Modeling Communicative Purpose with Functional Style: Corpus and Features for German Genre and Register Analysis

Thomas Haider
Max Planck Institute for Empirical Aesthetics
Frankfurt am Main, Germany
thomas.haider@ae.mpg.de

Alexis Palmer
University of North Texas
Denton, Texas, USA
alexis.palmer@unt.edu

Abstract

While there is wide acknowledgement in NLP of the utility of document characterization by genre, it is quite difficult to determine a definitive set of features or even a comprehensive list of genres. This paper addresses both issues. First, with prototype semantics, we develop a hierarchical taxonomy of discourse functions. We implement the taxonomy by developing a new text genre corpus of contemporary German to perform a text based comparative register analysis. Second, we extract a host of style features, both deep and shallow, aiming beyond linguistically motivated features at situational correlates in texts. The feature sets are used for supervised text genre classification, on which our models achieve high accuracy. The combination of the corpus typology and feature sets allows us to characterize types of communicative purpose in a comparative setup, by qualitative interpretation of style feature loadings of a regularized discriminant analysis. Finally, to determine the dependence of genre on topics (which are arguably the distinguishing factor of sub-genre), we compare and combine our style models with Latent Dirichlet Allocation features across different corpus settings with unstable topics.

1 Introduction

Language users exhibit a high degree of variability at all levels of the linguistic system and language use. In this paper, we focus on variation at the level of text (or discourse). Texts vary along numerous parameters such as *medium* (spoken, written), *topic / domain* (e.g. art, science, religion, government), *rhetorical mode* (e.g. narration, argumentation, description, exposition), or *communicative purpose* (e.g. persuade, report, entertain, edify, instruct, express opinion).

Such variational aspects, captured under the terms *register* and *genre*, have been central to previous investigations of discourse and textual variation. Both terms have been used to refer to language variety associated with particular situations of use and, lacking a clear differentiation between the two terms, many studies simply adopt one and disregard the other (cf. Biber et al., 2007, 1.4).

For Biber and Conrad (2009), though, *genre*, *register* and *style* are different perspectives on a single text. Each dimension can describe the others, e.g. a *commentary* voices an *opinion* that is *inclusive*, *angry* and *aloof* – it refers to non-specific entities, but avoids deixis and possession.

The cornerstone of our approach is to model textual variation via stylistic features, which we argue is the level at which both genre and register variation can be convincingly modeled.

Following Lee (2001), we consider *register* as variation according to use in broad societal situations. It describes a functional adaptation to the immediate situational parameters of contextual use, as different situations 'require' appropriate configurations of language. *Genre* views text by consensus within a culture, as artifacts categorized by purposive goals, distinguished by conventionally recognized criteria and hence subject to change as conventions are challenged and revised over time. In short (see table 1): *genre* is described by a **conventional label**, while *register* is described through its **pervasive features** (cf. Biber and Conrad, 2009).

A comprehensive typology of texts at the same level of generality is a research prerequisite for any comparative register analysis. Because current multi-genre text corpora do not easily ad-

Proceedings of the Workshop on Stylistic Variation, pages 74–84
Copenhagen, Denmark, September 7–11, 2017.

Genre	Purpose / Function
scientific texts	inform
advertising	persuade
legal texts	instruct
...	...

Table 1: Sample genres, with dominant purpose.

mit to functional analysis of types (Section 2), we turn instead to the theoretical framework of Steen (1999), which promises a general taxonomy of discourse. We operationalize the core of Steen's theory for corpus design, modeling register variation top-down with prototype semantics to develop a comparative genre taxonomy (Section 3.1). The taxonomy is then implemented in a general genre/register corpus of contemporary German. (Section 3.2).

We employ a wide range of stylistic features for the classification of text, (Section 3.3), going beyond previous computational stylometric genre analysis, that has often relied on shallow lexico-syntactic patterns such as function words, surface forms, character / part-of-speech n-grams, etc., (Karlgren and Cutting, 1994; Stamatatos et al., 2000a,b; Koppel et al., 2003; Gries and Shaoul, 2011; Sharoff, 2007; Kanaris and Stamatatos, 2007), extending beyond linguistically motivated features (Biber and Conrad, 2009; Santini, 2005) with a fine-grained morphology, psycholinguistic word norms, and topic models. With these feature sets and corpus, we perform supervised genre classification (Section 4), showing that results remain high and stable across shifting sets of categories.

A major problem with relying on surface level features - particularly lexical features - is that they tend to capture topical information. Petrenz and Webber (2011) make a strong case that a genre classification system should not be susceptible to changes in topic/domain. We therefore test topic distributions learned with Latent Dirichlet Allocation (LDA) (Blei et al., 2003) against lexico-syntactic features in such a scenario (Section 4.4). Finally, we identify functional dimensions for characterizing communicative function (register) by examining the features most prominently associated with different communicative purposes. (Section 5).

