. The choice has changed from being abstract, colligational or that of a semantic preference, to verbatim, i.e. collocational. Such lexicalised realisations of the construction can further develop into separate multi-word units instead of being merely variants of the construction.

In search of further evidence of fixing in Josef's use, I will look at an extended pattern of the construction and in particular at the patterning of the adv. + adj. combinations within it. Again, in principle, allowing for a combination with an adverb should result in an even higher type-token ratio since the chances for recurrence of specific combinations become lower. As Table 3 shows, this is indeed the case for the Non24 corpus: the TTR becomes much closer to 1, meaning that there are a few combinations which occur more than once but there are much fewer of them. Interestingly, in Josef's case the picture is different: his TTR indeed becomes a bit higher than for single adjectives, but it still remains very low implying a lot of reuse of identical combinations.

Table 3: Type-token distribution of adv.+adj. combinations in *it* BE ADV+ADJ *that* in Josef_1750k and Non24_1750k

Corpus	Adv.+adj.: Types	Adv.+adj.: Tokens	TTR
Josef_1750k	167	635	0.26
Non24_1750k	110	180	0.61

Table 4 showing frequency ranked lists of 10 most frequent adv.+adj. combinations used in Josef_1750k and Non24_1750k confirms the conclusions drawn based on calculations of TTRs. There is clearly more divergence between what is common at the communal level and what Josef prefers. But perhaps the biggest difference is the fact that in contrast to Josef's use, there are no exceptionally frequent adv.+ adj. combinations in the Non24 corpus. In other words, Non24 does not exhibit any specific uses which have become conventional or

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

popular at the communal level, and, therefore, *it* BE (ADV+) ADJ *that* indeed can be said to function at an abstracted level as a grammar pattern. In Josef's corpus, while there are some adverbs which combine with adjectives in the construction relatively freely, as a long tail of one-off occurrences suggests (n = 115), others form pretty fixed combinations. For example, *it's still true that* occurs 104 times which is almost as often as the most frequently chosen single adjective *clear* (*it's clear that*, n = 139). It is interesting to mention that both *it's still true that* and *it's clear that* in an overwhelming number of cases (n = 92 and n = 105, respectively) occur with the contracted form *it's*. This preference for a contracted over a non-contracted form which becomes set can also be regarded as an effect of the process of fixing the unit is undergoing. Such settling of a preference for a contracted form within a multi-word unit seems to be very similar to observations of phonetic reduction due to frequency effects (see e.g. Bybee 2006), but at the individual level.

Table 4: 10 Most frequent adv.+adj. combinations used in *it* BE ADV+ADJ *that* in Josef_1750k and Non24_1750k

Non24_1750k		**Josef_1750k**	
Adv.+adj.	**Freq.**	**Adv.+adj.**	**Freq.**
not true	11	**still true**	104
very clear	9	not true	82
more likely	7	very clear	65
not surprising	5	**very likely**	21
pretty clear	5	**also true**	18
certainly true	4	pretty clear	15
extremely unlikely	4	more likely	14
not obvious	4	**not shocking**	14
almost certain	3	not surprising	12
also clear	3	**totally obvious**	10

There are other trends which become visible in Table 4. First, the frequency difference between the most common adv.+adj. combinations on Josef's and Non24 lists becomes even more pointed suggesting indeed a qualitative difference between a realisation which happens to be used by several people in the case of Non24 and a fixed realisation which has in fact become a separate multi-word unit in Josef's case. And second, while Josef's individual patterns not shared with the Non24 are spread out on the frequency list and can be both very frequent as *it's still true that* and relatively infrequent, distinct patterns on the Non24 list are all at the end of the list, just as it was the case with single adjectives. This again suggests that the main determinant of the frequency distribution

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

in Non24 is the popularity of the pattern which means that with the decrease in frequency a certain pattern is less and less likely to be popular with a specific speaker, like Josef in this case. In contrast, the main determinant of the frequency distribution in Josef's list is the cognitive strength of association, which in extreme cases can lead to the development of a separate multi-word unit (with presumably a separate cognitive representation).

Obviously, so far we have dealt with just one speaker who is also a non-native speaker. How generalisable are the observations? For this purpose, I will take 150k samples from five other speakers, in this case NSs of English, as well as 150k samples from Non24 and Josef for comparison purposes. Since it has become apparent from the analysis of Josef's use that a combination of adv.+adj. filling the slot in the *it* BE (ADV+) ADJ *that* construction can be just as frequent as a single adjective and form a fixed holistic pattern, like one "big word" (Ellis 1996: 111; Wray 2002:7), I will collapse adjectives and adv.+adj. combinations in one frequency ordered list for each corpus. Table 5 presents the results.

