Differentiating Social Media Texts via Clustering

Hannah Seemann*, Tatjana Scheffler

Germanistisches Institut, Ruhr-Universität Bochum, Germany

Abstract

We propose to use clustering of documents based on their fine-grained linguistic properties in order to capture and validate text type distinctions such as medium and register. Correlating the bottom-up, linguistic feature driven clustering with text type distinctions (medium and register) enables us to quantify the influence of individual author choice and medium/register conventions on variable linguistic phenomena. Our pilot study applies the method to German particles and intensifiers in a multimedia corpus, annotated for register. We show that German particles and intensifiers differ across both register and medium. The clustering based on the linguistic features most closely corresponds to the medium distinction, while the stratification into registers is reflected to a lesser extent.

Keywords

clustering, social media, register, media, German modal particles, intensifiers

1. Introduction

In this pilot study we investigate the use of clustering to capture macro-level distinctions between texts. We construct a bottom-up view of textual similarities via clustering based on their specific linguistic features. We compare the results with annotations of the medium and register of the texts to see to what extend effects of the used medium and register can be differentiated from individual author's variability.

It is known that the linguistic phenomena (e.g., word choice, use of tenses, punctuation, etc.) found in a text are shaped by many factors. In particular, highly variable phenomena such as discourse particles are known to be influenced by a wide range of aspects. Such aspects of text level variation can be sociolinguistic factors like author demographics and identity, author persona, or simply individual style, as investigated by sociolinguists [35, 20, 27, 30, 1] and corpus linguists [21, 12, 28]. Furthermore, writers also adapt to external circumstances of the utterance situation, such as the mode, medium, topic, or register (the situational context of language use) [4, 2]. For example, the "conceptual orality" theory proposes that a conceptual mode (spoken or written) is realized by a language producer by using different linguistic means in informal (conceptually spoken) vs. formal (conceptually written) language [17]. Different media are located in different places on the conceptual orality scale from typical spoken interaction to written text. Other research proposes that the register of a text influences the

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

*Corresponding author.

hannah.seemann@rub (H. Seemann); tatjana.scheffler@rub.de (T. Scheffler)

https://tscheffler.github.io/ (T. Scheffler)

0000-0001-8568-2124 (H. Seemann); 0000-0001-7498-6202 (T. Scheffler)

linguistic features that can be found in it, to the extent that linguistic features can be used to distinguish between different registers [5, 8].

So while both the author as well as various external aspects are known to influence linguistic variables, it is difficult to pinpoint to what extent each linguistic feature depends on each of the influences. The reason for this is that natural corpus data typically only covers a single medium or register, or conflates all categories: individual authors only contribute in one medium, each medium contains different registers or wildly different topics than the others, or the corpus is balanced for genre but it is not possible to track individual authors.[1]

In this paper, we make use of a social media corpus containing data from two different media (blogs and tweets), but covering the same set of 44 authors, the same topic (parenting and family life), and the same three registers (more detail below). We cluster the texts in our corpus using the relative frequency of two highly variable linguistic features, German modal and intensifying particles, found in each user's texts, divided by medium and register. We then compare the resulting clustering with the groupings based on register or medium to assess whether the linguistic features reflect these external aspects of the utterance situation.

We find that both medium and register are positively correlated with the clustering of documents based on linguistic features, where the alignment is better for the medium distinction than for register. We argue that our method makes it possible to tease apart the individual influence of medium, register, as well as individual author properties on the linguistic features studied.

The tables and scripts used in this paper can be accessed via the Open Science Framework.[2]

2. Categorizing texts: Register and medium

Various concepts have been used to characterize the situational circumstances in which a discourse is produced, as these directly or indirectly influence the way the discourse is shaped: text type, genre, topic, register, and others [4, 19]. In this study, we focus on the dimensions of medium and register.

