Recognising Occupational Titles in German Parliamentary Debates

Johanna Binnewitt
German Federal Institute for Vocational Education and Training, Bonn
johanna.binnewitt@bibb.de

Abstract

The application of text mining methods is becoming more and more popular, not only in Digital Humanities (DH) and Computational Social Sciences (CSS) in general, but also in vocational education and training (VET) research. Employing algorithms offers the possibility to explore corpora that are simply too large for manual methods. However, challenges arise when dealing with abstract concepts like occupations or skills, which are crucial subjects of VET research. Since algorithms require concrete instructions, either in the form of rules or annotated examples, these abstract concepts must be broken down as part of the operationalisation process.

In our paper, we tackle the task of identifying occupational titles in the plenary protocols of the German Bundestag. The primary focus lies in the comparative analysis of two distinct approaches: a dictionary-based method and a BERT fine-tuning approach. Both approaches are compared in a quantitative evaluation and applied to a larger corpus sample. Results indicate comparable precision for both approaches (0.93), but the BERT-based models outperform the dictionary-based approach in terms of recall (0.86 vs. 0.77). Errors in the dictionary-based method primarily stem from the ambiguity of occupational titles (e.g., *baker* as both a surname and a profession) and missing terms in the dictionary. In contrast, the BERT model faces challenges in distinguishing occupational titles from other personal names, such as *mother* or *Christians*.

1 Text Mining in VET research

Thanks to the spread of the internet in general and social media in particular, more and more communication is now taking place in written form. And communication that takes place outside the internet also finds its way into web archives and repositories, for example in the form of minutes or transcripts. Text documents can therefore be a valuable source for (Social Science) research, from which conclusions about social processes and contexts can be derived. The challenge often lies in building bridges between what is said or written and the extra-textual phenomena, i.e. the actual objects of research (see Krippendorff (2019)). In qualitative content analysis, this is done by including a small number of texts in the manual analysis, in which the theoretical concepts are then related to specific text passages, for example using a hermeneutic approach. However, as soon as larger volumes of text are to be analysed, manual methods reach their limits.

Like other disciplines from the social science or humanities, research on the labour market and vocational training deals with fuzzy concepts, such as occupations, skills or job tasks, which are often characterised by the fact that they are defined in an abstract and complex way (Rodrigues et al., 2021, p. 4f.) (Alexopoulos, 2020). As interdisciplinary research field, VET research connects many disciplines, like Social Sciences, Education Sciences, Economics or Psychology, and deals with research questions on apprenticeships, learning designs or labour market needs. Within the intersection of VET research and text mining, many studies analyse skills, competencies or qualifications based on job advertisements as data source (Buchmann et al., 2022; Stops et al., 2020). Recent studies also include training regulations and curricula in the quantitative text analysis (Fischer et al., 2021). In many cases, skills extraction is operationalised by using dictionary-based approaches. Initial approaches are also testing the use of machine learning to recognise skills in job advertisements (Zhang et al., 2022). Khaouja et al. (2021) provide an extensive overview on methods for skills extraction.

Alongside skills, a central concept in VET research is the one of occupation. At first glance, this concept seems easy to grasp, but it harbours various challenges when it comes to (automated)

Proceedings of LaTeCH-CLfL 2024, pages 221–230

identification of references to occupations in texts. We will use the example of identifying occupational titles automatically in the plenary protocols of the German Bundestag to demonstrate these challenges. The plenary protocols are well suited to this demonstration as they contain around 900,000 speeches, which would be difficult to process manually. The corpus is briefly described in the following section. Section 3 then introduces the concept of occupational titles. Section 4 is devoted to the concrete implementation by describing two different approaches to operationalisation and comparing their results on a small data basis. Section 5 finally describes the application of both operationalisations on a larger data basis in order to compare the varying results. The code is available on GitHub: https://github.com/johannabi/ProfRec_Bundestag

2 Plenary Debates as research object

Parliamentary debates have been in focus for recent research in multiple disciplines, like Political Sciences, Corpus Linguistics or Computational Linguistics. Research interest ranged from network analysis (Padó et al., 2019) over rhetorical analysis (Rehbein et al., 2021) to argument mining (Eide, 2019) or sentiment analysis (Abercrombie and Batista-Navarro, 2020). To our best knowledge, there has not yet been any research in the context of parliamentary debates that has dealt with the identification of personal nouns, i.e. common nouns referring to (groups of) persons, in general or occupational titles in particular. For VET research, the identification of occupational titles in plenary debates could provide valuable insights into how politicians – as representatives of society – talk about certain professions. For example, it could be analysed to what extent the interests of certain professional groups are addressed and asserted in the speeches. Or, for research into inequality and stereotypes, occupational titles could be analysed to determine the extent to which the mention of certain professions within debates on inequality is characterised by stereotypes.