2 Selected related work

There are a number of genre-aware corpora for English, but none for contemporary German that go beyond web-genre, or are freely available. Early examples for English include the Brown corpus (Francis and Kučera, 1964/79) and the Lancaster-Oslo/Bergen (LOB) corpus (Johansson et al., 1978). Both were sampled according to library classification systems and contain relatively small numbers of samples distributed over various genre classes of different granularity. MASC[1] (Ide, 2008) also balances genre classes over number of tokens. To analyze the variety across texts, one needs to arbitrarily split its documents (to 2000 tokens, as done by Passonneau (2014)). There is an extensive collection of web-genre corpora (Santini, 2007; Meyer zu Eißen and Stein, 2004; Rehm et al., 2008; Santini et al., 2010). See Sharoff and Markert (2010) for an overview and the success of Char-4-bin features (later found to be unstable by Petrenz and Webber (2011)). *GECCo* is a bilingual (English-German) corpus for investigating cohesion across register (Lapshinova-Koltunski et al., 2012). It is not freely available. The DWDS 'Kernkorpus' for super-genre of 20th century texts is also not available.[2]

The Hierarchical Genre Corpus (HGC) (Stubbe and Ringlstetter, 2007) and the British National Corpus (BNC) [3] are designed to offer representative samples across different genres in a hierarchical fashion. However, the categories of HGC are not clear-cut and focus on web-genre. The BNC is highly imbalanced.

Some additional related work uses features from systemic functional grammar in the tradition of Halliday for text genre classification (Argamon and Koppel, 2010; Argamon et al., 2003; Argamon and Koppel, 2012; Argamon et al., 2007).

3 Method

We present a methodology for corpus driven analysis of situated language use. We achieve this by: 1) building a corpus, and 2) classifying and characterizing situationally-defined text categories, aiming at a comparative register analysis.

3.1 A taxonomy for discourse

Genre follows a categorical paradigm, such that it assigns labels to text. A problem with genre labels is that they can have many different levels of generality, e.g. the genre "academic discourse" is very

[1]Manually Annotated Sub-Corpus of American English

[2]http://194.95.188.16/ressourcen/kernkorpus/

[3]http://www.natcorp.ox.ac.uk/

broad, and texts within such a high-level genre category will show considerable internal variation in their use of language, as Biber (1989) has shown. On a lower level, different genres can be based on many different criteria (domain, topic, participants, setting, form, etc.), e.g. 'Western' vs. 'Romance' novels[4] or 'Elegy' vs. 'Ballad'.[5]

Steen (1999) develops a solution for this by applying prototype theory (Rosch, 1973) to the conceptualization of genre (and hence to the formalisation of a taxonomy of discourse). A prototype is the most typical instance of a more encompassing and varied, fuzzy conceptual category – some instances are more central than others – e.g. the basic-level concept *chair* is a prototypical instance of the superordinate concept *furniture*. Functionally, basic-level concepts are maximally informative (easily recognized, remembered, and learned), whereas subordinate concepts are less richly differentiated from their respective alternatives (e.g. *dentist chair* vs. *recliner*).[6] Taylor (1995) finds that "terms above the basic level are sometimes deviant in some way (e.g. furniture is morphosyntactically unusual in that it is uncountable, i.e. one cannot say 'a furniture' or 'furnitures')".

Steen proposes that we can recognize genres by their cognitive basic-level status: True genres, being basic-level, are maximally distinct from one another. He analyzes the distance of genres in terms of specific attributes (parameters). Biber (1993, table 1) introduces situational parameters as sampling strata for corpora, which we combine with the parameters of Steen (1999).

For our corpus design, we use the following parameters, that our features aim to cover, to distinguish genre: **medium / discourse channel** (written, spoken, scripted), **factuality** (imaginative), **purpose / discourse function** (persuade, entertain, report, edify, inform, instruct, explain, keep records, reveal self, express attitudes, opinions, etc.), **rhetorical mode / discourse type** (narration, argumentation, description, exposition), **participants** (plurality, interactiveness, shared knowledge, demographic), **topic / domain** (art, science, religion, government, etc.), **content** (topics, themes, keywords). We do not use **setting**, **formality, format, form**.

[4]Distinguised by topic, protagonists, and purpose.

[5]Distinguished by topic, form, and purpose.

[6]Steen (1999) also claims superordinates to be less differentiated.

3.2 Corpus Design

Genre corpora are faced with the problem of finding an operationalizable definition for each genre and avoiding meaningless miscellaneous categories, i.e. choosing the right granularity of classes. The multitude of possible genre categories makes it impractical to determine a fixed set of classes for a corpus that is representative for all genre. However, for a corpus to be useful for analysis, it needs to include a representative range of classes. We focus on written language that allows us to model types of communicative function through genre.

We design our genre corpus in a top-down hierarchical fashion as a taxonomy, where supergenre categories are based on the *broad social embedding* of text. The four super-level categories for written language are taken from the DTA (Deutsches Textarchiv) (Geyken et al., 2011): *Wissenschaft (science), Belletristik (literature), Zeitung (press) and Gebrauchstext (operative text)*. We add a *Gesprochen (spoken)* variety to also test our model on a different medium of communication.

We subdivide each super-category into functionally dichotomous basic-categories, i.e. maximally distinct prototypical instances, mainly relying on *communicative purpose/function* as the distinctive attribute for written language. Then we assign a basic level-genre to each function, as found in DeReKo[7] (Kupietz et al., 2010). The genre annotation in DeReKo was delivered by the publishers and is not evaluated on annotators, consequently only being a 'silver standard'. Table 2 illustrates our taxonomy.

