Investigating the Morphological Complexity of German Named Entities: The Case of the GermEval NER Challenge

Bettina Klimek Markus Ackermann Amit Kirschenbaum Sebastian Hellmann
AKSW/KILT Research Group
InfAI, University of Leipzig
{klimek,ackermann,amit,hellmann}@informatik.uni-leipzig.de

Abstract

This paper presents a detailed analysis of Named Entity Recognition (NER) in German, based on the performance of systems that participated in the GermEval 2014 shared task. It focuses on the role of morphology in named entities, an issue too often neglected in the NER task. We introduce a measure to characterize the morphological complexity of German named entities and apply it to the subset of named entities identified by all systems, and to the subset of named entities none of the systems recognized. We discover that morphologically complex named entities are more prevalent in the latter set than in the former, a finding which should be taken into account in future development of methods of that sort. In addition, we provide an analysis of issues found in the GermEval gold standard annotation, which affected also the performance measurements of the different systems.

1 Introduction

Despite initiatives to improve Named Entity Recognition (NER) for German such as in challenges as part of CoNLL 2003[1] and GermEval 2014[2], a noticeable gap still remains between the performance of NER systems for German and English. Pinpointing the cause of this gap seems to be an impossible task as the reasons are manifold and in addition difficult to realize due to their potentially granular (and subtle) nature as well as their inter-relatedness. However, we can name several aspects that might have an influence: (1) lack of linguistic resources suitable for German, (2) less demand (and interest) for improving the quality of NER systems for German, (3) variance of annotation guidelines and annotator consensus, (4) different NER problem definitions, (5) inherent differences between both language systems, (6) quality of provided data and source material, (7) etc. Studying the degree of impact for each of these factors as a whole revokes any attempt to apply scientific methods for error analysis. However, a systematic investigation of linguistic aspects of proper nouns, i.e. named entities in technical terms[3], in German can reveal valuable insights on the difficulties and the improvement potential of German NER tools. Such an aspect is the morphological complexity of proper nouns. Due to its greater morphological productivity and variation, the German language is more difficult to analyze, offering additional challenges and opportunities for further research. The following list highlights a few examples:

- More frequent and extensive compounding requires correct token decompounding to identify the named entity (e.g. ***Bibel****forscherfrage* - 'bible researchers' question').
- Morphophonologically conditioned inner modifications are orthographically reflected and render mere substring matching ineffective (e.g. *außer****europä****isch* (*Europa*) - 'non-European').
- Increased difficulty in identifying named entities which occur within different word-classes after derivation (e.g. ***luther****ischen*, an adjective, derived from the proper noun *Martin Luther*).

[1] CoNLL 2003 Challenge Language-Independent Named Entity Recognition, http://www.cnts.ua.ac.be/conll2003/ner/

[2] GermEval 2014 Named Entity Recognition Shared Task, https://sites.google.com/site/germeval2014ner/, see also (Benikova et al., 2014a)

[3] From a linguistic perspective *named entities* are encoded as *proper nouns*. In this paper both terms are treated synonymously.

Sentence	NE type
1951 bis 1953 wurde der nördliche Teil als Jugendburg des **Kolpingwerkes** gebaut.	OTH
Beschreibung Die **Kanadalilie** erreicht eine Wuchshöhe von 60 bis 180 cm und wird bis zu 25 cm breit.	LOCpart
Um 1800 wurde im ehemaligen **Hartung'schen** Amtshaus eine Färberei eingerichtet.	PERderiv
1911 wurde er Mitglied der **sozialistischen Partei**, aus der er aber ein Jahr später wieder austrat.	ORG

Table 1: Example of reference data from the GermEval provided annotated corpus.

These observations support the hypothesis that morphological alternations of proper nouns constitute another difficulty layer which needs to be addressed by German NER systems in order to reach better results. Therefore, this paper presents the results of a theoretic and manual annotation and evaluation of a subset of the GermEval 2014 Corpus challenge task dataset. This investigation focuses on the complexity degree of the morphological construction of named entities and shall serve as reference point that can help to estimate whether morphological complexity of named entities is an aspect which impacts NER and if it should be considered when creating or improving German NER tools. During the linguistic annotation of the named entity data, issues in the GermEval gold standard (in the following "reference annotation") became apparent and, hence, were also documented in parallel to the morphological annotation. Even though an analysis of the reference annotations was originally not intended, it is presented as well because it effects the measures of tool performance.

