Automatic Extraction of Nominal Phrases from German Learner Texts of Different Proficiency Levels

Ronja Laarmann-Quante, Marco Müller, Eva Belke
Ruhr University Bochum, Faculty of Philology, Department of Linguistics
{ronja.laarmann-quante, marco.mueller-z3b, eva.belke}@rub.de

Abstract

Correctly inflecting determiners and adjectives so that they agree with the noun in nominal phrases (NPs) is a big challenge for learners of German. Given the increasing number of available learner corpora, a large-scale corpus-based study on the acquisition of this aspect of German morphosyntax would be desirable. In this paper, we present a pilot study in which we investigate how well nouns, their grammatical heads and the dependents that have to agree with the noun can be extracted automatically via dependency parsing. For six samples of the German learner corpus MERLIN (one per proficiency level), we found that in spite of many ungrammatical sentences in texts of low proficiency levels, human annotators find only few true ambiguities that would make the extraction of NPs and their heads infeasible. The automatic parsers, however, perform rather poorly on extracting the relevant elements for texts on CEFR levels A1-B1 ($<$ 70%) but quite well from level B2 onwards (~90%). We discuss the sources of errors and how performance could potentially be increased in the future.

Keywords: learner corpora, L2, nominal inflection, German, dependency parsing, automatic extraction

1. Introduction

One of the most challenging areas when learning German as a second or foreign language (L2) is nominal inflection (e.g. Hopp, 2013, 2018; Ruberg, 2013; Wegener, 1995). Determiners and adjectives in noun phrases must agree with the noun in case, number and grammatical gender, as shown in (1), where case is dative (Dat), number is singular (Sg) and gender is masculine (M):

(1)	mit	dem	Hund
	with	the.Dat.Sg.M	dog.Dat.Sg.M

This easily results in errors as in (2), where the choice of the article *der* could be interpreted as an error in case (nominative) or gender (feminine):

	mit	*der	Hund
(2)	with	the.**Nom**.Sg.M	dog.Dat.Sg.M
	with	the.Dat.Sg.**F**	dog.Dat.Sg.M

While gender and number is determined by the noun and governs the type of inflection used (Sg.M, as in (1), Sg.F, Sg.N or Pl), case depends on a noun's thematic role or is contextually constrained, e.g. by prepositions, as in (1), see e.g. Eisenberg (2020) for details.

With more and more learner corpora being available, it is possible to study the acquisition of nominal inflection empirically based on free productions across large numbers of learners with diverse language backgrounds, proficiency levels etc. To date, such corpus-based studies have typically been based on a single corpus with an idiosyncratic annotation scheme (e.g. Spinner and Juffs, 2008; Studinger, 2010). For a broader analysis across corpora, including corpora that are as yet unannotated, the first step is to extract the relevant instances from a text, i.e. *nominal phrases* (NPs) comprising a noun and a determiner and/or adjectives that have to agree with the noun (see Section 3.1 for details about the definition of an NP we use for the purposes of our annotation). We also need to identify the grammatical head of the noun and its syntactic relation with the noun in order to determine the required case marking. In principle, this can be achieved via automatic dependency parsing. It has been shown for several languages, though, that the performance of parsers decreases when applied to learner language as compared to standard language. Apparently, performance varies across dependency relations and depends on the types of errors present (e.g. Ott and Ziai, 2010; Köhn et al., 2016 for German; Volodina et al., 2022 for Swedish).

In this paper, we investigate the feasibility of automatically extracting NPs in German as a prerequisite for studying nominal inflection in learner texts on a large scale. Our ultimate goal is a corpus-based study of the development of nominal agreement in learners of German. We want to use the results and insights from this to develop computer-assisted language learning (CALL) applications to support learners in acquiring this challenging aspect of the German language.