To measure human agreement on assigning these categories, we randomly selected 20% of the test set of our 8-way typology for written basic-genre (10 documents per class) for manual annotation. The three raters were (under)graduate students, native speakers of German, with backgrounds in linguistics (R1,R3) and psychology (R1,R2), employed at the MPIEA[8]. They were given minimal instruction on text genre, communicative functions and the purpose of the study. The first eight texts covered all types to make them familiar with the variety.

Inter-rater agreement is measured with Cohens κ and shown in table 4. We compare each rater to

[7]Deutscher Referenzkorpus: German Reference Corpus

[8]Max Planck Institute for Empirical Aesthetics

Super-Genre	**Genre**	Dominant purpose	Ger. label	Comment
Science	Academic Popular science	research educate	Wissensch. Pop. Wiss.	Linguistik Online crawl Spektrum d. Wiss.
Literature	Novel (epic) Drama	narrate perform	Roman Drama	
Press	Report Commentary Reportage	report opinion coverage	Bericht Kommentar Reportage	
Operative Text	Advertising Pharma leaflets	persuade instruct	Anzeigen Pack.beilage	From newspapers Rote Liste crawl
Spoken	Speech Interview	asymmetric symmetric	Rede Interview	German Bundestag

Table 2: DeGeKo Genre Taxonomy translated to English

	advertising	report	novel	commentary	leaflets	pop.sci.	reportage	academic
document_length	486.7	736.6	1404.4*	788.4	2689.4	933.4	2042.4	3631.6**
avg_sentence_length	12.70	18.77	27.25	19.22	19.04	21.41	17.80	15.83
avg_word_length	5.25	5.38	4.98	5.29	5.66	5.48	4.91	5.24
type_token_ratio	0.317	0.265	0.230	0.270	0.269	0.240	0.219	0.294

Table 3: DeGeKo written document stats

	R1	R2	R3	Silver
R1	-	.79	.62	.84
R2		-	.58	.78
R3			-	.61

Table 4: Inter-rater agreement, 8-way typology (κ)

the others, and to the silver standard. R1 and R2 show a high level of agreement with each other (κ of .79) and with the silver standard (κ of .84 and .78, respectively). R3 shows lower agreement, often confusing academic writing with popular science.[9] A common difficulty for all raters was to distinguish among the press varieties (report, commentary, coverage), as we will also encounter in our experiments.

We propose that a fine-grained topic annotation at document level acts as viable proxy for sub-genre distinction, e.g. advertising text can be subcategorized to *Leisure_Entertainment:Travel* ads or *Economy_Finance:Banking* ads. Topic annotation in DeReKo was assigned by a Naive-Bayes classifier trained on the opendirectory[10] taxonomy as described by Weiß (2005). Where this annotation is not consistent, we use the existing domain annotation to examine genre-internal variation.[11]

In the press genres, some topics were overly represented in the original population (e.g. reports on sports clubs). While it can be argued that those are the most prototypical instances of a given genre, we balance those topics in the population to achieve a more 'natural' topic distribution through sampling, so there is no bias towards certain content. The target is the mean size of topic classes plus one standard deviation.

Table 2 illustrates our taxonomy. For classes with insufficient material in DeReKo to satisfy our sampling criteria (below), we crawl the web (academic & leaflets). Where we still did not retrieve enough documents (academic & drama), we employ an *upsampling* technique: we chop documents evenly by three-sentence chunks and disperse them according to their original position in the document (i.e., beginning, middle and end are still intact). Due to this upsampling, we cannot use document length as a feature for classification.

Genre collections are often relatively small and / or imbalanced. We implement a modular corpus balancer tool able to fine tune the selection of documents. In line with our focus on 'register by genre', we balance the corpus by documents, attaining 500 documents for each of the eleven genre classes, randomly split to 400 docs for training, 50 for development and 50 for testing. With synchronic analysis in mind, we take no documents published before 1950. To retrieve a prototypical size of the documents, we restricted the max_doc_size to one standard deviation over the mean. For min_doc_size, we used $\frac{mean\_size}{2}$ or 120 tokens, as they would be too small for stylistic

[9] R3 complained of having had a stressful day.
[10] http://dmoztools.net/
[11] Domain here is equivalent to the newspaper section in which the text originally appeared (ger.: *ressort*).

analysis otherwise. Biber (1989, 1993) argues that a text 'sample' should be 2000 tokens large. This is not an issue in our setup, as each class is itself as large as the whole LOB corpus.

As you can see in table 3, on average, advertisements are the shortest documents and academic articles (*wissenschaft*) are the longest. Superscript ** documents have been upsampled. Also * signifies that the size for novels is not entirely trustworthy, because this category includes both shortened novels and short stories, skewing the document length distribution. Still, novels have the longest sentences by far. Reports (*berichte*) dominate in average word length. Advertising (*anzeigen*) has the highest type-token ratio.

