R. Reussner, A. Koziolek, R. Heinrich (Hrsg.): INFORMATIK 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021 1327

Exploring the Use of the Pronoun *I* in German Academic Texts with Machine Learning

Melanie Andresen[1], Dagmar Knorr[2]

Abstract: The use of the pronoun *ich* ('I') in academic language is a source of constant debate and a frequent cause of insecurity for students. We explore manually annotated instances of *I* from a German learner corpus. Using machine learning techniques, we investigate to what extent it is possible to automatically distinguish between different types of *I* usage (author *I* vs. narrator *I*). We additionally inspect which context words are good indicators of one type or the other. The results show that an automatic classification is not straightforward, but the distinctive features are in line with previous research. The results of the automatic classification are not perfect, but would greatly facilitate manual annotation. The distinctive words are in line with previous research and indicate that the author *I* is a more homogeneous class.

Keywords: annotation; academic language; German; machine learning; classification

1 Introduction

We present an exploratory study about the use of *ich* ('I')[3] in German academic texts by students. The main focus is on a quantitative exploration of different types of *I* using machine learning techniques, namely principal component analysis (PCA) and classification with a support vector machine (SVM). The task can be roughly understood as a case of word-sense disambiguation, even though the types of *I* differ functionally rather than semantically. Our aim is not to achieve full automation, but to deepen the understanding of the use of *I* from a humanities point of view by modeling its uses quantitatively.

The use of references to the author in academic language, most often realized by the pronoun *I*, is a source of constant debate and a frequent cause of insecurity for students. Authorial identity and self-reference in academic writing have therefore been a popular research topic ([Hy05], [Äd06], [Kr12]). We want to highlight two typologies of first person references that are suitable for empirical application: [TJ99] examine English academic essays written by students and suggest six types of first person references that form a continuum from low to high authorial power: The first two types, a) *I* as the representative (of the general public or the discourse community), and b) *I* as the guide through the essay, are low in

[1] Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung, Pfaffenwaldring 5b, 70569 Stuttgart, Germany, melanie.andresen@ims.uni-stuttgart.de
[2] Leuphana Universität Lüneburg, Schreibzentrum/Writing Center, Universitätsallee 1, 21335 Lüneburg, Germany, dagmar.knorr@leuphana.de
[3] For brevity, we will subsequently refer to the target word as *I*, even though we always mean the German *ich*.

doi:10.18420/inf2020_124

authorial power and mostly realized by *we* rather than *I*. The more powerful types include c) *I* as the architect of the essay, d) *I* as the recounter of the research process, e) *I* as the opinion-holder, and f) *I* as the originator.

Steinhoff [St07] explores the use of *I* in a corpus of German research articles and distinguishes between 1) the author *I* that comments on the text and guides the reader, corresponding to type c) by [TJ99] (*In the following chapter, I will present my results about ducks*), and 2) the researcher *I* refers to the research object, terminology, or claims by other researchers (*I use the term 'duck' as referring to the waterbird*), roughly summarizing d), e), and f) by [TJ99]. He further includes 3) the narrator *I* that gives subjective, often auto-biographic information (*I originally wanted to write about birds, but then I learned there are so many different kinds of birds*). While the author *I* and the researcher *I* are considered acceptable in academic writing, the narrator *I* is mostly deemed inappropriate.We decided to apply the simpler typology by Steinhoff to our data. To the best of our knowledge, there has been no empirical application of the model beyond the original study.

2 Data

Corpus. The texts we use for our experiments are taken from the learner corpus *KoLaS* ('Kommentiertes Lernendenkorpus akademisches Schreiben', [AK17]). *KoLaS* comprises of academic texts written by students that visited the *Writing Center Multilingualism* at the Universität Hamburg between 2011 and 2016. The corpus is very diverse with respect to text types, disciplines, language skills, and the progression in the writing process. For many texts, several versions and comments by writing tutors are available. For this study, we reduced the corpus to the first version of every text and excluded text types with a focus on personal reflection. This resulted in a corpus size of 330 texts. We use a learner corpus, because we expect to find a substantial number of the narrator *I* not common in published academic texts. Doing so, however, can lead to the possibility of there being language errors and it affecting the results.

Annotation. Four annotators classified the instances of *I* in the corpus. The annotators were students trained as writing tutors. Before partaking in the annotation, they participated in a workshop about Steinhoff's *I* types where they reflected on textual indicators for these categories (see [AK17] for a description of the workshop concept). In order to cover all instances of *I* in the corpus, we extended the tagset with categories for *I* in example sentences, *I* referring to the general public, and *I* in comments by writing tutors.