Regarding NP extraction, we have the following hypotheses: (1) Texts of low-proficiency learners feature many ungrammatical sentences that even human annotators struggle to annotate with respect to NPs and the nouns' heads, resulting in lower agreement scores compared to those for texts of high-proficiency learners. Accordingly, this will be reflected in worse automatic extraction performance. (2) Since in NPs, nouns, determin-

LREC-COLING 2024, pages 1925–1931

Level	# sents	avg. sent. length	# sents with NPs	# NPs	avg. NP length	# NPs addr.
A1	40 (14%)	7.8	32	70	1.67	2 (3%)
A2	40 (1.8%)	8.7	32	71	1.76	10 (14%)
B1	40 (1.1%)	11.7	35	98	1.70	19 (19%)
B2	40 (0.9%)	16.2	40	131	1.73	21 (16%)
C1	40 (6.8%)	18.7	39	150	2.08	3 (2%)
C2	40 (73%)	19.1	40	149	1.98	0 (0%)

Table 1: Dataset statistics based on the gold standard annotation of nominal phrases (NPs). '# NPs addr.': number of NPs in address or place/date lines in a letter which are not of interest for our study.

ers and adjectives occur in close proximity, their automatic extraction should work rather reliably even if the sentence as a whole is not fully grammatical. The automatic identification of the NPs' heads, by contrast, should be more error-prone, as they are often more distant from the nouns.

We test these hypotheses on random samples of sentences from the MERLIN Corpus across levels A1-C2 of the Common European Reference Frame (CEFR).

2. Data

The data for this study is taken from the German section of the freely-available MERLIN Corpus (Wisniewski et al., 2018) that includes 1,033 L2 German texts in total, distributed across CEFR levels A1-C2. We use the CEFR level that was assigned to the text by the re-rating (fair rating) that comes with the corpus. The texts were produced as part of an official language proficiency test and comprise different text types, mostly formal and informal letters and e-mails but also essays and reports for higher proficiency levels. Although we only use one corpus in this pilot study, our approach is corpus-independent because we use the plain text files to ensure that our pipeline works on unannotated learner corpora as well. Hence, in principle, we could run the same pipeline for identifying NPs on any available learner corpus without being dependent on pre-existing annotations.

For each of the six proficiency levels, we randomly sampled 40 sentences, yielding 240 sentences in total. We used the sentence boundaries assigned by the ParZu parser (Sennrich et al., 2009, 2013) as the basis for sampling sentences from the corpus. We do not evaluate sentence segmentation in this pilot study. Table 1 shows some statistics of the resulting dataset based on the gold standard annotations (see Section 3.2). The samples of 40 sentences per level constitute highly variant proportions of sentences available per level (see percentage under '# sents'). Descriptively, we see that sentence length, number and length of the NPs (measured in tokens) increase with proficiency, as one would expect. Especially in the lower proficiency levels, not every sentence contains an NP. Nevertheless, we keep all sentences in the study because the automatic parsers may identify NPs in these sentences by mistake. One particularity of this dataset is that, since many of the texts are letters, several address lines and lines containing place and date are present. They mostly consist of NPs but these are not of interest with regard to studying agreement. The last column of Table 1 ('# NPs addr.') shows how many of the annotated NPs under '# NPs' come from address or place/date lines.

3. Methods

3.1. Annotation Scheme

As to our knowledge there are no existing annotation schemes targeting nominal agreement in German learner language, we created a new scheme. This was done in a bottom-up rather than a top-down fashion in order to impose as few constraints as possible on the analyses. In particular, we wanted to avoid using syntactic theories that take native speakers as their standard. Learners of German as a foreign language typically acquire nouns as part of their vocabulary training and learn the corresponding gender subclass information by heart. This is why we put the noun into focus as the carrier of semantic information from which grammatical features of the NP have to be derived. Following the descriptive grammar of Eisenberg (2020), an NP in our scheme consists of the noun and, if applicable, a determiner and adjective(s) that have to agree with the noun. Furthermore, we follow the framework of a dependency analysis as described in the grammar by Foth (2006) for parsing German. We acknowledge that a full nominal phrase may also include post-nominal modifiers, which are not considered here, and that, depending on the syntactic tradition, the phrases of interest to us may also be referred to as determiner phrases rather than nominal phrases.