The rest of the paper is structured as follows. Section 2 presents an overview of related work in German NER morphology and annotation analysis. The corpus data basis and the scope of the analysis are described in Section 3. The main part constitutes Section 4, where in Section 4.1 the morphological complexity of German named entities is investigated and in Section 4.2 the distribution of morphologically complex named entities in the dataset is presented. Section 5 then explains and examines six different annotation issues that have been identified within the GermEval reference annotation. This part also discusses the outcomes. The paper concludes with a short summary and a prospect of future work in Section 6.

2 Related Work

The performance of systems for NER is most often assessed through standard metrics like precision and recall, which measure the overall accuracy of matching predicted tags to gold standard tags. NER systems for German are no exception in this respect. In some cases the influence of difference linguistic features is reported, e.g. part of speech (Reimers et al., 2014) or morphological features (Capsamun et al., 2014; Schüller, 2014). The closest to our work, and the only one, to the best of our knowledge, which addresses linguistic error analysis of NER in German is that of Helmers (2013). The study examined different systems for NER, namely, TreeTagger (Schmid, 1995), SemiNER (Chrupała and Klakow, 2010), and the Stanford NER (Finkel and Manning, 2009) trained on German data (Faruqui and Padó, 2010). Helmers (2013) applied these systems to the German Web corpus CatTle.de.12 (Schäfer and Bildhauer, 2012) and inspected the influence of different properties on NER in a random sample of 100 true positives and 100 false negatives. It reports the odd-ratios for false classification for each of the properties. It was found that, e.g., named entities written exclusively in lower case were up to 12.7 times more likely to be misidentified, which alludes the difficulty of identifying adjectives derived from named entities. Another relevant example was named entities labelled as "ambiguous", i.e. which have a non-named entity homonym as in the case of named entities derived from a common noun phrase. In this case three out of four NER systems were likely to not distinguish named entities from their appellative homonyms with an odd-ratio of up to 13.7. Derivational suffixes harmed the identification in one classifier but inflectional suffixes seemed not to have similar influence. In addition, abbreviations, special characters and terms in foreign languages

were features which contributed to false positive results. In comparison with this study, ours addresses explicitly the effect of the rich German morphology on NER tasks.

Derczynski et al. (2015) raise the challenges of identifying named entities in microblog posts. In their error analysis the authors found that the errors were due to several factors: capitalization, which is not observed in tweets; typographic errors, which increase the rate of OOV to 2-2.5 times more compared to newsire text; compressed form of language, which leads to using uncommon or fragmented grammatical structures and non-standard abbreviations; lack of context, which hinders word disambiguation. In addition, characteristics of microblogs genre such as short messages, noisy and multilingul content and heavy social context, turn NER into a difficult task.

Benikova et al. (2015) describe a NER system for German, which uses the NoSta-D NE dataset (Benikova et al., 2014a) for training as in the GermEval challenge. The system employs CRF for this task using various features with the result that word similarity, case information, and character n-gram had the highest impact on the model performance. Though the high morphological productivity of German was stressed in the dataset description as well as in the companion paper for the conference (Benikova et al., 2014a), this method did not address it. What is more, it excluded partial and nested named entities which were, however, used in the GermEval challenge.

As this overview shows, linguistic error analysis is of great importance for the development of language technologies. Error analysis performed for NER tasks has been mostly concentrated on the token level, since this is the focus of most NER methods. However, our analysis differs in that it investigates specifically the role that morphology plays in forming named entities given that German is a language with rich morphology and complex word-formation processes.

3 Data Basis and Approach

3.1 GermEval 2014 NER Challenge Corpus

In order to pursue the given research questions we decided to take the Nosta-D NE dataset (Benikova et al., 2014b) included in the GermEval 2014 NER Challenge as the underlying data source of our investigations. The GermEval challenges were initiated to encourage closing the performance gap for NER in German compared to similar NER annotations for English texts. GermEval introduced a novelty compared to previous challenges, namely, additional (sub-) categories have been introduced indicating if the named entity mentioned in a token is embedded in compounding. Altogether, the named entity tokens could be annotated for the four categories *person*, *location*, *organisation* and *other* together with the information if the token is a compound word containing the named entity (e.g. LOCpart) or a word that is derived from a named entity (e.g. PERderiv). In addition it highlights a second level of 'inner' named entities (e.g. the person "Berklee" embedded in the organisation "Berklee College of Music"). Though the latter was addressed earlier, e.g. in Finkel and Manning (2009), it has been generally almost neglected. For detailed information about the GermEval NER Challenge, its setup, and the implemented systems we refer to (Benikova et al., 2014a). Out of the eleven systems submitted to the challenge, only one considered morphological analyses (Schüller, 2014) systematically. The best system, however, albeit utilizing some hand-crafted rules to improve common schemes of morphological alterations, did not model morphological variation systematically.