3.3 Feature Design

We model style features that are (a) able to distinguish particular usage situations, and (b) based on sufficiently robust linguistic annotation tools. Therefore, we focus on the engineering of fine grained morpho-syntactic features, linguistic lexicons, word norms and surface forms. To test the topic sensitivity of genre, we also generate topic distributions for documents with Latent Dirichlet Allocation (LDA). Our feature-groups are organized as a nested hierarchy, shown in Table 5. Individual features are described below. We implemented our feature extraction pipeline in python. Each feature is normalized relative to its own individual group (e.g. pos with pos) per text. Before classification, we use the sklearn StandardScaler.

Preprocessing for feature extraction. We use the Julie Lab Segmenter (Tokenization, Sentences) (Hahn et al., 2016) and the RF-Tagger (Lemmatization, STTS pos-tags, SMOR morphological tags) (Schmid and Laws, 2008).

Part-of-Speech Tags We use the Stuttgart-Tübingen Tagset (STTS)[12] with 47 tags.

Verb Classes German verb classes are retrieved from GermaNet (Hamp et al., 1997; Henrich and Hinrichs, 2010). The GermaNet scheme contains 9,382 unique verbs (including particles and affixes) across 15 groups, where a verb can be a member of several groups, totaling 15,327 tokens. For each verb token that we detect, we count every relevant class with equal weight.

[12] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

Surface Cues This is a heterogenous feature-group of linguistic surface cues.

1. *Avg. word length* in # of characters.
2. *Avg. sentence length* in # of words.
3. *Type-Token-ratio*: The ratio of unique types and tokens thereof. Always between 0 and 1.
4. *Alliteration*: Two subsequent words share the same first character (***b**itter **b**utter*).
5. *Assonance*: Two subsequent words share the same first vowel (*l**oo**se g**oo**se*).
6. *Repetition*: Minimum four character words recur within a 20 word context. + variant without proper names to exclude speaker roles in drama.

We do not use document length, as we want to learn linguistic information only.

Morphology RF-Tagger (Schmid and Laws, 2008) annotates very fine-grained (767) morphological tags according to SMOR (Schmid et al., 2004). One such feature would be "VFIN.Full.2.Pl.Pres.Ind" for a *full finite verb in second person plural present indicative*.

WWN word norms Lahl et al. (2009) crowd-sourced ratings for *concreteness, valency* and *arousal* for 2,654 German nouns. We draw the mean for each dimension (0 - 10) per document.

LIWC - word norms The English Linguistic Inquiry and Word Count (Tausczik and Pennebaker, 2010; Pennebaker et al., 2015) contains 6400 words and stems (and select emoticons). The German version (Wolf et al., 2008) includes 7510 entries. It provides a hierarchical annotation of 68 linguistic and psychological categories, e.g. the word *cried* is part of five categories: *sadness, negative emotion, overall affect, verbs* and *past focus*. Hence, all five will be counted for the document.

Connectives The *HDK* list of 312 discourse connectives is described in (Versley, 2010). We match connectives by iterating over word n-grams. For connectives with a gap ("entweder ... oder"), we look ahead 20 words. If the right side element returns a match, we include the whole (gapped) connective, otherwise we only count the left side.

Stopwords Our German stopword list is by solariz,[13] containing 996 inflected wordforms (of which 4 do not occur in the corpus).

[13] https://solariz.de/de/deutsche_stopwords.htm

Feat.set	Features
POS	Part-of-speech tags (47)
BASIC	POS + verb classes (15), surface cues (7)
SELECT	BASIC + SMOR morphology (767), LIWC (62), WWN (3), connectives (231)
FULL	SELECT + POS-bigrams (1822), morph-single (81), stopwords (992), punctuation (13)
POS3	POS-trigrams (51473)
LDA200	LDA topics (200), trained on whole corpus
	- CONTENT: only content words - STOP: only stopwords

Table 5: Nested hierarchy of feature sets; numbers quantify individual features.

Latent Dirichlet Allocation - LDA We train gensim (Řehůřek and Sojka, 2010) LDA (Blei et al., 2003) models on word lemmas, to model semantic domain. We train on the whole corpus (incl. the test set) and derive the topic distribution for each document (as probabilities). We experimented with 50, 100 and 200 topic dimensions, the latter giving best results. For feature generation, a relatively large number of topics is preferred.

3.4 Classification algorithms

For classification, we use Linear Discriminant Analysis (LinDA), a Naive Bayes Multinomial classifier, Random Forest ensemble classifiers (FOREST) and Support Vector Machines (SVM). We train one SVM on 10 dimensions (ordered by explained covariance) of a Principal Component Analysis (PCA), one SVM vanilla version, and lastly, with a feature selection based on ANOVA, selecting the (3-20 percentile) best performing features. All models were optimized for several parameters with a grid search.[14] We used the API of scikit-learn 0.18 (Pedregosa et al., 2011). The algorithms were selected based on their success in the related literature on genre classification. The use of Random Forests and LDA is novel however.