The presented experiments are based on the *Open Discourse* corpus (Richter et al., 2020), but the experiments could also be adopted to other corpora of German parliamentary debates, such as GermaParl (Blaette and Leonhardt, 2023). The Open Discourse corpus was developed by using the transcriptions of plenary sessions that are published by the German parliament in an XML format. In addition to the German-language speech content, the corpus contains metadata such as the date, the legislative period as well as various information on the speaker such as their name or fraction. Documents are separated by speakers and interjections are stored in a different structure. The *Open Discourse* corpus contains 907,644 different speeches from the period 1949 to 2021. Sections 4 and 5 use samples from the *Open Discourse* corpus to show how occupational titles can be automatically identified in the plenary transcripts and what challenges arise in the process. So, in order to better describe these challenges, the following section will first introduce the concept of occupational title.

3 Occupational Titles as research object

From a linguistic point of view, occupational titles are common nouns that represent a subgroup of personal nouns. These in turn are linguistic expressions that refer to individual persons or groups of persons. In this context, both personal and occupational nouns stand in contrast to proper names. While proper names refer to concrete persons, personal and occupational nouns refer to more abstract concepts. In the case of occupational titles, the terms are characterised by the fact that the person or group of persons performs a profession. While other research have already focused on identifying personal nouns in German texts automatically (Sökefeld et al., 2023), the focus on occupational titles as a separate group was only set for the task of classifying Spanish tweet to whether they contain occupational titles. (Miranda-Escalada et al., 2021). Occupational titles can refer to a specific person or group (examples 1 & 2), they can be more generic (example 4 & 5), or they can be attributes of (specific or generic) individuals (example 3).

(1) Mein *Friseur* hat ein Wunder vollbracht.
My *hairdresser* worked a miracle.

(2) Die *Lehrer* an unserer Schule bereiten das Schulfest vor.
The *teachers* of our school are preparing for the school festival.

(3) Der neue Nachbar ist *Tischler*.
The new neighbour is a *carpenter*.

(4) Als *Krankenschwester* musst du heute sehr belastbar sein.
As a *nurse* today, you have to be very resilient.

(5) Ich spreche hier im Namen aller *Busfahrer*.
I am speaking here on behalf of all *bus drivers*.

Further characteristics of occupational titles, which are mainly related to the German language, concern the formation of compounds and gender forms. Since compounding is very productive in German, occupational titles can either be part of other lexemes that do not refer to occupations (as in *Bauernverband* (farmers' association)) or the occupational title can be a specification of another occupational title (as *Versicherungsmaklerin* (insurance broker) being a specification of *Maklerin* (broker)). Productivity of occupational titles in compounds is a first indicator that dictionary-based approaches might be insufficient to identify all occupational title within a text. With regard to the gender forms of occupational titles, dictionary-based approaches can be adapted to the extent that the word list is enriched with masculine, feminine and neutral forms (see section 4.2). As the use of gendered forms and gender-inclusive language can itself be of research interest (see Carla Sökefeld (2021); Damelang and Rückel (2021); Hodel et al. (2017); Horvath et al. (2015)), the correct extraction of all mentioned categories is crucial to avoid bias in subsequent analysis.

Occupational titles can semantically cover different aspects of the profession. In some cases, the activity is emphasised, as in *Dachdecker/in* (roofer), *Lehrer/in* (teacher) or *Verkäufer/in* (salesperson). Other occupational titles originate from the subject that the person deals with during work (*Stahlarbeiter/in* (steelworker), *Immobilienmakler/in* (real estate agent)) or the place where the profession is usually practised (*Grundschullehrer/in* (primary school teacher), *Bankkaufmann/frau* (bank clerk)) (Stooß and Saterdag, 1979; Schierholz et al., 2018). Like many other expressions, occupational titles can also be affected by ambiguities. On the one hand, many (German) surnames originate from traditional professions, such as *Bäcker* (baker), *Müller* (miller) or *Fischer* (fisherman). In example 6, the two meanings of the lexeme *Müller* can be clearly separated from each other, as it belongs to the group of proper names mentioned above and not to the profession of miller. Another type of ambiguity arises particularly with occupational titles that are derived from the verb that describes the activity, such as *Pfleger/in* (caregiver) or *Verkäufer/in* (salesperson) (see example 7). Here, the resolution of the ambiguity between professional and non-professional activity on the basis of a sentence is not always clear. In example 7, however, the lexeme *Verkäufer* refers to the legal meaning of a seller, so that all persons who sell something are meant here, regardless of whether they do so professionally or non-professionally[1]. Finally, another type of ambiguity has emerged from a rather metonymic use of occupational titles. In example 8, the occupational title is used to refer to the business type rather than the person because *beim Bäcker* could also be replaced by *bei der Bäckerei* without changing the meaning of the sentence.