Besides a considerable volume of manual ground truth (31300 annotated sentences), the challenge data favourably was based upon well-documented, pre-defined guidelines[4]. This allowed us to create our complimentary annotations and to (re-)evaluate a subset of the original challenge ground truth along the same principles as proposed by the guidelines. Table 1 shows example sentences annotated for named entities (which can also be multi-word named entities consisting of more than one token) and their expected named entity types according to the provided GermEval reference annotation.

3.2 GermEval 2014 System Predictions

In order to obtain insights on the distribution of morphological characteristics of ground truth named entities which were successfully recognized by the systems (true positives) compared to ground

[4]The guidelines describing the categorization choice and classification of named entity tokens can be consulted in the following document: `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5` (revision 1.6 effective for GermEval is referenced in `https://sites.google.com/site/germeval2014ner/data`)

truth named entities which were not recognized or categorized correctly[5] (false negatives), we requested the system prediction outputs of GermEval participants from the challenge organizers[6].

Based on the best predictions[7] submitted for each system, we computed (1) the subset of ground truth named entities that all systems recognized (i.e. the true positive intersection, TPi; 1008 named entities) and (2) analogously the subset of ground truth named entities that none of the systems was able to recognize correctly (false negative intersection, FNi; 692 named entities). As performance of participating systems varied widely, we also analyzed (3) the false negatives of Hänig et al. (2014) (FN ExB; 1690 named entities).

3.3 Scope of the Analyses

The three mentioned data subsets were created to pursue two analysis goals: first, to investigate to what extent German named entities occur in morphologically altered forms and how complex these are and second, to report and evaluate issues we encountered in the GermEval reference annotations. The first investigation constitutes the main analysis and targets the question of whether there is a morphological gap in German NER. The second examination evolved out of annotation difficulties during the conduction of the first analysis. Even though not intended, we conducted the analysis of the reference annotation issues and present the results because the outcomes can contribute to the general research area of evaluating NER tools' performances.

The three data subsets build the foundation for both examination scopes. To obtain insights into the morphological prevalence and complexity of German named entities, the annotation was conducted according to the following steps: First, the annotator looked at those named entities in the datasets, which deviated from their lexical canonical form (in short LCF) which is the morphologically unmarked form. From gaining an overview of these named entities, linguistic features have been identified that correspond to the morphological segmentation steps which were applied to these morphologically altered named entities (see Section 4.1 for a detailed explanation). These linguistic features enable a measurement of the morphological complexity of a given named entity token provided by the reference annotation (i.e. the source named entity, in short SNE), e.g. "Kolpingwerkes" or "Kanadalilie" in Table 1. This measurement, however, required a direct linguistic comparison of the SNEs to their corresponding LCF form (i.e. their target named entity, in short TNE, e.g. "Kolpingwerk" and "Kanada"). Since the reference annotations provided only SNE tokens but no TNE data, a second annotation step was performed in which, all TNEs of the three subsets were manually added to the morphologically altered SNEs respectively[8]. In the third and last step the SNE has been annotated for its morphological complexity based on the numbers of different morphological alterations that were tracked back.

During the second and the third step of the morphological complexity annotation, problematic cases occurred in which a TNE could not be identified for the SNE given in the reference annotation. The reasons underlying these cases have been subsumed under six different annotation issues (details on these are explained in Section 5.1), which can significantly affect the performance measure of the tested GermEval NER systems. Therefore, if a SNE could not be annotated for morphological complexity, the causing issue was annotated for this SNE according to the six established annotation issues.

All three created GermEval data subsets have been annotated manually by a native German speaker and linguist and have been partially revised by a native German Computer Scientist while the code for the import and statistics was developed[9].

4 Morphological Complexity of German NE Tokens

4.1 Measuring Morphological Complexity

Morphological variation of named entity tokens has been considered as part of the GermEval annotation guidelines. I.e. next to the four named entity types, a marking for SNEs being compound words

[5] We adopted the criteria of the official Metric 1 of (Benikova et al., 2014a).

[6] We kindly thank the organizers for their support by providing these and also thank the challenge participants that agreed to have them provided to us and shared with the research community as a whole.

[7] according to F_1-measure

[8] The choice of a TNE included also the consideration of the four classification labels PER, LOC, ORG and OTH provided together with the SNE.

