Probabilistic Register Modelling

Roland Schäfer
Germanistische Sprachwissenschaft,
Friedrich-Schiller-Universität
Fürstengraben 30, 07745 Jena
roland.schaefer@uni-jena.de

Felix Bildhauer
Abteilung Grammatik,
Leibniz-Institut für Deutsche Sprache
R5 6-13, 68161 Mannheim
bildhauer@ids-mannheim.de

Pauline Reiß
Germanistische Sprachwissenschaft,
Friedrich-Schiller-Universität
Fürstengraben 30, 06645 Jena
pauline.reiss@uni-jena.de

Elizabeth Pankratz
Centre for Language Evolution,
University of Edinburgh
3 Charles St, Edinburgh, EH8 9AD
e.c.pankratz@sms.ed.ac.uk

Stefan Müller
Deutsche Sprache und Linguistik,
Humboldt-Universität
Dorotheenstraße 24, 10117 Berlin
st.mueller@hu-berlin.de

Abstract In this paper, we propose a theory of register variation in probabilistic grammar. We assume register (situational-functional intra-individual variation in lexicon and grammar) to be a probabilistic phenomenon: First, speakers assign a probability distribution to the set of known registers based on their assessment of the communicative situation, and they choose a register from that distribution. Second, the chosen register provides the information required to update the probability distribution of all linguistic signs, and speakers subsequently select register-appropriate signs with a high probability. The theory is tested in the analysis of a large, unstructured, and partly noisy collection of documents (a Web corpus), using Latent Dirichlet Allocation—a method that implements our theoretical model—to discover register candidates based on the distribution of lexical and grammatical forms within the documents. Those candidates are then validated in a large-scale annotation experiment, wherein annotators classified with high reliability the situational-functional properties of the documents. We conclude that our work supports the assumption that register variation should be modelled as a probabilistic phenomenon.

Keywords: register, probabilistic grammar, Latent Dirichlet Allocation, corpora, German

1 Registers as probabilistic categories

1.1 Recovering registers from written texts

Many linguists call cross-situational variation within a speaker's grammar *register* (see Section 1.2 for details). Under this definition, any theory of register has to incorporate a description of grammatical and lexical means (the overt expressions of register), a description of situational-functional factors, as well as a mechanism by which they are connected. Many such approaches have been proposed, and many of them have been used to analyse corpus data. In this paper, we propose a generative model of register variation, and we show how it can be applied in the analysis of large unstructured (and partially

noisy) collections of text such as data from the Web. Collections like this pose a specific problem inasmuch as the situation in which the documents were written as well as their purpose is mostly opaque. Therefore, our approach is specifically designed to recover as much information about the registers of the documents as can be reliably recovered without getting lost in either pre-defined (and usually insufficient) or ever-growing (and thus less and less informative) taxonomies of register labels. It works by first discovering distributions of grammatical features that could be registers (*potential registers* or *pregisters*) with an unsupervised method, then assigning situational-functional properties to them through human annotation. We suggest that whatever information ultimately cannot be recovered using this method cannot be recovered by naive but competent readers either, to the effect that it is actually lost. Thus, our approach has the potential to delineate the register information encoded in texts themselves once the immediate context has been taken away.

The paper proceeds as follows. The remainder of this section discusses existing definitions of registers and approaches to form-based classification (Section 1.2), views on the situational-functional side of registers (Section 1.3), and finally positions on registers as probabilistic categories (Section 1.4). Section 2 describes a formal model of register variation and derives a method of discovering registers as latent dimensions in documents from large unstructured text corpora. In Section 3, the potential registers discovered in the previous step are connected to a set of low-level situational-functional parameters (education, interactivity, narrativity, and proximity) via a demonstrably reliable human annotation. In Section 4, we derive clear desiderata for future work, both in terms of the methods used and the conceptual-theoretical underpinnings of register research.

1.2 Register and grammar

Variation in speakers' and writers' speech and writing depending on the context of use has long been observed. This kind of variation has been analysed in terms of *style*, *genre*, or *register* (among other terms) with definitions varying wildly between different research traditions (see, e. g., Lee 2001 for a rather early overview). We focus on the notion of *register*, which Lüdeling et al. (2022) (building on Biber & Conrad 2009) define as those "aspects of socially recurring intra-individual variation that are influenced by situational and functional settings". In our view, Halliday & Hasan (1976) and subsequent publications in Systemic-Functional Linguistics (SFL) also suggest a definition of register that is most obviously compatible with grammatical theory as we understand it.

> The linguistic features which are typically associated with a configuration of situational features [...] constitute a register. The more specifically we can characterize the context of situation, the more specifically we can predict the properties of a text in a situation. (Halliday & Hasan 1976: 22)

Some authors (prominently, for example, Agha 2007) focus specifically on register as a component of speaker behaviour that also recurs in the behavior of many other individuals and is thus culturally agreed upon (see also the definition by Lüdeling et al. 2022 above). While this connects register to sociolinguistics (which is a necessary connection to make), we approach the subject from a cognitive grammatical side and focus on the mechanism by which speakers and writers chose linguistic forms depending on situational and functional factors (see especially Section 2).

Our empirical approach is strongly quantitative and computational in that it starts by analysing the distribution of grammatical features in documents coming from large unstructured text corpora and then adds an analysis of the situational and functional correlates of the grammatical features in an independent subsequent step. A number of previous quantitative approaches have sought to analyse registers (or related categories like style and genre) based on the occurrence of linguistic features in documents.[1] While many of these quantitative approaches are primarily concerned with finding and evaluating features for automatic document classification (e. g.Karlgren & Cutting 1994; Stamatatos & Kokkinakis & Fakotakis 2000; Ferizis & Bailey 2006; Levering & Cutler 2009; Kim & Ross 2011; Santini 2011; Egbert & Biber & Davies 2015; Sharoff 2018; Ortmann & Dipper 2019, and so on), others have used lexico-grammatical feature counts as means for linguistic register analysis proper.

A very prominent approach in this regard is Douglas Biber's Multidimensional Analysis (MDA, see Biber 1988; 1989; 1995; Biber & Egbert 2018), which we briefly describe here because it may look similar to our approach while, in fact, it is not. Its aim is to analyse a set of registers in terms of a small number of underlying *dimensions of variation*. The first step in obtaining the relevant dimensions of variation is to apply factor analysis to linguistic feature counts from a large number of text samples, taken from a variety of the assumed registers. Factor analysis finds co-occurrence patterns among the features and uses this information to assign individual features to one or more of a small number of so-called factors along with a *loading* that represents a feature's prominence on a factor. Subsequently, the researcher interprets the resulting factors, thereby trying to find a communicative or functional interpretation for the joint prominence of particular linguistic features within a particular factor. Where this succeeds, a dimension of variation has been found (for example, Biber 1989 identifies five such dimensions of variation), and where it fails, the researcher may ignore the respective factor. In MDA, dimensions are regarded as scalar, where the end points are characterised by opposite communicative or functional descriptions, such as "overtly argumentative" or "not overtly argumentative" (Biber 2009a: 840). An individual text may then be located on a particular scale by calculating a score from its feature counts and the loadings of features on that scale/factor. Likewise, a register may be located on a particular scale by calculating an average score from all the texts belonging to that register. An individual text, or a register, is exhaustively characterised by its location on each one of the dimensions of variation thus identified. Additionally, texts may be clustered by their location in the resulting n-dimensional space (where n is the number of dimensions of variation), yielding what Biber (1989: 5) called "linguistically well defined" text types, a category assumed to be distinct from register and intended to account for linguistic variation *within* predefined registers.

Our approach will appear to some readers to be similar to Multi-Dimensional Analysis, and it is true that it partially builds upon work by Biber and his collaborators (see Biber 1995; 2009b, and a great many other publications). However, one major epistemological difference between MDA and our approach is that MDA assumes both a set of registers and an unambiguous mapping from individual texts to registers as *a priori knowledge* on the part of the researcher. On this basis, MDA sets out to *analyse* the assumed registers. By contrast, the approach presented in this paper does not operate on an assumed inventory of registers but *uncovers* a set of potential registers and seeks to determine whether they are actual registers by applying linguistic and functional analysis. There are fur-

[1] There is also a strand of research concerned with not just analysing but also generating text in a given register (Argamon 2019: 109). For space reasons, we will not discuss this prominently.

ther important differences between MDA and our approach regarding assumptions about probabilism in registers, which we discuss in Section 1.4.

1.3 Situations and functions

As will become clear in Sections 2 and 3, our approach begins with an analysis of the distribution of grammatical features in corpus documents, then adding annotation for the situational aspects of the documents. We do not use any specific previously suggested taxonomy of such aspects, but we draw from many of them, which range from minimalistic to highly complex and specific. A relatively minimalistic approach to situational parameters is taken by Paolillo (2000). In a strongly sociolinguistically-informed approach, Paolillo proposes only four different "communicative attitudes" (for speakers of Sinhala) that affect the distribution of register variables. These attitudes are the extents to which speakers wish to index *editedness*, *interactivity*, *correctness*, and *publicness* Paolillo (2000: 239). In contrast to Paolillo (2000), Biber & Conrad (2009) create a much more granular taxonomy. They present an extensive list in their Table 2.1 (Biber & Conrad 2009: 40) of what they consider to be major situational characteristics that are important for defining register. These characteristics are grouped into seven overall categories: *participants*, *relations among participants*, *channel*, *production circumstances*, *setting*, *communicative purposes*, and *topic*. Their inventory of parameters is applicable to both spoken and written language, but several of the categories in this taxonomy cannot be inferred from textual data alone or simply do not make sense to apply. For instance, under *participants*, they ask "[whether there are] on-lookers", which might be a virtually impossible question to solve for texts posted on the Web (our source of data). Thus, we turn now to parameters developed specifically with Web texts in mind.

According to Sharoff (2018: 68), Sinclair & Ball (1996) were the first to propose that factors affecting language use in text be split into two types: text-external and text-internal. Text-external parameters include properties like characteristics of the author, or the communicative aims that they intended. Text-internal parameters have to do with the linguistic features that appear within the texts – the variants that are chosen in order to instantiate a particular register. In essence, what we have aimed to do is use the given text-internal parameters to infer probable text-external parameters. Here, we consider the text-external parameter sets used by three studies: Egbert & Biber & Davies (2015); Sharoff (2018); Biber & Egbert & Keller (2020a).

The goal in Egbert & Biber & Davies (2015) is to classify Web texts into discrete registers. They created an annotation task in which annotators move through a decision tree, selecting the most appropriate value for a handful of situational parameters at each step. This procedure leads to a hierarchical conceptualisation of registers (or genres) in which, e. g., travel blogs and short stories are both sub-categories of *narrative*. The situational distinctions they include in their decision tree are *mode* (whether the text was originally written vs. originally spoken); if originally written, then *participants* (if the text was produced by a single author or co-authors vs. multiple participants); and if single author or co-authors, then *purpose* (to narrate events vs. describe information vs. express opinion vs. explain information vs. express lyrically; see their Table 5 (Egbert & Biber & Davies 2015: 1824). Inter-annotator agreement was measured using Fleiss' κ, and the overall score was 0.47 (Egbert & Biber & Davies 2015: 1825), which signals moderate agreement.

Sharoff (2018) takes a different approach to the classification of Web texts. He focuses nominally on genre, but his definition (Sharoff 2018: 68) makes *genre* almost synonymous with *register* under our definition. Unlike the approach of Egbert & Biber & Davies

(2015), who use atomic labels, Sharoff (2018) introduces the idea of annotating more basic functional parameters in Web texts using scales. In Sharoff's approach, genres are equated to what the author calls "functional text dimensions". Twenty such dimensions are presented, including *informative reporting* and *argumentation*. These dimensions are diagnosed using test questions, such as, for *argumentation*: "To what extent does the text contain explicit argumentation to persuade the reader (for example, argumentative blogs, opinion pieces or discussion forums)?" (Sharoff 2018: 70). The inter-annotator agreement was measured using Krippendorff's α, and a good overall agreement score of 0.76 was obtained. The difference in success between Egbert & Biber & Davies (2015) and Sharoff (2018) encouraged our decision to use more elementary parameters rather than complex register categories.

Finally, Biber & Egbert & Keller (2020a) propose 23 scalar situational parameters that can be grouped into four categories: features of the text, features of the author/speaker, purpose of the text, and the basis of information within the text (Table 2, Biber & Egbert & Keller 2020a: 591). For example, features of the authors/speakers may be the extent to which they are experts, focus on themselves, assume technical background knowledge, and so on. Presumably partly because no coding rubric or annotation guidelines were provided, inter-annotator agreement was middling: Cohen's κ of 0.46 (Biber & Egbert & Keller 2020a: 592).