(6) Frau *Müller* kennt sich damit aus.
Ms. Müller knows all about it.

(7) Damit wird ein Vertrag zwischen Käufer und *Verkäufer* geschlossen.
This concludes a contract between buyer and seller.

(8) Beim *Bäcker* um die Ecke gibt es die besten Brötchen.
The bakery around the corner has the best bread rolls.

In addition to these ambiguities, where the textual context – i.e. the sentence – determines whether a particular lexeme is an occupational title, there are other personal nouns where the decision whether it is an occupational title depends heavily on the definition of occupation. As with many concepts in the humanities and social sciences, this definition varies greatly depending on the discipline and the research question being asked. Sailmann (2018) provides a detailed overview of the conceptual history of the profession. Sombart (1959), for example, divides the term into an objective meaning, which focuses on the social function, and a subjective meaning, which emanates from the individual. He defines the social function of occupations very broadly, so that, for example, husband also becomes a profession. In comparison, two other criteria are decisive for Weber: the gainful character of work and a minimum level of qualification required (Weber and Winckelmann, 1985). According to this definition, voluntary activities, such as working as a lay judge, would be excluded from the definition of an occupation. As Weber's definition often forms the basis for occupational statistics today, the operationalisation here will also be based on it.

[1] see also https://www.dwds.de/wb/Verk%C3%A4ufer

4 Operationalising for automatic language processing

After introducing the concept of occupations, this next section will now compare two ways to identify occupational titles automatically, namely a dictionary-based approach and a machine-learning approach. As we have stated above, occupational titles can be affected by ambiguities and lexical productivity, which might make it difficult to apply a simple keyword search (see Widmann and Wich (2023)). Nevertheless, dictionary-based analysis are often applied in DH and CSS, because they are feasible if a dictionary already exists (see Calanca et al. (2019); Stops et al. (2020); Djumalieva et al. (2018)). So, we aim to evaluate such an approach in order to investigate error sources of word lists in a more detailed way. For both approaches, the identification of occupational titles takes place at token level, i.e. for each word in a sentence, the algorithm decides whether it is an occupational title. It is also possible for an occupational title to be represented by a multiword expression, such as *medical assistant*.

4.1 Data Annotation

Since we aim to evaluate both approaches, we need test data as gold standard. Additionally, for the BERT fine-tuning, we also need training data, so this next section describes the annotation process. Since we assume that most occupational titles can be identified at sentence-level, we apply a sentence tokenizer and build our annotation corpus on sentences rather than on whole debates. Since professions are seldom the main topic of debates in the Bundestag and occupational titles are therefore rather rare, a seed list of keywords [2] was used to search for sentences that could be considered for the gold standard. The keywords were selected to include words that are not an occupational title in every context, such as *Bauer* (differentiation from surname) or *Sportler* (differentiation from non-professional). For each keyword, random sentences were selected from the entire corpus, so that the annotation corpus finally comprises 817 sentences, whereby a sentence can also contain several keywords from the list.

The selected sentences were then annotated using the INCEpTION software (Klie et al., 2018). In order to enhance the annotation process, the keywords found were initially marked as potential occupational titles. These annotations could then be corrected during the annotation process and supplemented with additional occupational titles that were not included in the initial word list. Compounds in which an occupational title is not the root of the compound were not annotated (e.g. *Bauernverband* (farmers' association)). The annotation process was carried out by a single person, so no inter-annotator agreement could be formed. The distribution of tokens in train and test set as well as the share of annotated tokens can be seen in table 1. In addition to the initial list of keywords mentioned above, the training and test sentences also contain less common occupational titles, such as *Textilingenieurin* (textile engineer) (textile engineer) or *Agrarsoziologe* (agricultural sociologist). The test sentences contain 54 occupational titles that do not appear in the train sentences.