[9] The entire annotations of the morphological complexity of the named entities as well as the identified reference annotation error types can be consulted in this table including all three data subsets: https://raw.githubusercontent.com/AKSW/germeval-morph-analysis/master/data/annotation_imports/compl-issues-ann-ranks.tsv

or derivates of a TNE has been introduced (e.g. LOCderived or ORGpart). While this extension of the annotation of named entity tokens implies that German morphology impacts NER tasks, it does not indicate which morphological peculiarities actually occur. The linguistic analysis investigating morphologically altered SNEs revealed that SNEs exhibit a varying degree of morphological complexity. This degree is conditioned by the morphological inflection and/or word-formation steps that have been applied to a SNE in order to retrace the estimated TNE in its LCF. The resulting formalization of these alternation steps is as follows:

$$L \in \{\mathcal{C}_k\mathcal{D}_l \mid k,l \in \mathbb{N}\} \times \mathcal{P}(\{c,m,f\}) \text{ where}$$

$\mathcal{C}_k$ denotes that k compounding transformations were applied

$\mathcal{D}_l$ denotes that l derivations were applied

c denotes that resolving the derivation applied to the SNE resulted in a word-class change between SNE and TNE

m denotes that the morphological transformation process applied encompasses an inner modification of the TNE stem compared to its LCF

f denotes that the SNE is inflected

For convenience, we will omit the tuple notation and simplify the set representation of c and f: $\mathcal{C}_1\mathcal{D}_2 f$, $\mathcal{C}_1\mathcal{D}_1 cmf$, $\mathcal{C}_3\mathcal{D}_0 \in L$. In order to obtain the differing levels[10] of morphological complexity for named entities, we went through the identified morphological transformation steps always comparing the the given SNE in the test set with the estimated TNE in its LCF. It is defined that all named entities annotated with a complexity other than $\mathcal{C}_0\mathcal{D}_0$ are morphologically relevant and all named entities with a complexity satisfying $\mathcal{C}+\mathcal{D} \geq 1$ (i.e. involving at least one compounding relation or derivation) are morphologically complex, i.e. these require more than one segmentation step in the reanalysis of the SNE to the TNE in its LCF.

Thus, the SNE token can be increasingly complex, if it contains the TNE within a compound part of a compound or if the TNE is embedded within two derivations within the SNE. An example illustrating the morphological segmentation of the SNE "Skialpinisten" is given in Figure 1. It shows each segmentation step from the SNE back to the TNE in its LCF in detail and illustrates how deeply German named entities can be entailed in common nouns due to morphological transformations. Overall, the annotation of the three subsets revealed 27 levels of morphological complexity for German named entities. The appendix holds a comprehensive listing in Table **A** of these levels together with examples taken from the corpus[11].

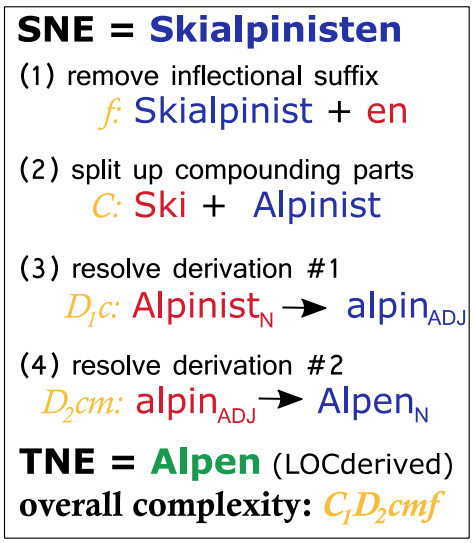

Figure 1: Example segmentation for annotating the SNE "Skialpinist" with the estimated TNE "Alpen".

4.2 Distribution of Morphologically Complex NE Tokens

Based on our systematization of complexity, we defined more focused complexity criteria such as $\mathcal{C} > 0$ and 'has m' (i.e. inner modification occurred) to complement the criteria morphologically relevant and morphologically complex introduced in section 4.1. Figure 2 shows comparative statistics of the prevalence of named entities matching these criteria for the TPi, FNi and FN ExB[12]. In general, morphologically relevant and morphologically complex named entities are much more prevalent among the false negatives. With respect to

[10] Although, we use the term level to simplify formulations, no strict ordering between the different possible configurations for the aforementioned formalization of complexity is presupposed.

[11] Note, that more levels can be assumed but no occurrences were found in the annotated subsets.