3.5 Characterization algorithms

For the characterization of communicative functions, we work with a Linear Discriminant Analysis (LinDA) and a Stochastic Gradient Descent (SGD). A linear model allows us to easily interpret feature loadings for each class, as each class is characterized by the linear combination of its feature weights. Also, it can be easily evaluated with a F1 score or a confusion matrix. The general form (1) means that it is easy to see the relative importance and contribution of each feature and to sanity check the model. The equation is solved by calculating a Bayesian objective, i.e. fitting a Gaussian density distribution.

$$C_k = C_{k0} + C_{k1}X_1 + C_{k2}X_2 + ... + C_{kn}X_n \quad (1)$$

where C_k is the classification score for group k and C_{kn} are the coefficients for the features X_n.

The main problem of a linear model is posed by strongly collinear features from different feature groups (PTKZU vs. Part.ZU) that consequently dominate the objective function (they become important for many classes). So we need to apply regularization techniques that allow a noise-free interpretation. But penalizing (e.g. setting variables to zero) with L1 or L2 makes the model less interpretable. This may ignore relevant information from the dataset. Consequently, we regularize LinDA with a PCA (with 150 dimensions), so that we "align" (near) identical features that load into opposing directions by their covariance. A side-effect is that this also avoids overfitting.[15]

4 Experiments

This section presents supervised classification experiments for labeling texts with communicative function, as construed in our corpus by genre labels. First, we classify basic-level genre for written language only (Section 4.1). Second, we add spoken varieties to the set of genres, changing the range of variation (Section 4.2). The third experiment changes the granularity of classification, instead targeting super-genre classes (Section 4.3). Finally, to ensure that our models learn genre rather than simply capturing differences in topics, we create an expanded sub-corpus of press documents, allowing us to keep the set of topics present in training data distinct from those represented in the test data (Section 4.4). Details of models and settings appear in Sections 3.4 and 3.5.

[14] Most notably for SVM: C and kernel method. For Forest: Number of trees and their depth.

[15] SGD with an ElasticNet consistently delivers somewhat similar results, but due to its nature it only "approximates" results, making it less preferable. On a small dataset (which ours arguably is), the closed-form-solution LinDA is to be preferred, as it delivers more consistent results.

Featureset	POS	BASIC	FULL	POS3	SELECT.	LDA200 STOP	LDA200 CONTENT	SELECT.+LDA200
	F1 score	F1 score	F1 score	F1 score	F1 score	F1 score	F1 score	F1 score
$LinDA$	.70	.77	.30	.28	.80	.73	.79	.86
$BAYES_{multinom}$	?	.73	.75	.51	.76	.73	.78	.81
$FOREST_{entropy}$	.74	.81	.86	.80	**.88**	.81	**.90**	**.92**
$FOREST_{gini}$	.75	.81	**.88**	.82	**.87**	.82	**.89**	**.92**
SVM_{PCA10}	.68	.75	.85	.55	.82	.77	.86	time
$SVM_{VANILLA}$	time	.79	.83	.72	.83	time	**.92**	.88
SVM_{ANOVA}	time	.70	**.88**	.77	.86	.	.	.

Table 6: Supervised classification on DeGeKo's eight written classes.

4.1 Written Basic-Level

In our corpus, the basic-level written genres are academic, popular science, novel, report, commentary, reportage, advertising, and leaflets.

Table 6 shows the classification results for written genres. Results shown are for the test set; performance is similar (± 2 points) for the dev set. 'time' means that the classifier did not finish in a reasonable time frame (a day).

For all classifiers, SELECTED and LDA200CONTENT feature sets show the best results. The FOREST classifiers appear to be the most robust to changing the feature set. Overall, the best result is obtained by a vanilla SVM on LDA200CONTENT, on par with FOREST on SELECTED+LDA200CONTENT. Also, the smaller SELECTED set compares well to the larger FULL set, making it the best model for a characterization of communicative function (FULL contains POS2-grams).[16] The main confusion between classes is caused by the press varieties, mostly because reports and commentaries are confused for each other, and commentaries confused with many other classes.

Most strikingly, LDA200CONTENT outperforms SELECTED by 2 - 4 points. This raises the important question of how strongly the genre of a document is influenced by its topics. Petrenz and Webber (2011) show that some genre classification models suffer heavily when the topics present in a given genre during testing are different from those seen in training.

4.2 Including Spoken Classes

Next, we enrich the written basic-genre classes with the spoken varieties *symmetric speech*, *asymmetric interviews*, and *drama*, which is written to be spoken. The main difference is that *drama* does not contain spontaneous speech, indicated by monologues. It is also arguable that political speeches – as used here – were prepared in written form to be performed in spoken form.

Experiment	Written+Spoken		Super-Level	
Feature set	BASIC	SEL.	BASIC	SEL.
	F1: test	F1: test	F1: test	F1: test
$LinDA$	.74	.80	.89	.91
$BAYES$	.68	.76	.83	.89
$FOREST_{ent}$	**.78**	.85	.91	**.96**
$FOREST_{gini}$	.77	**.86**	.91	.95
SVM_{PCA10}	.	.82	.86	.94
SVM_{VAN}	.	.80	.91	.94

Table 7: Written+spoken (L), Super-genres (R).

The left-hand side of Table 7 shows classification results for the BASIC and the SELECTED feature sets. The richer feature set clearly outperforms the simpler one. Interestingly, even though we added three classes of spoken material, we do not lose any accuracy over the corpus with only written varieties.

