KOJAK: A New Corpus for Studying German Discourse Particle *ja*

Adil Soubki[1†], Owen Rambow[2†], Chong Kang[2]
[1]Department of Computer Science, Stony Brook University
[2]Department of Linguistics, Stony Brook University
[†]Institute for Advanced Computational Science
{adil.soubki,owen.rambow,chong.kang}@stonybrook.edu

Abstract

In German, *ja* can be used as a discourse particle to indicate that a proposition, according to the speaker, is believed by both the speaker and audience. We use this observation to create KOJAK, a distantly-labeled English dataset derived from Europarl for studying when a speaker believes a statement to be common ground. This corpus is then analyzed to identify lexical choices in English that correspond with German *ja*. Finally, we perform experiments on the dataset to predict if an English clause corresponds to a German clause containing *ja* and achieve an F-measure of 75.3% on a balanced test corpus.

1 Introduction

Predicting an author's belief, also called Event Factuality Prediction (EFP), has been studied extensively in the last decade. However, in addition to modeling their own beliefs, discourse participants develop a model of their audience's beliefs as well. It is well known since at least Grice (1975) that a speaker or writer must be modeling the addressee's cognitive state in order to communicate felicitously, and the notion of "common ground" has received increased attention in cognitive science (Brennan and Clark, 1996; Brennan et al., 2010) and philosophy (Stalnaker, 2002).

The task of predicting if a speaker believes a listener already knows a proposition, either because it has been established as common ground or connected to some shared reality, has been studied markedly less. This is at least in part due to a lack of corpora annotated for the task.

This paper makes the following contributions:

1. We develop a distantly-labeled dataset in English for studying when a speaker believes their audience already believes what they are saying.
2. We perform a statistical analysis to identify which tokens in English correspond to German discourse particle *ja*.
3. We perform experiments to predict for an English sentence whether its German translation equivalent has a *ja* or not. On a balanced test corpus, we achieve an F-measure of 75.3% on the *ja* sentences.

The paper is structured as follows. We start out by describing German discourse particles and *ja* in particular (Section 2). We then present a detailed discussion of the procedure used to create KOJAK (Section 3) as well as some statistical analysis of the corpus (Section 4). Finally, we use KOJAK to train models for predicting if an English sentence corresponds to a German sentence containing *ja* (Section 5) and conclude with a discussion of results (Section 6) and future work (Section 7).

2 German Discourse Particles and the Common Ground in Discourse

German has a closed class of discourse particles, including *ja*, *doch*, *wohl*, and *etwa*. These discourse particles have cognates in other parts of speech; for example, *ja* is also the equivalent of English *yes*, occurring sentence-initially. We can distinguish discourse particles from homonyms by their syntax: Egg (2011) points out that they cannot be questioned, they cannot function as one-word answer to a question, they cannot be coordinated or modified, and they occur in the so-called "Mittelfeld" (between the finite verb and any non-finite verbal elements). In terms of their function, Abraham (2017) suggests that "the speaker uses modal particles to negotiate the truth value of a proposition with the addressee. (...) The speaker distinguishes between the source of evidence and the judger of the source of evidence in the sense of Theory of Mind" (our translation). Put differently, modal particles serve to indicate that the speaker is distinguishing between her cognitive state and her "theory" of the addressee's cognitive state.

Proceedings of 3rd Workshop on Computational Approaches to Discourse, pages 1–6
October 16 – 17, 2022.

Turning specifically to *ja*, Gast (2008) characterizes its function as follows: "What is indicated by *ja* is that the state of affairs reported is unquestionable." Egg (2011) provides a sharper characterization: "The particle *ja* expresses (roughly) that the information conveyed by the sentence is already part of the common ground". Thus, the reason that the state of affairs is unquestionable is that the hearer already believes it (or has in the past). This is echoed by Döring and Repp (2016): "*ja* is generally taken to indicate (roughly), that the speaker assumes that the proposition *ja* scopes over is already part of the common ground, i.e. that it is not new (see many of the references above)".

We see that German discourse particle *ja* interacts crucially with how the discourse participants manage the common ground. However, most languages do not have such discourse particles. For example, English speakers may turn to phrases like *you know* or *of course* in the absence of such a particle. This paper suggests that by looking at translation equivalents of German *ja*, we can learn about how other languages manage the common ground.

3 Corpus Creation

Our corpus is called KOJAK, which stands for "Korpus für *ja* in Kontext" (or "Corpus for *ja* in context" in English). It contains roughly 3,000 examples of English sentences corresponding to German *ja*.