	All Tokens	Tokens PROF	Types PROF
Train	20,176	1,094 (5.4%)	201
Test	4,941	269 (5.4%)	91

Table 1: Distribution of Labels in Train and Test

4.2 Dictionary-based approach

The basis for the dictionary-based approach was a search term list from the Federal Employment Agency (BA), which was initially developed for processing search queries on the BA portals, such as BERUFENET[3]. It contains a total of 179,002 different German-language terms, which are grouped into male, female and neutral search terms. Neutral terms include both valid occupational titles, such as *Bürokaufleute* (office clerks), but also expressions such as *Pflanze* (plant), which can support the search for occupational information but are not occupational titles. In the case of neutral occupational titles, these were in some cases also assigned to the male and female group, such as *Archivfachkraft* (archivist).

We used SpaCy's PhraseMatcher module to search for all occupational titles from the list (Honnibal et al., 2020). The module enables the rule-based search of words and multi-word expressions in a text. Since the dictionary is divided into male,

[2]The following keywords were used: Bauer, Bäuerin, Landwirt, Landwirtin, Arzt, Ärztin, Mediziner, Medizinerin, Sportler, Sportlerin, Verkäufer, Verkäuferin, Krankenpfleger, Krankenpflegerin, Krankenschwester, Pfleger, Pflegerin.

[3]see https://download-portal.arbeitsagentur.de/

ID	name	level	groups	#keywords
1	mfn_{lemma}	lemma	male female neutral	179,002
2	mf_{token}	token	male female	107,020
3	mf_{lemma}	lemma	male female	107,020

Table 2: Configuration for rule-based operationalisation

female and neutral search terms, we ran different experiments with different subsets of the keyword list. The different configurations are summarised in table 2. Firstly, the length of the keyword list was varied by excluding the neutral terms from the keyword list in Experiment 2 & 3. In addition, we varied whether the phrase matching was applied on token (experiment 2) or on lemma level (experiment 1 & 3).

4.3 BERT-based approach

For the machine learning approach, we decided to apply a fine-tuning on a pre-trained BERT model (Devlin et al., 2019). To our best knowledge, there does not exist any BERT model that is trained on German parliamentary debates, so we chose bert-base-german-cased [4] as base model. All hyperparameters of the finetuning itself were left on default values (see table 3). But since compounds, such as *Lehrerverband* (teachers' association), might pose a challenge for the identification of occupational titles, particular attention was paid to aggregation strategies within the configuration. Since BERT-based models divide tokens in further subtokens, the labels of subtokens can be aggregated in different ways. We varied these aggregation strategies to evaluate their effect on token classification result (experiment 4 to 7).

4.4 Evaluation

In order to assess the quality of the operationalisations described above, all algorithms are applied to the same 163 sentences from the test data. Table 4 summarises the evaluation results for all seven experiments. The metrics are computed at token-level, i.e. multi-word expressions are split into tokens and then counted separately. All configurations of BERT fine-tuning achieve the best recall of 0.86, showing that the aggregation strategy has no effect on identifying more true positives within the test set. The fewest occupational titles were found by the PhraseMatcher at the tokenised level. The BERT model also performs best in terms of precision. Here, the aggregation strategy **max** achieved the best results, closely followed by the PhraseMatcher at the lemmatised level with all male and female occupational titles. The difference to all other aggregation strategies is also marginal, at one percentage point. The PhraseMatcher, which also includes neutral terms, achieved the most false positives, leading to a precision of 0.26. This result is not surprising, as the keyword list also contained terms such as *Steuerwesen* (taxation), which describe the topic of a professional activity, but are no occupational titles.