The medium is the specific communication channel via which an utterance is made and reaches its addressee, such as television, phone, oral speech, Twitter, or Facebook. This notion is helpful in distinguishing between different communication situations specifically related to different so-called social media, as each medium carries its own affordances. The affordances a medium offers its users determine in which way the user and medium can interact [13, 36], and have subtle influences on the linguistic behavior of users (e.g., whether a post will be publicly visible or only to my friends might influence whether I will use a swear word). In our work, we study the two media blog posts and tweets. Both are written, but exhibit many informal and variable linguistic properties. They occupy different locations in the conceptual orality space [17]. The data will be described in more detail in §3.

[1] One example is the Ontonotes corpus, which contains a range of text types in both spoken and written language, but no overlap between medium and register, or individual authors: https://catalog.ldc.upenn.edu/LDC2013T19. A notable exception, pointed out to us by a reviewer, is the Early Modern Multiloquent Authors (EMMA) Corpus, which tracks changes of authors' language use over their lifetime in different spoken and written registers: https://www.uantwerpen.be/en/projects/mind-bending-grammars/emma-corpus/.

[2] https://osf.io/kjnsu/

While the notion medium is based on the technical implementation of a discourse, register takes various aspects of the extralinguistic context into account, such as whether a specific discourse is interactive, what the relation of the disourse participants is like, whether it is emotionally charged or its purpose is merely the exchange of information, etc. [4]. Due to this interplay of contextual properties, register has, following Biber [3], frequently been characterized as multidimensional. Some researchers even propose to do away with a language-external inventory of register altogether [6], and want to instead represent registers as combinations of linguistic features present in the text. We do not follow this approach here, since we want to specifically investigate the influence of register on linguistic features – and therefore the registers themselves must be delineated independently.

In this work, we distinguish the registers Informative, Narrative, and Persuasive, based on situational properties such as the purpose of conversation (passing on information, reporting on life events, argumentation, respectively), the interactivity with the addressee, and the author involvement (both ranging from low for Informative to high for Persuasive). Linguistic features, with the exception of pronouns, were not used to distinguish between register dimensions. All registers are present in each of the two media.

3. Data

To compare whether register, the medium or individual authorship has the most influence on text similarity, it is necessary that we look at texts from the same author in different media and registers. To this end, we collected a corpus of German language blog posts and tweets from the same 44 individuals, but in a single domain: parenting. The community of parenting bloggers is relatively coherent and writes about similar topics both in their blogs, as well as on Twitter.

Blogs are a long-form text format with limited interactivity, while tweets are short posts (all our tweets are still under 140 characters) which allow direct responses; both media are public. Thus, the two media offer different types of communicative situations, but they are both available for all three registers introduced in §2, depending on the individual usage.

We constructed the corpus using the Twitter API and the user's corresponding blog's RSS feed. The initial data collection was carried out in February, 2017, and the data used here comprises the 500 most recent tweets and the 5 or 10 (depending on availability) most recent blog posts. A more detailed description of the corpus and data collection can be found in [26] and on its website.[3] The resulting corpus consists of data from 44 authors, comprising 390 blog posts (~350k tokens) and 20,131 tweets (~300k tokens). All data has been manually pseudonymized.

We manually annotated each blog post with one register (Informative, Narrative, or Persuasive). Since the tweets are often too short to be assigned a clear register, we grouped them together and assigned one register to the entire tweet collection from one author, capturing the main usage of Twitter by that author. For tweet collections, we additionally allowed the intermediate registers Narrative-Informative and Narrative-Persuasive, denoting a mix between

[3] http://staff.germanistik.rub.de/digitale-forensische-linguistik/forschung/textkorpus-sprachliche-variation-in-sozialen-medien/

these registers.

In addition, all modal and intensifying particles were manually identified and disambiguated in the corpus, with the help of word lists, annotation guidelines, and additional trained linguistics students. We define these phenomena in the next section.

4. Variable linguistic features

German modal and intensifying particles are used differently in different communicative situations (spoken vs. written and formal vs. informal communication) and also depending on the author using them. Both types of particles are noninflected and modify the element in their scope, and both are generally assumed to be more frequent in speech or conceptually spoken language [14, 34, 32].