1.4 Registers in probabilistic grammar

We treat register as a probabilistic phenomenon with respect to its two main aspects: (i) the mapping from situational-functional parameters (SFPs) to registers; (ii) the mapping from registers to lexico-grammatical features (LGFs).

As we have encountered many fuzzy views of the concept, we first provide an exact definition of the term *probabilistic*. All grammars define constraints on possible utterances, thus defining the set of these utterances. Even given all the wildly diverging views of the nature of grammars held by linguists, it is still difficult to imagine reasonable disagreement over the fact that a language's grammar constrains in some way the forms that belong to that language. Whether a grammar is deterministic or probabilistic is a question of how it restricts possible outputs. Given a particular meaning that is to be expressed, and once all constraints (for example syntactic, semantic, pragmatic, contextual) have applied, a deterministic grammar predicts a set (possibly singleton) of equally well-formed alternatives. On the other hand, a probabilistic grammar assigns a probability distribution to possible outputs, and concrete outputs are chosen from that distribution. Which form is actually chosen can only be predicted in terms of lower or higher probability, wherein the probability of a form is equal to its long-run relative frequency (at least in the limit).[2]

Probabilistic effects in grammar have been explored extensively for over twenty years in various ways (Bresnan 2007; Divjak & Dąbrowska & Arppe 2016; Gries 2017; Grafmiller et al. 2018; Wolk & Szmrecsanyi 2018). Like Engel et al. (2021), we see a need to investigate register as a probabilistic phenomenon, regardless of whether the register-dependent choice of forms is seen as part of grammar in a narrow sense or as driven by

[2] From a probabilistic perspective, a deterministic grammar that allows several equally grammatical forms without specifying any probabilities actually predicts that all forms should have the same probability. Further selection criteria may be attributed to external factors or performance effects. However, systematically resorting to external factors would require the very disciplined setup of delineation criteria to avoid convenient ad-hoc arguments. See Elman (2009) for a high-level discussion of related problems and empirical evidence against a strong separation of grammar from external systems.

external factors. It would be highly implausible to assume that a register licenses sets of equally acceptable or unacceptable forms with no differences in probability.

In Systemic-Functional Linguistics (SFL), register has always been treated as a probabilistic phenomenon (for example Halliday & Hasan 1989; Matthiessen 1993; more recently Neumann & Evert 2021). Halliday (1991) already commits to a probabilistic model, discussing the prediction of grammatical features across different registers as a process of identifying general patterns. However, Halliday (1991: 32) states that "[i]t is clear that the significance of such probabilities is not that they predict single instances. What is predicted is the general pattern." Taken literally, this statement is formally quirky inasmuch as one cannot predict the general pattern without also making predictions for single instances, albeit only with a certain probability (usually below 1), and not with certainty. Overal, Halliday nevertheless make it very clear that "'a register' is a tendency to select certain combinations of meanings with certain frequencies, and this can be formulated as the probabilities attached to grammatical systems" (Halliday 1991: 33) and that "register variation is the resetting of the probabilities in the grammar" (Halliday 1991: 37). While we do not commit to the specific semiotic framework of SFL, we consider our work to stand in a wider tradition that has its theoretical roots in this work.

In regard to multi-dimensional analysis (MDA), there are three important differences (in addition to the major epistemological difference discussed above) between Biber's and our approach stemming from our commitment to probabilism. First, we treat lexicogrammatical register-driven variation as fully probabilistic whereas in Biber's approaches it is commonly assumed that registers are deterministic with only the distribution of LGFs within the register being (implicitly, through the statistical methods used) treated as probabilistic. With Biber & Egbert & Keller (2020b), this has begun to shift, but no formal probabilistic model has been adopted in MDA. Second, and as an intrinsic element of a probabilistic model, we allow multiple registers to be instantiated in a document (or conversation, etc.). Even from every-day experience, it should be evident that speakers or writers cannot always be absolutely certain what the socially and functionally appropriate mode of expression is, and formal models should therefore take into account the possibility that texts or even single utterances instantiate multiple registers with different probabilities. Third, in accordance with the appropriate formal models (see Section 2.1.1) and given our applied computational linguistic focus, we stress that registers are latent dimensions of texts. We work with an unstructured collection of documents (large Web corpora) and attempt to discover those latent dimensions based on LGFs. Only then do we attempt to assign situational-functional characteristics to them. As will become obvious, this is in many regards a very different approach, even compared to recent developments in the same direction such as Biber & Egbert & Keller (2020b).[3]

2 Discovering registers

2.1 A probabilistic generative model of register variation

2.1.1 Production model

Any theory of registers should first and foremost specify a theoretical model of how speakers make decisions to use linguistic signs (words, morphological patterns, syntactic constructions, etc.) based on a situational-functional setting. Such a model must establish causal links between the situational-functional setting (as perceived by the speaker)

[3] It is not clear to us what exactly is meant by reference to a "continuous quantitative space of variation" (Biber 2019: 44), although it sounds like at least a partial commitment to probabilism.

and linguistic signs produced in the output, making any such theory generative. As already argued in Section 1.4, Halliday's model in SFL does indeed specify a quite complex generative model, while Biber's work in Multidimensional Analysis (Biber 1988; 2009a) has always been more or less unspecific with respect to the underlying generative mechanism.

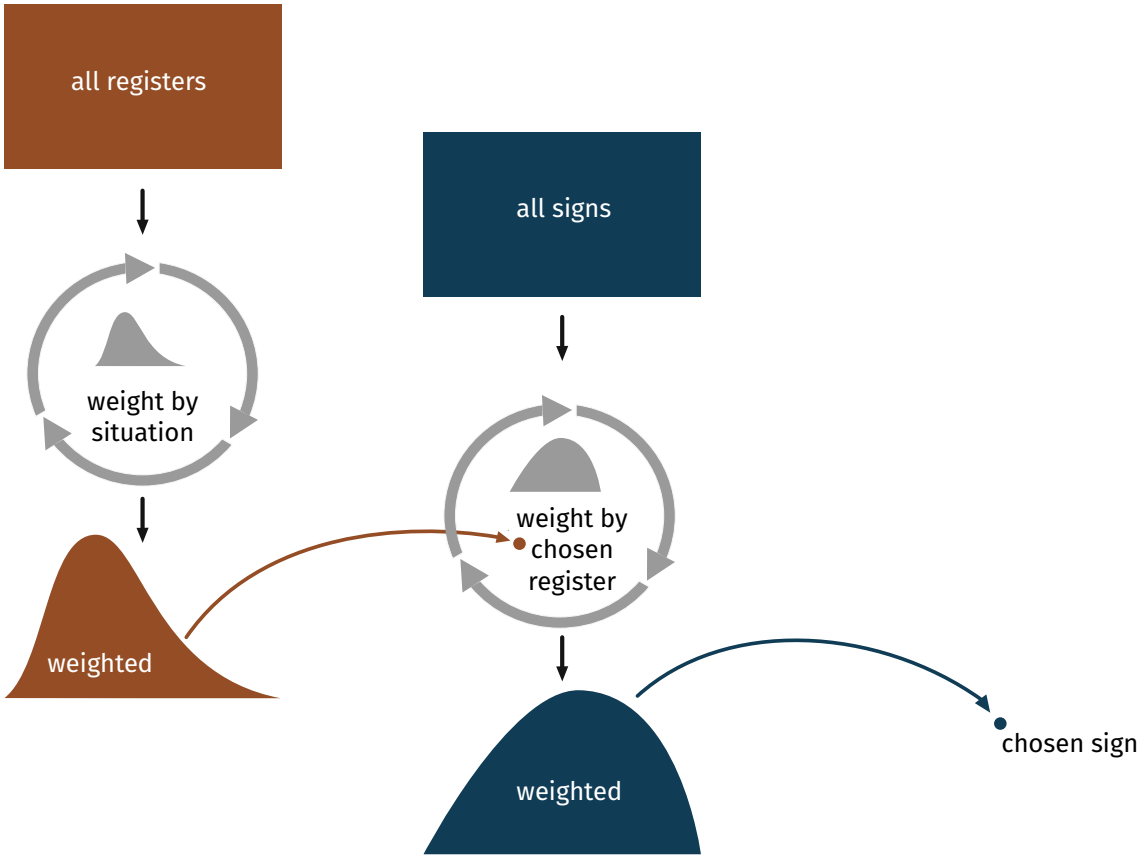

Figure 1: How speakers choose linguistic signs once the situational-functional properties have been set; from the weighted set of all registers, one is chosen probabilistically; its associated distribution over all signs is used to weight those signs, such that one specific sign can be chosen

In the following elaboration, please notice that we use the term *sign* here to denote any linguistic item, including words, morphological forms, constructions, etc. A sentence is thus a collection of signs (with a hierarchical structure) while also being a sign itself, and speaking and writing are just a matter of chosing the appropriate collections of signs to convey meaning or, more generally, communicate whatever the speaker wants to communicate. This view is customary in theories like Head-Driven Phrase Structure Grammar (HPSG), where all levels of linguistic description use a unified formalism (Pollard & Sag 1987; 1994; Müller et al. 2021). Figure 1 shows how we assume register-specific signs are chosen. Speakers start by assuming a uniform probability distribution over all registers (shown at the top left). Drawing on their assessment of situational-functional properties (and the previous discourse), they update this distribution, assigning some registers a higher probability and some a lower probability. From this weighted distribution, a single register is chosen probabilistically. We would like to stress that this does not mean that the register with the highest probability is chosen, but that this register merely has the best chance of being chosen. This effectively means that a mixture of registers is

determined even if the probability distribution over all registers is fixed for a whole text or conversation. Speakers and writers might chose a different register from the given distribution for two sentences in sequence.[4]

The chosen register consists of a probability distribution over all signs, which is used to update the probabilities of the set of those signs the speaker has available. In this simplified model, the speaker then choses a sign probabilistically from the updated distribution.[5] While this model is overly simplified (see immediately below), it also accounts for speakers' uncertainties in determining the appropriate register and in determining the concrete signs given that register.

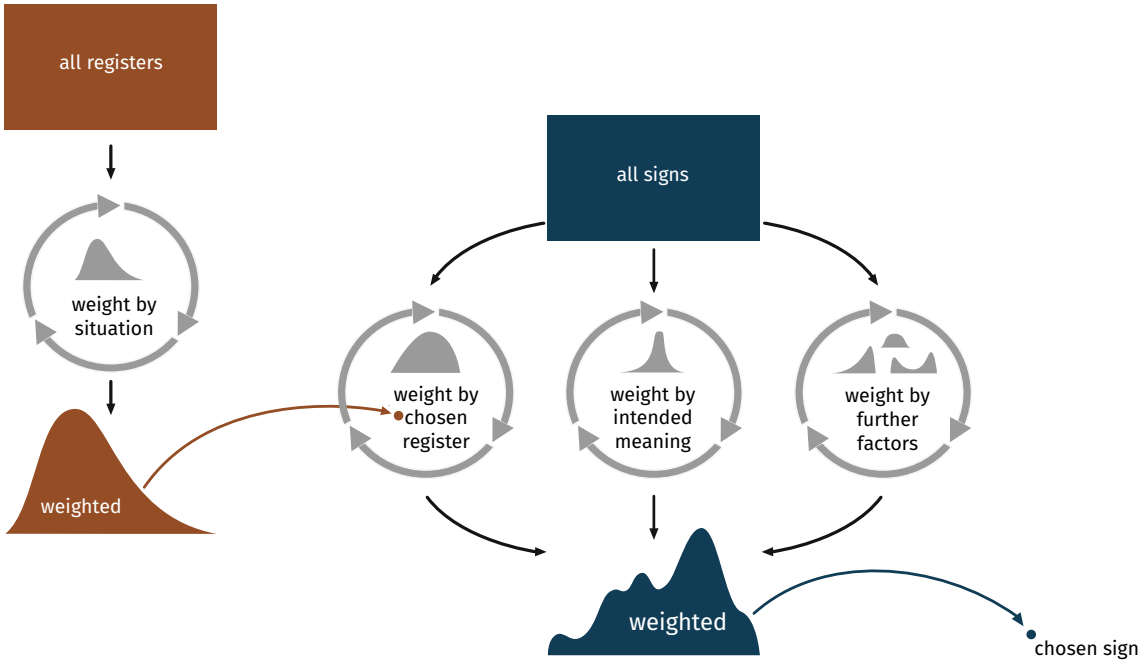

Figure 2: A general generative model of how speakers make lexical and grammatical choices

The process described above leaves out many aspects of chosing signs, and it focusses exclusively on the role played by register. Figure 2 suggests an actually much higher complexity. The intended meaning and a multitude of other factors related to morphology, syntax, pragmatics, etc. as well as the linguistic context also serve to update the probability distribution of all signs. The generative model proposed here thus does not serve just as a model of register variation but as a general model for alternation research in probabilistic linguistics.[6]

2.1.2 Model implementation

We now turn to the implementation of the model introduced above, specifiying the components to be used in the corpus-based inductive procedure. When the generative process

[4] To us, it seems as if even this was hinted at in Halliday (1991: 37–38).