3.2. Manual Annotation

Two of the authors of this paper, who both have a background in linguistics, manually marked all NPs and each noun's head in the sampled sentences using the annotation tool *INCEpTION* (Klie

Level	whole dataset # NPs	% NP agree	# overlap NPs	% head agree	cleaned dataset # NPs	% of total	% NP agree	# overlap NPs	% head agree
A1	71	.76	55	.84	57	.80	.95	55	.84
A2	75	.71	62	.87	63	.84	.83	61	.87
B1	101	.63	69	.91	80	.79	.80	69	.91
B2	133	.77	107	.86	109	.82	.94	107	.86
C1	151	.85	147	.86	148	.98	.87	147	.86
C2	149	.91	149	.90	149	1.0	.91	149	.90

Table 2: Inter-annotator agreement for the whole and the cleaned dataset (ignoring address lines, place/date lines and names in greetings and closing formulas). '#NPs': total number of NPs that are annotated by at least one of the annotators. '% NP correct': number of completely identical NPs divided by the total number of NPs. '# overlap NPs': number of overlapping NP annotations as the basis for the head annotation. '% of total': proportion of NPs in the cleaned dataset compared to the whole dataset.

et al., 2018). They followed the guidelines in Foth (2006), which are also underlying the ParZu parser. Each NP is marked as a span of tokens, based on the tokenization of the ParZu parser. Also bare nouns, i.e. nouns without any dependents, are marked. Determiners may be articles (engl. *the, a*) or attributive pronouns (e.g. possesive pronouns such as engl. *my, their*). A marked span may also include adverbs (as in a phrase like *der* ***wirklich*** *schöne Strand* 'the **really** beautiful beach'). After the annotation process, the two annotators discussed their disagreements and jointly decided on a final gold standard.

3.3. Automatic Extraction

The dependency parsers we chose for the automatic extraction were ParZu (Sennrich et al., 2009, 2013)[1] and, for comparison, *spaCy* (v3.6.1, Honnibal et al. 2020)[2], which both can be used off-the-shelf. We store the parsing results in CoNLL format. Ortmann et al. (2019) found ParZu to be the best performing parser when used on German data from several non-standard domains such as movie subtitles and sermons. Likewise, Adelmann et al. (2018) found it to work well on out-of-domain test data, especially when it comes to exact matches. To our knowledge, it has not yet been tested on learner data. The spaCy parser follows annotation principles that are similar to ParZu for the elements of interest to us (e.g. preposition as head of the noun).

NPs are extracted by collecting elements tagged as nouns (including proper nouns) and their dependents (comprising determiners and adjectives as well as adverbs depending on adjectives), as shown in Figure 1. Modifiers that follow the noun are not considered further here. We also extract the head of the head because this is relevant

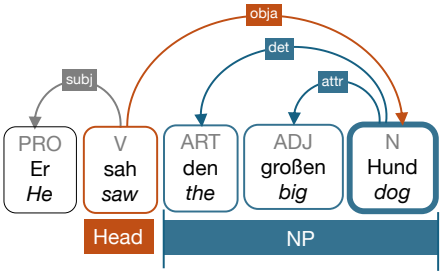

Figure 1: Automatic NP extraction and head identification from a dependency parse.

e.g. for certain prepositions that require a different case depending on the context but we do not evaluate this in this pilot study.

4. Evaluation

To determine the agreement between the two human annotators, the annotated NPs were semi-automatically aligned and the percentage of exact matches, i.e. identical NPs, was calculated. Spans that overlap but are not identical count as a disagreement, as do spans that are only present in one of the annotations. The automatically extracted NPs and the gold standard are compared accordingly. To determine the agreement/accuracy for identifying the nouns' heads, we only consider those NPs where the NP annotations are identical or at least overlap because missing NP annotations automatically result in missing head annotations.