4.1. Modal particles

German modal particles are used to express the author's attitude towards a proposition or to make assumptions concerning the "common ground", the shared knowledge of author and reader [29], but they do not affect the truth conditions of a sentence [37]. (1) is an example of this function: In (1-a), the modal particle 'ja' is used to indicate that the fact that the author is writing is known from the sentence before. In (1-b), 'doch' is used to express the author's (negative) attitude towards the idea that only fathers who are in fact not able to pay refrain from paying.

(1) a. Wenn ich schreibe, kann ich immerhin nicht einschlafen wie gestern beim Staatsanwalt. Allerdings bekomme ich trotzdem nicht mit was passiert, weil ich ja schreibe.
'While I'm writing I can not fall asleep, as it happened yesterday at the prosecutor's office. But I still don't get what's happening because I'm JA writing.'
(blogposts-5487-3)[4]

b. @[USERNAME] Zu denken, nur diejenigen Väter würden nicht zahlen, die es nicht können, ist doch völlig weltfremd.
'@[USERNAME] It is DOCH naive to think that only fathers who are not able to pay won't do so.' (tweets-1123)

Due to the possibility of expressing multiple functions with one modal particle, the meaning of one modal particle can vary in different contexts. Additionally, not all modal particles have an exact match in other languages and they can not be directly translated to English [10]. Kratzer shows two examples of 'ja' that include a translation to English. In both cases, there is no word that matches the meaning of 'ja' exactly, it is rather the function of the modal particle that is translated:

(2) Ich bin ja ein Einzelkind.
'As you know, I am an only child.' [18]

[4]If not indicated otherwise, examples are from our corpus.

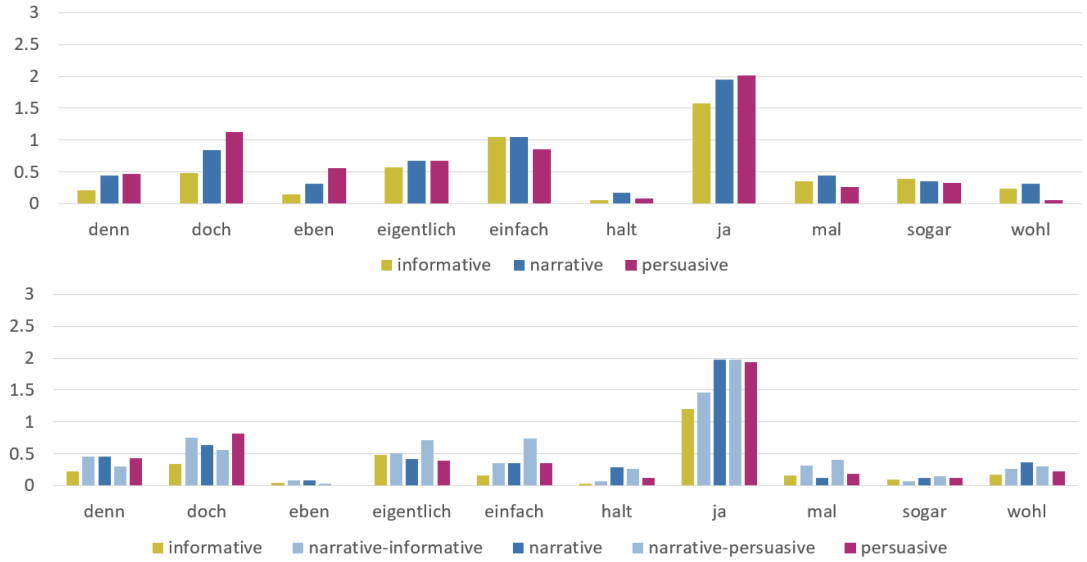

Figure 1: Top 10 modal particles in blog posts (top) and tweet collections (bottom) in each register. Counts are relative to the number of sentences in this medium and register

(3) Du hast ja ein Loch im Ärmel!
'Look, you have a hole in your sleeve!' [18]

The use of modal particles varies between individuals and linguistic modes [34, 11]. We therefore expect to see differences in particle use between different authors, but also between different media and registers. Figure 1 shows the distribution of the ten most frequent modal particles in our corpus divided by our register dimensions. As expected, there are differences in how frequently modal particles are used in different media and registers. With $\chi^2 = 2188$ and $p < 0.01$, it can be assumed that there is a dependency between modal particle count and medium/register. This indicates that modal particles can be used as a linguistic feature to cluster documents by medium and register.