Table 5: 5 most frequent adj./adv.+adj. combinations across cognitive and communal corpora

C1/Freq.		**C2/Freq.**		**C3/Freq.**		**C4/Freq.**		**C5/Freq.**	
true	16	obvious	5	interesting	7	true	7	**quite conceivable**	12
likely	5	**quite possible**	5	unfortunate	5	**amazing**	3	true	8
clear	4	true	4	**apparent**	3	unfortunate	3	**best**	4
obvious	3	possible	3	clear	2	**too bad**	3	**conceivable**	4
possible	3	ironic	2	likely	2	clear	2	not true	4
TTR: 43/74 = 0.58		24/41 = 0.59		24/42 = 0.57		37/50 = 0.74		39/67 = 0.58	
Non24_150k		Non24_1750k		Josef150_1		Josef150_5		Josef_1750k	
obvious	5	true	52	likely	13	clear	20	clear	139
true	5	clear	41	clear	6	**still true**	15	**still true**	104
clear	3	obvious	31	**plausible**	6	**plausible**	10	not true	82
possible	3	interesting	16	true	6	not true	7	**plausible**	69
unlikely	3	possible	16	conceivable	4	obvious	6	obvious	65
28/45 = 0.62		171/468 = 0.37		43/85 = 0.51		40/117 = 0.34		236/1220 = 0.19	

Analysis of the data presented in Table 5 provides the following observations. First, just like Josef, most other commenters (4 out of 5) have individual preferences, which are not popular at the communal plane or with any other commenter in the table (non-overlapping slot fillers are marked in bold), which, together with the different ordering of preferences overall, makes their profiles distinct from the communal average. Also, very often in individual language use combinations of adv.+adj. are among the five most frequent slot fillers which supports

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

the assumption that such combinations can become quite fixed and treated holistically as "one word". This does not happen at the communal level which means that such combinations usually stay as individual preferences. In fact it appears possible that when a new/separate multi-word unit is developing in an individual's language use, it attracts new components,[11] in this case a modifying adverb. Frequency lists of slot fillers taken from chronologically different samples of Josef's use, the first and the fifth 150k words, as well as the total 1750k words point to the possibility of such diachronic development. Josef's preferences in the fifth sample of 150k words are in general closer to his overall preferences counted for 1750k words. If we look at the development of preferences for the adjective *true* in particular, we will see that in the first 150k words it appears among the first five most frequent slot fillers for the construction. In the fifth sample, it already appears in combinations *still true* and *not true* which together make the occurrence of *true* almost four times as frequent as in the first sample. This gives tentative evidence that with increase in frequency of use, the pattern becomes more fixed and attracts new associations.

What does this lead us to? The interim conclusions we can draw so far is that individual languages seem to exhibit distinct preferences in their lexical patterning. This lexical patterning at the individual/cognitive plane also seems to be more fixed than the patterning at the communal plane. At the extreme, certain patterns at the cognitive plane appear to be lexicalised while corresponding patterns at the communal plane are schematic or grammatical. Therefore, the more abstracted patterning at the communal plane can be seen as emergent from the more specific patterns of the individual languages. That is, the patterning at the communal plane is not the sum of individual patterns, but is qualitatively different from them. This in fact is not surprising since the processes leading to lexical patterning at the two planes are different: convergence of individual speakers on certain popular patterns in the first case, and cognitive propensity to chunking and forming progressively stronger associations between chunk components with increase in frequency of use, in the second.

At the same time, there is indisputable similarity in the overall shape of the patterning between the two planes. This similarity can be viewed as evidence of the fractal structure of language, another property of complex systems, as it was put forward in Section 1. In other words, we can distinguish between the communal and the cognitive/individual planes of language representation but the relationship between them is fractal, i.e. individual language is a "reduced-size

11 Acquisition of new components/associations, such as new collocations, colligations or semantic preferences, in a unit of meaning was discussed as part of the process of fixing in Vetchinnikova 2014.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

copy" of language as represented at the communal plane. The processes at work at the communal and the cognitive planes are different but they result in similar patterning. Conformity of this patterning at both planes to Zipfian frequency distribution can be regarded as a symptom of such fractality.

In fact, as we know from previous research, Zipfian distributions seem to be pervasive in language. Zipf's law holds for frequency distribution of words in any single text in general (Zipf 1935; Manning and Schütze 1999 on Mark Twain's *Tom Sawyer*) and in any corpora (see Manning and Schütze 1999 on the Brown corpus), for type-token frequency distribution of verbs in verb argument constructions (Ellis and Ferreira-Junior 2009; Ellis and O'Donnell 2012; Ellis et al. 2014; see also Goldberg 2006) and for other types of frequency distributions at different levels of language representation (see e.g. Kretzschmar 2009). Researchers of language as a complex adaptive system have already referred to Zipf's law as an indication of the fractality of language when it is split into different levels or time scales. For example, Larsen-Freeman 1997 writes that: "[a]n example of the fractality of language can be seen in Zipf's power law connecting word rank and word frequency for many natural languages" (150). Ellis (2006) mentions Mandelbrot's fractal geometry when discussing language as emergent at all its levels, starting from neurological and physical. De Bot et al. (2013) argue for the "fractal approach to time and change" (207). In a similar vein, it seems reasonable to suggest then that the relationship between the communal and the cognitive can also be described by fractality.