Data Extraction. Since the annotation project was originally aimed at a qualitative analysis, the annotation was carried out with the support of the *MAXQDA* tool[4]. Unfortunately, this

[4] VERBI Software, `https://www.maxqda.de/`.

tool does not offer a direct export of annotated spans with their context. While a large part of the data could be extracted from the database format *MAXQDA* offers, not all annotations could be extracted for technical reasons.[5] In total, 2784 instances of *I* could be extracted. For this analysis, we were only interested in the three types of *I* by Steinhoff. We therefore filtered the data set for those instances, where at least two annotators agreed on one of those three categories. The resulting data set (n=360) comprises of 213 instances of the author *I* and 122 instances of the narrator *I*. As there were only 25 instances of the researcher *I*, we decided to exclude this type from the analysis. The data set is publicly available at Zenodo.[6]

Inter-Annotator Agreement. The inter-annotator agreement was calculated using Krippendorff's alpha [Kr80]. The agreement on the full data set (n=2784) that could be extracted from the annotation files is 0.76, which is a substantial agreement, following the (rather generous) scale by [LK77]. However, the three types by Steinhoff are among the more difficult categories. The agreement on the data set filtered for these categories (n=360) drops to 0.56 (moderate agreement). This indicates that further refinement of the guidelines could be beneficial. However, we can reasonably assume that the phenomenon as such shows ambiguities that cannot always be resolved. Under these conditions, we cannot expect excellent classification results.

Feature Extraction. In our experiments, the frequency of words in the immediate context before and after the *I* served as features. We did experiments based on words in a context window of three, five, and seven words left and right of the *I*. The smallest context of three words turned out to be the most helpful. Larger context windows led to an increased impact of idiosyncrasies of individual texts. To further reduce this effect, we exclude all word forms that occur only once. This results in a vocabulary of 183 words, whose frequencies serve as our features.

3 Unsupervised Experiment: PCA

As an initial exploration of the data, we use an unsupervised experiment[7] to learn about patterns emerging from the data without enforcing our ideas about the typology. For this purpose, we use a principal components analysis (PCA). The PCA takes the original data set with many variables or dimensions (one for each context word type in our data) as input and transforms them into new dimensions that capture as much variation in the data as possible.

Figure 1 shows the distribution of all instances of the author *I* and the narrator *I* across the first two dimensions of the PCA. While the samples of the two classes overlap, there

[5] More specifically, it was not always possible to unambiguously identify the position of the annotated span in the full text.

[6] https://doi.org/10.5281/zenodo.3999304

[7] For the implementation, we use *Python 3* with *pandas* [Th20], *scikit-learn* [Pe11], and *matplotlib* [Hu07].

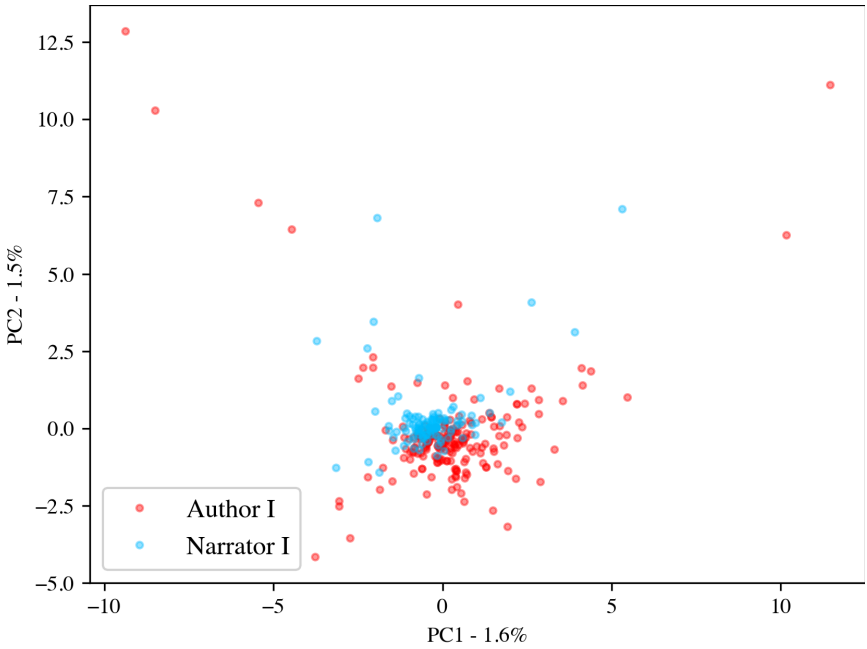

Fig. 1: Distribution of the author *I* and the narrator *I* in the first two dimensions of the PCA