We use the Corrected and Structured Europarl Corpus (CoStEP), released by Graën et al. (2014), as a base for constructing KOJAK. Initially created for machine translation tasks, the Europarl corpus contains roughly 30 million words parallel-translated to 11 languages including English and German (Koehn, 2005). They are sourced from proceedings of the European Parliament starting as early as 1996 and contain additional languages as time moves on.

There is a notable asymmetry in the realization of *ja* depending on the direction of translation. If the sentence containing *ja* was translated from German to some other language then we can be sure that the conception of common ground being expressed is that of the original speaker. However, if the German *ja* sentence was translated from some other language then the *ja* may be expressing the translator's belief regarding the speaker's belief of what the common ground is. We ignore the distinction in this paper but discuss possible improvements in Section 7.

	Train		Dev		Test	
Ver.	Nat.	Bal.	Nat.	Bal.	Nat.	Bal.
Ja	2,021	2,012	286	293	591	593
Na	370,052	2,045	52,867	286	105,716	567
Total	372,073	4,057	53,153	579	106,307	1,160

Table 1: Summary of the clause-extracted dataset.

Our corpus is compiled in two steps. First, we create a filter for identifying sentences containing uses of *ja* as a discourse particle. We then use a heuristic for extracting only the clause containing the *ja* in question (i.e., the clause over which *ja* scopes). The latter step is motivated by an interest in the proposition which *ja* is modifying.

3.1 Filtering

CoStEP data is provided in an xml format with untokenized text for each speaker's turn. We used SpaCy to segment sentences from turns in both English and German (Honnibal and Johnson, 2015). To ensure the segmentation lines up, turns where the number of sentences does not match are discarded. We then filter the remaining sentence pairs by searching for ones where,

1. The German text contains *ja*.
2. The *ja* is not sentence-initial.
3. The English text does not contain *yes*.

If these three checks are successful then the sentence is considered to contain a use of *ja* as a discourse particle. Conversely, the sentence is not considered to contain a use of *ja* as a discourse particle if any of these checks fail. This creates two categories of sentences – JA sentences where this filter succeeded and, affectionately called, NA sentences for everything else.

3.2 Clause Extraction

Since the data comes from parliament meetings, sentences can be long with many nested clauses. When this is the case, the task of predicting what the speaker believes is muddled since the proposition we wish to predict is unclear. To address this we develop a heuristic for extracting the clause *ja* is modifying. SpaCy is again used to tokenize and parse the sentences along with SimAlign from Jalili Sabet et al. (2020) to align the German and English. The end result is a dependency parse for both the English and German sentences along with a mapping from one to the other.

2-grams	3-grams	4-grams
after all	after all ,	, of course ,
of course	, of course	, after all ,
, on	of course ,	is , after all
of the	, after all	is , of course
it is	of the european	, in fact ,
all ,	the committee on	of the european union
, but	it is not	at the same time
course ,	, it is	in the european union
, after	have voted for	-
the european	is , after	-

Table 2: The top 10 n-grams from the train/dev splits.

	1-grams		2-grams		3-grams		4-gr
Num.	>100		>100		73		7
Cutoff	20	100	20	100	20	73	7
Good	20%	8%	35%	16%	30%	12%	63%
Ntrl	75%	67%	65%	74%	45%	60%	13%
Bad	5%	25%	0%	10%	25%	27%	25%

Table 3: Analysis of top-100 and top-20 (where applicable) n-grams by significance for detecting *ja*-sentences; Num = number of such n-grams, Ntrl = Neutral

To extract the corresponding English clause we find the location of *ja* in the German sentence and then travel up the parse tree until a VERB or AUX tag is found or it reaches the root. If the head is a clausal object, we probably have a problem with the parse. This is because German *ja* typically does not appear in embedded (object) clauses. We have found that we get better results if we move up one more level to the matrix clause in case we find ourselves in a subordinate clause. (Note that we do not do this with other types of embedded clauses, such as relative clauses or parentheticals.) The subtree rooted at this node is the candidate clause in German that now must be extracted from English.

It is possible that after alignment there are multiple tokens in English that correspond to the head word in German. For each English token corresponding to the German head word, if it is a leaf and tagged with AUX we move up one level and take the subtree. This process results in a set of, often overlapping, English subtrees. The leaves of these subtrees are then naively arranged in order to yield the final English clause.

A similar process is repeated for NA sentences to make them comparable and avoid sentence length being a strong indicator for the model. Instead of starting at the *ja*, a random token is selected and the algorithm described above is applied. This yields a dataset of English clauses labeled JA if the corresponding German clause contains *ja*, and labeled NA if not.