4.2. Intensifiers

Intensifying particles (for short, 'intensifiers') can be used to boost or tone down the intensity of a gradable expression or utterance [22]. Similar to modal particles, there is inter-individual variation in the use of intensifiers, but intensifiers are subject to much more rapid change of use [15]. Even though they are assumed to be more frequent in speech than in written language, it has been shown that they are frequently used in written social media [25] and they can be found in our social media corpus, as well (see (4)).

(4) a. @[USERNAME] wieso kann ein Tattoo so brillante Farben haben? Wo hast du das machen lassen?
'@[USERNAME] How can a tattoo have such brilliant colours? Where did you get it?' (tweets-4677)

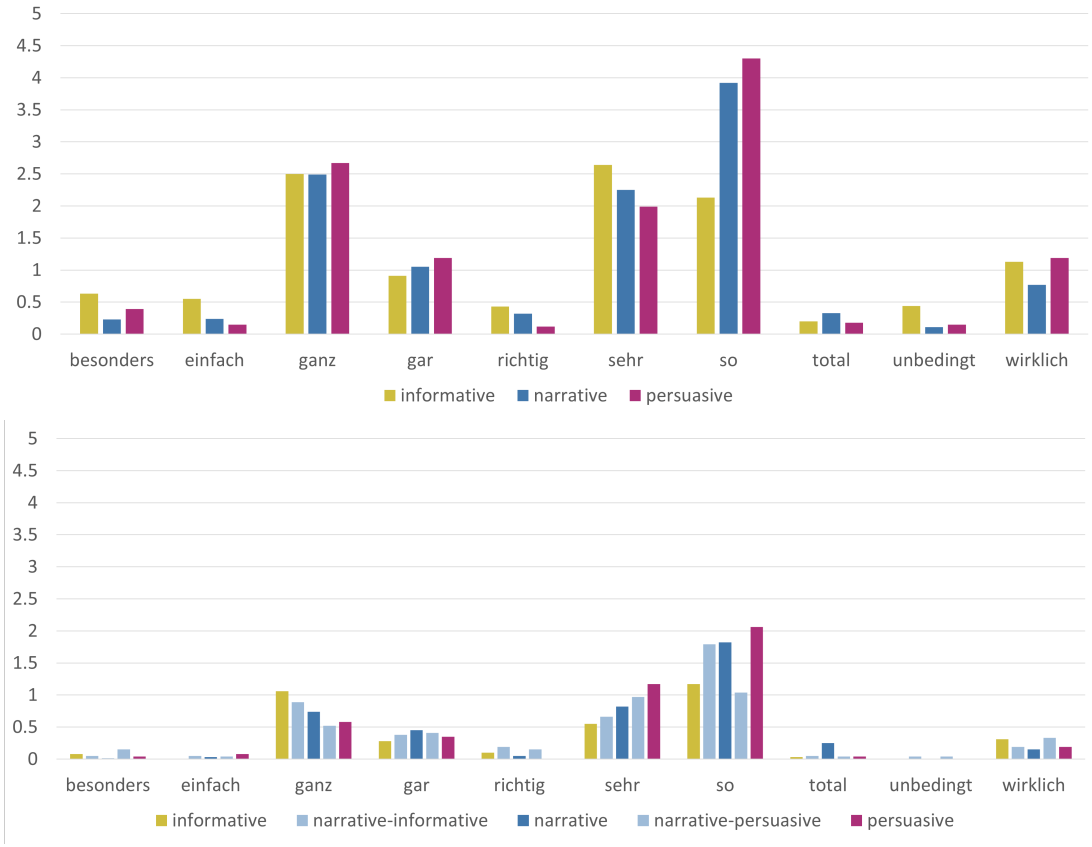

Figure 2: Top 10 intensifiers in blog posts (top) and tweet collections (bottom) in each register. Counts are relative to the number of sentences in this medium and register

b. Da gibt es wirklich tolle Sachen - @[USERNAME]
'There you can find really great things - @[USERNAME]' (tweets-7846)

An overview of German intensifiers can be found in [9], Breindl discusses (issues with) the categorization of German intensifiers [7]. A large-scale corpus study of intensifiers in spoken German was conducted by Stratton, showing that intensifiers are used quite frequently in spoken language and that the use of intensifiers varies by individual demographic characteristics [30]. This was previously shown for English intensifiers, as well [33].