is a tendency for the narrator *I* to score higher than the author *I* in the second dimension of the PCA. Notably, the variance in the data set that is explained by this dimension is rather low (1.5%). We conclude that a clear distinction between the types will not be possible. Nevertheless, the word frequency data analyzed using the PCA clearly contain some information that is related to the distinction of *I* types. The next section examines this information in greater detail.

4 Supervised Experiment: SVM

In addition to the PCA, we present a supervised experiment using a linear support vector machine. This type of classifier allows us to inspect the contribution of the individual features to the classification. These are important for our interpretation beyond classification scores.

Classification Success. Table 1 shows the results of a 5-fold crossvalidation. We include two baselines: One for a classifier that makes random choices and one that always votes for the majority class, the author *I*. We report precision, recall and f1-scores for the two classes together with a macro average and a weighted average that takes the number of instances

		Precision	Recall	F1-Score
SVM	Narrator *I*	0.72	0.77	0.74
	Author *I*	0.86	0.83	0.84
	Mean (macro)	**0.79**	**0.80**	**0.79**
	Mean (weighted)	**0.81**	**0.81**	**0.81**
Baseline (random)	Narrator *I*	0.50	0.50	0.50
	Author *I*	0.50	0.50	0.50
	Mean (both)	**0.50**	**0.50**	**0.50**
Baseline (majority class)	Narrator *I*	0.00	0.00	0.00
	Author *I*	0.62	1.00	0.77
	Mean (macro)	**0.31**	**0.50**	**0.39**
	Mean (weighted)	**0.39**	**0.62**	**0.48**

Tab. 1: Classification results for the SVM and two baseline models (random choice and majority class)

per class into account. The weighted means are always (equal or) higher as they give more weight to the majority class which also scores higher.

Our classifier clearly outperforms both baseline models in all metrics—with the obvious exception of recall for the majority class in the majority class baseline. For the author *I*, we achieve very good scores with a precision of 0.86 and a recall of 0.83. Both metrics are lower for the narrator *I*. This might be due to the fact that we have less examples of narrator *I*s in our data. To sum all scores up, our classifier achieves a mean f1-score of 0.79. This does not allow for full automation but is a very good result that could, for instance, serve as a useful pre-analysis to facilitate manual classification.

Features. In Figure 2, we can see the features that contributed most to the distinction between the two classes based on the coefficients of the SVM. Words on the left are indicators of the narrator *I* and words on the right are indicators of the author *I*. The best indicators for the author *I* have higher coefficients,[8] i. e. they are more helpful for the classifier. The words can be clearly interpreted with respect to the function of the author *I*: the verbs *werde* ('will'), *möchte* ('want to'), and *kann* ('can') are used to cataphorically announce what follows in the text. The word *Arbeit* ('work') is frequently used to refer to the text (analogous to *In this paper, I will. . .*). The indicators for the narrator *I*, on the other hand, have slightly lower coefficients and their interpretation is less straightforward. This can be due to the smaller sample for this type. Another explanation is that while the author *I* is limited to a rather fixed set of possible expressions, the narrator *I* has more freedom in wording as well as topic.

[8] Scores can be read from the side of the words that faces the center of the plot.

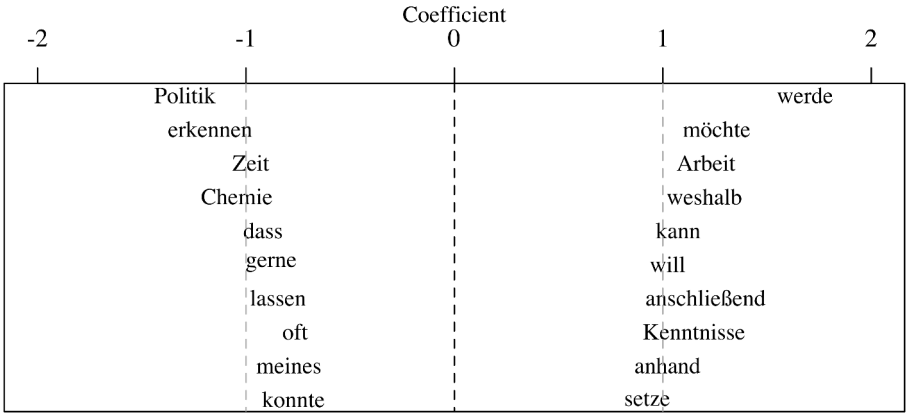

Fig. 2: Distinctive features for narrator *I* (left) and author *I* (right) based on SVM coefficients