[12] The Scala and Python source code used to prepare the annotations, gather statistics and generate the plots is available at: https://github.com/AKSW/germeval-morph-analysis

the more focused criteria, the strongest increases occur for $\mathcal{C} > 0$, $\mathcal{D} > 0$ and 'is inflected'. In line with the definition of the criterion c, we observe $P(D > 0 \mid c) = 1$. I.e. the occurrence of c in a complexity assignment strictly implies that at least one derivation was applied. The observation of a strong association between inner modification and derivation processes ($P(D > 0 \mid m) = 0.86$) also is in line with intuitive expectations for German morphology.

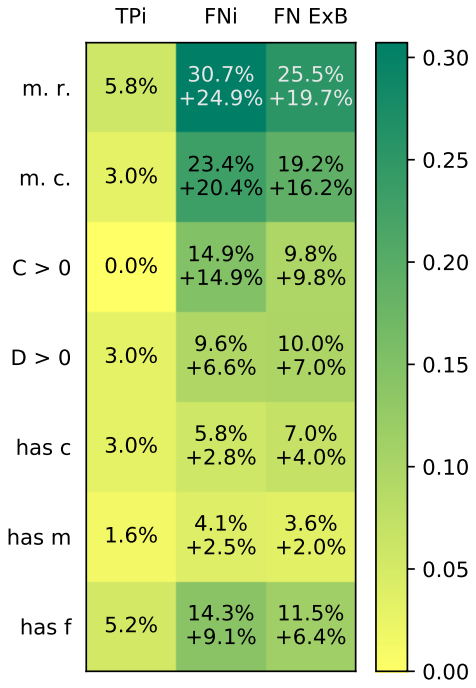

Figure 2: Prevalence of morphological complexities satisfying specified criteria. Colors encode magnitude of increase of the FN subset compared to the TPi. (m.r. = morph. relevant, m.c. = morph. complex).

Figure 3 presents the same comparative statistics between TPi and FNi for the named entities grouped according to their reference classification. In general morphological alteration is more common in named entities annotated with the types PER and LOC. Further, we find lower variance of increase of $C > 0$ across the classes compared to $D > 0$, which is much more common in LOC named entities (+20.9%) and PER named entities (+12.8%) than in named entities classified ORG and OTH (increase $\leq 2\%$). The statistics partitioned by named entity type also reveal that the only types morphologically complex named entities in the TPi subset are LOC named entities with derivations. Analogous statistics between TPi and FN ExB showed similar trends and were omitted for brevity[13].

4.3 Morphological Complexity in Context of NER System Errors

Interestingly, the LOC and PER named entities, that were found to be morphologically complex most often on the one hand are, conversely, the ones covered best by the top GermEval systems according to Benikova et al. (2014a). However, these classes were also deemed more coherent in their analysis, a qualitative impression we share with respect to variety of occurring patterns for morphological alterations. Also, since the morphological complexity of named entities is also one of many factors determining its difficulty to be spotted and typed correctly (besides e.g. inherent ambiguity of involved lexcial semantics), this might indicate that these two categories might still simply be the ones potentially benefiting most from more elaborate modelling of effects of morphological alteration, as the reported F1 of approx. 84 % for LOC and PER still indicates space for improvements.

Further, 19 morphologically complex named entities in FNi could be found, whose TNE was identical with a TNE from the TPi. For example, all systems were able to correctly assign LOC-deriv to 'polnischen' (TNE='Polen'), however no system was able to recognize 'austropolnischen' (same TNE). Analogously, there is 'Schweizer' in TPi, but 'gesamtschweizerischen' in FNi (common TNE: 'Schweiz'). There were 38 additional morphologically complex named entities in FN ExB with a corresponding TPi named entity sharing the TNE, e.g. 'Japans' (TP) vs. 'Japan-Aufenthaltes' (FN). For all of these pairs, it appears plausible to assume that the difficulty for the corresponding false negative can be attributed to a large extend to the morphological complexity, as simpler variants posed no hindrances to any of the tested systems[14]. For the ExB system, these kind of false negatives constitute 3.4 % of all false negatives, which could be viewed raw estimation of potential increase in recall if hypothetically morphological complexity of named

[13]The corresponding plot is available at: https://github.com/AKSW/germeval-morph-analysis/blob/master/plots/phrase-partitioned-stats-FalseNegExB.pdf

[14]Still we also acknowledge that several factors of lexical semantics, syntax etc. influence how challenging it is to spot a specific NE occurrence in context and more systematic analysis of these factors would be needed to attribute the error to morphological causes with certainty.