Based on the previous sociolinguistic results, we expect that the use of intensifiers in social media varies by individual demographic factors of the authors (as shown for speech), but may also vary by medium and register. Figure 2 shows the distribution of the ten most frequent intensifiers in our corpus divided by our register dimensions. Similar to modal particles, different intensifiers are used more or less frequently in different media or registers. With χ^2 = 1062.2 and $p < 0.01$, it can be assumed that there is a dependency between intensifier count and medium as well as register. This indicates that intensifiers, as well, can be used as a linguistic feature to cluster documents by medium and register.

5. Clustering

It is our hypothesis that register influences the low-level linguistic choices in addition to the medium or the author style. Starting from this hypothesis, we carry out a pilot study to cluster texts in a data-driven way based on their linguistic features. We want to find out whether these features enable us to distinguish registers from each other, e.g. rather than clustering each user's tweet collection with their blog posts (as would be expected if the features reflect only individual linguistic style). The features we use are the per-sentence frequency of the top 10 modal particles and intensifiers found. We use the relative frequency of every feature to take the different lengths of the blog posts/tweet collections into account. Each document is represented by a vector containing the relative frequencies of the particles and intensifiers (see Table 1).

Table 1
Example of vectors used for clustering.

Document	Vector (top 10 modal particles and top 10 intensifiers)
b_1095_I	[0, 0, 1.13, 0.38, 0.38, 1.13, 1.5, 0, 0, 0, 0, 0, 0, 0, 0, 0.39, 0.39, 0.78, 0, 0]
b_1095_N	[0, 0, 0, 0, 0, 1.75, 0, 0, 0, 0, 0, 3.57, 0, 0, 1.79, 0, 0, 0, 1.79, 0]
t_1095_I	[0, 0, 1.59, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.64, 0, 0, 0]

Each user's texts are split by medium (blogs and tweets) as well as register into a minimum of 2 (and a maximum of 4) documents per user. For example, the user 1095 is represented in three documents: b_1095_I (containing all Informative blog posts), b_1095_N (Narrative blog posts), and t_1095_I (containing all tweets, which were annotated as Informative). Document names reflect the medium, user id, and register, in order.

We used the agglomerative clustering algorithm implemented in Python's scikit-learn package [23]. Agglomerative clustering successively groups the most similar documents together, until all clusters have been merged. Figure 3 shows the clustering results.

Out of nine groups of clusters, three contain mainly data points labelled as using Narrative register (document names shown in blue), two contain mainly data points labelled as Narrative or Informative (yellow) and one cluster contains data points labelled as Narrative or Persuasive (red). The last three groups of clusters contain data points from all of the register dimensions in equal amount. Out of these nine groups of clusters, six contain data points labelled as coming from blog posts and of the remaining three clusters, only one contains more data points labelled as coming from tweet collections than from blog posts.

Out of 64 pairs of documents that were clustered directly together, 22 had the same register label. 18 out of these 22 cases are nodes where blog posts and tweet collections were clustered together, 23 other cases are nodes which contain data points from the same medium. In one node, neither medium and register nor author are the same. There are only two cases where the same user's blog posts and tweets were clustered together. For one author, both documents were also labelled as Informative (blog post and tweet collection), for the other author, one was labelled as Informative and one as Narrative.