In this study, Zipfian power law relationship was found to apply to the type-token frequency distribution of adjectives in the *it* BE ADJ *that* construction for Josef_1750k (cognitive plane) and Non24_1750k (communal plane). The distribution of adj./adj.+adv combinations in the construction in the rest of the cognitive corpora (C1 to C5) do not really fit Zipfian profile (see Appendix 2 for the log-log plots), but the reasons for this may be different. It is possible that there is simply not enough data in these individual cognitive corpora to produce clearly Zipfian distributions for specific constructions (yet, see the log-log plot for Josef's fifth 150k sample which yields a well-fitting Zipfian distribution). It is also possible that commenters C1 to C5 do not use the construction often enough for it to develop very fixed and frequently occurring preferences. It is clear though that the smaller the TTR, i.e. the less diversity the profile shows, the more the frequency distribution fits Zipfian power law relationship. Out of all the profiles, those of Non24_1750k, Josef1750k and Josef150k_5 provide best fits to Zipfian linear distributions (Josef150_5: $y = -1.05$; $r^2 = 0.96$; Josef1750k: $y = -0.88$; $r^2 = 0.94$; Non24_1750k: $y = -1.16$, $r^2 = 0.93$). It remains to be clarified in future studies which factors warrant a neat Zipfian distribution.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

6 Chunking at communal vs. cognitive planes: evidence of fixing?

The analysis of the *it is* ADJ *that* construction has revealed two clear trends: the lack of a one-to-one relationship between the cognitive and the communal levels and a tendency for fixing in individual preferences. In this section, I would like to further test the hypothesis that lexical patterning at the cognitive plane is overall more fixed than at the communal plane. This hypothesis entails that individual language use should contain more chunks or lexical patterns, which also exhibit less variability, than observed at the communal plane. Therefore, to gauge the level of fixedness of a corpus, it is convenient to focus on lexical bundles (Biber et al. 1999) or immediate co-occurrences of four words.[12] In the following, I compare a cognitive corpus, Josef_1750k, to a communal corpus, Non24_1750k, in terms of the number of types and tokens of lexical bundles which occur at least 17 times, which approximates a commonly used threshold of 10 instances per million (Biber et al 1999; Conrad and Biber 2004; Biber 2009). To retrieve lexical bundles from the corpora, I use AntConc's Clusters/N-grams Tool (Anthony 2014). Table 6 presents the results of this comparison.

Table 6: Number of types and tokens of lexical bundles (freq. threshold = 17) in Josef_1750k and Non24_1750k

Corpus	Types (N)	Tokens (N)
Josef_1750k	1351	50,292
Non24_1750k	550	17,747

As Table 6 shows, there are 2.5 times more different types of bundles in Josef_1750k compared to Non24_1750k. Such frequent bundles also occur 2.8 time more often in Josef's use than at the communal plane. These results seem to support the hypothesis: there are more fixed lexical patterns at the cognitive plane, as Josef's use suggests, than at the communal plane. Yet, it is certainly possible that this is an idiosyncratic feature of Josef's use. To test this possibility, I take five more individual samples of comments data each equalling 150k words (C1 to C5) and retrieve lists of lexical bundles occurring at least five times from them. For these much smaller corpora, a frequency threshold of 10 per million would

12 Please note that in this paper I am not interested in the properties of lexical bundles as such but the phenomenon of verbatim co-occurrence. Thus, I will use the term only in the methodological sense, as a convenient tool.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

mean that a lexical bundle needs to occur only once. Thus, a frequency threshold of 5 was chosen since it is in the mid-range between resulting in too many bundles as frequency thresholds of 2 and 3 do, and too few, as frequency thresholds of 9 and higher do. I also include Josef and Non24 in the comparison. To do this, I take five 150k samples from Non24_1750k and one sample from Josef_1750k, Josef_1, with the mean length of comments closest to the C1–C5 corpora (n = 114) (see Section 4 for details). Josef_1 also dates back to the very beginning of the blog which means that these are the first 150k words he wrote for this blog in comments making the selected sample closer to the samples of comment data from other contributors. Figure 4 shows the number of types of lexical bundles occurring at least five times across five communal (Non24_1 to Non24_5) and six cognitive corpora (Josef and Commenters 1 to 5).

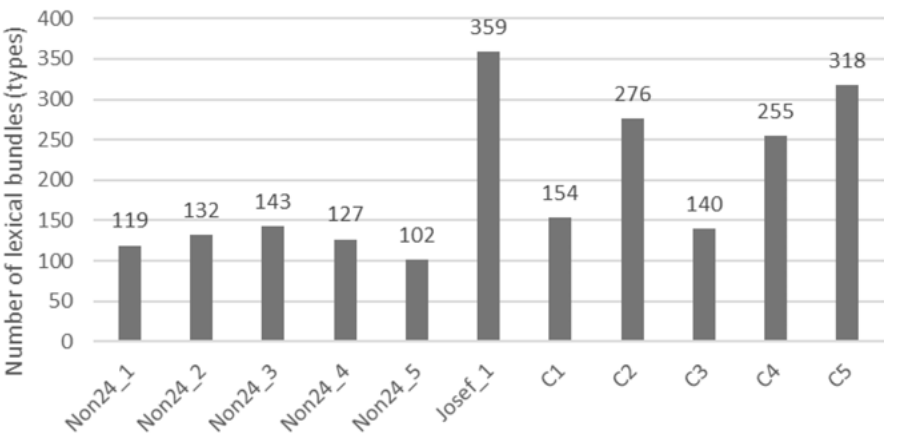


Figure 4: Types of lexical bundles occurring at least 5 times in communal and individual corpora

Figure 5, in turn, shows the number of tokens of lexical bundles occurring at least five times across five communal (Non24_1 to Non24_5) and six cognitive corpora (Josef and Commenters 1 to 5).