[5] Again, this is compatible with many views expressed by Halliday: "Diatypic variation, or register, is variation in the probabilities of the grammar; this is the major resource for systematically construing variation in the environment (the 'situation'). Systematically, the grammar of a language can be represented paradigmatically as a network of choices, each choice consisting of a small number of alternatives related to probability; these probabilities appear to form a pattern related to the construing of 'information'." (Halliday 1991: 41)

[6] For the multitude of potentially relevant influencing factors see Gries (2017).

described in Section 2.1.1 generates documents in a given situational-functional setting (mixture of registers), we expect the probabilities assigned to signs under this setting to influence the occurrence frequencies of the signs in the documents. Thus, it is our task from a corpus analysis perspective – especially with large unstructured corpora – to infer the probabilities involved from the distribution of signs in documents.

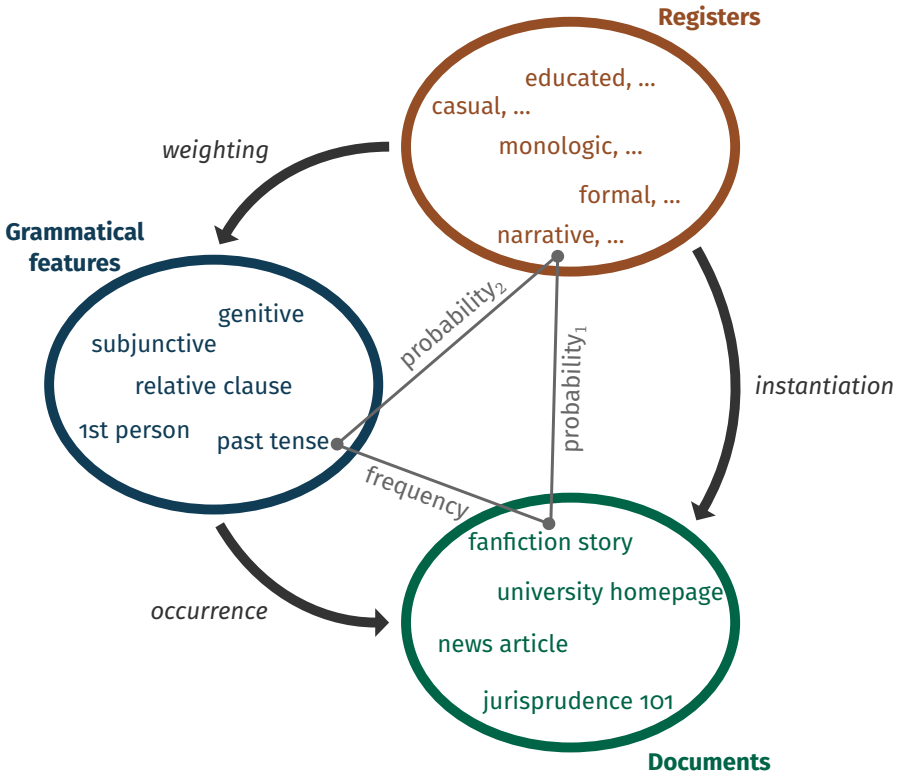

Figure 3: Illustration of a probabilistic model of the distribution of registers, grammatical features, and documents

Figure 3 illustrates the probabilistic mapping of registers, grammatical features, and documents.[7] A given register is assumed to be characterised by certain situational-functional parameters such as formality, narrativity, etc. Since there is no way of inferring the actual situation in which a document was written, we assume the corresponding weighting process to have taken place for given corpus documents such that each document is written under a given probability distribution of registers. A narrative register would, for example, be instantiated with a high probability – named *probability*$_1$ in Figure 3 – in a fanfiction story. At the same time, the register comes with a probability distribution over grammatical features, and each feature's probability is weighted accordingly (see *probability*$_2$ in Figure 3). For example, past tense might be assigned a high probability in a narrative register. Such a high proability leads to an increased frequency in the respective document. The fanfiction story should contain a high number of past tense forms by virtue of instantiating a narrative register which comes with a high probability for past tense forms.

As we will show, this is an appropriate model of the distribution of registers and grammatical features within documents. The situational-functional parameters (SFPs) are not a part of this implementation, but we will deal with them separately in Section 3. First,

[7] As we will motivate further below, we now speak of *grammatical features* rather than *signs*.

since we work with corpus documents written by many and sometimes multiple writers instead of single utterances by individual speakers, we assume at least some homogeneity within the speech community with respect to the probability distributions involved. In other words, we assume that speakers/writers in a community have acquired compatible register knowledge. Given both the situational-functional and conventionalised nature of language and especially register, this appears to be a justified assumption (see Schmid 2020; for a corresponding theory of the acquisition of conceptual and linguistic knowledge see Barsalou 2016). Second, we assume a fixed distribution of registers for each document for the following reasons. Mixtures of registers are expected under our model, simply because our probabilistic approach allows writers to choose registers at each point in the communication from the probability distribution and choose grammatical features accordingly. From a document perspective, there is no distinction between a register shift in the middle of the document and a probability distribution for registers that is set appropriately for the whole document to produce the observed forms. Similarly, for documents authored by multiple people, there is no distinction between each speaker having their own probability distribution for registers or a wholistically set distribution that produces similar outcomes. Distinguishing between these cases is reserved for future research.[8] Third, as indicated, we work with the notion of *grammatical feature* rather than *sign*. A full set of signs would include all lexemes of the language as well as many fine-grained construction types [the complexity of the resulting sortal hierarchies for German is evident even in overviews such as (Müller 1999; 2013). Including all kinds of lexemes would blur the distinction between registers and topics. While we assume a strong link between registers and topics (as it is assumed prominently, for example, in the field/tenor/mode model by Halliday, see Halliday 1978; Halliday & Hasan 1989; Lukin et al. 2008), it seems wise to separate the two for the time being. Including numerous fine-grained linguistic construction types, on the other hand, would make the model highly dependent on a specific morphological and syntactic framework, and it could potentially harm the inductive method chosen for our corpus analysis.[9]

Expressed more formally, the set of registers R_k and the probabilities from registers to documents $p_{R_{kj}}$ (***probability***$_1$ in Figure 3) as well as the probabilities from registers to grammatical features $p_{G_{ki}}$ (***probability***$_2$ in Figure 3) are intrinsically unknown in corpora of written language. The only observables are the grammatical features G_i, the documents D_j, and the occurrence frequencies of features within documents $f_{G_{ij}}$. Thus, our aim is to infer the sets of R_k, $p_{R_{kj}}$, $p_{G_{ki}}$ from the observable sets of G_i, D_j, and $f_{G_{ij}}$. Certain Bayesian models are indeed able to infer these sets from the observables. Among these models is Latent Dirichlet Allocation as popularised in topic modelling. In topic modelling, it is assumed that documents instantiate mixtures of topics with given probabilities. The topics themselves are characterised by the probabilities they assign to lexical words, which occur with measurable frequency in documents. The situation is parallel to ours: registers correspond to topics, grammatical features to lexical words, topics to registers, and the only observables are the words, the documents, and the occurrence frequency of words in documents. The topics and the probabilities of words given topics and topics given documents have to be inferred. This is not the place to discuss how LDA works mathematically and algorithmically. An accessible introduction for readers

[8] Register shifts have also been discussed in SFL, for example in O'Donnell (2021).

[9] See Section 2.3 for the relatively weak informativity of some low-level dependency features, which supports our decision to primarily adopt more high-level grammatical features.

without a background in maths is Blei (2012), and the maths is introduced in Blei & Ng & Jordan (2003).[10]

There are two related caveats to consider: one conceptual, and one technical. Conceptually, the latent dimensions inferred by LDA via the grammatical features are not guaranteed to represent true registers, which is why we call them *potential registers*, *pregisters* for short. We will perform additional steps to filter out those pregisters which represent spurious results rather than actual registers. On the technical side, due to the way the LDA algorithm works, the number of latent dimensions to be discovered has to be pre-specified. Tweaking the number is a balancing act between the performance of the algorithm and the interpretability of the results. However, the process of filtering out spurious pregisters should take care of this element of arbitrariness as well.

2.2 Feature extraction and corpus choice

Models like LDA treat documents as bags of features. In topic modelling, the features are the lexical words occurring in the document.[11] For first-generation register modelling, we suggest treating documents as bags of grammatical features (or higher-level signs). Extracting these features requires a deeper linguistic annotation as is required for lexical words. Furthermore, the LDA algorithm is not optimised to work with very small feature sets. This section provides a brief description of the feature engineering and the 1,613 features used.

We extracted two subsets of features. First, features similar to those used in MDA were annotated using a dedicated piece of software mostly written in Python.[12] It reads in a corpus consisting of one or more documents and produces a reduced rendering of the document as a sequence of tags standing for grammatical features without the actual text.[13] The tags correspond to standard linguistic concepts, such as occurrences of particular parts-of-speech, periphrastic passive and perfect constructions, non-standard morphological forms, etc. Currently, the software extracts 69 such features. Most of the underlying linguistic annotations used to create these bags of tags were not performed by the software directly. Rather, it relies on freely available tools, which it conveniently wraps in Python code, creating a fully automated toolchain. Such tools are used to annotate part-of-speech tags (STTS, Schiller et al. 1999), morphological features (MarMoT, Müller & Schmid & Schütze 2013), compounds (based on SMOR, Schmid & Fitschen & Heid 2004; Faaß & Heid & Schmid 2010), named entities (Stanford NER tagger, Finkel & Grenager & Manning 2005), topological structures (Berkeley parser, Petrov & Klein 2007; Cheung & Penn 2009; Telljohann et al. 2012), and dependency structures (Mate-Tools, Bohnet 2010).

Second, the feature set was augmented considerably by including a large number of dependency bigrams derived from the aforementioned syntactic dependency parses. Representing syntactic dependencies as a graph with tokens as nodes and dependency relations as edges, we can traverse the tree and record sequences of the encountered relations. Such

[10] LDA has been used in a wide variety of fields beyond standard topic modelling such as genetics (Pritchard & Stephens & Donnelly 2000), and clinical psychology (Chiu & Clark & Leigh 2022). See Jelodar et al. (2019) for one of many overview articles.

[11] Usually, they are filtered by frequency, stemmatised or lemmatised, or pre-processed in some other way to optimise the results of the algorithm.

[12] The code will be released under an open license after the paper is accepted for publication.

[13] Therefore, documents cannot be reconstructed from these renderings. Not even the topic or subject of the document can be recovered, which makes some results presented in Section 3.2 quite impressive. For example, LDA found a register strongly related to topics from jurisprudence and law exclusively based on grammatical features.

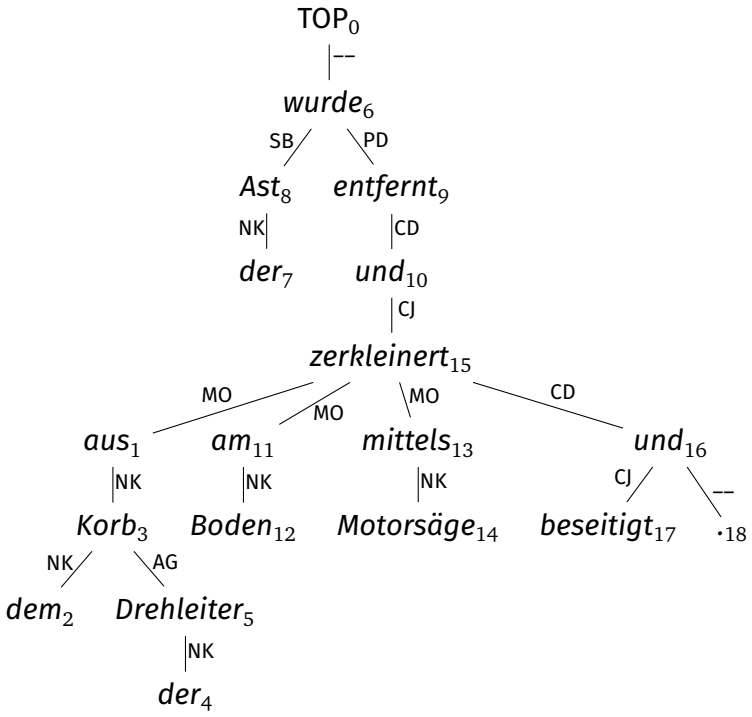

Figure 4: A dependency tree from DECOW16B

sequences are dependency bigrams if their length is 2.[14] For instance, take sentence (1) with its dependency tree shown in Figure 4.[15]

(1) Aus dem Korb der Drehleiter wurde der Ast entfernt und am
Out of the basket of the turnable ladder was the branch removed and on the
Boden mittels Motorsäge zerkleinert und beseitigt.
floor by chainsaw chopped and disposed of
The branch was removed from the basket of the turnable ladder and then chopped into pieces on the floor by chainsaw and finally disposed of.