In the qualitative error analysis, the dictionary-based approaches also show that ambiguous occupational titles (sportsperson, salesperson, trader) lead to false positives. This problem also affects surnames, as the context is not taken into account. Two causes can be identified with regard to the

[4] https://huggingface.co/bert-base-german-cased

ID	name	Aggregation strategy
4	$bert_{simple}$	whole token is annotated as PROF if at least one subtoken is annotated as PROF
5	$bert_{first}$	label of first subtoken is taken as label for the whole token
6	$bert_{average}$	highest mean probability of all labels for whole token
7	$bert_{max}$	highest probability of all labels

Table 3: Configuration for BERT fine-tuning (further hyperparameters: learning rate: 5^{-5}; epochs: 5; optimizer: $adamw_{torch}$)

ID	name	pre	rec	f1
1	mfn_{lemma}	0.261	0.784	0.391
2	mf_{token}	0.925	0.463	0.617
3	mf_{lemma}	0.935	0.769	0.844
4	$bert_{simple}$	0.926	**0.863**	0.893
5	$bert_{first}$	0.932	**0.863**	0.896
6	$bert_{average}$	0.932	**0.863**	0.896
7	$bert_{max}$	**0.936**	**0.863**	**0.898**

Table 4: Evaluation results for all experiments (the best scores are bold)

low recall of the PhraseMatcher: on the one hand, incorrect lemmatisation sometimes impedes the identification of occupational titles. Initial qualitative analyses indicate that female plural forms, such as *Erzieherinnen* (female educators), are more frequently affected by this problem than other lexemes, but quantitative analyses on this are still pending. This would not fulfil the requirement that the method for identifying occupational titles would not include any systematic bias regarding gendered forms. Secondly, valid occupational titles, such as *Agrarsoziologe* (agricultural sociologist), are not included in the keyword list and therefore cannot be found. This clearly shows that even a comprehensive word list with over 100,000 terms can be inadequate, as language is productive and new occupational titles are constantly being added.

The model-based experiments also reveal various causes for false extractions. False positives are often other personal nouns, such as *Bürgerin* (citizen) in example 9[5], or cases in which the distinction between profession and company is blurred, as in example 10. In addition, the aggregation strategy influences the extractions, especially in the case of compound nouns such as in example 11. False negatives mainly affect expressions such as *Beamte* (civil servants), *Angestellte* (employees) or *Beschäftigte* (staff) (see example 12). In addition, multi-word expressions such as in example 13 are often annotated in abbreviated form. Here, *Verkäufer* was extracted correctly, but without the specialisation. Depending on the subsequent downstream task, a distinction between different specialisations of salespersons would be crucial.

(9) Als Medizinerin, als Politikerin, aber auch als **Bürgerin** sage ich: [...]
As a doctor, as a politician, but also as a **citizen**, I say: [...]

(10) Wenn einer Diplom-Landwirtin [. . .] dort empfohlen wird, Bäuerinnen in den alten Bundesländern einmal zum **Friseur** oder zum Einkaufen zu fahren, dann muß man zumindest beachten, daß ihr vorgeschlagen wird, Bäuerinnen zu fahren.
If a qualified farmer [...] is recommended there to drive women farmers in the old federal states to the **hairdresser** or to the shops, then one must at least note that she is suggested to drive women farmers.

(11) Wir haben die **Junglandwirteförderung** und in der Familienpolitik das Erziehungsgeld für die Bäuerin durchgesetzt.
We have pushed through **young farmers support** and, in family policy, the child-raising allowance for female farmers.

(12) Es ist mit den staatlichen und den Hoheitsaufgaben des *Beamten* einfach nicht vereinbar, daß er sich nebenbei noch als Verkäufer betätigt.
It is simply not compatible with the state and sovereign tasks of the *civil servant* that he also works as a salesman.

(13) Um die notwendige Aufklärung und Beratung sicherzustellen, sind Regelungen über die fachlichen Kenntnisse der *Verkäufer im Einzelhandel* zu treffen.
In order to ensure that the necessary information and advice is provided, regulations on the specialist knowledge of *sales staff in the retail sector* must be put in place.

5 Application on further parliamentary protocols

As the test corpus is comparatively small at 163 sentences, we applied the best dictionary-based and the best BERT-based model[6] to further 2,000 speeches from 2010 in order to compare the extractions where both models disagree. Since we do not have manually annotated sentences here, we can not report evaluation scores in this section. The most common types that were annotated by each model are reported in figures 1 & 2. In total, the PhraseMatcher identified 3,285 tokens (397 types) and the BERT model 2,509 tokens (866 types) as occupational titles. Of these, both models agree with their decision for 1,185 tokens (272 types). The diverging type-token-ratio indicates that the PhraseMatcher mostly annotated more frequent types, such as *Präsident/Präsidentin* (president), which are usually part of the salutation at the beginning of the speech. Meanwhile, BERT annotates more different types, which indicates that the context is taken into account and the lexical variance of the occupational titles themselves is not an obstacle for the model.