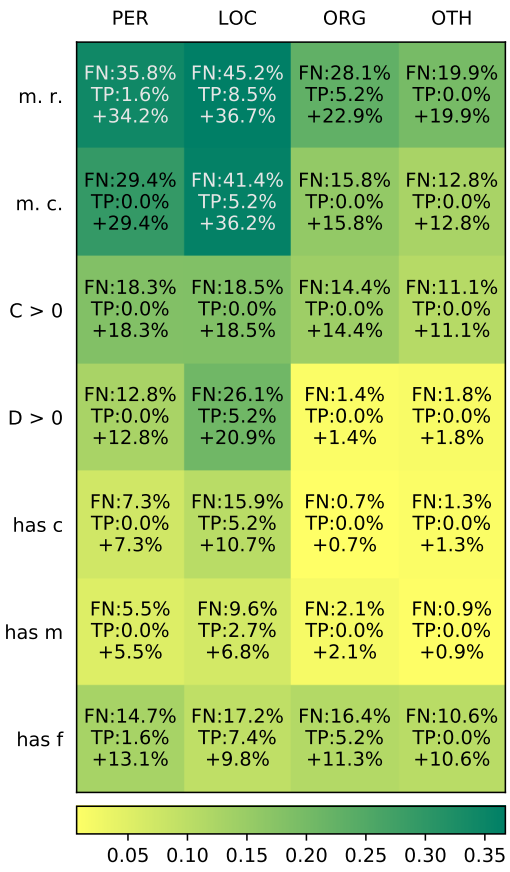

Figure 3: Prevalence of morphological complexities satisfying specified criteria, grouped be named entity type. Each cell presents ratios in the FNi, the TPi and respective increase. Colors encode magnitude of increase. (m.r. = morph. relevant, m.c. = morph. complex).

entities would be mitigated entirely. It should also be noted that the reported occurrence counts of these pairs for ExB are lower bounds, since not all of its true positives had been annotated at the time of writing.

5 Reference Annotation Related Issues

5.1 Reference Annotation Issue Types

During the annotation for morphological complexity issues arose with regard to the GermEval reference annotations which led to various difficulties.

Overall, six reference annotation issues have been identified and all three subsets have been annotated for these issues (also cf. Table 2):

Issue #1 NOT DERIVED: A significant number of SNEs with the type LOCderived is morphologically not derived from the location TNE but from the inhabitant noun, e.g. "Kirgisisch" is not derived from "Kirgistan" but from "Kirgise".

Issue #2 WRONG NE TYPE: This issue refers to SNEs which are correctly identified, but are assigned to the wrong named entity category.

Issue #3 WRONG SPELLING: SNEs annotated with this issue are either incorrectly spelled or tokenized.

Issue #4 NO NE: This issue holds for SNEs, which turn out to be only common nouns in the sentences they occur.

Issue #5 INVALID REFERENCE: SNEs referring to book/film titles, online references or citations which are incomplete, wrong or the online reference is a title for a website given by some person but not the real title or URL.

Issue #6 TNE UNCLEAR: This issue summarizes reasons for preventing a TNE of being identifiable form a given SNE, i.e. it is not possible to morphologically decompose the SNE to retrieve the TNE or there are more than one TNEs included in the SNE.

If NOT DERIVED, NO NE, INVALID REFERENCE or TNE UNCLEAR occur for a named entity, assignment of a morphological complexity level becomes impossible. Consequently, the corresponding named entities (189) were excluded from the complexity statistics presented in sections 4.2 and 4.3. WRONG NE TYPE and WRONG SPELLING, on the other hand, albeit also implying difficulties for NER systems, do not interfere with identifying the TNE (and thus the complexity level). Hence, such named entities were not excluded.

5.2 Distribution and Effects of Annotation Issues

Table 2 provides, in addition to examples for the aforementioned categories of annotation issues, their total prevalence across TPi and FN ExB (subsuming FNi). Table 3 additionally indicates the distribution of issue occurrences in comparison between the subsets. Overall, occurrence of annotation issues are about three times more likely in the false negative sets compared to TPi, a trend in a similar direction as for the occurrence of morpholoically complex named entities.

It appears questionable to count named entities with WRONG NE TYPE, NO NE and INVALID REFERENCE that have not been recognized by any NER system as a false negative, as these named entities do not actually constitute named entities as defined by the guidelines (analogously for true positives).

issue	example	prevalence
NOT DERIVED	SNE = *Kirgisische* (LOC-deriv) with TNE = *Kirgistan*	94 (31.5 %)
WRONG NE TYPE	SNE = *barocker* (ORG-deriv) with TNE = *Barock*, "Baroque" is an epoch, it should have been annotated as OTH-deriv	62 (20.8 %)
WRONG SPELLING	SNE = *Freiburg/31:52* with TNE = *Freiburg*	51 (17.1 %)
NO NE	SNE = *Junta* - "Junta" is a common noun, there is no TNE	18 (6.0 %)
INVALID REFERENCE	SNE = *Was ist theoretische Biologie ?* - this is a HTML link label, which is not related to any NE	7 (2.4 %)
TNE UNCLEAR	SNE = *Köln/Weimar/Wien* - TNE is unclear, unknown to which of the three named entities is referred to	66 (22.2 %)

Table 2: Encountered issues pertaining to GermEval reference annotations.