4.1. Human-Human Agreement

Table 2 shows the inter-annotator agreement (IAA) between the two human annotators. The left part of the table refers to the whole dataset. The number of NPs ('# NPs') comprises the total number of NPs that were annotated by at least one of the annotators, which is why the numbers are slightly higher than in the gold standard in Table 1.

[1] https://github.com/rsennrich/ParZu/

[2] https://github.com/explosion/spaCy/releases/tag/v3.6.1; we used model `de_core_news_sm` for spaCy.

	# NPs		% NP correct		# overlap NPs		% head correct		
Level	ParZu	spaCy	ParZu	spaCy	ParZu	spaCy	ParZu	spaCy	**TH1Diffs**
A1	81	91	.58	.54	60	63	.67	.68	.47
A2	74	80	.62	.56	61	62	.72	.74	.29
B1	94	96	.65	.68	72	77	.71	.81	.22
B2	113	111	.89	.86	109	109	.84	.88	.12
C1	153	158	.86	.78	150	148	.85	.89	.09
C2	157	153	.85	.88	144	149	.87	.90	.11

Table 3: Performance of automatic NP extraction and head identification for both parsers. '#NPs': total number of NPs annotated in the gold standard or/and by the parser. '% NP correct': number of completely identical NPs divided by the total number of NPs. '# overlap NPs': number of overlapping NP annotations as the basis for the head annotation. 'TH1Diffs': Measure of ungrammaticality, see Sec. 4.2.

	span		miss		extra		other		total	
Level	ParZu	spaCy	ParZu	spaCy	ParZu	spaCy	ParZu	spaCy	ParZu	spaCy
A1	12	14	9	5	12	23	1	0	34	42
A2	13	14	2	1	11	17	2	3	28	35
B1	10	11	8	3	14	16	1	1	33	31
B2	5	10	2	1	2	1	3	3	12	15
C1	9	17	0	1	3	9	9	8	21	35
C2	3	9	8	1	5	3	7	6	23	19

Table 4: Categorization of errors in NP extraction for both parsers. 'span': same NP but not exactly the same span, 'miss': missing NP, 'extra': extra NP not in the gold standard, 'other': special cases, e.g. regarding tokenization or punctuation.

We found unclear guidelines to be one consistent source of disagreement, as one of the annotators did not mark names in address lines or greeting and closing formulas. This is why we also measured the IAA in a cleaned version of the dataset, excluding address lines, place and date lines and NPs in greeting formulas. The result is shown in the right part of Table 2, giving a more accurate view on the actual annotation difficulties.

Our hypothesis was that the agreement for identifying NPs increases with increasing proficiency level but the results show that the highest agreement (in the cleaned dataset) can be found for level A1. It turns out that these texts are rather easy to annotate because they often feature common simple nouns and simple structures. For higher proficiency levels, agreement partly drops because structures become more complex and more ambiguous due to errors. For example, the phrase *ohne etwas zum essen* (level B1) can be interpreted in different ways, shown in (3), postulating different kinds of errors (printed in bold). Only in interpretation (a), the sentence would contain an NP.[3] Most disagreements, however, resulted from not following the guidelines perfectly.

(3) a. ohne etwas zum **Essen**
without sth. to.PREP.ART eat.NOUN
'without something to eat'

b. ohne etwas **zu** essen
without sth. to.PART eat.VERB
'without eating something'

Ambiguities resulting from ungrammatical and incomprehensible sentences were very rare even in the low proficiency levels. One such example from level A2 is shown in (4):

(4) möchtest du mit einem Zug Fahrkarte
want you with a train ticket
'do you want with a train ticket'

This might be an error concerning lexical choice (5a) or inflection (5b; note that this reading requires a question word or another verb to render it grammatical). Hence, there is no unambiguous interpretation of NPs and heads.

(5) a. mit einem Zug **fahren**
with a train ride.VERB
'[Do you want (to)] ride with a train'

b. mit **einer** **Zugfahrkarte**
with a train ticket
'[Do you want (to)] with a train ticket.'