The set of dependency bigrams corresponding to this sentence is `--` < `SB`, `SB` < `NK`, `--` < `PD`, `PD` < `CD`, `CD` < `CJ`, ….[16] Each type of bigram is treated as an individual feature, totalling 1,562.

The software was used to annotate a subcorpus of 630,899 documents from the German DECOW16B web corpus (Schäfer & Bildhauer 2012; Schäfer 2015). Documents were chosen if they contained more than 999 tokens, and if they were likely to be of good text quality, which was operationalised as follows. Each document in the corpus contains a *Badness* score (Schäfer & Barbaresi & Bildhauer 2013), and we only used documents from the top five quality strata. Still, the data can be expected to be noisy to a certain extent, which might seem like a disadvantage. Compare the implied superiority of hand-picked and carefully designed corpora in the following quote from Halliday (1991).

[14] The concept extends to n-grams for an arbitrary n instead of 2.

[15] Sentence from DECOW16B, ID: dc45880f042c40e1d484adc52abf30573095:72, source: *www.feuerwehr-badgrund.de*.

[16] The relations are encoded in their standard abbreviations in the tree. `SB` stands for *subject*, `NK` for *noun kernel*, `PD` for *predicative*, `CD` for *coordinating conjunction*, `CJ` for *conjunct*, and so forth. In later sections, these abbreviations are given in human-readable form. We use `A < B` to indicate that `A` is above `B` in the tree.

	Preg 0	Preg 1	Preg 2	Preg 3	Preg 4	Preg 5	Preg 6	Preg 7	Preg 8	Preg 9	...
Doc 1	0	0.07	0.01	0	0	0.13	0.03	0	0.41	0	...
Doc 2	0	0	0	0	0.03	0.03	0	0	0.06	0.02	...
Doc 3	0	0	0.03	0.05	0.01	0.06	0	0.05	0.14	0.02	...
Doc 4	0	0	0.05	0.02	0.08	0.07	0	0.02	0.03	0.05	...
Doc 5	0	0	0.01	0	0.02	0.04	0	0	0.02	0	...
Doc 6	0	0.03	0	0	0.01	0.02	0	0	0.01	0	...
Doc 7	0	0.02	0	0.01	0.06	0.03	0	0.02	0.04	0.03	...
Doc 8	0	0.06	0.09	0.02	0	0.04	0	0	0.04	0	...
Doc 9	0	0.03	0.01	0.01	0.07	0.09	0.01	0.01	0.06	0.16	...
Doc 10	0	0.01	0	0.01	0.07	0.09	0	0	0.04	0.02	...
...	...	...	...	...	...	...	...	...	...	...	...

Table 1: First ten rows and first ten columns of the document-pregister-matrix

> A corpus which is organised by register, as all the great first-generation ones have been, makes it possible to study such external conditioning of probabilities, and to show how the grammar of doing science differs quantitatively from that of telling stories, advertising and so on. (Halliday 1991: 38)

To us, working with noisy data represents a welcome challenge as we have found more than once that large and slightly noisy corpora contain relevant material that is not usually found in cleaner corpora. Also, we would like to stress that one disadvantage of using pre-annotated intentionally stratified corpora (designed in the spirit of Biber 1993) means that one can only research those register distinctions which the corpus designers deemed relevant. As laid out in Section 1.1, however, it is also an important test of a theory of register whether it provides a means of recovering all register distinctions that can possibly be recovered from unknown texts. In Section 2.3, we report the results of applying LDA to achieve this.

2.3 LDA run and results

In this section, we describe the results of the LDA and the discovered pregisters. Then, Section 3 will describe the additional step in the analysis ensuring that we found actual registers and not artefacts. The LDA algorithm terminated with the number of latent dimensions (pregisters) set to 25.[17] Additional runs with 50 and 100 pregisters led to obvious redundancy, where very similar pregisters were found, as was determined by qualitative inspection, comparing the most typical documents for each pregister. At 50 or 100 pregisters, there was a significant overlap in the lists of typical documents between many pregisters.[18]

The output of LDA can be understood as two two-dimensional matrices. Both put the latent features (in our case the 25 pregisters) on one dimension. The other dimension is that of the observable features (1,631 grammatical features) in one matrix and that of the documents (in our case 630,899 documents from DECOW16B) in the other matrix. Table 1 illustrates the upper-left corner of the document-pregister matrix (actual

[17] We used the LDA implementation in the Gensim library Řehůřek & Sojka (2010). Configurable hyperparameters (other than the number of pregisters to infer, mostly α and β) were left at their default after a series of experiments which did not seem to improve the results.

[18] The results of the alternative LDA runs are included in the data package.

values for the first ten documents and the first ten pregisters). Each coordinate gives the probability that the respective document belongs to the respective pregister. In terms of Section 2.1.1, this is $p_{R_{kj}}$. The other (superficially similar) matrix gives the probability for each feature in each pregister. In terms of Section 2.1.1, this is $p_{G_{ki}}$. Thus, we have obtained a set of potential registers and the corresponding probabilities that link the registers to the grammatical features and the documents.

A first step in the interpretation is to look at the features associated with each of the pregisters at a probability above 0. By way of visualisation, we provide Figures 5 through 7. For three selected pregisters (numbered 6, 8, and 24), these figures plot clouds of the corresponding feature labels, and the font size corresponds to the feature's probability.

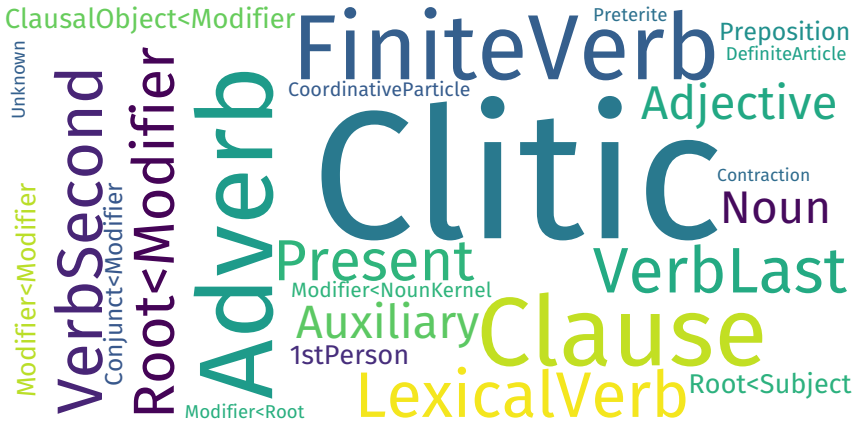

Figure 5: Feature cloud for pregister 6

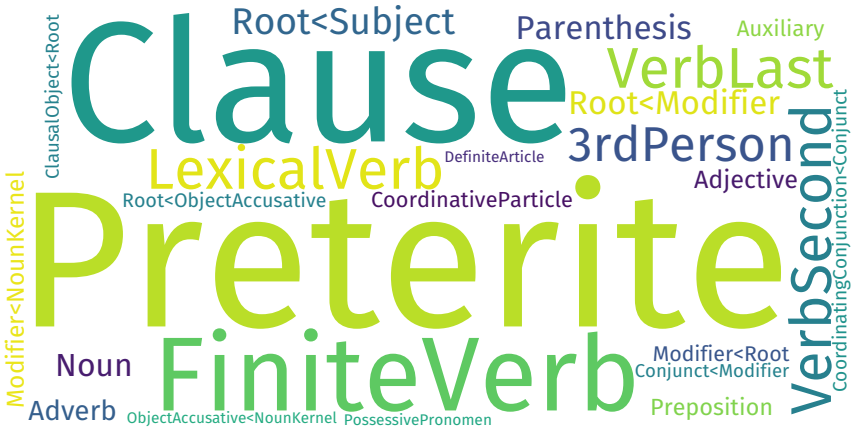

Figure 6: Feature cloud for pregister 8

Pregister 6 is dominated by clitics (which are strongly associated with informal writing in German, Schäfer & Sayatz 2014), adverbs, finite verbs, as well as present tense forms, and it contains a relatively high number of clausal constructions. Pregister 8 differs most prominently in that it is characterised even more strongly by finite verbs, and they often take preterite tense. Pregister 24 appears to be different altogether in that it is characterised by nouns, especially in the genitive and with the definite article. While it is tempting to speculate about such registers and why they might contain the grammatical features they contain, we defer the analysis to Section 3, where we describe the rigid annotation process that led us to a full analysis of the LDA results.

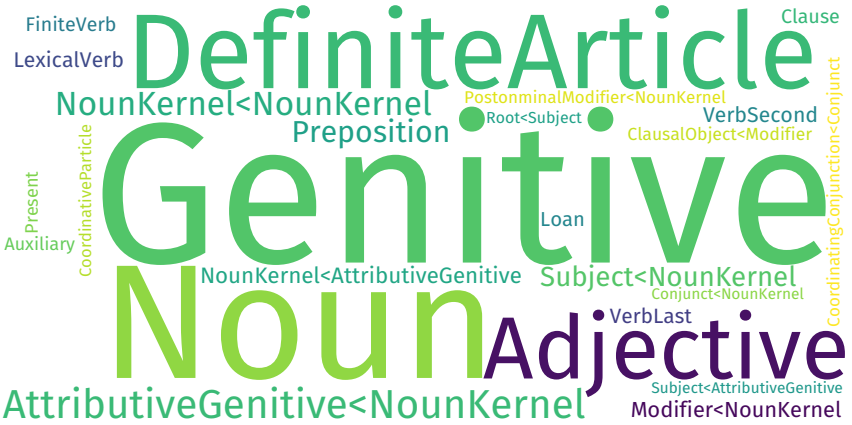

Figure 7: Feature cloud for pregister 24

The low-level features from the dependency parser do not appear prominently in the visualisation. However, they are not completely lost. Figure 5 shows that sentential modifiers are overproportionally frequent in pregister 6 in the form of the high probability for the feature `Root < Modifier`. In Figure 7, features like `NounKernel < NounKernel` (nouns modifying nouns) and `AttributiveGenitive < NounKernel` (genitive attributes) support the impression that complex NP syntax plays an important role in pregister 24. Overall, the mean probability assigned to dependency features is 0.019 compared with 0.035 for the high-level features.

3 Situational-functional parameters

3.1 Annotation scheme

As explained in Section 1.3, instead of describing registers as macro-categories, we break down their situational-functional aspect into parameters (situational-functional parameters, SFPs). This section briefly describes the annotation scheme and the inter-annotator agreement achieved in our extensive test and production runs.[19] It cannot be stressed enough that the annotation is completely independent of and agnostic towards the LDA results and LGFs in general. It represents a completely separate step in which we looked at documents (not registers), assigning SFPs to the documents. Therefore, combining the LDA results with the SFP annotation (Section 3.2) represents a validation of both independent steps as well as our theoretical assumptions.

The annotation scheme used in the main production works with a limited set of parameters, which is tailored specifically for the kind of Web data we encountered in our data set. The parameters and their possible values are as follows:

- Education: Yes, No
- Interaction: Yes, No
- Proximity: Yes, Default
- Narration: Yes, No

We will discuss each of these parameters presently. However, we like to add a brief excursion on how this reduced set of parameters came to being used. Originally, we started

[19] The full guidelines will be made available as part of the data package for this paper.

out with a much more detailed annotation scheme inspired by Halliday (1978), Lukin et al. (2008), Biber & Conrad (2009), and others. It comprised twelve SFPs, namely the number of – and interaction between – the interactors (see below), the identifiability of an audience (in the sense of Bell 1984), whether the communication was cooporative, potential gradience of power among the writers or the reader and the writer (Gotzner & Mazzarella 2021), attempted manipulation (such as in advertisements), identifiability of a beneficiary of the communication, task-orientation, politeness/formality of the tone (see below), the level of emotionality, requirement of an elevated educational background (see below), metalinguistic discussion, narrativity (see below) and/or performed interactor roles (theatre plays, TV scripts, etc.). However, many of them turned out to be impossible to retrieve from large unstructured corpora for two reasons. First, this taxonomy was too detailed in that the total number of documents instantiating one or more of many parameters' values was too low for samples of typical size. Second, we could not achieve a high enough inter-rater agreement (measured as Fleiss' κ for three to four raters) on many parameters in a total of four annotation test runs.[20] Specifically, we required each parameter value to be instantiated in at least 5 of 100 documents and inter-rater agreement to be at $\kappa \geq 0.5$ after two iterations of refining the guidelines. Instead of chasing categories that are virtually non-existing in Web corpora or cannot be operationalised with any reasonable precision, we therefore focused on the four categories named above, and which we now discuss in turn. The annotation guidelines to be released as part of the data package go into more detail with respect to the operationalisation.[21]

3.1.1 Education

This parameter with its possible values *Yes* and *No* encodes whether a document requires an elevated educational background (EEB) on the side of the reader. Existing research on *Bildungssprache* ('educated language') from the German linguistic tradition provided the motivation for this parameter (Gogolin & Lange 2011; Feilke 2012). The notion of *Bildungssprache* is related (but not identical) to Cummins' *Cognitive academic language proficiency* (Cummins 2008), and it defines a communicative purpose, typical usage situations, and linguistic means. EEB is required in situations where complex matters involving non-factual or counter-factual or hypothetical information, causal or temporal sequences, decontextualised and depersonalised states of affairs have to be expressed, where intensional definitions have to be given, where arguments have to be made, and so on. EEB and the corresponding registers are not exclusively tied to situations in academia or schools, although those contexts often require EEB. While such communicative situations can be easily identified by concrete lexical and grammatical features (such as complex NP syntax, complementation, modal verbs, etc.), we did not specify these in the annotation guidelines to avoid circularity between the form-based LDA and the situational-functional annotation. Instead we named four communicative functions (according to Feilke 2012) which are typical of EBB:

i. Explication: Does the situation require expressions of clarification, specificity, and disambiguation?

[20] The set of annotators contained the first, second, and fourth authors of this paper and one student assistant. Not all annotators were active in each test run.

[21] In Halliday & Hasan (1976)'s terms, Education and Narration should be part of the *field* of discourse, Proximity contributes to the *tenor* of the discourse, whereas Interaction is a component of *mode*.

ii. Condensation: Does the situation require complex propositional content to be expressed?
iii. Generalisation: Does the situation require expressions of interpersonal and universally relevant matters detached from the present situation temporally, spatially, or by epistemic status?
iv. Discussion: Does the situation require writers to evaluate pros and cons or react to other people's opinions?

These functions usually require EEB, and the means to express them are acquired at a relatively late age in primary education. In case of doubt, annotators were therefore instructed to consider whether preschoolers could in principle follow the documents (a clear case of *Education = No*) or whether the cognitive and linguistic proficiency typically acquired in at least primary education was required (a clear case of *Education = Yes*).

3.1.2 Interaction

Interaction with its possible values *Yes* and *No* appears to be easy to annotate. This parameter is not supposed to encode an interactive attitude (as, for example, in Paolillo 2000: 239) but merely whether there is more than one interlocutor involved in producing the document (more like *purpose* in Egbert & Biber & Davies 2015). Hence, we classify a document as interactive if multiple agents are directly engaged in verbal interaction with one another. Both parties have to be actively involved in the verbal interaction, and both have to be cognitive agents (humans, groups of humans, or institutions led by humans). Typically, this is the case for interviews, forum threads with multiple contributors, blogs with a discussion section, minutes that actually record people's interactions and not just summaries of the results (for example from parliaments or conventions), and so on.

3.1.3 Proximity

Proximity (which we annotated with the values *Yes* and *Default*, see below) and distance are related to informality and formality, where formality as an important aspect of register was already deemed relevant by Labov (1966). However, they are not the same, and we consider *proximity* and its opposite *distance* to be more holistic notions than formality encompassing formality, politeness, and related concepts.[22] Based on Koch & Oesterreicher (1985); Ágel & Hennig (2006); Feilke & Hennig (2016); Hennig & Feilke (2016); Oesterreicher & Koch (2016) and many others, we define a proximal situation as one where the interlocutors are close to each other, prototypically friends and family or friendly co-workers. We also included imagined interactions, for example with gods or deities, in sermons, etc. Such situations can often also be described as proximal, and interlocutors usually make use of linguistic and extralinguistic markers of proximity. Proximity in German is often (albeit not exclusively) recognisable by phrases that are distinctly marked for proximity, such as the second-person pronoun *du* instead of *Sie*. Since this is very obvious, we included it in the annotation guidelines despite the fact that it introduces a blurring between formal features and situational parameters. Many corpus documents are not proximal but at the same time not distal, especially if they do not address the reader and are, for example, just informational or narrative (see also

[22] Formality has received a lot of attention in applied computational linguistics, and it has been suggested that it can be measured on a continuous scale even at the sentence level, see Eder & Krieg-Holz & Wiegand (2023) and references cited therein. We reserve experiments with annotations on continuous scales for the future.

Neumann 2014: 67). Hence, we described the Proximity parameters to annotators as having the values *Yes* and *Default* (instead of *No*).

3.1.4 Narration

The final parameter is Narration with its possible values *Yes* and *No*. Narration has been dealt with as a complex category in register research and related fields (Zeman 2018; Egbert & Mahlberg 2020), and a detailed scheme for annotating narration has recently been proposed in Lehmann et al. (2023). The guidelines by Lehmann et al. (2023) allow fine-grained decisions and, in principle, the annotation of changes within a document from narrative to non-narrative passages and vice versa. In our data set, it appeared to be relatively easy to determine whether a document was predominantly narrative, and we did not encounter documents with significant back-and-forth between narrative and non-narrative passages. Based on the aforementioned work, we therefore used simplified guidelines based on a few mostly noncontroversial properties of narrative texts. As it turned out, providing only the definition "The document's main purpose is to tell a story or stories." results in a substantial inter-rater agreement. To catch some borderline cases typical of our data set, we excluded certain types of Web documents explicitly, such as on-the-fly story building by multiple authors via role play or similar activities, reports on actual past events, telling the author's life's story, etc. We specified further that the narrated events should be put in temporal, logical, or causal sequence (as opposed to free and random association). The document should be coherent and have a recognisable theme that structures the narrative text. Finally, narrative texts should be *about* something at the macro level that transcends the particular events.

We would like to stress that this inventory of four parameters is the subset from the set of parameters which we would have liked to annotate based on theoretical assumptions and which can be operationalised and successfully annotated for large collections of Web texts. We measured good to very good inter-rater agreement on the final data, which provides evidence that the annotation scheme was consistent. In Section 3.2, we present the results of the annotation.

3.2 Annotation process and results

3.2.1 Goal and overall approach

The LDA gave us 25 pregisters characterised by linguistic signs or rather LGFs. To map these pregisters onto SFPs and interpret them as actual registers, human annotation of those parameters is required. However, it is not feasible for human raters to annotate whole pregisters by looking at the contained documents. The LDA assigns a probability to each pregister in each document. Hence, we let raters annotate individual documents for Education, Interaction, Proximity, and Narration (see Section 3.1). The annotated categories for each document were then used to calculate a score for each pregister, using both the categories assigned to the documents by the raters and the probabilities of the pregister for the document. We now report the details of the annotation process and the results.

3.2.2 Final dataset and quality of annotation

The final dataset for manual annotation consisted of the 30 LDA-top-ranked documents for each pregister ($25 \cdot 30 = 750$ documents) and 600 randomly selected additional documents. The additional random selection of documents was included to avoid overtraining

	Education	Interaction	Proximity	Narration
R1, R3	0.91	0.97	0.92	0.88
R1, R2, R3	0.73	0.95	0.87	0.82
R1, R3, R4	0.74	0.95	0.89	0.78
R3, R4	0.69	0.94	0.92	0.72
R2, R3	0.64	0.93	0.87	0.80
R1, R2, R3, R4	**0.67**	**0.94**	**0.87**	**0.76**
R1, R2	0.64	0.93	0.82	0.77
R2, R3, R4	0.62	0.93	0.87	0.72
R1, R4	0.59	0.93	0.84	0.74
R1, R2, R4	0.59	0.92	0.83	0.71
R2, R4	0.52	0.91	0.83	0.63

Table 2: Inter-rater agreement (Fleiss' κ) for the final dataset and all possible permutations of raters, sorted by their mean agreement across all four annotated parameters; the highlighted row is the agreement between all four raters

on highly typical documents. Thus, they increase the external validity of the findings. The dataset was annotated by four annotators (the first, second, and fourth author and a student assistant). Each rater coded 250 documents exclusively and 350 documents which were annotated by all four raters for quality control, resulting in the total of 1,350 documents. The documents in the final sample had not been annotated previously by any of the four raters, the raters did not communicate about the guidelines or specific documents during the annotation process, and raters were unaware as to which documents were annotated by only themselves or the whole group.

Inter-rater agreement was measured on the 350 documents as Fleiss' κ, see Table 2. The group of four raters reaches Education $\kappa = 0.67$, Interaction $\kappa = 0.94$, Proximity $\kappa = 0.87$, and Narration $\kappa = 0.76$. These are all acceptable, good, or very good. It should be noticed that R1 and R3 agree very well on Education ($\kappa = 0.91$), where R2 and R4 agree only moderately ($\kappa = 0.52$). The involvement of R2 and R4 in any of the pairwise values for Education significantly lowers agreement (from R3 and R4 with $\kappa = 0.69$ to R1 and R4 with $\kappa = 0.59$). Apparently, R2 and R4 diverged in their understanding of the guidelines for Education without substituting their own systematic interpretation of the category. Further strengthening of the guidelines is therefore required for future work, while the obtained result in this study is still acceptable. In sum, the annotation was successful at least in being consistent. However, Section 3.2.3 projects the results onto the LDA results, showing that it is more than just consistent.

3.2.3 Mapping SFPs onto pregisters

All document-level annotations are binary (Yes/No or Yes/Default).[23] To arrive at a score for Education, Interaction, Proximity, and Narration, we proceeded as follows for each pregister. If a document was annotated as Yes for a parameter, we multiplied the

[23] For the documents annotated by all four raters, the final dataset contains majority decisions or, if there was a tie between two and two raters, a random decision.

pregister's probability for the document by 1, else by 0, adding up all resulting values for the pregister.[24]

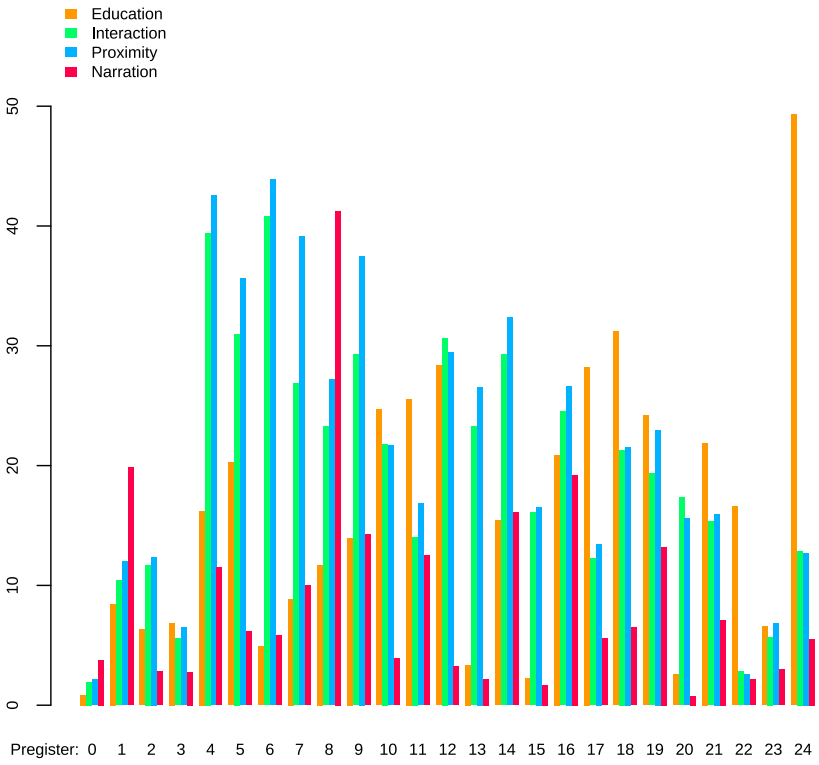

Figure 8: Scores for each SFP in each pregister

Figure 8 visualises the results, and a table with the numeric results can be found in the data package. Each pregister is assigned four scores, and they range from well below 10 to around 50. Clear patterns emerge.[25] Many pregisters have a strong correlation between high scores for Interaction and Proximity (both around 30 or higher for pregisters 4, 5, 6, 9, 12, 14), which is likely due to the many forum documents in these pregisters. Pregisters 4, 9, and 14 contain a lot of forums with emotional storytelling, including travelogues (14). Pregisters 6 and 12 contain forums with casual information exchange and discussions. Pregister 8 is less interactive but still proximal, and it contains forum documents where opinions and advice are shared rather than discussed. Pregisters with low scores for Interaction and Proximity are 1 (containing among other things biographies and sports reports), 22 (technical information), and 24 (laws and legal discussion). Education appears to be mostly independent of the other parameters. High scores (above approximately 30) are assigned to 12 (a lot of philosophy forums), 17 (business

[24] These scores are not probabilities. However, note that our model (Section 2.1.1) is agnostic with respect to how the SFPs are weighted. We leave a more precise formulation for future research and use the score calculated here as a first exploratory approximation.