4.3 Written Super-Level

Next, written-language classes are mapped to four coarse-grained super-genres: *Presse, Wissenschaft, Belletristik* and *Gebrauchstext*.

The right-hand side of Table 7 shows these results. We see that basic-level genre classes are quite robust concerning their super-class. The score improves somewhat over basic-genre, partly because the task is simplified from 8 classes to 4. Prototype theory (and consequently Steen (1999)) would hypothesize that super-genre cannot be as richly distinguished as basic-genre. However, given the machine learning context of fewer classes and more data, the results are what you would expect. In a production system, this coarse set of classes can be used to predict text genre with a fair amount of certainty with most classifiers.

[16] The bad performance of LinDA_POS3, LinDA_Full, Bayes_POS3 and SVMPCA10_POS3 is likely attributable to a skewed distribution of pos-n-grams.

Topic Class		Politik	Freizeit_Unterh.	Kultur	Sport	Wirtsch._Finanz.	Staat_Gesell.	Wissensch.
Bericht	train	147	65	88	-	-	-	-
Kommentar	train	95	-	180	25	-	-	-
Reportage	train	176	-	-	-	-	118	6
Bericht	test	-	-	-	31	19	50	-
Kommentar	test	-	19	-	-	14	67	-
Reportage	test	-	89	8	3	-	-	-

Table 8: DeGeKo Presse Topic Distinct Set # of documents

...	Featureset Classifier	**Basic** F1 score	**Full** F1 score	**Selected** F1 score	**LDA Cont retrain** F1 score	**LDA Stop full** F1 score	**LDA Cont full** F1 score
original	$LinDA$	.68	.65	.54	.56	.67	.56
	$FOREST_{entropy}$	.75	.78	**.79**	**.70**	**.69**	**.82**
	$SVM_{vanilla}$	time	.70	.73	.68	.70	.79
distinct	$LinDA$	.63	.61	.48	.37	.63	.65
	$FOREST_{entropy}$	.68	.69	**.68**	.54	**.68**	**.70**
	$SVM_{vanilla}$	.63	.65	.65	**.61**	.66	.69

Table 9: DeGeKo Topic Stability Compared Results

4.4 Topic Distinct Set

Theoretically, a text from any given genre can be about any given topic, yet it is clear that covariances exist between genre and topic, with some genre/topic combinations more likely than others. Because both exploit low-level features to make predictions, a feature indicative of topic benefits a genre classifier through correlations in the training corpus. However, if the topics addressed in a genre can change unpredictably over time, such correlated features can harm performance. Petrenz and Webber (2011) found that neither character-4-grams nor bag-of-words models actually learn genre, but drop from 98% F1 to 38% (with char4) on three classes when topic is not held stable.

To test whether LDA topics are stable over a changing topic distribution, we create a subcorpus with the three press genre, where the topic annotations in our corpus are most reliable. Crucially, the distributions of topics for training data vs. test data are distinct. This yields two corpora: *Original & Distinct*. See Table 8 for distribution of documents over topics and genre. See Table 9 for classification results over changing topics.

We retrain LDA on the subcorpora and compare classification results to LDA trained on the full corpus, and against our style features. We find that each model compares unfavorably in the unstable topic setting, e.g. the FOREST&SELECTED model loses 11 F1 points. In the unlikely case that we have a huge genre corpus available for training LDA, the model is comparable to the style feature set (which would be theoretically possible if we feed new documents to our gensim model). The retrained LDA model compares badly for all models. This shows that (a) LDA needs as much training data as it can get, and (b) LDA is not robust against changing topics.

5 Characterizing register

A major advantage of our corpus is that we do not need sophisticated covariance metrics for the analysis of stylistic variation. In our setup, we can interpret class feature loadings, and we can validate our linear classifier with a simple F1 metric. We achieve .81 F1 score. The error stems mostly from press variety. The details of our register characterization approach are described in Section 3.5.

For each class, we retrieve the 80 features with the largest coefficient (40 negative & 40 positive) and use them for a qualitative analysis based on hypotheses formed on prior investigations (Breuer and Eroms, 2009) and to identify feature agglomerations that are apparent in a comparative setup (e.g. scientific text uses lots of connectives, particularly contrastive connectives). Figures 1 and 2 show such coefficient plots for advertising and academic writing. We next discuss, for four representative registers, the features most strongly associated, according to the method just described.

Gebrauchstext / Advertising (persuasion) Advertising often features *repetition, named entities, proper nouns* with the according *compositional parts* and *adjectives*, *plural pronouns* of *first and third person*, and also *attributive possessive pronouns*. We rarely find verbs or articles. So ads feature *object reference* and *blunt language* (nominal style but rarely articles). We find a *simple syntax, but lexical diversity* (high type/token ratio, short sentences, no sub. conj.) and *overt persuasion* (Positive sentiment, Certainty).

Presse / Bericht (report) Reports feature most prominently *present tense, passive voice, indirect speech* (subjunctive), *facts* (indicative) and *information* (num., art., NN, NE, ADJ). Also, by a positive loading of *prepositions, adverbs, reflexive pronouns* and negative loading of sub. conj., we conjecture a *balanced, compact style*.