If we look at the cases where only the dictionary-based model has marked an occupational title, it is noticeable – as in the previous section described – that these are often words that are not occupational titles in the respective context, such as surnames, or where an incorrect lemmatisation has led to the

[5] FP are shown in bold, FN are shown in italics.

[6] https://huggingface.co/johannabi/german_tc_professions_debates

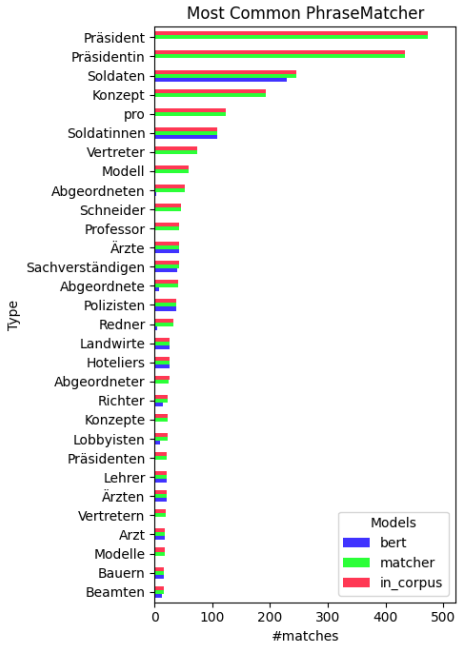

Figure 1: Most Common Types found by best PhraseMatcher in 2000 debates. *in corpus* refers to how often the type appeared.

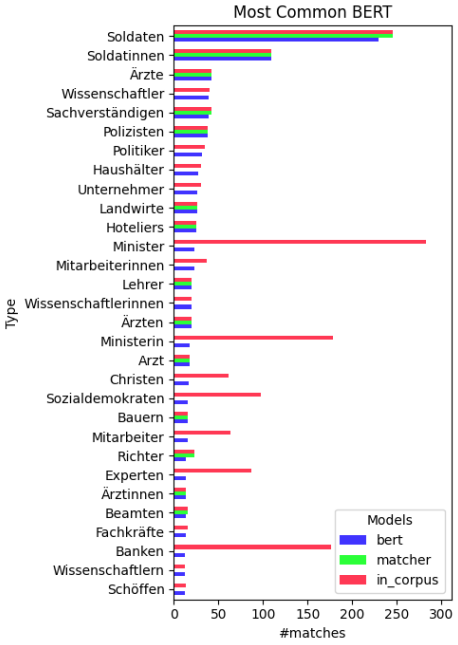

Figure 2: Most Common Types found by best BERT model in 2000 debates. *in corpus* refers to how often the type appeared.

annotation. For example, *Koch* (cook) is marked seven times by the PhraseMatcher, although the occupation is only meant once and the surname in all other cases. In contrast, the BERT model was able to correctly distinguish between the two meanings. There are also some valid occupational titles, such as *Milchbauer* (dairy farmer), *Banker* (banker) or *Pfleger* (carer), that were only found by the BERT model because they were missing from the keyword list.

In contrast, the BERT-based model often annotates other personal nouns that do not refer to professions, such as *Christen* (Christians), *Spekulant* (speculator), *Rentner* (pensioner) or *Schöffen* (lay judge) (see figure 1). In addition, in some cases, company types, such as *Banken* (banks), are annotated as occupations. To prevent these false extractions, the training data for a future model should include more personal nouns or company types as negative examples, so that the distinction between occupational groups and other personal nouns or types of organizations becomes clearer. Finally, the BERT model annotates many generic terms such as *Spezialist* (specialists) or *Akademiker* (academics) as occupational titles. These terms are (for the most part) not included in the keyword list because they do not reflect the specialised nature of an occupation like the dictionary of the Federal Employment Agency does. Nevertheless, these terms could be a starting point for subsequent analysis, since they group together various professions. In this case, the decision whether these terms are occupational terms again depends heavily on the occupational definition and the subsequent content analysis.