Thus, we projected the M1 performance measures on the test split for the ExB system disregarding these named entities[15]. The adjustment results in discounting five false positives and 44 false negatives, result in an increase in recall by 0.48 % and F1 by 0.34 %. Although, this change is not big in absolute magnitude, it can still be viewed relevant considering that the margin between the to best systems at GermanEval was merely 1.28 % for F1 as well (Benikova et al., 2014a).

issue	TPi	FNi	FN ExB
#1	41 (4.07 %)	18 (2.60 %)	53 (3.14 %)
#2	0 (0.00 %)	30 (4.34 %)	62 (3.67 %)
#3	1 (0.10 %)	24 (3.47 %)	50 (2.96 %)
#4	1 (0.10 %)	10 (1.45 %)	17 (1.01 %)
#5	0 (0.00 %)	4 (0.58 %)	7 (0.41 %)
#6	0 (0.00 %)	19 (2.75 %)	66 (3.91 %)
All	43 (4.27 %)	105 (15.17 %)	255 (15.09 %)

Table 3: Frequencies of occurrence of annotation issues by category and subset. Percentages in parentheses are relative frequencies for the corresponding subset.

6 Conclusion

This study presented an analysis of German NER as reflected by the performance of systems that participated in the GermEval 2014 shared task. We focused on the role of morphological complexity of named entities and introduced a method to measure it. We compared the morphological characteristics of named entities which were identified by none of the systems (FNi) to those identified by all of the systems (TPi) and found out that FNi named entities were considerably more likely to be complex than the TPi ones (23.4% and 3.0% respectively). The same pattern was detected also for the system which achieved the best evaluation in this shared task. These findings emphasize that morphological complexity of German named entities correlates with the identification of named entities in German text. This indicated that the task of German NER could benefit from integrating morphological processing.

We further discovered annotation issues of named entities in the GermEval reference annotation for which we provided additional annotation. We believe that the presented outcomes of this annotation can help to improve the creation of NER tasks in general.

As a future work, we would like to extend our annotation to analyze how these issues affect the evaluation of the three best performing systems more thoroughly. In addition, a formalization to measure the variety of occurring patterns of morphological alteration (used affixes/affix combinations, systematic recurrences of roots...) as a complementary measure for morphological challenges seems desirable. We will further have multiple annotators to morphologically annotate the named entities of the GermEval reference, in order to estimate the confidence of our observation by measuring inter-annotator agreement.

Acknowledgment This paper's research activities were funded by grants from the H2020 EU projects ALIGNED (GA-644055) and FREME (GA-644771) and the Smart Data Web BMWi project (GA-01MD15010B).

[15] Due to lack of complete screening of all true positives of ExB for annotation issues we linearly interpolated the exemption of one true positive according to TPi to the exemption of five true positives for all true positives of that system.

References

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. 2014a. Germeval 2014 named entity recognition shared task: companion paper. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 104–112, Hildesheim, Germany.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. Nosta-d named entity annotation for german: Guidelines and dataset. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Darina Benikova, Seid Muhie Yimam, Prabhakaran Santhanam, and Chris Biemann. 2015. Germaner: Free open german named entity recognition tool. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 31–38, Duisburg-Essen, Germany. German Society for Computational Linguistics and Language Technology.

Roman Capsamun, Daria Palchik, Iryna Gontar, Marina Sedinkina, and Desislava Zhekova. 2014. Drim: Named entity recognition for german using support vector machines. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 129–133, Hildesheim, Germany.

Grzegorz Chrupała and Dietrich Klakow. 2010. A named entity labeler for german: Exploiting wikipedia and distributional clusters. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150. Association for Computational Linguistics.

Lea Arianna Helmers. 2013. Eigennamenerkennung in Web-Korpora des Deutschen. Eine Herausforderung für die (Computer)linguistik. Bachelor Thesis, Humboldt-Universität zu Berlin.

Christian Hänig, Stefan Bordag, and Stefan Thomas. 2014. Modular classifier ensemble architecture for named entity recognition on low resource systems. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 113–116, Hildesheim, Germany.