The two figures show that 1) cognitive corpora exhibit a greater variety of types and a larger number of tokens of lexical bundles than the communal corpora and 2) variation in the cognitive corpora is wider than in the communal (for types: SDcom = 15.34; SDcog = 87.77; for tokens: SDcom = 101.72; SDcog = 835.05). Yet, if the unequal variance t test (the Welch t test) is applied, the difference between the means of the two sets of data, communal and cognitive, is statistically significant (for types: two-tailed p = 0.0183; for tokens: two-tailed

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

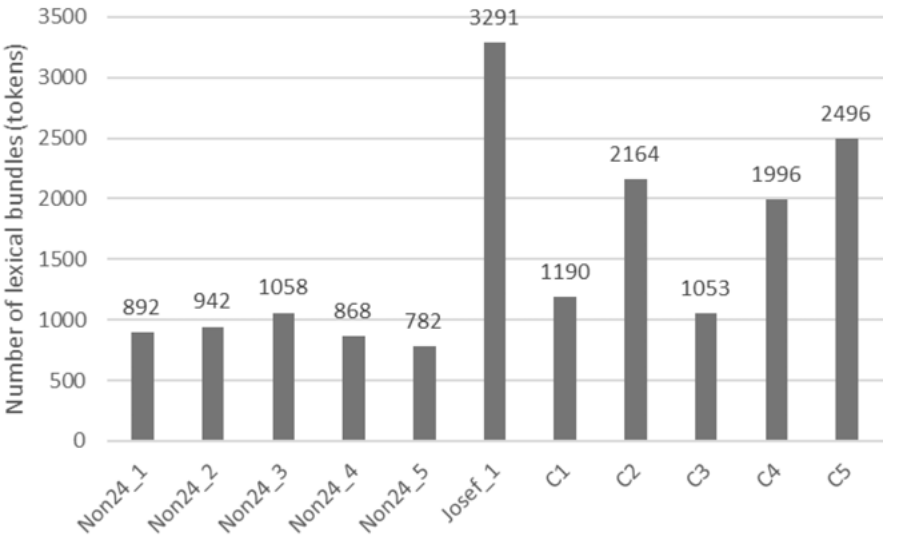


Figure 5: Tokens of lexical bundles occurring at least 5 times in communal and individual corpora

p = 0.0223). Therefore, we can conclude that indeed lexical patterning at the cognitive plane seems to be more fixed than at the communal plane.

There are other observations that can be made. Commenters 1 and 3 show substantially fewer lexical bundles both in terms of types and tokens. Based on SLA literature, we would expect these two commenters to be non-native speakers of English for many SLA studies on second language phraseology have been persistently demonstrating that "learners' phraseological skills are severely limited" (Granger 1998: 158) and that "the non-native speaker, however accurate in grammar and knowledgeable at the level of words, would always be a potential victim of that lesser store of formulaic sequences" (Wray 2002: 210). Yet, among the six individuals examined in this study, Josef is the only non-native speaker, and he is also the one showing the greatest variety and number of lexical bundles in his writing. Should we then make a conclusion that actually the situation with the use of multi-word units is the other way round: that it is the NNSs "who have a greater store of formulaic sequences"? Probably not, since in fact there is another variable involved which is rarely taken into account: the amount of practice in a certain genre. I have taken the first set of 150k words from Josef's comments, so what must make a difference is not the amount of writing he has done for the blog in total but the period of time in which he wrote his 150k words, the density of practice. For Josef this period

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

of time is 15 months, for Commenters 1 and 3, who had smallest numbers of bundles, 40 and 72 months respectively, for Commenters C4 and C5, who had more bundles than C1 and C3 but less that Josef, 27 and 31 months respectively. Commenter C2 wrote his 150k words in the period of only 7 months. Pearson test shows that indeed there is a negative correlation between the time span of writing and the number of lexical bundles (for types, $r = -0.7752$; for tokens, $r = -0.7439$). Thus, a more plausible explanation of the numbers seems to be that lexical patterns get fixed with regular practice and do not so much depend on the native/non-native speaker status.

Section 6 has demonstrated further evidence to suggest that lexical patterning at the cognitive plane, i.e. in individual languages, is more fixed that at the communal. It is proposed that this happens due to the process of fixing in which associations between components of multi-word units strengthen in frequent use, up to becoming verbatim, and attract further associations which can also become entrenched with time.

7 Conclusions

In this paper, I have attempted to separate the communal and the cognitive/individual representation of language and examine the relationship between them. Modelling the two planes using complexity theory suggests that the relationship between them can be characterised by the properties of perpetual dynamics (continuous interaction), emergence and fractality or scale-free self-similarity. In other words, according to this view, the two planes are in constant interaction with each other, the communal plane emerges from the interaction of individual languages and is therefore qualitatively different from them, and the processes underlying language representation at each of the planes are different but lead to similar overall patterning.

Examination of the properties of chunking at the two planes seems to corroborate these predictions. It was observed that multi-word unit patterning at the communal plane is likely to result from averaging, while at the cognitive plane it is determined by the cognitive propensity to chunking and strengthening of internal associations with frequent use. In this way, multi-word unit patterning at the communal plane can be seen as emergent from individual preferences. At the same time, the frequency distribution of preferences is similar at both planes and seems to conform to Zipf's power law: it holds for both planes that while there are only a few very frequent preferences, the number of one-off occurrences is very large. This similarity of frequency distributions can be described as fractal.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

The argument that language patterning at each of the two planes is qualitatively different is further supported by the evidence that lexical patterning at the cognitive plane seems to be more fixed than at the communal: all individuals examined in this paper use a greater variety and a higher number of lexical bundles, which are in essence immediate verbatim co-occurrences of four words, than identified at the communal plane. The reason why we see more four-word bundles at the cognitive plane might be that it contains quite many patterns which are very fixed. Since, presumably, each speaker has his/her own preferences for a particular variant of a pattern, cumulatively these preferences result in variation observable at the communal plane which also leads to fewer lexical bundles: thus, we can say, an *average* form of a pattern is less often fixed than a *cognitive* form.