Literature / Novel (storytelling) Storytelling stands out through the use of the *past tense and the third person (V.3.past, 'damals')*. We also find quite *long sentences* (almost 30 words on average), consequently many commas, and an aesthetic feature: *alliteration*.

Wissenschaft / Academic texts (Linguistik online) Academic writing (unsurprisingly) shows *complex exposition* and *argumentation* with many *(contrastive) connectives* (dass, sowohl, einerseits, hinsichtlich, bzw., also), *diverse punctuation* (parentheses, slashes) and the LIWC classes *insight, causation, communication*. Furthermore, this text genre uses fairly *abstract language*, as we find no concreteness and no arousal. We find a lot of *foreign material* (we use linguistics papers), and a prominent *focus on the future* (liwc). Apparently, academic writing is assonant.

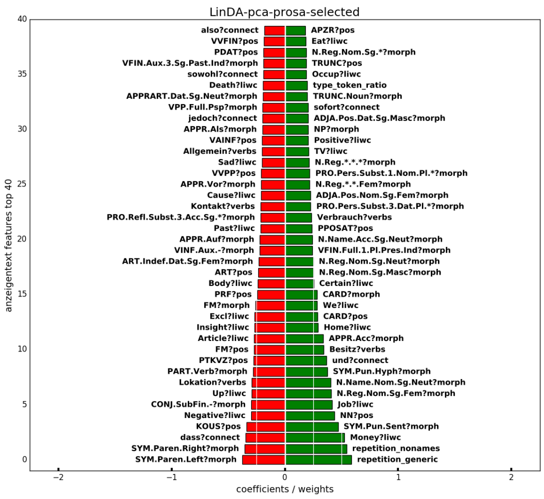

Figure 1: Feature loading for advertising

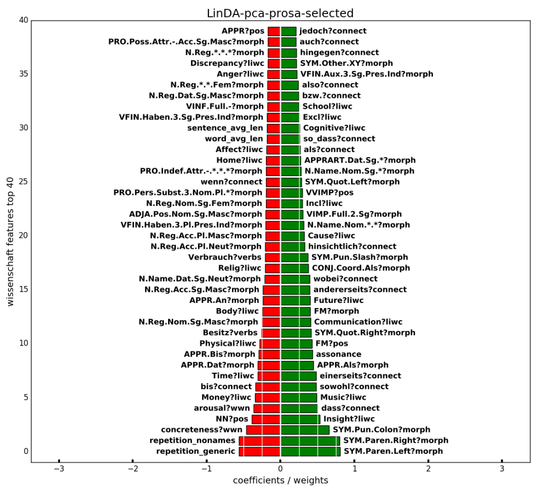

Figure 2: Feature loading for academic text

6 Conclusion

We have developed a genre taxonomy (for German) based on prototype semantics that can be used for a comparative register analysis, modelling a central aspect of situative text use: communicative purpose of text.

We find that fine grained morphology, surface cues and psycholinguistic word norms allow us to reason about situational text embedding, while – given enough training data – Latent Dirichlet Allocation can approximate genre distinctions, seeing that certain topics are prevalent in most genre categories. However, LDA is not stable over changing topic distributions under constant genre.

Future work should look at the communicative/situative function of constituency tree features, as they have proven to be useful e.g. for authorship attribution or deception detection. Also, the dimension of aesthetic style features (foregrounding) has typically been ignored in register research, as those are not necessarily functional. Given the abundance of material, we should look at press variety only. We have seen that report, commentary and reportage are prone to be confused, particularly by linear models. As humans also have a problem here, we have to conclude that they are not as clearly distinguished as other genre. Furthermore, press includes genre categories that are not as prototypical as the ones selected here (Dossier, Portrait, Feuilleton, Leitartikel). There are promising results (Sharoff, 2016) to view genre as topology, not as typology.

Finally, future research might benefit from word embeddings and particularly morphological embeddings to model stylistic variation.

Acknowledgements

The bulk of this work was conducted while the first author was a Master's student at the University of

Heidelberg, with support of the Leibniz Science Campus (LiMo). We thank Yannick Versley, Katja Markert, Josef Ruppenhofer and Stefan Blohm for helpful discussions and our annotators Elena Vaporova, Sebastien Nicolay and Christian Dueck. Thanks also to the anonymous reviewers for helpful comments.

References

Shlomo Argamon and Moshe Koppel. 2010. The rest of the story: Finding meaning in stylistic variation. In *The Structure of Style*, Springer, pages 79–112.

Shlomo Argamon and Moshe Koppel. 2012. Systemic functional approach to automated authorship analysis, a. In *JL & Pol'y 21*, page 299.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. WALTER DE GRUYTER & CO, volume 23, pages 321–346.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58(6):802–822.

Douglas Biber. 1989. A typology of english texts. *Linguistics 27.1* pages 3–44.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and linguistic computing* 8(4):243–257.

Douglas Biber, Ulla Connor, and Thomas A Upton. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*, volume 28. John Benjamins Publishing.

Douglas Biber and Susan Conrad. 2009. *Register, genre, and style*. Cambridge University Press.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Ulrich Breuer and Hans-Werner Eroms. 2009. *Stil und Stilistik. Eine Einführung*. Grundlagen der Germanistik 45.