Regarding the annotation of each noun's head, the IAA remains stable across proficiency levels and is rather high. We found that here, disagreements

[3]Note that, put somewhat simplified, nominalizations, like nouns in NPs, are capitalized in German.

in the lowest proficiency levels partly result from true ambiguities due to ungrammatical sentence structures as in (6):

(6)

Schreiben	Sie	am	e-mail
Write.IMP	you	at the	e-mail

While one annotator interpreted *am e-mail* as an NP, with *am* being erroneously used for the indefinite article *eine* (not shown in (6)), and the verb *schreiben* as the head of the noun, the other interpreted it as a prepositional phrase (as shown in (6)) with *am* as the head of the NP *e-mail*, corresponding to *an dem*, i.e. a preposition + article used here with the wrong grammatical gender.

In the higher proficiency levels, disagreements arise mostly because learners use complex sentence structures that render the application of the dependency annotation guidelines more difficult, e.g. with regard to complex verbs or coordinations.

4.2. Automatic Extraction Performance

NP Extraction For evaluating automatic NP extraction, we only report the parsers' performance on a cleaned dataset without NPs occurring in address lines or place/date lines.[4] Table 3 shows the performance of the automatic NP extraction and head identification based on the two parsers, ParZu and spaCy, compared to the gold standard. '#NPs' shows the total number of NPs that are annotated in the gold standard or by the parser (or ideally both). '% NP correct' is the proportion of completely identical NPs out of these NPs.

We see that with regard to NP extraction, the performance of the two parsers is comparable. Descriptively, ParZu outperforms spaCy in four out of six proficiency levels but the difference is marginally statistically significant in level C1 only ($\chi^2(1) = 3.19$, $p < 0.08$). For both parsers, we see a significant increase in performance with increasing proficiency level, confirming our hypothesis (both $\chi^2(5) > 57$, both $p < .001$). There is a clear cutoff point: For levels A1-B1, performance remains $< 70\%$, whereas for levels B2-C2 it is close to 90%. For both parsers, the performance difference between B1 and B2 was statistically significant (both $\chi^2(1) > 9$, both $p < .01$).

To analyze the errors made by the parsers further, we categorized them according to four categories: **span** (same NP but not exactly the same span), **miss** (missing NP), **extra** (extra NP that is not in the gold standard) and **other**, comprising some special cases, e.g. regarding tokenization or punctuation marks. The result is shown in Table 4. We see some clear differences between the parsers: Compared to ParZu, spaCy more often assigns a different span, especially with regard to adverbs, and also tends to annotate more NPs than there are in the gold standard. Mostly these are misspelled words and words that the learner capitalized but that are not nouns. On the other hand, ParZu misses more NPs than spaCy, mostly because the nouns were not capitalized.

Head Identification Contrary to our expectation, the performance for head identification is slightly better than for NP extraction (recall, however, that head annotations are only regarded when the NP was at least partly correctly identified by the parser in the first place). In both parsers, there is a similar performance increase with increasing proficiency level. For head identification, spaCy descriptively outperforms ParZu in all proficiency levels but the differences are not statistically significant ($\chi^2(1) < 1.41$, $p > .2$).

Role of Ungrammaticality Finally, we analyzed whether there is a relationship between parser performance and (un)grammaticality. The texts in the MERLIN corpus come with a minimal target hypothesis (TH1), which corrects for grammatical errors (including spelling and punctuation), and an annotation of the differences between the original text and the TH1 (TH1Diffs). For our (cleaned) dataset, the number of TH1Diffs per proficiency level (normalized by the total number of tokens) is shown in the last column of Table 3. As we see, the number of TH1Diffs decreases with increasing proficiency, most notably between levels B1 and B2, which is where we had found parser performance to increase markedly. We found the point-biserial correlation between correct NP extraction and the number of TH1Diffs in the respective sentence to be highly significant for ParZu (r = -0.32) and spaCy (r = -0.3) but not for human-human agreement (r = -0.07, p = 0.1).