[25] The corpus of documents is part of the data package such that the following characterisations can be verified.

information), 18 (technical information, user guides, etc.), and 24 (laws and legal discussion). Like Education, Narration is also highly independent of the other parameters. The prominent narrative pregisters (above 20) are 1 (biographies, sports reports), 8 (all sorts of storytelling), and 16 (short tales and soap opera summaries).

As expected with LDA on noisy data, some pregisters (like topics in topic modelling) are associated with documents containing non-text or incoherent text like lists, and some have picked up documents mostly from only a small set of servers. These are not interpreted, in our case 0, 2, 3, 15, 20. The remaining 20 pregisters are interpretable as registers having clear lexico-grammatical and situational-functional characteristics. The three LDA pregisters illustrated as feature clouds in Section 2.3 are among the interpretable ones.[26] Figure 5 corresponds to pregister 6 (completely informal forum discussions), which is highly interactive and proximal, does not require EEB, is not narrative, and it is characterised by writers using a lot of clitics, adverbs, modifiers, present tense. Figure 6 illustrates pregister 8 (storytelling, mostly in forums) which combines a low EEB score with medium Interaction and Proximity, a very high Narration score on the situational-functional side with verb-second sentences, finite verbs (parataxis), and preterite forms on the lexico-grammatical side. Figure 7 represents pregister 24 (legal material) requiring EEB, being neither interactive nor proximal, and contains highly typical complex NP syntax and definite articles.

3.2.4 Further aggregation of the results

We have shown that unsupervised form-based register induction combined with a fully independent annotation for SFPs leads to plausible and interpretable results. This demonstrates that our model introduced in Section 2 provides a fruitful novel approach to discovering and analyzing registers, especially in large unstructured and partially noisy document collections. Since our model allows for mixtures of registers and because it is robust against feature dependencies, pregisters can be similar to each other in terms of LGFs and SFPs. For example, pregisters 12, 13, and 14 (see the data package for the LGF clouds and SFP scores) are all characterised by a high probability of adverb use, and they rank high on Interaction and Proximity. Pregister 12 has a high Education score, and 14 has a high Narration score, whereas 13 has low scores for both SFPs. These similarities and differences have a number of potential sources. It could be that the LDA overdifferentiated in some cases, such that some pregisters could and should be merged. Alternatively, registers could be hierarchically structured into super- and sub-registers. Other options include missing SFPs in the analysis as well as other factors (style, genre, topic) interacting with register or influencing the same set of LGFs as registers and creating spurious categories.[27]

[26] The data package provides feature clouds and situational-functional scores side-by-side for all pregisters except 0, 2, 3, 15, 20.

[27] Importantly, missing SFPs could be due to shortcomings of our annotation guidelines or to a general inaccessibility of some SFPs in written documents.

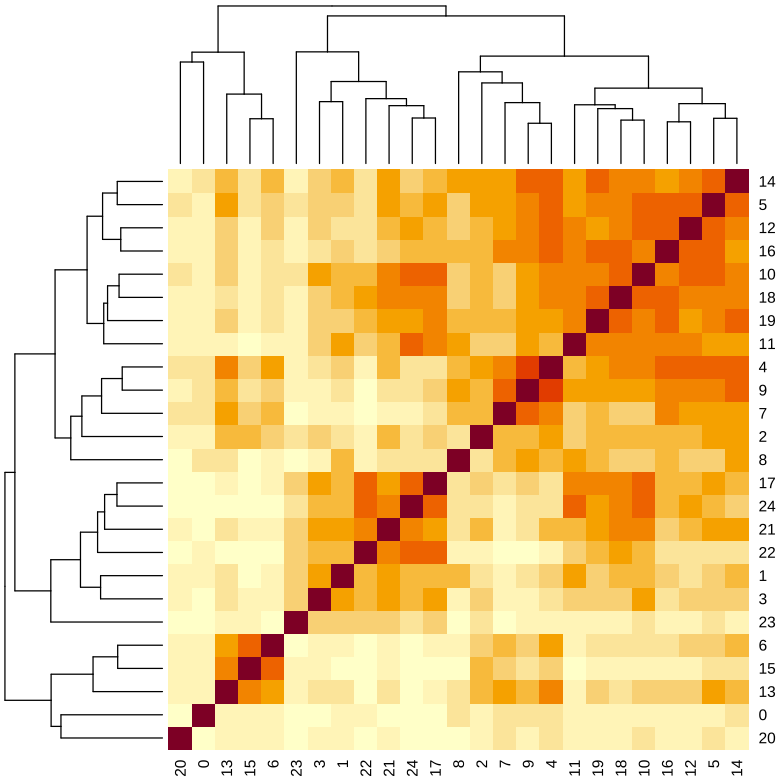

Figure 9: Similarity of pregisters in the LDA document space

Therefore, in order to aggregate the data further, we cluster the pregisters. The document-pregister-matrix (as illustrated in Table 1) allows us to use the probabilities for the 630,899 documents as features to arrive at clusters of pregisters which are similar in terms of their instantiation in concrete documents. By assumption, two pregisters are more similar to each other if they have similar (higher or lower) probabilities in the same documents. Using their cosine similarity within the document probability space, we arrive at a clustering of the pregisters shown with a heatmap in Figure 9.

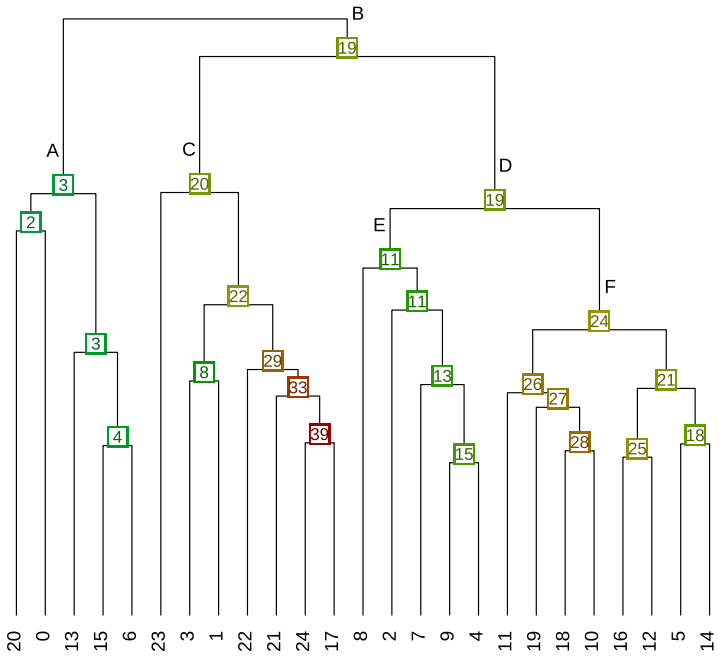

Figure 10: Mapping Education SFP scores onto clusters of pregisters

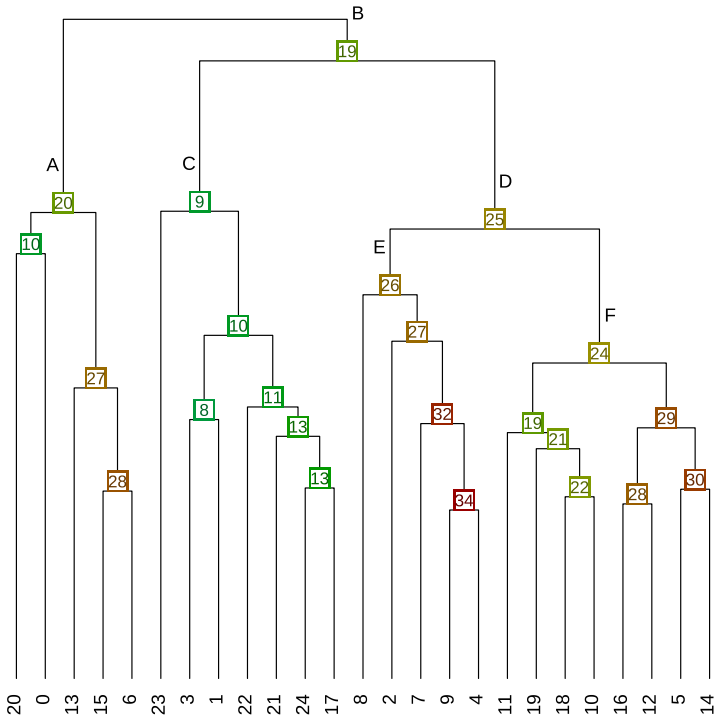

Figure 11: Mapping Interaction SFP scores onto clusters of pregisters

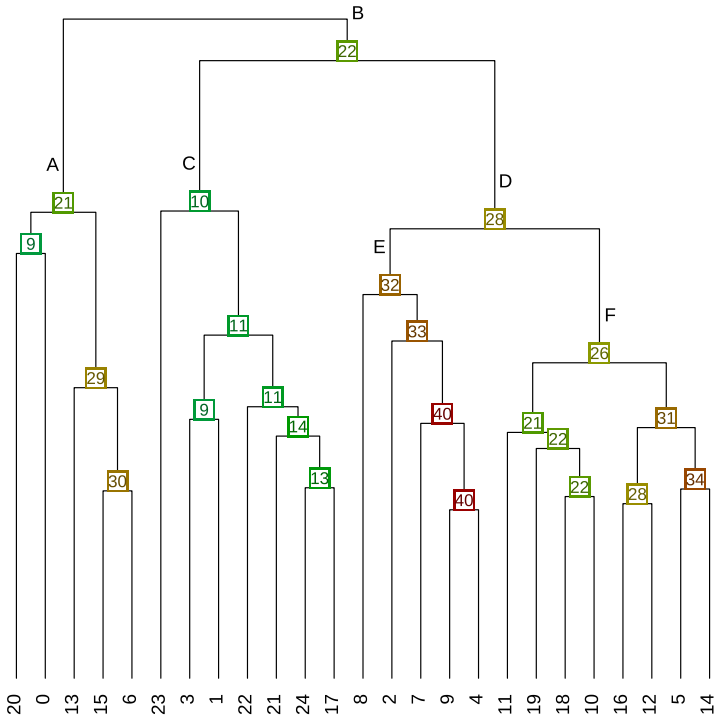

Figure 12: Mapping Proximity SFP scores onto clusters of pregisters

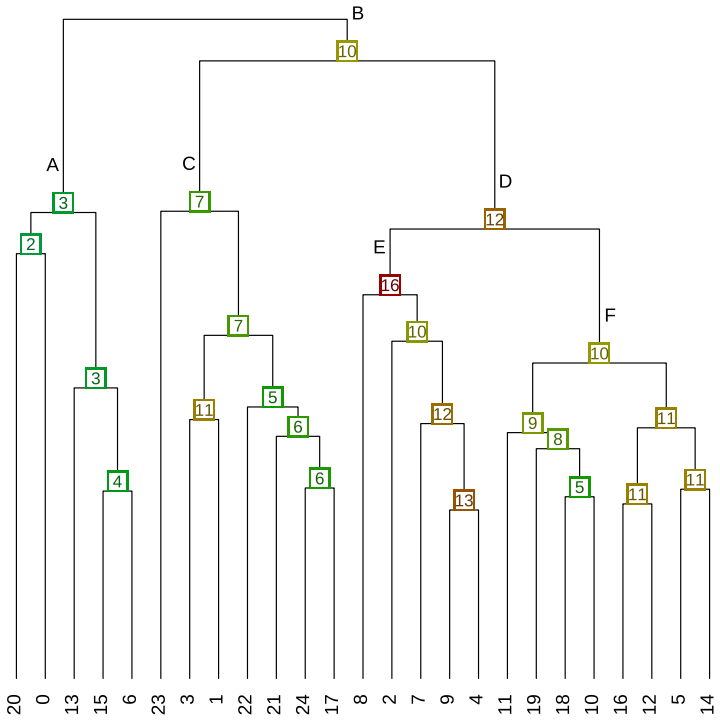

Figure 13: Mapping Narration SFP scores onto clusters of pregisters

The three highest splits produce six major clusters of pregisters labelled A through F in Figures 10–13. These show the cluster dendrogram with an overlay of the SFP scores. It must be kept in mind that the clustering is an aggregation of the LDA results, and that no data regarding SFPs has contributed to the formation of the clusters. Each node in the dendrogram was then annotated with the mean of the SFP scores of the pregisters contained in the cluster. Education scores are neatly correlated with the clusters: Cluster A has very low Education scores (mean 3), C has high scores (mean 20), and D (mean 19) is mixed. Its sub-cluster E is assigned mediocre Education scores (mean 11) and sub-cluster F has the highest Education scores (mean 24). Cluster A is also mostly interactive (mean 20), and again cluster C behaves very differently with respect to Interaction (mean 9). Interaction scores are also high for cluster D (mean 25) and both of its sub-clusters. The clusters also have clear preferences for Proximity and Narration.