Winthrop Nelson Francis and Henry Kučera. 1964/79. *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University, Department of Lingustics.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das deutsche textarchiv: Vom historischen korpus zum aktiven archiv. *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland, 20./21. September 2010, Köln. Beiträge der Tagung, 2., ergänzte Fassung* pages 157–161.

Stefan Th John Newman Gries and Cyrus Shaoul. 2011. N-grams and the clustering of registers. *Empirical Language Research Journal 5.1* .

Udo Hahn, Franz Matthies, Erik Faessler, and Johannes Hellrich. 2016. Uima-based jcore 2.0 goes github and maven central — state-of-the-art software resource engineering and distribution of nlp pipelines.

Birgit Hamp, Helmut Feldweg, et al. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. pages 9–15.

Verena Henrich and Erhard W Hinrichs. 2010. Gernedit-the germanet editing tool. In *ACL (System Demonstrations)*. Citeseer, pages 19–24.

Nancy et al. Ide. 2008. Masc: The manually annotated sub-corpus of american english. In *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

S Johansson, G Leech, and H Goodluck. 1978. Manual of information to accompany the lancaster-olso/bergen corpus of british english, for use with digital computers .

Ioannis Kanaris and Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Tools with Artificial Intelligence, 2007. ICTAI 2007*. 19th IEEE International Conference on. Vol. 2. IEEE.

J. Karlgren and D. Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *In Proc. of the 15th. International Conference on Computational Linguistics (COLING 94)*. Kyoto, Japan, page 1071 – 1075.

Moshe Koppel, Navot Akiva, and Ido Dagan. 2003. A corpus-independent feature set for style-based text categorization. In *IJCAI-2003 Workshop on Computational Approaches to Text Style and Synthesis, Acapulco, Mexico*.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The german reference corpus dereko: A primordial sample for linguistic research. In *LREC*.

Olaf Lahl, Anja S Göritz, Reinhard Pietrowsky, and Jessica Rosenberg. 2009. Using the world-wide web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 german nouns. *Behavior Research Methods* 41(1):13–19.

Ekaterina Lapshinova-Koltunski, Kerstin Kunz, and Marilisa Amoia. 2012. Compiling a multilingual spoken corpus. In *Proceedings of the VIIth GSCP*

International Conference: Speech and Corpora. Firenze: Firenze University Press.

D. Lee. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the bnc jungle. In *Language Learning and Technology*, page 5(3):37–72.

Sven Meyer zu Eißen and Benno Stein. 2004. Genre Classification of Web Pages: User Study and Feasibility Analysis. In Susanne Biundo, Thom Frühwirth, and Günther Palm, editors, *Advances in Artificial Intelligence. 27th Annual German Conference on AI (KI 04)*. Springer, Berlin Heidelberg New York, volume 3228 LNAI of *Lecture Notes in Artificial Intelligence*, pages 256–269.

R. J. Ide N. Su S. an Stuart J. Passonneau. 2014. Biber redux: Reconsidering dimensions of variation in american english. In COLING, pages (pp. 565–576).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works* .

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. In *Computational Linguistics 37.2*. pages 385–393.

Georg Rehm, Marina Santini, Alexander Mehler, Pavel Braslavski, Rüdiger Gleim, Andrea Stubbe, Svetlana Symonenko, Mirko Tavosanis, and Vedrana Vidulin. 2008. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *LREC*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology* 4(3):328–350.

Marina Santini. 2005. Itri-05-02 linguistic facets for genre and text type identification: A description of linguistically-motivated .

Marina Santini. 2007. Automatic identification of genre in web pages. diss. .

Marina Santini, Alexander Mehler, and Serge Sharoff. 2010. Riding the rough waves of genre on the web. In *Genres on the Web. Springer Netherlands*, pages 3–30.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. Smor: A german computational morphology covering derivation, composition and inflection. In *LREC*.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging.

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*.

Serge Sharoff. 2016. Functional text dimensions for annotation of web corpora .

Serge Zhili Wu Sharoff and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *LREC*.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000a. Automatic text categorization in terms of genre and author. In *Computational linguistics 26.4*, pages 471–495.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000b. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics*.

Gerard Steen. 1999. Genres of discourse and the definition of literature. *Discourse Processes* 28(2):109–120.

A. Stubbe and C. Ringlstetter. 2007. Recognizing genres. In *In Santini, M. and Sharoff, S., editors, Proc. Towards a Reference Corpus of Web Genres*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

John Taylor. 1995. Linguistic categorization: Prototypes in linguistic theory. *and Categorization, JR Linguistic Clarendon: Oxford University Press* .

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. pages 83–82.

Christian Weiß. 2005. Die thematische erschließung von sprachkorpora. *Mannheim: Institut für Deutsche Sprache.(= OPAL-Online publizierte Arbeiten zur Linguistik, 1/2005)* .

Markus Wolf, Andrea B Horn, Matthias R Mehl, Severin Haug, James W Pennebaker, and Hans Kordy. 2008. Computergestützte quantitative textanalyse: Äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica* 54(2):85–98.