4 Statistical Analysis

With the relevant sentences now separated, our attention turns to how we can identify what items in English relate to the discourse particle usage of German *ja*. If the lexical choices between JA and NA sentences are different, then we would expect certain sequences to appear significantly more frequently in the JA corpus than elsewhere.

For any token sequence of length n, we can count the number of times that n-gram appears in the JA and NA sentences respectively. This is similar to creating two sets of samples and asking whether they are likely to be from different populations. Intuitively, n-grams which are unique to JA sentences are probably related to the presence of the discourse particle.

To investigate this we compute the counts of every n-gram for $2 \leq n \leq 4$ in each population. These samples are used to perform a t-test at the 95% confidence level and then sorted by p-value. To ensure sequences which are reasonably robust, we discard any that did not appear more than 10 times in the data. The result, seen in Table 2 for the train and dev splits, is a list of n-grams most unique to JA sentences according to this metric.

To roughly evaluate the quality of the extracted n-grams, the second author, a native speaker of both German and English, performed an error analysis on the n-gram lists. We used the following categories:

- **Good**: This is clearly an n-gram that on its own or in conjunction with some predictable missing words carries the same pragmatic meaning as German *ja*.
- **Neutral**: These n-grams contain no evidence of being either **Good** or **Bad**. Typically, these are sequences of function words without content words.
- **Bad**: This is clearly an n-gram which does not carry the same pragmatic meaning as German *ja*. A typical example is *of the European Union*. In fact, almost all examples refer explicitly to the European Union or its political procedures including those of the parliament (e.g., *Madam President*).

The rating was performed only on the n-gram types

rather than on occurrences of the n-grams, and no further context was provided. The goal is to provide a sense of the quality of the extracted n-grams, and we acknowledge the limitations of this study. The error analysis we present in Section 6, in contrast, was based on actual full phrases.

Results are shown in Table 3. The percentage of **Good** n-grams is much higher among the top-20 n-grams as opposed to top-100 (or top-73 in the case of 3-grams). Similarly, the percentage of **Bad** n-grams is lower among the top-20 compared to the top-100 (top-73). These two observations support the claim that the ranking by p-value is meaningful. The **Neutral** n-grams among the top-20 decrease with increasing n, which makes sense as shorter token sequences are more likely to be impossible to judge. Correspondingly, the percentage of **Good** n-grams (both top-20 and top-100 for 1- and 2-grams) increases from $n = 1$ to $n = 2$, though $n = 3$ does not continue the trend. For **Bad** n-grams, we first see a decrease with n and then an increase again, as longer token sequences are more likely to contain content words. For the *Bad* category, we find basically the same examples at all n-gram levels.

Overall, our simple statistical approach has extracted good n-grams, with a small number of bad ones. The results support the claim that the discourse meaning of German discourse particle *ja* is often preserved in translation equivalents.

5 Modeling & Experiments

We perform machine learning experiments to predict whether an English clause is the equivalent of a German *ja* clause or not.

5.1 Transformer-Based Model

We start by preparing a balanced version of the dataset such that JA and NA sentences appear equally often and use this as input to a transformer model. The model is fine-tuned on top of multilingual BERT for text classification using the transformers library from Hugging Face (Wolf et al., 2020). Training is performed for three epochs with a learning rate of 2e-5.

The results are promising with the model achieving an F-measure of 75.3% on JA clauses. Though it is difficult to determine exactly what features the model is using, this result is much better than would be expected if the clauses were randomly selected.

	Transformer		**Statistical**	
Strategy	Nat.	Bal.	Nat.	Bal.
Precision	50.0	76.7	7.5	48.7
Recall	0.7	74.1	5.2	56.2
F-measure	1.4	75.3	6.2	52.1

Table 4: Model performance achieved on *ja* examples.

In reality, *ja* events occur much less frequently than half of the time. As can be seen from Table 1, JA sentences are a tiny minority class, appearing in only ~0.5% of sentences. To emulate this, we also examine the performance of multilingual BERT on a dataset which contains a "natural" proportion of JA clauses. On this highly imbalanced dataset, the model achieves an F-measure of 1.4% on JA clauses. In other words, it performs extremely poorly.

5.2 Statistical Model

The results on the natural proportion were so low it seemed like a more simplistic model based on the analysis in Section 4 could possibly outperform multilingual BERT. We investigate this by performing the same t-test using only the training and dev splits to get a ranked list of 2-grams, 3-grams, and 4-grams (See Table 2). The model then selects some number of the top n-grams from each list and naively classifies a clause as JA if it case-insensitively contains any of those phrases.