As a possible avenue for further research, it seems pertinent to mention that the process of fixing in an individual's language use seems to be remarkably similar to the evolutionary processes leading to emergence of phraseological patterns, conventionalisation and grammaticalization observed in language change at the communal plane. Certainly, an item in an individual's use cannot go all the way toward becoming grammatical since this stage of grammaticalization requires diffusion and acceptance of a language community. However, if the process of grammaticalization is viewed as a continuum, then an item in an individual's use seems to be able to move along this continuum at least to the point of becoming non-compositional. The case of *it's still true that* is a good example of this. So the proposition that it is frequency which drives grammaticalization (Bybee and Hopper 2001; Bybee 2003) seems to have a lot of explanatory power. Growing frequency leads to chunking or, in other words, a switch to processing on the idiom rather than the open-choice principle, which in its turn leads to structural reanalysis (see e.g. Beckner and Bybee 2009) or in Sinclair's terms, delexicalisation and meaning-shift (see Cheng et al. 2009), which may or may not in the end lead to grammaticalization. This way, the processes observed in an individual's language use can be seen as micro-processes of language change which feed into macro-processes of language change observed at the communal plane. The fact that such micro-processes look remarkably similar to the macro-processes can be again described as scale-free self-similarity or fractal structure.

With respect to changing English, the importance of the processes underlying individual language use means that recent dramatic increase in second language use of English and especially ELF must inevitably have a visible impact at the communal plane. For example, if second language users are prone to approximation (Mauranen 2005, 2012), or communicatively unproblematic but slightly non-standard use of multi-word units, it is possible that these approxi-

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

mative uses can also get fixed in their repertoires (see also Mauranen this volume). As a hypothesis, this might lead to more distinct individual preferences resulting in more divergence between idiolects or wider inter-individual variation and eventually in more variability at the communal plane. Such variability will be an emergent property since it clearly does not characterise language use at the individual plane: both native and non-native individual languages were more fixed in their lexical patterning than the communal representation in this study and the extent of their fixedness depended on the density of practice rather than native or non-native speaker status. This fits well with previous research on ELF which describes it as highly variable language use.

For the discussion of the global and the local in changing English, this study therefore suggests two tentative hypotheses. First, if indeed current language use can be characterised by wider inter-individual variation, then idiolects gain more weight in language variation and change and thus present an interesting object of further research. Second, it seems promising to continue modelling language as a CAS on the social dimension, or that of spread. In this study, I have chosen the relatively safe option of exploring the relationship between individual languages and the communal representation of discourse in which these individual languages participate. Modelling Global English as the communal plane and identifying the component parts from whose interaction it emerges certainly present more challenges as well as intriguing possibilities for further research. I hope this study can serve as a step in this direction.[13]

References

Anderson, Richard C. & Peter Freebody. 1981. Vocabulary knowledge. In John T. Guthrie (ed.), *Comprehension and teaching: Research reviews*, 77–117. Newark, DE: International Reading Association.

Anthony, Laurence. 2014. AntConc (Version 3.4.3w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora* 5(1). 1–27.

Barlow, Michael. 2013. Individual differences and usage-based grammar. *International Journal of Corpus Linguistics* 18(4). 443–478.

13 I would like to thank the author of the blog used in this study for generously providing his blog data for research purposes. I am also grateful to Ray Carey† and Nina Mikusova for their help in preparing the data for corpus linguistic analysis. Special thanks to all the readers for their valuable comments on different versions of this paper.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Beckner, Clay & Joan Bybee. 2009. A usage-based account of constituency and reanalysis. *Language Learning* 59(1). 27–46.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language learning* 59(s1). 1–26.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3). 275–311.

Biber, Douglas et al. 1999. *Longman grammar of spoken and written English*. London: Longman.

Bock, Kathryn J. & William F. Brewer. 1974. Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology* 103(5). 837–843.

Bybee, Joan & Paul Hopper. 2001. *Frequency and the emergence of linguistic structure*. Amsterdam, Philadelphia: John Benjamins.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertram Malle (eds.), *The evolution of language from pre-language*, 109–32. Amsterdam: John Benjamins.

Bybee, Joan. 2003. Mechanisms of change in grammaticalization: The role of frequency. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 602–623. Malden (MA): Blackwell.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Bybee, Joan. 2012. Domain-general processes as the basis for grammar. In Maggie Tallerman & Kathleen R. Gibson (eds.), *The Oxford Handbook of Language Evolution*, 528–536. Oxford: Oxford University Press.

Carey, Ray. 2013. On the other side: Formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca* 2(2). 207 – 228

Carter, Ronald & Michael McCarthy. 2006. *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

Charles, Maggie. 2004. *The construction of stance: A corpus-based investigation of two contrasting disciplines*. University of Birmingham. Unpublished PhD thesis,

Cheng, Winnie, Chris Greaves, John McH. Sinclair & Martin Warren. 2009. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics* 30(2). 236–252.

Conrad, Susan & Douglas Biber. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* 20. 56–71.