5. Conclusion and Outlook

We investigated how well nominal phrases and their heads can be automatically extracted from German learner texts based on dependency parsing. We found that human annotators can identify most NPs and their heads unambiguously, even in texts with many ungrammatical structures. Automatic extraction, however only works well from level B2 onwards. In lower proficiency levels, the parsers are often misled by capitalization errors and spelling errors, which should be easy to fix. Our goal is to achieve a reliable automatic extraction of NPs for a large-scale corpus-based study of the development of nominal inflection in learner texts. The results presented here provide a baseline for the extraction performance that we will try to improve in future work, e.g. by some automatic normalization of the texts prior to parsing (see Volodina et al., 2022).

[4] On the whole dataset, performance is slightly lower for both parsers but the patterns are similar.

6. Code and Data Availability

We make the code and data from this study available under https://github.com/rlaarmqua/Automatic-NP-Extraction-from-German-Learner-Texts.

7. Acknowledgments

We would like to thank the students of the seminar *Corpus-based analyses of nominal inflection in German* in summer semester 2023 Ekram Abdalla, Christopher Chandler, Niels Lange, Raha Musavi, Georg Stin, Jay Suchardt, Johanna Wrede and Imge Yüzüncüoglu for kick-starting the analyses. Furthermore, we thank the anonymous reviewers for their constructive and helpful comments and ideas.

8. Bibliographical References

Benedikt Adelmann, Wolfgang Menzel, Melanie Andresen, and Heike Zinsmeister. 2018. Evaluation of Out-of-Domain Dependency Parsing for its Application in a Digital Humanities Project. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 121–135.

Peter Eisenberg. 2020. *Grundriss der deutschen Grammatik: Der Satz*. J.B. Metzler, Stuttgart.

Kilian A. Foth. 2006. *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. University of Hamburg.

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

Holger Hopp. 2013. Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research*, 29(1):33–56.

Holger Hopp. 2018. Kasus- und Genusverarbeitung des Deutschen im Satzkontext. In Sarah Schimke and Holger Hopp, editors, *Sprachverarbeitung im Zweitspracherwerb*, pages 141–168. De Gruyter.

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. Place: Santa Fe, USA.

Christine Köhn, Tobias Staron, and Arne Köhn. 2016. Parsing Free-Form Language Learner Data: Current State and Error Analysis. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 135–145.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating Off-the-Shelf NLP Tools for German. In *Proceedings of the Conference on Natural Language Processing (KONVENS 2019)*, pages 212–222, Erlangen, Germany.

Niels Ott and Ramon Ziai. 2010. Evaluating Dependency Parsing Performance on German Learner Language. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, volume 9, pages 175–186. NEALT Tartu.

T. Ruberg. 2013. Problembereiche im kindlichen Zweitspracherwerb. *Sprache · Stimme · Gehör*, 37(04):181–185.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Patti Spinner and Alan Juffs. 2008. L2 grammatical gender in a complex morphological system: The case of German. *IRAL - International Review of Applied Linguistics in Language Teaching*, 46(4):315–348.

Hanna Studinger. 2010. *Kasusfehler in Nominalphrasen von Lernern des Deutschen als Fremdsprache*. Magisterarbeit, Humboldt-Universität zu Berlin.

Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Lauriala, and Daniela Piipponen. 2022. Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts. In *Selected papers from the CLARIN Annual Conference 2021*, pages 151–167.

Heide Wegener. 1995. *Die Nominalflexion des Deutschen - verstanden als Lerngegenstand:*. De Gruyter.

9. Language Resource References

Wisniewski, Katrin and Abel, Andrea and Vodičková, Kateřina and Plassmann, Sybille and Meurers, Detmar and Woldt, Claudia and Schöne, Karin and Blaschitz, Verena and Lyding, Verena and Nicolas, Lionel and Vettori, Chiara and Pečený, Pavel and Hana, Jirka and Čurdová, Veronika and Štindlová, Barbora and Klein, Gudrun and Lauppe, Louise and Boyd, Adriane and Bykh, Serhiy and Krivanek, Julia. 2018. *MERLIN Written Learner Corpus for Czech, German, Italian 1.1*. [link].