Using only the top ranked sequence from each list, this simple model outperforms BERT, achieving an F-measure of 6.2% for JA clauses on the natural proportion test set. Use of additional n-grams did not improve performance on the natural proportion dataset. However, on the balanced dataset including every 2-gram, 3-gram, and 4-gram achieved the best results with an F-measure of 52.1% on JA clauses. While significantly worse than BERT, this is again a large improvement from the imbalanced dataset.

6 Discussion

We also investigated the use of sentence-level data, i.e. a version of KOJAK generated without extracting only the clauses over which *ja* scopes, but instead using the whole sentences in which *ja* occurs. When including this additional context both models performed worse, which supports the intuition for including only the clauses in scope.

The use of *ja* is one way German speakers indicate they believe a proposition is already common ground, but it is not the only way. It is possible that the systems above are correctly identifying sentences in which this occurs but they correspond to a German sentence which does not contain *ja*. We analyzed 70 false positive errors of the statistical model, and found that 71% could plausibly be cases in which the speaker believes the hearer already believes the content of the clause, despite the absence of *ja* in the German clause. Interestingly, another 11% look like cases in which the speaker is pretending as if the audience shares his or her beliefs, even though they probably do not (*we both know you will clean the dishes now*).

7 Future Work

In the relatively near future we hope to make improvements to KOJAK. While the corpus can currently only be used to study English, its underlying source provides data in many more languages. Using a methodology similar to that which was described in Section 4, we hope to expand KOJAK to support every language offered by Europarl. On a similar note, CoStEP also includes information about the original language for each utterance. If this were propagated, we could investigate the issue of translation direction mentioned in Section 3 more closely by partitioning data along these lines.

While the n-gram analysis discussed in Section 4 roughly identifies sequences which correspond to *ja*, many artifacts (E.g. *of the european union*) persist in the output. One way to reduce these might be to perform a similar analysis but on the German text and discard sequences that correspond directly to the English n-gram list.

These enhancements open up several directions for continuing work, the most conspicuous of which might be investigating the effectiveness of multitask learning, in which we exploit multiple languages, or related tasks such as factuality (Saurí and Pustejovsky, 2009). It could also be interesting to use the n-grams identified in English and search for their German counterparts, which likely include more than just *ja*. We have only just scratched the surface of what is possible here.

8 Access to KOJAK

The natural and balanced preparations of KOJAK are made available on GitHub.[1] Additional tooling used for parsing and filtering CoStEP, which might be useful in its own right, is also available.[2]

9 Acknowledgements

We would like to thank our three anonymous reviewers for their insightful comments and suggestions. This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract HR001122C0034. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Werner Abraham. 2017. Modalpartikel und Mirativeffekte. In Shin Tanaka, Elisabeth Leiss, Werner Abraham, and Yasuhiro Fujinawa, editors, *Grammatische Funktionen aus Sicht der japanischen und deutschen Germanistik*, Linguistische Berichte Sonderheft 24, pages 75–107. Buske, Hamburg.

Susan Brennan, Alexia Galati, and Anna Kuhlen. 2010. *Chapter 8 - Two Minds, One Dialog: Coordinating Speaking and Understanding*, volume 53, pages 301–344.

Susan E. Brennan and Herbert H. Clark. 1996. Lexical choice and conceptual pacts in conversation. *Journal of Experimental Psychology: Learning, Memory And Cognition*, pages 1482–93.

Sophia Döring and Sophie Repp. 2016. The modal particles *ja* and *doch* and their interaction with discourse structure: Corpus and experimental evidence. In S. Featherston, R. Hörnig, S. von Wietersheim, and S. Winkler, editors, *Information Structure and Semantic Processing*. De Gruyter.

Markus Egg. 2011. Discourse particles between cohesion and coherence. In *Proceedings of the Workshop on Constraints in Discourse*, Agay, France.

Volker Gast. 2008. Modal particles and context updating: The functions of german 'ja', 'doch', 'wohl' and 'etwa'. In H. Vater and O. Letnes, editors, *Modalverben und Grammatikalisierung*, pages 153–177. Wissenschaftlicher Verlag.

Johannes Graën, Dolores Batinic, and Martin Volk. 2014. Cleaning the europarl corpus for linguistic applications. In *KONVENS*.

Herbert Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and semantics, vol 3*. Academic Press, New York.

[1] https://github.com/cogstates/kojak

[2] https://github.com/cogstates/costep

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268. 10.1007/s10579-009-9089-9.

Robert C. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5-6):701–721.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.