5 Discussion and Future Work

Our exploration of *I* types in German academic language with machine learning shows promising tendencies, though the classification results are far from perfect. We consider the results highly beneficial for the understanding of the phenomenon from a humanities perspective. In particular, the features of the SVM allow for interpretations that build on existing research. One important result is that the author *I* appears to be a more homogeneous class than the narrator *I*. This is suggested by the higher scores in the classification, the higher SVM coefficients for features indicating the author *I*, and a more straightforward interpretability of these features. This result indicates that the author *I* type is restricted to a very specific function and a limited number of possible realizations at the text surface.

In order to obtain more stable and generalizable results, more data would be beneficial. This might also allow for the inclusion of the third type of *I*, the researcher *I*. A potential problem of the utilized data set is that it includes more than one instance of *I* per text. Consequently, these data points are not independent and their shared topic can have distorting effects. A larger data set would reduce this effect or allow us to include only one *I* per text. However, we have to take into account that the categories are, to some extend, ambiguous, as reflected in the inter-annotator agreement. Refinement of the guidelines might result in some improvement, but there is an upper limit as to what can be meaningfully disambiguated.

In the future, we intend to broaden our database by annotating more academic data. The inclusion of reviewed and published academic texts could be helpful in identifying the effects of limited language skills in our learner corpus. The narrator *I* indicates that a comparison to literary data could be beneficial: Do the narrator *I* type in academic texts and the narrator *I* type in literary texts have commonalities? In terms of features, we would like to refine our approach by using linguistic annotations and including, for instance, verbal morphology, which is known to be a good indicator for narration.

6 Acknowledgements

Melanie Andresen's work on this paper was funded by the *Landesforschungsförderung Hamburg* in the context of the project *hermA* [Ga17] (LFF-FV 35) at Universität Hamburg.

References

[Äd06] Ädel, A.: Metadiscourse in L1 and L2 English. Benjamins, Amsterdam, 2006.

[AK17] Andresen, M.; Knorr, D.: KoLaS – Ein Lernendenkorpus in der Schreibberatungsausbildung einsetzen. Zeitschrift Schreiben/, pp. 10–17, 2017, URL: https://zeitschrift-schreiben.ch/2017/#andresen.

[Ga17] Gaidys, U.; Gius, E.; Jarchow, M.; Koch, G.; Menzel, W.; Orth, D.; Zinsmeister, H.: hermA: Automated Modelling of Hermeneutic Processes. Hamburger Journal für Kulturanthropologie/7, pp. 119–123, 2017.

[Hu07] Hunter, J. D.: Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9/3, pp. 90–95, 2007, DOI: 10.1109/MCSE.2007.55.

[Hy05] Hyland, K.: Metadiscourse: Exploring Interaction in Writing. Continuum, London, 2005.

[Kr12] Kruse, O.: Wissenschaftliches Schreiben mehrsprachig unterrichten: Was ist möglich, was ist nötig? ÖDaF-Mitteilungen/2, pp. 9–25, 2012.

[Kr80] Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage, Beverly Hills, California, 1980.

[LK77] Landis, J. R.; Koch, G. G.: The Measurement of Observer Agreement for Categorical Data. Biometrics 33/1, pp. 159–174, 1977.

[Pe11] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.: Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research 12/, pp. 2825–2830, 2011.

[St07] Steinhoff, T.: Zum ich-Gebrauch in Wissenschaftstexten. Zeitschrift für germanistische Linguistik 35/1-2, pp. 1–26, 2007.

[Th20] The pandas development team: Pandas 1.0.3, Mar. 18, 2020, DOI: 10.5281/zenodo.3715232.

[TJ99] Tang, R.; John, S.: The 'I' in Identity: Exploring Writer Identity in Student Academic Writing through the First Person Pronoun. English for Specific Purposes 18, Supplement 1/, S23–S39, 1999, DOI: 10.1016/S0889-4906(99)00009-5.