Even though particles and intensifiers are not enough data to cluster the same register to-

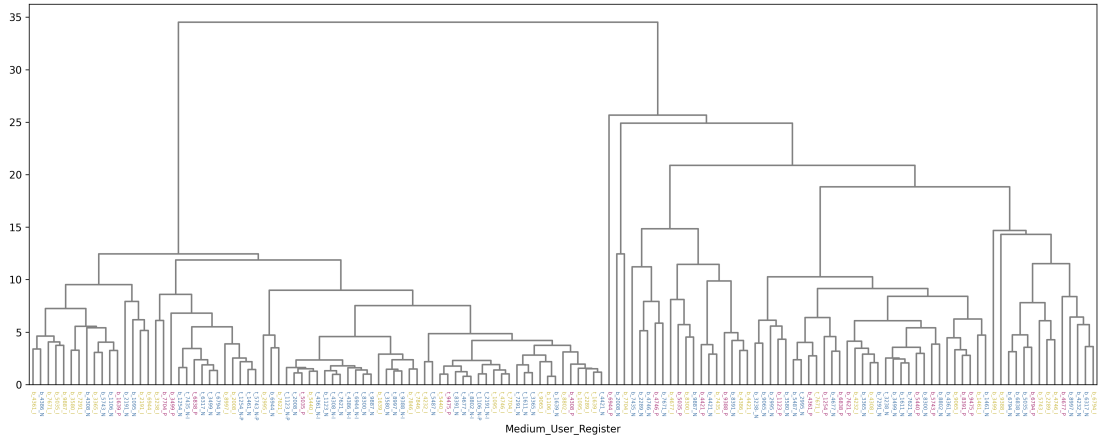

Figure 3: Agglomerative clustering of our data

gether in all cases, the algorithm still tends to cluster documents from the same register together, as opposed to grouping the same user's blog posts and tweet collection. This indicates that medium and register influence how users write and that writing in a specific register has an independent impact on linguistic choice from just the medium in which the user writes, and their personal linguistic style.

For evaluation, we compared the correlation of the clustering by linguistic features with the register distribution on the one hand, and the medium on the other. As a quality measure, we used the V-measure [24], which balances homogeneity (whether a cluster contains only documents of one class, i.e. belonging to one register/medium) with completeness (to what extent all documents from one class are put into the same cluster). We applied the V-measure implemented in Python's scikit-learn package to the comparison between a 20-cluster crosscut of the hierarchical clustering shown in Figure 3, and the grouping by register/medium as indicated by the document labels. The results show that the clustering corresponds more closely to the grouping by medium (V = 0.2246) than the grouping by register (V = 0.0839). Homogeneity of clusters is also higher for medium (0.5449) than for register (0.1419), though both show a positive correlation.[5]

6. Conclusion

We proposed clustering documents in a multi-media and multi-register corpus of German parenting bloggers by their usage of modal and intensifying particles. The method can be used to

[5]A reviewer suggests that Twitter's length restrictions for tweets lead to authors choosing shorter and, in general, less intensifiers than in blog posts. This effect can also be seen for modal particles, though it is less strong. Thus, the question arises whether conflating both types of particles to cluster the document is reasonable. In fact, using only intensifiers for clustering leads to a slightly better V-measure for grouping by register (V = 0.1293), but does worse for grouping by medium (V = 0.1628). Using only modal particles leads to worse V-measure results than using only intensifiers. The prevalence of the medium probably arises from this differing use of particles in both media, possibly due to length restrictions in tweets.

generate bottom-up clusters of documents (based on linguistic features) and to compare these clusters to groupings of the same documents by medium and register. We show that both the medium and the register dimensions are reflected in the variation in our linguistic features. For our feature groups, modal and intensifying particles, the medium has a bigger effect than the register. We would like to argue that clustering enables us to determine the relative importance of individual author properties and text level properties (medium and register) on the linguistic expressions found in a text.

After having conducted this pilot study, we will apply this method to a different dataset to test the reproducibility of our results. Another natural next step would be to integrate other linguistic phenomena as features in the clustering. On the one hand, one could choose phenomena that have been argued to vary based on register or medium. On the other hand, linguistic variation in small-scale features has been used to account for individual author style, for example in authorship attribution or author profiling [16, 31]. If the features proposed in authorship analyses are integrated in our clustering account, it may be possible to tease apart influences based on medium or register from individual author style choices.