6 Conclusion

The evaluation using the test data (section 4.4) as well as the application of the best models on further debates (section 5) have shown that both approaches have various strengths and weaknesses. The weakness of dictionary-based approaches lies in the context-free consideration of keywords and the treatment of unknown occupational titles. With regard to the false negatives, the word list could first be enriched by searching for similar words in a type embedding, such as Word2Vec. Regarding false positives, some errors could be minimised by setting up additional rules. For example, surnames could be excluded by not annotating a potential token if the word *Kollege* (colleague) precedes it. However, it is clear that these rules reach their limits as soon as a distinction has to be made between profession and businesses, like for *Friseur* (hairdresser). The BERT model presented here also has problems with these distinctions, regardless of which aggregation strategy was chosen. In addi-

tion, the BERT model still seems to overgeneralise the concept of occupational title, as many of the annotated terms refer to other groups of people. As stated in the previous section, this overgeneralisation could be prevented by adding more examples of personal nouns to the training data in order to improve the distinction between occupational groups and other groups. In addition, the BERT base model could be compared to more specified models, such as jobBERT-de[7], which is domain-adapted on German-language job advertisements from Switzerland (Gnehm et al., 2022). This model may be able to depict occupational concepts more clearly than the BERT base model. Finally, other hyperparameters should be varied in training to determine the best possible configuration for fine-tuning.

One point that usually follows the identification of expressions in texts is the grouping of the expressions on the basis of existing classifications. This is because it is often not the term as such that is to be analysed, but the referenced concept or the terms in relation to the referenced concept. The classification of occupations (KldB) is often used for statistical analyses on occupational activities. However, Schierholz et al. (2018) have already shown that occupational titles are only suitable for coding according to the KldB to a limited extent, because occupational activities, as defined by the KldB, and occupational titles do not fully correspond to each other. For example, when a *Handwerker* (craftsman) is mentioned in the plenary protocols, it is often not recognisable which occupational activity is meant since labour market research often distinguishes between for example carpenters and mechanics. Alternatively, the extracted data could be used for analyses that are based on structuring features other than occupational specialisation. Initial ideas for groupings would be, for example, gender forms (*Lehrer vs. Lehrerin*) or industry affiliation (*employee in the automotive industry*).

Finally, lexical change of occupational titles could be examined at the level of occupational titles, for example by comparing whether a renaming of certain occupations or corresponding apprenticeships also led to a change in language usage or whether a preceding change in language usage has finally led to a change in official training regulations. In addition, occupational titles could be grouped according to profession and gender in order to investigate whether changes in the gender distribution in the profession have also led to a change in language (Oksaar, 1976).

[7] https://huggingface.co/agne/jobBERT-de

Limitations

All experiments presented in the paper were evaluated on a rather small test set (163 sentences). Furthermore, training and test data were annotated by a single person, which might have led to rather subjective annotations. Finally, most assumptions on compounds or gendered occupational titles refer to German-language data. This might not apply to other languages.

References

Gavin Abercrombie and Riza Batista-Navarro. 2020. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Panos Alexopoulos. 2020. *Semantic modeling for data: Avoiding pitfalls and breaking dilemmas*, first edition. O'Reilly, Beijing and Boston and Farnham and Sebastopol and Tokyo.

Andreas Blaette and Christoph Leonhardt. 2023. GermaParl Corpus of Plenary Protocols.

Marlis Buchmann, Helen Buchs, Felix Busch, Simon Clematide, Ann-Sophie Gnehm, and Jan Müller. 2022. Swiss Job Market Monitor: A Rich Source of Demand-Side Micro Data of the Labour Market. *European Sociological Review*.

Federica Calanca, Luiza Sayfullina, Lara Minkus, Claudia Wagner, and Eric Malmi. 2019. Responsible team players wanted: an analysis of soft skill requirements in job advertisements. *EPJ Data Science*, 8(1).

Carla Sökefeld. 2021. Gender(un)gerechte Personenbezeichnungen: derzeitiger Sprachgebrauch, Einflussfaktoren auf die Sprachwahl und diachrone Entwicklung. *Sprachwissenschaft*, 46(1).

Andreas Damelang and Ann-Katrin Rückel. 2021. Was hält Frauen von beruflichen Positionen fern? Ein faktorieller Survey zum Einfluss der Gestaltung einer Stellenausschreibung auf deren Attraktivitätseinschätzung. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 73(1):109–127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jyldyz Djumalieva, Antonio Lima, and Cath Sleeman. 2018. Classifying Occupations According to Their Skill Requirements in Job Advertisements: Economic Statistics Centre of Excellence (ESCoE) Discussion Papers.

Stian Rødven Eide. 2019. The Swedish PoliGraph: A semantic graph for argument mining of Swedish parliamentary data. In *Proceedings of the 6th Workshop on Argument Mining*, pages 52–57, Florence, Italy. Association for Computational Linguistics.