It might be tempting to assign further human-understandable labels to clusters. One could look at cluster A, for example, which has low scores for Education and Narration and mid to high scores for Interaction and Proximity. The obvious conclusion (knowing that the data come from a Web corpus) is that the registers in cluster A are typical forums where people discuss everyday matters. However, this would be unnecessary and potentially even doubtful from a theoretical viewpoint. All relevant information about the register should ultimately be encoded in LGFs and SFPs, and by adding labels (mostly by inspection and introspection on the side of the researcher) one risks introducing biases and overgeneralisations. Also, such high-level labels might blur the distinction between register on the one hand and genre, text type, style, topic, and so forth on the other hand.

In sum, the fact that LDA clusters based on LGFs and the SFP scores match so well is another result that corroborates our theoretical assumptions and the methods used. The results of the clustering can be taken as a good indicator that a hierarchical modelling of registers might be more adequate because registers are indeed organised hierarchically. We return to this presently in Section 4.

4 The future of registers as probabilistic categories

We have successfully introduced a method of discovering and modelling registers for any type of corpus of written data, including large, unstructured, and noisy corpora. The method is based on a simple but truly probabilistic generative theory of register variation.

Future further ways of validation as well as extensions of this approach suggest themselves. First, we made it clear that the set of SFPs might not be specific enough. On the other hand, we achieved a high quality of annotation (measured as inter-rater agreement) due to this reductionist approach. However, future annotation attempts should be directed towards annotating more operationalisable SFPs if possible. Second, as we have argued in Section 3.2.4, registers might have a hierarchical structure, so hierarchical variants of LDA could be usefully applied (Blei et al. 2003; Blei & Griffiths & Jordan 2010). Third, we pointed out in Section 2.1.2 that it is difficult to decide in principle whether the probability distribution over registers changes within a document, as there is always a global probability distribution that could lead to the same outcome. Examining this problem mathematically and empirically is a clear desideratum. Fourth, the influence of the topic or theme of a document on the distribution of LGFs might not be fully separable from the influence of the SFPs. In SFL, the field/tenor/mode model embraces this by including *field* in the set of determinants of registers. Performing topic modelling using LDA on the data already analysed with register modelling might provide empirical evidence to determine the degree and the nature of the correlation between register

and topic. Fifth, external validation of the results would lend support to the proposed model. Such validation could take the form of corpus comparison, where the goal is to see whether the distributions of LGFs and the associated SPFs can be retrieved from other similar or even less similar corpora. Also, corroborating the findings using behavioural experiments should be possible. We are not aware of experiments that could directly support the analysis performed here, but see Pescuma et al. (2023) for an overview of some experimental approaches to register phenomena. Sixth, the results of our analysis could be implemented in formal symbolic grammars (Machicao y Priemer et al. 2022) or text generation systems. Finally, the annotated corpus data created through the research presented in this paper can be used in alternation studies within the framework of probabilistic grammar. If interpretable results are obtained in such studies, it could further support the validity of register modelling as introduced here.

A Data Availability/Supplementary Files

All scripts and aggregated data used in this research can be found under the following DOI: DOI . Access to the register-annotated corpus can be obtained here: https://www.webcorpora.org

B Funding

Work on this paper was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

C Acknowledgments

We are indebted to Antonio Machicao y Priemer for discussions and feedback. We thank Sarah Böke, Johanna Kimmerl, and Pia Ortmann and for help with the annotation and research.

References

Ágel, Vilmos & Hennig, Mathilde. 2006. Theorie des Nähe- und Distanzsprechens. In Ágel, Vilmos & Hennig, Mathilde (eds.), *Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650-2000*, 3–31. Max Niemeyer.

Agha, Asif. 2007. *Language and social relations*. Cambridge UP. 446 pp. https://doi.org/10.1017/CBO9780511618284.

Argamon, Shlomo Engelson. 2019. Register in computational language research. *Register Studies* 1(1). 100–135. https://doi.org/10.1075/rs.18015.arg.

Barsalou, Lawrence W. 2016. Situated conceptualization offers a theoretical account of social priming. *Current Opinion in Psychology* 11(Supplement C). 6–11.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2). 145–204. https://doi.org/10.1017/S004740450001037X.

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

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27(1). 3–43.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257. https://doi.org/10.1007/978-0-585-35958-8_20.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge UP. https://doi.org/10.1017/cbo9780511519871.

Biber, Douglas. 2009a. Multi-dimensional approaches. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook*, vol. 2, 822–855. Berlin: Walter de Gruyter.

Biber, Douglas. 2009b. Multi-dimensional approaches. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook*, 822–855. Berlin: De Gruyter Mouton.

Biber, Douglas. 2019. Text-linguistic approaches to register variation. *Register Studies* 1(1). 42–75. https://doi.org/10.1075/rs.18007.bib.

Biber, Douglas & Conrad, Susan. 2009. *Register, genre, and style* (Cambridge textbooks in linguistics). Cambridge, UK: Cambridge UP. https://doi.org/10.1017/cbo9780511814358.

Biber, Douglas & Egbert, Jesse. 2018. *Register variation online*. Cambridge: Cambridge University Press.

Biber, Douglas & Egbert, Jesse & Keller, Daniel. 2020a. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory* 16(3). 581–616. https://doi.org/10.1515/cllt-2018-0086.

Biber, Douglas & Egbert, Jesse & Keller, Daniel. 2020b. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory* 16(3). 581–616. https://doi.org/10.1515/cllt-2018-0086.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4). 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, David M. & Griffiths, Thomas L. & Jordan, Michael I. 2010. The Nested Chinese Restaurant Process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* 57(2). https://doi.org/10.1145/1667053.1667056.

Blei, David M. & Jordan, Michael I. & Griffiths, Thomas L. & Tenenbaum, Joshua B. 2003. Hierarchical topic models and the Nested Chinese Restaurant Process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems* (NIPS'03), 17–24. Whistler, British Columbia, Canada: MIT Press.

Blei, David M. & Ng, Andrew Y. & Jordan, Michael I. 2003. Latent Dirichlet allocation. *J of Machine Learning Research* 3. 993–1022. https://dl.acm.org/doi/10.5555/944919.944937 (13 December, 2022).

Bohnet, Bernd. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *COLING '10: proceedings of the 23rd International Conference on Computational Linguistics*, 89–97. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Featherston, Sam & Sternefeld, Wolfgang (eds.), *Roots: linguistics in search of its evidential base* (Studies in Generative Grammar), 77–96. Berlin/New York: De Gruyter Mouton. https://doi.org/10.1515/9783110198621.75.

Cheung, Jackie Chi Kit & Penn, Gerald. 2009. Topological field parsing of German. In *Proc. of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp*, 64–72. Suntec, Singapore: Association for Computational Linguistics. https://doi.org/10.3115/1687878.1687889.

Chiu, Kenny & Clark, David M. & Leigh, Eleanor. 2022. Characterising negative mental imagery in adolescent social anxiety. *Cognitive Therapy and Research* 46. 956–966.

Cummins, Jim. 2008. BICS and CALP: Empirical and theoretical status of the distinction. In Hornberger, Nancy H. (ed.), *Encyclopedia of language and education*, 487–499. Springer US.

Divjak, Dagmar & Dąbrowska, Ewa & Arppe, Antti. 2016. Machine meets man: evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33. https://doi.org/10.1515/cog-2015-0101.

Eder, Elisabeth & Krieg-Holz, Ulrike & Wiegand, Michael. 2023. A question of style: a dataset for analyzing formality on different levels. In *Findings of the Association for Computational Linguistics: EACL 2023*, 580–593.

Egbert, Jesse & Biber, Douglas & Davies, Mark. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66(9). 1817–1831. https://doi.org/10.1002/asi.23308.

Egbert, Jesse & Mahlberg, Michaela. 2020. Fiction – one register or two? *Register Studies* 2(1). 72–101. https://doi.org/10.1075/rs.19006.egb.

Elman, Jeffrey L. 2009. On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cognitive Science* 33. 547–582.

Engel, Alexandra & Grafmiller, Jason & Rosseel, Laura & Szmrecsanyi, Benedikt & de Welde, Freek Van. 2021. How register-specific is probabilistic grammatical knowledge? a programmatic sketch and a case study on the dative alternation with give. In Seoane, Elena & Biber, Douglas (eds.), *Corpus-based approaches to register variation*, 51–84. Amsterdam: Benjamins.

Faaß, Gertrud & Heid, Ulrich & Schmid, Helmut. 2010. Design and application of a gold standard for morphological analysis: smor as an example of morphological evaluation. In Calzolari, Nicoletta & Choukri, Khalid & Maegaard, Bente & Mariani, Joseph & Odijk, Jan & Piperidis, Stelios & Rosner, Mike & Tapias, Daniel (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 803–810. Valletta, Malta: European Language Resources Association (ELRA). https://doi.org/10.3115/1220575.1220640.

Feilke, Helmuth. 2012. Bildungssprachliche Kompetenzen − fördern und entwickeln. *Praxis Deutsch* 233. 4–13.

Feilke, Helmuth & Hennig, Mathilde (eds.). 2016. *Zur Karriere von "Nähe und Distanz". Rezeption und Diskussion des Koch-Oesterreicher-Modells*. Berlin: de Gruyter.

Ferizis, George & Bailey, Peter. 2006. Towards practical genre classification of web documents. In *Proceedings of the 15th International Conference on World Wide Web* (WWW '06), 1013–1014. New York, NY, USA: ACM. https://doi.org/10.1145/1135777.1135991.

Finkel, Jenny Rose & Grenager, Trond & Manning, Christopher. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd annual meeting of the association for computational linguistics (acl 2005) 2005)*, 363–370. Association for Computational Linguistics.

Gogolin, Ingrid & Lange, Imke. 2011. Bildungssprache und durchgängige Sprachbildung. In Fürstenau, Sara & Gomolla, Mechtild (eds.), *Migration und schulischer Wandel*, 107–129. Wiesbaden: Springer VS.

Gotzner, Nicole & Mazzarella, Diana. 2021. Face management and negative strengthening: The role of power relations, social distance, and gender. *Frontiers in Psychology* 12, 602977. 1–13. https://doi.org/10.3389/fpsyg.2021.602977.

Grafmiller, Jason & Szmrecsanyi, Benedikt & Röthlisberger, Melanie & Heller, Benedikt. 2018. General introduction: a comparative perspective on probabilistic variation in grammar. *Glossa* 3(1). 94. https://doi.org/10.5334/gjgl.690.

Gries, Stefan Th. 2017. Syntactic alternation research. taking stock and some suggestions for the future. In Cuypere, Ludovic De & Vanderschueren, Clara & Sutter, Gert De

(eds.), *Current trends in analyzing syntactic variation*, vol. 31 (Belgian Journal of Linguistics), 7–27. Amsterdam: Benjamins.