Coulthard, Malcolm. 2004. Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics* 25(4). 431–447.

de Bot, Kees, Wander Lowie & Marjolijn Verspoor. 2007. A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition* 10(1). 7–21.

de Bot, Kees & Diane Larsen-Freeman. 2013. Researching second language development from a dynamic systems theory perspective. In Marjolijn Verspoor, Kees de Bot & Wander Lowie (eds.), *A dynamic approach to second language development: Methods and techniques*, 5–23. Amsterdam; Philadelphia: John Benjamins.

de Bot, Kees, Wander Lowie, Steven L. Thorne & Marjolijn Verspoor. 2013. Dynamic systems theory as a theory of second language development. In María del Pilar García Mayo, María

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Junkal Gutiérrez Mangado & María Martínez Adrián (eds.), *Contemporary approaches to second language acquisition*, 199–220. Amsterdam: John Benjamins.

Durrant, Philip & Alice Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2).

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition* 18. 91–126.

Ellis, Nick C. 2003. Constructions, chunking, and connectionism: The emergence of second language structure. In Catherine J. Doughty & Michael H. Long (eds.), *The handbook of second language acquisition*, 63–103. Oxford: Blackwell.

Ellis, Nick C. 2006. Cognitive perspectives on SLA: The associative-cognitive CREED. *AILA Review* 19(1). 100–121.

Ellis, Nick C. 2011. The emergence of language as a complex adaptive system. In James Simpson (ed.), *The Routledge handbook of applied linguistics*, 666–679. London: Routledge.

Ellis, Nick C. & Diane Larsen-Freeman. 2006. Language emergence: Implications for Applied Linguistics–Introduction to the special issue. *Applied Linguistics* 27(4). 558–589.

Ellis, Nick C. & Eric Frey. 2009. The psycholinguistic reality of collocation and semantic prosody (2): Affective priming. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley (eds.), *Formulaic language (vol. 2): Acquisition, loss, psychological reality, and functional explanations* [Typological Studies in Language, 83], 473–497. Amsterdam: John Benjamins.

Ellis, Nick C., Eric Frey, & Isaac Jalkanen. 2009. The psycholinguistic reality of collocation and semantic prosody (1): Lexical access. In Ute Römer & Rainer Schulze (eds.), *Exploring the lexis-grammar interface* [Studies in Corpus Linguistics, 35], 89–114. Amsterdam: John Benjamins.

Ellis, Nick C. & Fernando Ferreira–Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3). 370–385.

Ellis, Nick C. & Matthew Brook O'Donnell. 2012. Robust language acquisition: An emergent consequence of language as a complex adaptive system. In Patrick Rebuschat & John M. Williams (eds.), *Statistical learning and language acquisition, vol. 1* [Studies in Second and Foreign Language Education], 265–304. Berlin; Boston: De Gruyter Mouton.

Ellis, Nick C., Matthew Brook O'Donnell & Ute Römer. 2014. The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics* 25(1).

Fletcher, William H. 2002. *Phrases in English*. http://phrasesinenglish.org/ (last accessed 23 April 2015).

Francis, Gill, Susan Hunston & Elizabeth Manning. 1998. *Collins COBUILD Grammar Patterns: Nouns and adjectives*. London: HarperCollins.

Francis, Gill. 1993. A corpus-drive approach to grammar: Principles, methods and examples. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 137–156. Amsterdam: John Benjamins.

Gilquin, Gaëtanelle & Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics & Linguistic Theory* 5(1). 1–26.

Gleick, James. 1987. *Chaos: Making a new science*. New York: Viking.

Goldberg, Adele. 2006. *Constructions at work*. Oxford: Oxford University Press.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony P. Cowie (ed.), *Phraseology: Theory, analysis, and applications*, 145–160. Oxford: Clarendon.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Greaves, Chris. 2009. *ConcGram 1.0: A phraseological search engine* [Software]. Amsterdam: John Benjamins.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next. . . . *International Journal of Corpus Linguistics* 18(1). 137–166.

Gries, Stefan Th. & Anatol Stefanowitsch (eds). 2006. *Corpora in Cognitive Linguistics: Corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter.

Grondelaers, Stefan, Dirk Geeraerts and Dirk Speelman. 2007. A case for a cognitive corpus linguistics. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics*, 149–169. Amsterdam: John Benjamins.

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4(3). 257–277.

Gurevich, Olga, Matthew A. Johnson & Adele E. Goldberg. 2010. Incidental verbatim memory for language. *Language and Cognition* 2(1). 45–78.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hopper, Paul J. 1987. Emergent grammar. *Berkeley Linguistics Society* 13. 139–57.

Hopper, Paul J. 2011. Emergent grammar and temporality in interactional linguistics. In Peter Auer & Stefan Pfänder (eds.), *Constructions: Emerging and emergent*, 22–44. Berlin: De Gruyter Mouton.

Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2). 249–268.

Hunston, Susan. 2008. Starting with the small words: Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13(3). 271–295.

Hunston, Susan. 2010. Starting with the small words. In Ute Römer & Rainer Schulze (eds.), *Patterns, meaningful units and specialized discourses*, 7–30. Amsterdam; Philadelphia: John Benjamins.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar*. Amsterdam: John Benjamins.

Hunston, Susan & John McH. Sinclair. 2000. A local grammar of evaluation. In Susan Hunston & Geoff Thompson (eds.), *Evaluation in text: Authorial stance and the construction of discourse*, 74–101. Oxford: Oxford University Press.