Acknowledgments

We would like to thank Lesley-Ann Kern for discussion, and the annotators in preparing the corpus data. We are grateful to the anonymous reviewers for their helpful comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project ID 317633480, SFB 1287.

References

[1] J. Androutsopoulos. *Deutsche Jugendsprache: Untersuchungen zu ihren Strukturen und Funktionen*. Frankfurt a. M.: Peter Lang, 1998.

[2] J. Androutsopoulos. "Mediatization and Sociolinguistic Change. Key Concepts, Research Traditions, Open Issues". In: *Mediatization and sociolinguistic change*. Ed. by J. Androutsopoulos. Linguae & litterae v. 36. Berlin ; Boston: De Gruyter, 2014, pp. 3–48.

[3] D. Biber. "Using Register-Diversified Corpora for General Language Studies". In: *Computational Linguistics* 19.2 (1993), pp. 219–241.

[4] D. Biber and S. Conrad. *Register, Genre, and Style*. 2nd ed. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press, 2019. DOI: 10.1017/9781108686136.

[5] D. Biber and J. Egbert. "Register Variation on the Searchable Web: A Multi-Dimensional Analysis". In: *Journal of English Linguistics* 44.2 (2016), pp. 95–137. DOI: 10.1177/0075424216628955.

[6] F. Bildhauer, E. Pankratz, and R. Schäfer. *Corpus, Inference, and Models of Register Distribution*. Talk. 2021.

[7] E. Breindl. "Intensitätspartikeln". In: *Handbuch der deutschen Wortarten*. Berlin, New York: De Gruyter, 2007, pp. 397–422.

[8] I. Clarke and J. Grieve. “Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted between 2009 and 2018”. In: *Plos One* 14.9 (2019). Ed. by C. M. Danforth. DOI: 10.1371/journal.pone.0222062.

[9] U. Claudi. “Intensifiers of Adjectives in German”. In: *Language Typology and Universals* 59.4 (2006), pp. 350–369. DOI: 10.1524/stuf.2006.59.4.350.

[10] L. Degand, B. Cornillie, and P. Pietrandrea, eds. *Discourse Markers and Modal Particles: Categorization and Description*. Pragmatics & beyond new series volume 234. Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2013.

[11] G. Diewald. “Abtönungspartikel”. In: *Handbuch der deutschen Wortarten*. Berlin, New York: de Gruyter, 2009, pp. 117–141.

[12] L. Fonteyn and A. Nini. “Individuality in Syntactic Variation: An Investigation of the Seventeenth-Century Gerund Alternation”. In: *Cognitive Linguistics* 31.2 (2020), pp. 279–308. DOI: 10.1515/cog-2019-0040.

[13] J. J. Gibson. “The Theory of Affordances”. In: *The Ecological Approach to Visual Perception*. Psychology Press. New York London: Taylor & Francis Group, 2014.

[14] D. Hartmann. “Syntaktische Funktionen der Partikeln eben, eigentlich, einfach, nämlich, ruhig, vielleicht und wohl. Zur Grundlegung einer diachronischen Untersuchung von Satzpartikeln im Deutschen”. In: *Die Partikeln der deutschen Sprache*. Ed. by H. Weydt. Berlin, New York: De Gruyter, 1979, pp. 121–138.

[15] R. Ito and S. Tagliamonte. “Well Weird, Right Dodgy, Very Strange, Really Cool: Layering and Recycling in English Intensifiers”. In: *Language in Society* 32.2 (2003), pp. 257–279. DOI: 10.1017/s0047404503322055.

[16] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, and M. Potthast. “Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection”. In: *Clef*. 2018.

[17] P. Koch and W. Oesterreicher. “Sprache der Nähe - Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte”. In: *Romanistisches Jahrbuch* 36 (1985), pp. 15–43. DOI: 10.15496/publikation-20410.

[18] A. Kratzer. “Beyond ’Ouch’ and ’Oops’. How Descriptive and Expressive Meaning interact”. In: *Cornell Conference on Theories of Context Dependency* (1999), pp. 1–6.