Halliday, Michael Alexander Kirkwood. 1978. *Language as social semiotic: The social interpretation of language and meaning*. London: University Park Press.

Halliday, Michael Alexander Kirkwood. 1991. Corpus studies and probabilistic grammar. In Aijmer, Karin & Altenberg, Bengt (eds.), *English corpus linguistics*, 30–43. Mahwa, New Jersey, London: Taylor & Francis. https://doi.org/10.4324/9781315845890.

Halliday, Michael Alexander Kirkwood & Hasan, Ruqaiya. 1976. *Cohesion in English*. Longman.

Halliday, Michael Alexander Kirkwood & Hasan, Ruqaiya. 1989. *Language, context, and text: aspects of language in a social-semiotic perspective*. 2nd edn. Oxford: Oxford UP.

Hennig, Mathilde & Feilke, Helmuth. 2016. Perspektiven auf "Nähe und Distanz": Zur Einleitung. In Feilke, Helmuth & Hennig, Mathilde (eds.), *Zur Karriere von "Nähe und Distanz"*, 1–10. de Gruyter. https://doi.org/10.1515/9783110464061-002.

Jelodar, Hamed & Wang, Yeongli & Yuan, Chi & Feng, Xia & Jiang, Xiahui & Li, Yanchao & Zhao, Liang. 2019. Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78. 15169–15211.

Karlgren, Jussi & Cutting, Douglass. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of coling 94*, 1071–1075.

Kim, Yunhyong & Ross, Seamus. 2011. Formulating Representative Features with Respect to Genre Classification. In *Genres on the Web: Computational Models and Empirical Studies*. Mehler, Alexander & Sharoff, Serge & Santini, Marina (eds.). Dordrecht: Springer Netherlands. 129–147. https://doi.org/10.1007/978-90-481-9178-9_6.

Koch, Peter & Oesterreicher, Wulf. 1985. Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36(1). 15–43. https://doi.org/10.1515/9783110244922.15.

Labov, William. 1966. *The social stratification of English in New York City*. Center for Applied Linguistics.

Lee, David. Y. W. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and nevigating a path through the BNC jungle. *Language Learning and Technology* 5(3). 37–72.

Lehmann, Nico & Serova, Dina & Lukassek, Julia & Döring, Sophia & Goymann, Frank & Lüdeling, Anke & Akbari, Roodabeh. 2023. Guidelines for the annotation of parameters of narration. *REALIS: Register Aspects of Language in Situation* 2(5). 1–41. https://doi.org/10.18452/26437.

Levering, Ryan & Cutler, Michal. 2009. Cost-sensitive feature extraction and selection in genre classification. *Journal for Language Technology and Computational Linguistics* 24(1). 57–72. https://doi.org/10.21248/jlcl.24.2009.113.

Lüdeling, Anke & Alexiadou, Artemis & Adli, Aria & Donhauser, Karin & Dreyer, Malte & Egg, Markus & Feulner, Anna Helene & Gagarina, Natalia & Hock, Wolfgang & Jannedy, Stefanie & Kammerzell, Frank & Knoeferle, Pia & Krause, Thomas & Krifka, Manfred & Kutscher, Silvia & Lütke, Beate & McFadden, Thomas & Meyer, Roland & Mooshammer, Christine & Müller, Stefan & Maquate, Katja & Norde, Muriel & Sauerland, Uli & Solt, Stephanie & Szucsich, Luka & Verhoeven, Elisabeth & Waltereit, Richard & Wolfsgruber, Anne & Zeige, Lars Erik. 2022. Register: Language users' knowledge of situational-functional variation. *REALIS: Register Aspects of Language in Situation*. 1–58. https://doi.org/10.18452/24901.

Lukin, Annabelle & Moore, Alison & Herke, Maria & Wegener, Rebekah & Wu, Canzhong. 2008. Halliday's model of register revisited and explored. *Linguistics and the Human Sciences* 4. 187–213.

Machicao y Priemer, Antonio & Müller, Stefan & Schäfer, Roland & Bildhauer, Felix. 2022. Towards a treatment of register phenomena in HPSG. In Müller, Stefan & Winckel, Elodie (eds.), *Proceedings of the 29th International Conference on Head-Driven Phrase Structure Grammar, Online (Nagoya/Tokyo)*, 86–101. Frankfurt/Main: University Library. https://doi.org/10.21248/hpsg.2022.5.

Matthiessen, Christian M. I. M. 1993. Register in the round: diversity in a unified theory of register analysis. In Ghadessy, Mohsen (ed.), *Register analysis: theory and practice*, 221–292. Pinter.

Müller, Stefan. 1999. *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar für das Deutsche* (Linguistische Arbeiten 394). Tübingen: Max Niemeyer.

Müller, Stefan. 2013. *Head-Driven Phrase Structure Grammar: Eine Einführung*. 3rd edn. (Stauffenburg Einführungen 17). Tübingen: Stauffenburg Verlag.

Müller, Stefan & Abeillé, Anne & Borsley, Robert D. & Koenig, Jean-Pierre (eds.). 2021. *Head-Driven Phrase Structure Grammar: The handbook* (Empirically Oriented Theoretical Morphology and Syntax 9). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.5543318.

Müller, Thomas & Schmid, Helmut & Schütze, Hinrich. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 322–332. Seattle, Washington, USA: Association for Computational Linguistics.

Neumann, Stella. 2014. Cross-linguistic register studies: Theoretical and methodological considerations. *Languages in Contrast* 14(1). 35–57. https://doi.org/10.1075/lic.14.1.03neu.

Neumann, Stella & Evert, Stephanie. 2021. A register variation perspective on varieties of English. In Seoane, Elena & Biber, Douglas (eds.), *Corpus-based approaches to register variation*, 143–178. Amsterdam: Benjamins.

O'Donnell, Mick. 2021. Dynamic modelling of context: field, tenor and mode revisited. *Lingua* 261. 102952. https://doi.org/10.1016/j.lingua.2020.102952.

Oesterreicher, Wulf & Koch, Peter. 2016. 30 Jahre "Sprache der Nähe – Sprache der Distanz" Zu Anfängen und Entwicklung von Konzepten im Feld von Mündlichkeit und Schriftlichkeit. In Feilke, Helmuth & Hennig, Mathilde (eds.), *Zur Karriere von "Nähe und Distanz"*, 11–72. de Gruyter. https://doi.org/10.1515/9783110464061-003.

Ortmann, Katrin & Dipper, Stefanie. 2019. Variation between different discourse types: literate vs. oral. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, 64–79. Ann Arbor, Michigan: Association for Computational Linguistics.

Paolillo, John C. 2000. Formalizing formality: An analysis of register variation in Sinhala. *Journal of Linguistics* 36(2). 215–259. https://doi.org/10.1017/S0022226700008148.

Pescuma, Valentina N. & Serova, Dina & Lukassek, Julia & Sauermann, Antje & Schäfer, Roland & Adli, Aria & Bildhauer, Felix & Egg, Markus & Hülk, Kristina & Ito, Aine & Jannedy, Stefanie & Kordoni, Valia & Kühnast, Milena & Kutscher, Silvia & Lange, Robert & Lehmann, Nico & Liu, Mingya & Lütke, Beate & Maquate, Katja & Mooshammer, Christine & Mortezapour, Vahid & Müller, Stefan & Norde, Muriel & Pankratz, Elizabeth & Patarroyo, Angela G. & Pleșca, Ana-Maria & Ronderos, Camilo R. & Rotter, Stephanie & Sauerland, Uli & Schnelle, Gohar & Schulte, Britta & Schüppenhauer, Gediminas & Sell, Bianca Maria & Solt, Stephanie & Terada, Megumi & Tsiapou, Dim-

itra & Verhoeven, Elisabeth & Weirich, Melanie & Wiese, Heike & Zaruba, Kathy & Zeige, Lars Erik & Lüdeling, Anke & Knoeferle, Pia. 2023. Situating language register across the ages, languages, modalities, and cultural aspects: Evidence from complementary methods. *Frontiers in Psychology* 13. 1–31. https://doi.org/10.3389/fpsyg.2022.964658.

Petrov, Slav & Klein, Dan. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: the Conference of the North American Chapter of the Association for Computational Linguistics: proceedings of the main conference*, 404–411. Rochester, New York: Association for Computational Linguistics.

Pollard, Carl & Sag, Ivan A. 1987. *Information-based syntax and semantics, vol. 1: fundamentals*. Stanford: CSLI.

Pollard, Carl & Sag, Ivan A. 1994. *Head-driven phrase structure grammar*. Chicago: CSLI. https://doi.org/10.2307/416665.

Pritchard, Jonathan K & Stephens, Matthew & Donnelly, Peter. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155(2). 945–959. https://doi.org/10.1093/genetics/155.2.945.

Řehůřek, Radim & Sojka, Petr. 2010. Software Framework for Topic Modelling with Large Corpora. English. In *Proc. of the lrec 2010 workshop on new challenges for nlp frameworks*, 45–50. Valletta, Malta: ELRA.

Santini, Marina. 2011. Cross-Testing a Genre Classification Model for the Web. In *Genres on the Web: Computational Models and Empirical Studies*. Mehler, Alexander & Sharoff, Serge & Santini, Marina (eds.). Dordrecht: Springer Netherlands. 87–128. https://doi.org/10.1007/978-90-481-9178-9_5.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Bański, Piotr & Biber, Hanno & Breiteneder, Evelyn & Kupietz, Marc & Lüngen, Harald & Witt, Andreas (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Lancaster: IDS.

Schäfer, Roland & Barbaresi, Adrien & Bildhauer, Felix. 2013. The good, the bad, and the hazy: design decisions in web corpus construction. In Evert, Stefan & Stemle, Egon & Rayson, Paul (eds.), *Proceedings of the 8th web as corpus workshop (WAC-8)*, 7–15. Lancaster: SIGWAC. https://doi.org/10.3115/v1/w14-0402.

Schäfer, Roland & Bildhauer, Felix. 2012. Building large corpora from the Web using a new efficient tool chain. In Calzolari, Nicoletta & Choukri, Khalid & Declerck, Thierry & Dogan, Mehmet Ugur & Maegaard, Bente & Mariani, Joseph & Odijk, Jan & Piperidis, Stelios (eds.), *Proc. of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul: ELRA.

Schäfer, Roland & Sayatz, Ulrike. 2014. Die Kurzformen des Indefinitartikels im Deutschen. *Zeitschrift für Sprachwissenschaft* 33(2). 215–250. https://doi.org/10.1515/zfs-2014-0008.

Schiller, Anne & Teufel, Simone & Stöckert, Christine & Thielen, Christine. 1999. *Guidelines für das Tagging deutscher Textkorpora mit STTS (kleines und großes Tagset)*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart & Institut für Sprachwissenschaft, Universität Tübingen. Stuttgart & Tübingen.

Schmid, Hans-Jörg. 2020. *The dynamics of the linguistic system: usage, conventionalization, and entrenchment*. Oxford: Oxford UP. 432 pp. https://doi.org/10.1093/oso/9780198814771.001.0001.

Schmid, Helmut & Fitschen, Arne & Heid, Ulrich. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In Lino, Maria Teresa & Xavier, Maria Francisca & Ferreira, Fátima & Costa, Rute & Silva, Raquel (eds.),

Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 4), 1263–1266. Lisbon.

Sharoff, Serge. 2018. Functional Text Dimensions for the annotation of web corpora. *Corpora* 13(1). 65–95. https://doi.org/10.3366/cor.2018.0136.

Sinclair, John & Ball, John. 1996. Preliminary recommendations on text typology. *EAGLES (Expert Advisory Group on Language Engineering Standards)*.

Stamatatos, Efstathios & Kokkinakis, George & Fakotakis, Nikos. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26(4). 471–495. https://doi.org/10.1162/089120100750105920.

Telljohann, Heike & Hinrichs, Erhard W. & Kübler, Sandra & Zinsmeister, Heike & Beck, Kathrin. 2012. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Tech. rep. Universität Tübingen Seminar für Sprachwissenschaft.

Wolk, Christoph & Szmrecsanyi, Benedikt. 2018. Probabilistic corpus-based dialectometry. *Journal of Linguistic Geography* 6(1). 56–75.

Zeman, Sonja. 2018. What is a *Narration* – And why does it matter? In Hübl, Annika & Steinbach, Markus (eds.), *Linguistic foundations of narration in spoken and sign languages*, 173–206. Amsterdam: John Benjamins. https://doi.org/10.1075/la.247.08zem.