Kretzschmar, William A. 2009. *Linguistics of speech*. Cambridge: Cambridge University Press.

Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied Linguistics* 18(2). 141–165.

Larsen-Freeman, Diane. 2013. Complexity Theory/Dynamic systems theory. In Peter Robinson (ed.), *The Routledge encyclopaedia of Second Language Acquisition*, 103–106. New York: Routledge.

Larsen-Freeman, Diane & Lynne Cameron. 2008. *Complex systems and Applied Linguistics*. Oxford: Oxford University Press.

Mandelbrot, Benoit B. 1963. The variation of certain speculative prices. *The Journal of Business* 36(4). 394–419.

Mandelbrot, Benoit B. 1982. *The fractal geometry of nature. 1 edition*. San Francisco: W. H. Freeman and Company.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, Massachusetts: MIT Press.

Mauranen, Anna. 2005. English as a Lingua Franca – an unknown language? In Giuseppina Cortese & Anna Duszak (eds.), *Identity, community, discourse: English in intercultural settings*, 269–293. Frankfurt: Peter Lang.

Mauranen, Anna. 2009. Chunking in ELF: Expressions for managing interaction. *Journal of Intercultural Pragmatics* 6(2). 217–233.

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Mauranen, Anna. 2012. *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge: Cambridge University Press.

Mauranen, Anna. 2013. Hybridism, edutainment, and doubt: Science blogging finding its feet. *Nordic Journal of English Studies* 12(1). 7–36. (23 September, 2014).

Mauranen, Anna. Forthcoming. *Reflexively speaking: Uses of metadiscourse in ELF*. Berlin: De Gruyter Mouton.

Molenaar, Peter C. M. 2004. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives* 2(4). 201–218.

Molenaar, Peter C. M. 2008. On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology* 50(1). 60–69.

Molenaar, Peter C. M. & Cynthia G. Campbell. 2009. The new person-specific paradigm in psychology. *Current Directions in Psychological Science* 18(2). 112–117.

Mollin, Sandra. 2009a. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2).

Mollin, Sandra. 2009b. "I entirely understand" is a Blairism: The methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics* 14(3). 367–392.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*, 191–227. London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow: Longman.

Raumolin-Brunberg, Helena & Arja Nurmi. 2011. Grammaticalization and language change in the individual. In Heiko Narrog & Bernd Heine (eds.), *The Oxford Handbook of Grammaticalization*, 251–262. Oxford: Oxford University Press.

Read, John. 2004. Plumbing the depths: How should the construct of vocabulary be defined? In Paul Bogaards & Batia Laufer (eds.), *Vocabulary in a second language: Selection, acquisition, and testing*, 209–226. Philadelphia, PA, USA: John Benjamins.

Renouf, Antoinette & John McH. Sinclair. 1991. Collocational frameworks in English. In Karin Aijmer & Bengt Altenberg (eds.), *English corpus linguistics: Studies in honour of Jan Svartvik*, 128–143. London: Longman.

Schumann, John H. 2014. Foreword. In Dörnyei, Zoltán, Alastair Henry & Peter D. MacIntyre (eds.), *Motivational dynamics in language learning*, xv- xix. Bristol: Multilingual Matters.

Sinclair, John McH. 1987. Collocation: A progress report. In Ross Steele & Terry Treadgold (eds.), *Language topics: Essays in honour of Michael Halliday*, 319–331. Amsterdam: John Benjamins.

Sinclair, John McH. 1996. The search for units of meaning. *Textus* 9(1). 75–106.

Sinclair, John McH. 2004. *Trust the text*. London: Routledge.

Stubbs, Michael. 2007. An example of frequent English phraseology: Distributions, structures and functions. In Roberta Facchinetti (ed.), *Corpus Linguistics 25 years on*, 89–106. Amsterdam and New York: Rodopi.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge (MA): Harvard University Press.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. *Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on*

EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

Human Language Technology – Volume 1, 173–180. (NAACL '03). Stroudsburg, PA, USA: Association for Computational Linguistics.
van Geert, Paul. 2011. The contribution of complex dynamic systems to development. *Child Development Perspectives* 5(4). 273–278.
Vetchinnikova, Svetlana. 2014. *Second language lexis and the idiom principle*. Unigrafia: Helsinki.
Weinreich, Uriel. 1953. *Language in contact: Findings and problems*. New York: Linguistic Circle.
Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Wright, David. 2015. Testing the theory of idiolect. Paper presented at the BAAL Annual Meeting "Breaking theory: New directions in Applied Linguistics", Birmingham, September 3–5.
Zipf, George K. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Databases