Andreas Fischer, Patrick Hilse, and Sören Schütt-Sayed. 2021. Curricula, Ausbildungsordnungen und Lehrpläne – Spiegel der Bedeutung nachhaltiger Entwicklung.

Ann-Sophie Gnehm, Eval Bühlmann, and Simon Clematide. 2022. Evaluation of Transfer Learning and Domain Adaptation for Analyzing German-Speaking Job Advertisements. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3892–3901.

Lea Hodel, Magdalena Formanowicz, Sabine Sczesny, Jana Valdrová, and Lisa von Stockhausen. 2017. Gender-Fair Language in Job Advertisements. *Journal of Cross-Cultural Psychology*, 48(3):384–401.

Matthew Honnibal, Ines Montani, Sofie van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Lisa K. Horvath, Elisa F. Merkel, Anne Maass, and Sabine Sczesny. 2015. Does Gender-Fair Language Pay Off? The Social Perception of Professions from a Cross-Linguistic Perspective. *Frontiers in Psychology*, 6:1–12.

Imane Khaouja, Ismail Kassou, and Mounir Ghogho. 2021. A Survey on Skill Identification From Online Job Ads. *IEEE Access*, 9:118134–118153.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Klaus Krippendorff. 2019. *Content analysis: An introduction to its methodology*, fourth edition. SAGE, Los Angeles and London and New Delhi and Singapore and Washington DC and Melbourne.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Luis Gascó, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. 2021. The ProfNER shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 13–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Els Oksaar. 1976. *Berufsbezeichnungen im heutigen Deutsch: Soziosemantische Untersuchungen Mit deutschen und schwedischen experimentellen Kontrastierungen*, 1 edition. Schwann, Düsseldorf.

Sebastian Padó, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who Sides with Whom? Towards Computational Construction of Discourse Networks for Political Debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847, Florence, Italy. Association for Computational Linguistics.

Ines Rehbein, Josef Ruppenhofer, and Julian Bernauer. 2021. Who is we? Disambiguating the referents of first person plural pronouns in parliamentary debates. In *Proceedings of KONVENS 2021*, pages 147–158.

Florian Richter, Philipp Koch, Oliver Franke, Jakob Kraus, Fabrizio Kuruc, Anja Thiem, Judith Högerl, Stella Heine, and Konstantin Schöps. 2020. Open Discourse.

Margarida Rodrigues, Enrique Fernández-Macías, and Matteo Sostero. 2021. A unified conceptual framework of tasks, skills and competences.

Gerald Sailmann. 2018. *Der Beruf: Eine Begriffsgeschichte*, volume 147 of *Histoire*. transcript Verlag, Bielefeld.

Malte Schierholz, Miriam Gensicke, Nikolai Tschersich, and Frauke Kreuter. 2018. Occupation Coding During the Interview. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(2):379–407.

Carla Sökefeld, Melanie Andresen, Johanna Binnewitt, and Heike Zinsmeister. 2023. Personal noun detection for german. In *Proceedings of the 19th Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-19)*, pages 33–39.

Werner Sombart. 1959. Beruf. In Alfred Vierkandt, editor, *Handwörterbuch der Soziologie*, pages 25–31. Ferdinand Enke Verlag, Stuttgart.

Fridemann Stooß and Hermann Saterdag. 1979. Systematik der Berufe und der beruflichen Tätigkeiten. In Franz Urban Pappi, editor, *Sozialstrukturanalysen mit Umfragedaten*, Monographien sozialwissenschaftliche Methoden, pages 41–57. Athenäum-Verl., Königstein/Ts.

Michael Stops, Ann-Christin Bächmann, Ralf Glassner, Markus Janser, Britta Matthes, Lina-Jeanette Metzger, Christoph Müller, and Joachim Seitz. 2020. Machbarkeitsstudie Kompetenz-Kompass: Teilprojekt 2: Beobachtungen von Kompetenzanforderungen in Stellenangeboten.

Max Weber and Johannes Winckelmann. 1985. *Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie*, 5. rev. aufl. edition. Mohr, Tübingen.

Tobias Widmann and Maximilian Wich. 2023. Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*, 31(4):626–641.

Mike Zhang, Kristian Nørgaard Jensen, Sif Dam Sonniks, and Barbara Plank. 2022. SkillSpan: Hard and Soft Skill Extraction from English Job Postings.