Nils Reimers, Judith Eckle-Kohler, Carsten Schnober, and Iryna Gurevych. 2014. Germeval-2014: Nested named entity recognition with neural networks. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, pages 117–120, Hildesheim, Germany.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Peter Schüller. 2014. Mostner: Morphology-aware split-tag german ner with factorie. In *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 121–124, Hildesheim, Germany.

compl.	**TPi**	**FNi**	**FN ExB**	**example SNE**	**example TNE**
$\mathcal{C}_0\mathcal{D}_0$	910 (94.20%)	442 (69.28%)	1149 (74.47%)	Mozart	Mozart
$\mathcal{C}_0\mathcal{D}_0f$	27 (2.80%)	47 (7.37%)	98 (6.35%)	Mozarts	Mozart
$\mathcal{C}_1\mathcal{D}_0$	0 (0.00%)	62 (9.72%)	101 (6.55%)	Mozart-Konzert	Mozart
$\mathcal{C}_1\mathcal{D}_0f$	0 (0.00%)	15 (2.35%)	24 (1.56%)	Mozart-Konzerten	Mozart
$\mathcal{C}_1\mathcal{D}_0m$	0 (0.00%)	3 (0.47%)	5 (0.32%)	Pieterskirche	Pieter
$\mathcal{C}_1\mathcal{D}_0mf$	0 (0.00%)	3 (0.47%)	4 (0.26%)	Reichstagsabgeordneten	Reichstag
$\mathcal{C}_0\mathcal{D}_1$	0 (0.00%)	9 (1.41%)	20 (1.30%)	Donaldismus	Donald
$\mathcal{C}_0\mathcal{D}_1f$	0 (0.00%)	1 (0.16%)	4 (0.26%)	Donaldismusses	Donald
$\mathcal{C}_0\mathcal{D}_1m$	0 (0.00%)	7 (1.10%)	10 (0.65%)	Nestorianismus	Nestorius
$\mathcal{C}_0\mathcal{D}_1mf$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Spartiaten	Sparta
$\mathcal{C}_0\mathcal{D}_1c$	5 (0.52%)	16 (2.51%)	61 (3.95%)	japanisch	Japan
$\mathcal{C}_0\mathcal{D}_1cf$	9 (0.93%)	8 (1.25%)	14 (0.91%)	japanischen	Japan
$\mathcal{C}_0\mathcal{D}_1cm$	1 (0.10%)	1 (0.16%)	6 (0.39%)	europäisch	Europa
$\mathcal{C}_0\mathcal{D}_1cmf$	10 (1.04%)	8 (1.25%)	19 (1.23%)	europäischen	Europa
$\mathcal{C}_2\mathcal{D}_0$	0 (0.00%)	3 (0.47%)	5 (0.32%)	Bibelforscherfrage	Bibel
$\mathcal{C}_2\mathcal{D}_0mf$	0 (0.00%)	1 (0.16%)	1 (0.06%)	Erderkundungssatelliten	Erde
$\mathcal{C}_1\mathcal{D}_1$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Benediktinerstift	Benedikt
$\mathcal{C}_1\mathcal{D}_1f$	0 (0.00%)	2 (0.31%)	2 (0.13%)	Transatlantikflüge	Atlantik
$\mathcal{C}_1\mathcal{D}_1m$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Römerstrasse	Rom
$\mathcal{C}_0\mathcal{D}_2$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Geismarerin	Geismar
$\mathcal{C}_0\mathcal{D}_2f$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Hüttenbergerinnen	Hüttenberg
$\mathcal{C}_0\mathcal{D}_2m$	0 (0.00%)	0 (0.00%)	1 (0.06%)	Rheinländerin	Rheinland
$\mathcal{C}_0\mathcal{D}_2cf$	0 (0.00%)	1 (0.16%)	1 (0.06%)	austropolnischen	Polen
$\mathcal{C}_0\mathcal{D}_2cmf$	4 (0.41%)	0 (0.00%)	3 (0.19%)	transatlantischen	Atlantik
$\mathcal{C}_3\mathcal{D}_0$	0 (0.00%)	1 (0.16%)	1 (0.06%)	25-US-Dollar-Marke	US
$\mathcal{C}_1\mathcal{D}_2cf$	0 (0.00%)	2 (0.31%)	2 (0.13%)	gesamtschweizerischen	Schweiz
$\mathcal{C}_1\mathcal{D}_2cmf$	0 (0.00%)	1 (0.16%)	2 (0.13%)	Skialpinisten	Alpen
total	966	638	1543		

Appendix A: Distribution of the morphological complexities in the annotated subsets