[19] D. Y. Lee. “Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Nevigating a Path through the BNC jungle”. In: *Language Learning and Technology* 5 (3 2001), pp. 37–72. URL: https://ro.uow.edu.au/artspapers/598.

[20] L. MacKenzie. “Perturbing the Community Grammar: Individual Differences and Community-Level Constraints on Sociolinguistic Variation”. In: *Glossa: a journal of general linguistics* 4.1 (2019). DOI: 10.5334/gjgl.622.

[21] A. Nini. “An Authorship Analysis of the Jack the Ripper letters”. In: *Digital Scholarship in the Humanities* 33.3 (2018), pp. 621–636. DOI: 10.1093/llc/fqx065.

[22] C. v. Os. *Aspekte der Intensivierung im Deutschen*. Studien zur deutschen Grammatik 37. Tübingen: Narr, 1989.

[23] 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. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[24] A. Rosenberg and J. Hirschberg. "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 410–420. URL: https://aclanthology.org/D07-1043.

[25] T. Scheffler. "Conversations on Twitter". In: *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*. Ed. by D. Fišer and M. Beißwenger. Ljubljana: University Press, 2017, pp. 124–144.

[26] T. Scheffler, L.-A. Kern, and H. Seemann. "Individuelle linguistische Variabilität in sozialen Medien". In: *Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022*. Ed. by M. Kupietz and T. Schmidt. Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache (CLIP) 11. Tübingen: Narr, forthcoming.

[27] E. Schleef. "Individual Differences in Intra-Speaker Variation: T-Glottalling in England and Scotland". In: *Linguistics Vanguard* 7.s2 (2021). DOI: 10.1515/lingvan-2020-0033.

[28] H.-J. Schmid and A. Mantlik. "Entrenchment in Historical Corpora? Reconstructing Dead Authors' Minds from their Usage Profiles". In: *Anglia* 133.4 (2015), pp. 583–623. DOI: 10.1515/ang-2015-0056.

[29] R. Stalnaker. "Assertion". In: *Syntax and semantics 9: Pragmatics*. Ed. by P. Cole. Vol. 9. New York, NY, USA: Academic Press, 1978, pp. 315–332.

[30] J. M. Stratton. "Adjective Intensifiers in German". In: *Journal of Germanic Linguistics* 32.2 (2020), pp. 183–215. DOI: 10.1017/s1470542719000163.

[31] K. Sundararajan and D. Woodard. "What Represents "style" in Authorship Attribution?" In: *Coling*. 2018.

[32] S. A. Tagliamonte and D. Denis. "Linguistic Ruin? LOL! Instant Messaging and Teen Language". In: *American Speech* 83.1 (2008), pp. 3–34. DOI: 10.1215/00031283-2008-001.

[33] S. A. Tagliamonte. "So Different and Pretty Cool! Recycling Intensifiers in Toronto, Canada". In: *English Language and Linguistics* 12.2 (2008), pp. 361–394. DOI: 10.1017/s1360674308002669.

[34] M. Thurmair. *Modalpartikeln und ihre Kombinationen*. Linguistische Arbeiten 223. Tübingen: M. Niemeyer, 1989.

[35] W. Wolfram. "Variation and Language: Overview". In: *Encyclopedia of Language & Linguistics*. Elsevier, 2006, pp. 333–341. DOI: 10.1016/b0-08-044854-2/04256-5.

[36] Y. Zhao, J. Liu, J. Tang, and Q. Zhu. "Conceptualizing Perceived Affordances in Social Media Interaction Design". In: *Aslib Proceedings* 65.3 (2013), pp. 289–303. DOI: 10.1108/00012531311330656.

[37] M. Zimmermann. “Discourse Particles”. In: *Semantics*. Ed. by P. Portner, C. Maienborn, and K. v. Heusinger. Vol. 2. Handbücher zur Sprach- und Kommunikationswissenschaft HSK. Berlin: Mouton de Gruyter, 2011, pp. 2011–2038.