BNC: The British National Corpus. http://bncweb.lancs.ac.uk/

Appendix 1

Subsets of the Diachronic Blog Community Corpus (the DBCC) used in the study

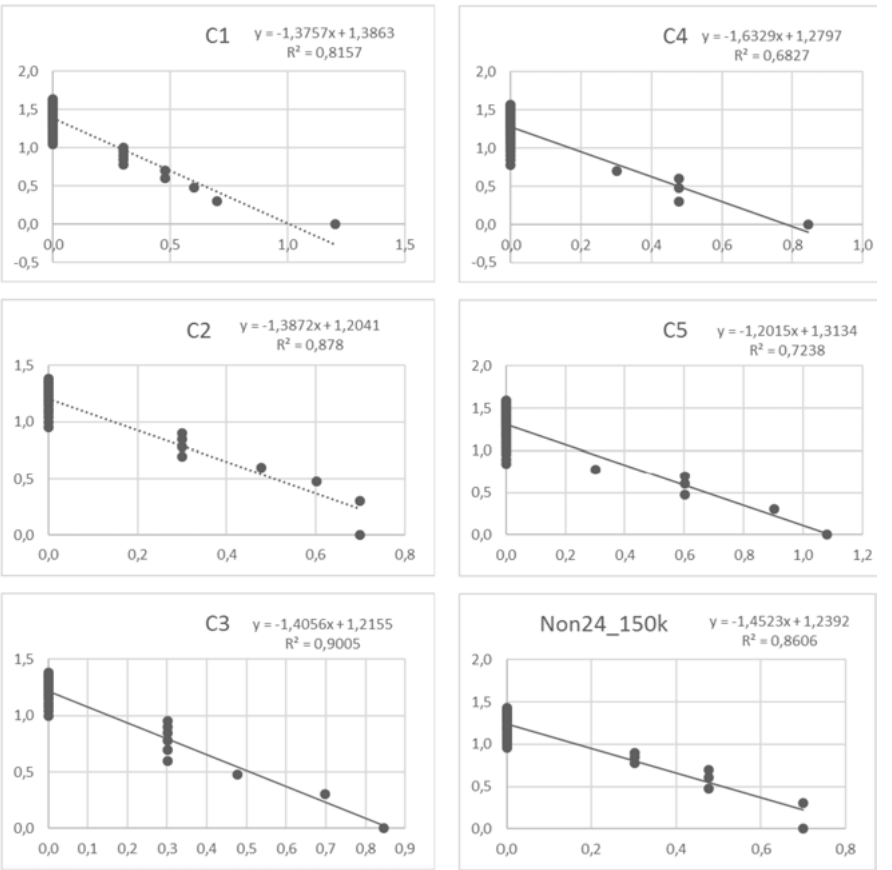
Corpus	Corpus size	N of comments	Mean length[14]	MAX length	N of short comments (< 5 words)		N of long comments (> 500 words)	
					N	%	N	%
Non24_1	150401	1673	90	525	36	2,2	10	0,6
Non24_2	150164	1598	94	554	37	2,3	16	1,0
Non24_3	150234	1556	97	1869	33	2,1	18	1,2
Non24_4	150488	1472	102	539	25	1,7	11	0,7
Non24_5	150212	1726	87	571	16	0,9	10	0,6
Josef_1	150287	1313	114	538	8	0,6	11	0,8
Josef_2	150299	1111	135	529	4	0,4	11	1,0
Josef_3	150962	906	167	1554	0	0,0	48	5,3
Josef_4	150786	551	274	1695	1	0,2	90	16,3
Josef_5	150486	959	157	520	4	0,4	4	0,4
C1	150156	1442	104	507	9	0,6	1	0,1
C2	150445	2215	68	504	46	2,1	1	0,0
C3	150162	2051	73	724	26	1,3	4	0,2
C4	150435	2126	71	817	43	2,0	5	0,2
C5	150157	1716	88	1171	2	0,1	4	0,2

14 Pearson test shows only a weak correlation between the mean length of a comment and the number of lexical bundles retrieved from a corpus (for types r = −0.0021; for tokens r = 0.0781).

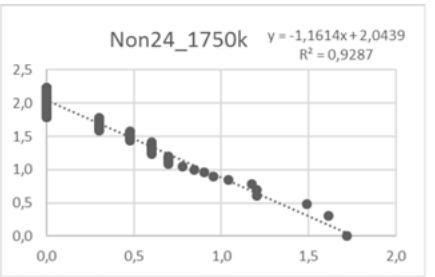
EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use

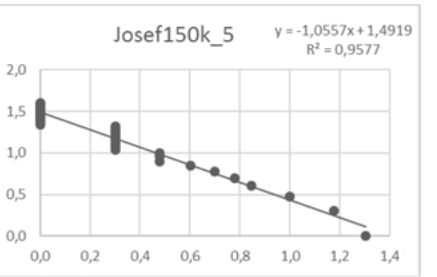
Appendix 2

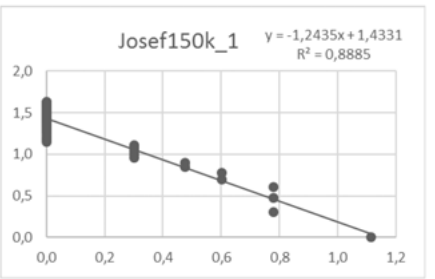
Logarithmic plots of type-token frequency distributions of adj./adv.+adj. combinations in the *it* BE ADJ (+ADV) *that* construction across cognitive and communal corpora

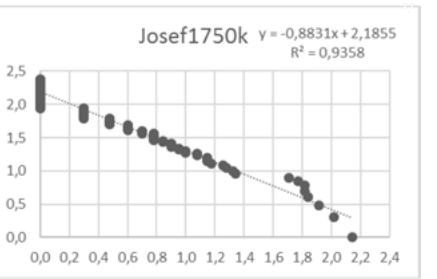


EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use









EBSCOhost - printed on 5/7/2020 7:19 AM via HELSINGIN YLIOPISTO - UNIVERSITY OF HELSINKI. All use subject to https://www.ebsco.com/terms-of-use