Sven Robert Storø

Annotation of modal particles in the GeWiss corpus

A syntactic and semantic-pragmatic analysis of the PTKMA annotation

1. Introduction

The development of spoken language corpora, i. e. the resources available therein and their usability both for research and for different user groups as well as user scenarios, is still in its early stages (Fandrych 2017). On the one hand, this is due to the fact that, compared to written language corpora, larger corpora on the spoken language of science have only been available for a short time. On the other hand, processing spoken language corpora with tools originally developed for written language corpora often produces unsatisfactory results and leads to a high error rate, which requires more extensive manual post-processing (Fandrych/Meißner/Wallner 2017). This includes, at least, the annotation of word types (POS tagging), which are relevant insofar as they enable word type- and lemma-specific access to the data and thus enable better and more targeted corpus queries.

This paper focuses on the automatic annotation of eight German modal particles (MPs) in two sub-corpora of the GeWiss corpus, which stem from examinations of L1 and L2 candidates in a German academic context. The MPs *ja*, *eben*, *halt*, *einfach*, *aber*, *mal*, *doch* and *denn* are manually checked for correctness using lists of criteria due to the low reliability of the automatic procedures for POS tagging in relation to spoken language data. In addition, the linguistic units *ja*, *eben*, *halt*, *einfach*, *aber*, *mal*, *doch* and *denn* that are not labelled as modal particles, but automatically receive another POS tag, are also examined for incorrect annotations and MP properties, and corresponding uses of these are annotated as modal particles. The POS tag system is subsequently evaluated with the current corpus software with regard to MP annotation in these dialogic, spontaneous language data.

The results of this corpus analysis of eight German modal particles (MPs) (*ja*, *eben*, *halt*, *einfach*, *aber*, *mal*, *doch* and *denn*) confirm the hypothesis that the currently available corpus software for automatic POS tagging in spoken language data is not reliable and should be further developed.

2. The modal particles

Based on previous pertinent research, the first section of this paper is mainly dedicated to MPs and their characteristic properties. These characteristics prove to be useful distinguishing criteria to differentiate MPs from their respective counterparts and thus serve as a basis for annotation. Common characteristic properties of MPs are their brevity, the fact that they are non-inflecting, that they cannot (normally) be stressed, their affinity to certain types of sentences, that they cannot be involved in word-formation processes, their sentence-integrated middle field position, as well as that they cannot be asked for and their wide scope (reference to the whole sentence) (Thurmair 1989; Hentschel/Weydt 2013; Müller 2014; Duden 2016).

3. Methodological considerations for the annotation of modal particles

The second section discusses methodological questions related to the annotation of MPs. Bochniak/Gräfe/Iliash (2017) already point out how the properties of MPs can be used as a tool when annotating MPs in spoken language corpora, but that some criteria should nonetheless be applied with reservation. Using specific examples, the section shows how some properties of the MPs are to be applied with reservations. It also presents the guidelines of Westpfahl et al. (2017) for annotation of POS tags for transcripts of spoken language using the example of *mal*; these guidelines provide support and determine rules in the delimitation and definition of a wide variety of word type classes.

4. The study

The third section is dedicated to the study. Here, we focus on the methodological approach of this research and then present the analysis and results. The data consists of German oral examinations with L1 and L2 speakers (DEU_L1/L2_PG) in the GeWiss corpus, taken from the database for spoken German (DGD). Of the 3834 tokens to be qualitatively examined in this study, 1559 tokens have the PTKMA tag and 2275 have another POS tag (non-PTKMA).

The results of the analysis show that the automatic POS tagging system has an error rate of 19.2% for the eight MPs *ja*, *eben*, *halt*, *einfach*, *aber*, *mal*, *doch* and *denn* in the GeWiss sub-corpus DEU_L1/L2_PG for the candidates in the DGD corpus. Figure 1 shows that the error rate ranges from 100% wrong (*aber*) to 100% correct (*einfach*). The POS tagging system seems to have considerable difficulties with the types *aber* (100%), *denn* (67.44%) and *mal* (50.72%), but the precision of the automatic POS tagging system is also far above the acceptable error rate for the types *doch* (20%) and *ja* (19.64%). By contrast, the POS tagging system seems to produce quite reliable results for the types *eben* (4.86%) and *halt* (1.09%) and error-free annotations for *einfach*.

Linguistic unit	**Automatically annotated PTKMA token**							
	einfach	*halt*	*eben*	*ja*	*doch*	*mal*	*denn*	*aber*
Error rate in percent	0%	1.09%	4.86%	19.64%	20%	50.72%	67.44%	100%
Error rate in tokens	0	2	16	119	10	35	29	175
PTKMA tokens total	175	185	330	609	60	70	43	87

Fig. 1: Annotated MPs of L1- und L2 candidates in the GeWiss sub-corpus DEU_L1/L2_PG per PTKMA token and share of erroneous tokens.

The results further show that of the 2275 non-PTKMA tokens *ja*, *eben*, *halt*, *einfach*, *aber*, *mal*, *doch* and *denn*, 30 tokens can be classed as MPs. Figure 2 shows that by manually post-processing the non-PTKMA types, 1 occurrence in 16 tokens (6.25%) had MP properties for *denn*, 5 occurrences in 13 tokens (38.46%) had MP properties for *eben*, 10 occurrences in 45 tokens (22.22%) had MP properties for *mal*, and 14 occurrences in 1546 tokens (0.91%) had MP properties for *ja*. In particular *ja* and *mal* exhibited a higher number of tokens with MP properties.

Linguistic unit	**The additional types with MP properties**			
	denn	*eben*	*mal*	*ja*
Token MP property	1	5	10	14
Error rate in percent	6.25%	38.46%	22.22%	0.91%
Non-PTKMA tokens total	16	13	45	1546

Fig. 2: The additional MP types in manual post-processing of the types *ja*, *mal*, *halt*, *einfach*, *aber*, *mal*, *doch* and *denn* for MP properties.

5. Conclusion and perspective

The manual post-processing of the data gives the impression that a further development of current corpus software for automatic annotation of MPs is required. In addition, the criteria also require further development and/or (more precise) adaptation with regard to the recognisability of MPs in spoken language corpora. Assuming that the high error rate in automatic annotation is caused by the pronunciation-oriented form of transcription and the syntactic-structural properties of spoken language (e. g. the problem of segmentability of spoken language with regard to syntactic units) (Fandrych 2017), the methods used would have to be further developed in this regard, because without the syntactic information, it is most likely problematic, if not impossible, for the corpus software to obtain information about the position of where a word is located (beginning, middle or end) in an utterance. The syntactic classification is of crucial importance for word type assignment, in particular for MPs, but also for other spoken language phenomena. The manual post-processing revealed that *ja*, *denn* and *aber*, for example, were found in syntactic positions other than in the middle field and were still annotated as PTKMA. But even MP types that were in the middle field and have "counterparts" in this syntactic position, such as *aber*, *denn* and *doch*, still proved to be problematic for the automatic annotation of PTKMA in the spoken language data. This seems to indicate that the MP types located in the middle field are also not easy to annotate as modal particles. In this respect, it would be interesting to find out whether the corpus software for automatic annotation has an unsuitable software setting with regard to the recognisability of MP types in the middle field (e. g. with regard to homonymous intertwining) or whether the incorrect annotation of MPs in the middle field is caused by a lack of further development. This should also be investigated in more depth. In order to enable a more precise retrieval of tokens such as modal particles, without significant error rates and need for correction in spoken language data, a further development of automatic annotation procedures that take into account the pronunciation-oriented form of transcription and the syntactic-structural properties of spoken language seems to be essential for certain user groups as well as user scenarios. This would require intensive discussions among linguists, corpus and software developers.

Reference

Bochniak, Kornelia/Gräfe, Karen/Iliash, Anna (2017): Zur Annotation von Modal-, Intensitäts- und Fokus-/Gradpartikeln im GeWiss-Korpus. In: Fandrych/Meißner/Wallner (eds.), pp. 79–106.

Duden (2016): Die Grammatik. Unentbehrlich für richtiges Deutsch. 9., completely revised and updated ed. (= Der Duden in zwölf Bänden, Vol. 4). Berlin: Dudenverlag.

Fandrych, Christian (2017): Gesprochene-Sprache-Forschung und Korpuserschließung am Beispiel von *GeWiss digital*. In: Fandrych/Cordula/Wallner (eds.), pp. 13–30.

Fandrych, Christian/Meißner, Cordula/Wallner, Franziska (eds.) (2017): Gesprochene Wissenschaftssprache – digital: Verfahren zur Annotation und Analyse mündlicher Korpora. (= Deutsch als Fremd- und Zweitsprache 11). Tübingen: Stauffenburg.

Hentschel, Elke/Weydt, Harald (2013): Handbuch der deutschen Grammatik. Berlin/Boston: De Gruyter.

Müller, Sonja (2014): Modalpartikeln. (= Kurze Einführungen in die germanistische Linguistik 17). Heidelberg: Winter.

Thurmair, Maria (1989): Modalpartikeln und ihre Kombinationen. Tübingen: De Gruyter.

Westpfahl, Swantje/Schmidt, Thomas/Jonietz, Jasmin/Borlinghaus, Anton (2017): Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS).