Erschienen in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) / Calzolari, Nicoletta et al. (Hrsg.). - Paris : ELRA, 2018. - S. 2715-2720

A Multilingual Approach to Question Classification

Aikaterini-Lida Kalouli, Katharina Kaiser, Annette Hautli-Janisz, Georg A. Kaiser, Miriam Butt
Department of Linguistics, University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

In this paper we present the Konstanz Resource of Questions (KRoQ), the first dependency-parsed, parallel multilingual corpus of information-seeking and non-information-seeking questions. In creating the corpus, we employ a linguistically motivated rule-based system that uses linguistic cues from one language to help classify and annotate questions across other languages. Our current corpus includes German, French, Spanish and Koine Greek. Based on the linguistically motivated heuristics we identify, a two-step scoring mechanism assigns intra- and inter-language scores to each question. Based on these scores, each question is classified as being either information seeking or non-information seeking. An evaluation shows that this mechanism correctly classifies questions in 79% of the cases. We release our corpus as a basis for further work in the area of question classification. It can be utilized as training and testing data for machine-learning algorithms, as corpus-data for theoretical linguistic questions or as a resource for further rule-based approaches to question identification.

Keywords: (non-)information-seeking questions, parallel multilingual corpora, question classification, Bible

1. Introduction

A central phenomenon in natural language as well as human-computer interaction is that of questions. Although this is a central phenomenon, it has been understudied in computational linguistics. Most of the existing work has concentrated on dealing with 'factoid' questions such as *When was Alan Turing at Bletchley Park?* This research has mostly been driven by the goal of building Question-Answering (QA) systems and finding intelligent, quick and reliable ways of matching a query to terms found in a given text collection, e.g., see Wang and Chua (2010). Comparatively less research has focused on understanding the structure of questions per se or on distinguishing different types of questions, i.e. information-seeking vs. rhetorical or discourse-structuring questions, among several other types. While a few approaches explicitly focus on non-information-seeking questions (Harper et al., 2009; Paul et al., 2011; Li et al., 2011), this work is either based on big data or on information gained from crowdsourcing. It does not tend to take recent theoretical linguistic work on questions into account, as has also been observed within the CReST project (Kübler et al., 2012), in which the existing PennTreebank (Marcus et al., 1993) annotation scheme was amended for yes-no and back-channeling questions.

This paper makes use of theoretical linguistic insights for automatically classifying questions as information-seeking or non-information-seeking. We devise a rule-based system with a heuristic scoring methodology that uses linguistic indicators to classify questions into information-seeking (ISQ) or non-information seeking questions (NISQs) across four different languages: German, French, Spanish and Koine Greek.[1] One result of our work has been the creation of a new resource: KRoQ (Konstanz Resource of Questions), a first dependency-parsed, parallel multilingual corpus of ISQs vs. NISQs.

The work described in this paper makes the following contributions to the field: For one, we provide a novel parsed and annotated corpus for question classification that can be used as a first reliable and improvable source for further theoretical and computational linguistic research. For another, we assign scores to determine how likely a question is information-seeking or not – these scores are also incorporated into the resource to make the classification more transparent and to be used for further research. Finally, our multilingual, rule-based technique can be applied as-is on further, multilingual data for question classification and can also be improved and adapted depending on the task.

Section 2. provides an overview of related work. In Section 3. we describe the corpus data and the linguistic indicators we made use of. Section 4. provides a description and evaluation of our multilingual system. Section 5. briefly describes the corpus we make available and in the final section we discuss our findings and our goals for future work.

2. Background

Everyday conversation frequently contains questions – in a randomly sampled 2-million tweets corpus compiled by Efron and Winget (2010), 13% of phrases are questions. But questions are far from being a homogeneous group. One type of question is posed to elicit information and get an answer — these are canonical, information-seeking questions (ISQs). Questions where the speaker does not expect an answer but instead triggers a certain type of speech act (Dayal, 2016) are treated as non-canonical, non-information-seeking questions (NISQs). The latter type is in itself a heterogenous class, including various subtypes.

2.1. Theoretical Linguistic Viewpoint

Perhaps the most well-known and well-recognized type of NISQ is a rhetorical question. This has the syntactic structure of a canonical question, but the pragmatic value of a declarative (Sadock, 1971; Han, 2002) and is often used to make a sarcastic comment or statement (*Have you ever even touched a computer?*). Another type, echo questions, are

[1] Dialect of Greek, also known as Alexandrian dialect, common Attic, Hellenistic or Biblical Greek; spoken and written during the Hellenistic and Roman Antiquity and the early Byzantine era.

used when the listener does not hear or understand properly what is being said, or when the listener wants to express incredulity or surprise (*She said what?*). Other types of NISQs are deliberative questions (*When shall we three meet again?*) (Wheatley, 1955) or self-addressed questions (*Where have I left my keys?*) (Ginzburg et al., 2013). A further well-known type are the ability/inclination questions, which are used as directives, requests or orders (*Can you pass the salt?*) (Dayal, 2016). Suggestive questions are used to imply that a certain answer should be given in response (*Don't you think that eating chips is unhealthy?*). In contrast, tag questions (*He is not coming, is he?*) are used when the speaker asserts something while also seeking confirmation for the assertion (Cattell, 1973). Looking at natural language corpora we also detect less-researched cases of NISQs, such as quoted questions (*She always asks "When will we meet?"*) or discourse-structuring questions (*What have we learned from this? We have learned that we need a better education system*). For the current work we do not employ a finer-grained distinction between these types but treat all of them as NISQ. Initial experimentation with a more detailed annotation scheme has shown that finer-grained subclassifications are difficult to be achieved consistently given the current state of our understanding of question types. We thus leave a finer-grained classification for further research.

2.2. Computational Approaches

In computational linguistics, one body of work uses social media data to classify ISQs and NISQs, training models on a limited set of manually annotated data (Harper et al., 2009; Li et al., 2011; Zhao and Mei, 2013; Ranganath et al., 2016). Paul et al. (2011) use crowdsourcing techniques to collect human classifications for a large amount of Twitter questions. While social media data has its own set of problems (e.g., length of the turn, ungrammaticality of sentences, spelling mistakes), the data is enriched with information like usernames, hashtags and urls, which helps in identifying the type of the question. Bhattasali et al. (2015) develop a machine-learning mechanism to identify rhetorical questions in the Switchboard Dialogue Act Corpus; Zymla (2014) uses a rule-based approach to heuristically identify rhetorical questions in German Twitter data.

The challenges for this type of work are manifold. First, distinguishing ISQs from NISQs based on syntactic properties is difficult because they are mostly structurally indistinguishable. Instead, context and intonation play a much bigger role (Bhatt, 1998; Zymla, 2014). Secondly, only some languages have special lexical markers that might indicate the type of question, e.g. German tends to use discourse particles in NISQs (Maibauer, 1986). Thirdly, expressions such as *give a damn*, *lift a finger* or *even*, which have been identified as generally conveying NISQs (Bhatt, 1998), are not frequent enough in real texts for computational purposes.

From a data perspective, it is not trivial to find suitable resources for looking into questions. Whereas real data in large quantities, e.g., Twitter, contains many questions of both types (Wang and Chua, 2010), the context in which they are found is either limited or lacking altogether and thus it is hard even for humans to decide on the two categories. On the other hand, corpora with well-edited text such as newspapers, books and speeches are generally less suitable since questions, in particular NISQs, tend to appear more often in spontaneous, unedited communication.

In order to overcome some of the challenges above, we developed our own rule-based system. This system leverages linguistic cues in one language for the scoring and classification of questions in other languages. This multilingual approach helps us in the classification because even if there are no indicators for the type of the question in one language, it is probable that there will be some in the other languages. The multilingual approach allows the pooling of information from multiple sources: the language of the question itself and the other three. The motivation to use such a multilingual approach goes back to Gale et al. (Gale et al., 1992), who used parallel corpora for word-sense-disambiguation (WSD). The logic behind such a technique is simple: get the things you cannot get from the current language from another language. In the WSD field this means that a polysemous word in one language can be looked up in parallel corpora and its translation in the other languages conveys the correct sense of the word in the current language. We developed our approach along similar lines: we identify linguistic cues in each of the languages and then use those cues to help question classification of the translations in the other languages.

3. The System

3.1. Data

Collection The data underlying our system is a parallel Bible corpus[2] in four languages, namely German, French, Spanish and Koine Greek. The choice of the four languages is deliberate. All four languages allow for the use of specific linguistic markers to indicate NISQs (Maibauer, 1986; Escandell, 1999; Bonifazi et al., 2016). Moreover, Koine Greek is the original language of the New Testament and the first language into which the Old Testament was translated and thus the language of the primary biblical text from which the Bible was translated into Latin (Tov, 2011). This means that it is the most suitable to be used as the prototypical version, which is crucial for our implementation.

We also deliberately decided on the Bible as a corpus. It is available in many languages, is inherently aligned, is for

[2]The Bible was crawled from online resources: German (translation of 1980, 73 books, `https://www.die-bibel.de/bibeln/online-bibeln/einheitsuebersetzung/bibeltext/`, Einheitsübersetzung der Heiligen Schrift © 1980 Katholische Bibelanstalt GmbH, Stuttgart), French (translation of 1997, 71 books, Textes bibliques tirés de la Bible en français courant © Société biblique française – Bibli'O, 1997 Avec autorisation. La responsabilité de la Société biblique française – Bibli'O est engagée uniquement sur les textes bibliques cités dans cet ouvrage. `http://lire.la-bible.net`), Spanish (translation of 1995, 66 books, `https://www.unitedbiblesocieties.org`, © 2018 United Bible Societies), Koine Greek Bible (Septuagint translation of the Old Testament of the 3rd century BCE and the Original New Testament, 77 books, `http://www.bibles.gr/`.

the most part written in prose and contains a large amount of narration and dialogues (Kaiser, 2015; de Vries, 2007).

Preprocessing The question extraction is rule-based in that we look for sentences ending with question marks of each of the four languages ('?' for French, German and Spanish and ';' for Greek). This provides us with 3,081 questions for German, 2,960 for French, 3,164 questions in Spanish and 3,300 in Koine Greek.[3] In addition, all texts except for the Koine Greek version are parsed with the Mate Parser (Bohnet and Nivre, 2012; Bohnet and Kuhn, 2012) and converted to the CoNLL-U format. We should note that the parsing is not used further in our approach; our approach solely uses the surface forms of the questions. Nevertheless, we wished to provide the corpus in a state-of-the-art, parsed format so that it is directly usable for further research.

3.2. Linguistic Indicators

Our system builds on language-specific markers that serve as possible indicators of ISQs vs. NISQs, based on ongoing theoretical linguistic work in the area. The core methodology of our system involves combining these insights with a multilingual corpus approach where markers in one language serve to classify questions across languages. It needs to be stressed that these cues do not represent absolute markers of ISQs vs. NISQs, but need to be taken as possible indicators of an ISQ vs. an NISQ. The higher the number of these indicators that can be found, particularly across languages, the higher the likelihood that the classification will be correct. As detailed below, we supplement information about the linguistic cues (originating in theoretical linguistic work) with heuristics we have developed ourselves based on observations of language structure in this and other corpora.

German Discourse particles are frequently found in German questions and in particular in NISQs (Maibauer, 1986; Zymla, 2014). These particles make subtle pragmatic contributions to an utterance and often convey a speaker's stance towards a proposition, situating that proposition in the web of information that comprises the discourse. We use the following as indicators for NISQs: *denn* 'lit. then', *schon* 'lit. already', *denn schon* 'lit. then already', *jemals* 'ever' and *niemals* 'never', following Maibauer (1986). The lexical item *ob* 'if' at the beginning of the question and its co-existence with *wohl* 'probably' as well as the tag element *oder?* 'right?' at the end of the question further serve as NISQ indicators.

French For French, we use *vraiment* 'really' and the tag phrase *n'est-ce pas?* 'isn't it?' at the end of the question as relevant NISQ indicators, in accordance with our own observations. We additionally use the presence of a negated predicate structure as indicative of NISQs (Maibauer, 1986; Sadock, 1971). Moreover, we make use of the fact that the French translation we used encodes direct speech with quotation marks. This allows us to track the dialogue turns with the assumption that a speaker who continues an utterance after posing a question is rather not seeking information, but is using the question to structure the dialog or to promote some other underlying meaning.

Spanish For Spanish we use *acaso* 'really' at the beginning of a question and *verdad?* 'true?' at the end of the question (Escandell, 1999). Both are considered good indicators for NISQs. The same goes for the tag element *no?* 'no?' at the end of a question. We also use the existence of negated predicate structures as NISQ indicators (Maibauer, 1986; Sadock, 1971).

Koine Greek For Koine Greek we also make use of particles, in particular some of those presented by Bonifazi et al. (2016): ἆρά γε 'maybe?', ποτε 'ever' and ἆρα 'maybe?'. We also take negated predicates as markers for NISQs (Maibauer, 1986; Sadock, 1971). In addition, we use our own observation for Koine Greek: If a question in Koine Greek is not translated as a question in the other languages, but as a declarative, we assume that this question is most probably a NISQ. In other words, if the translators chose to not translate it as a question, then the question is not asking for information but is rather accomplishing other communicative goals. This is also why it was important to choose Koine Greek as one of our languages — this assumption can only be made if we know the source language of the translation.

3.3. Scoring

The core methodology of the system is to use a scoring mechanism that indicates how strongly a question belongs to the group of ISQs or NISQs as no absolute markers are available. The higher the score of the question, the more likely it is to be an NISQ rather than an ISQ. The scoring is done in two steps, first an *individual scoring* where each question is analyzed individually for each language, and then a *cross-linguistic scoring* where we take into account the question and its translations.

Individual scoring We assign a score of either 1 or 2 to each of the linguistic indicators discussed in §3.2.. The score is based on the theoretical literature about how strongly each indicator correlates with being a NISQ. Table 1 provides an overview of the different indicators and their scores.

Heuristics	**Score**
German particles, *ob...wohl* and *oder?*	2
Spanish negated predicate	1
Spanish *no?*, *acaso* and *verdad?*	2
French *vraiment?* and negated predicate	1
French *n' est-ce pas?*	2
French dialogue turns	1
Greek particles	2
Greek negated predicate	1
Greek question not found in language X	2

Table 1: Scores of the heuristics used.

Then, we look for these indicators within each question and, if present, we add up their scores so that each question

[3] In Koine Greek 70 questions have no translation in any of the other languages (the questions of the books *1 Esdras*, *3 Maccabees*, *4 Maccabees* and *Sosana*) and thus only 3,230 were aligned and annotated.

is assigned an overall score (assuming that the initial score of each question is 0). This means that the more of the described patterns present in the question, the higher the question is scored. The following examples are meant to make the individual scoring clearer. If we have the question *Herr, mein Herr, was willst du mir schon geben?* 'Lord, my Lord, what do you *schon* ('really') want to give to me?', it is assigned a score of 2 because it contains the particle *schon* which has a score of 2. If the question would also include a further marker, e.g. *jemals* 'ever', then its score would have been increased by another 2 and the overall score of the question would be 4. Another example is the question *N'as-tu qu'une seule bénédiction?* 'Do you have only one blessing?'. The question gets an initial score of 1 because it starts with a negated predicate — the first part of the split expression *ne ... que'* ('only') — and another 1 because of the dialogue turns; if we look at the text following, we will see that the same person goes on speaking which we can tell because French conveniently marks direct speech with quotation marks as explained in section 3.2.. Thus, this question will get a total score of 2. This scoring has the benefit of showing tendencies – giving us something like a "quantified tendency" that a question belongs to one type or the other. The higher score means higher probability that the question will be NISQ rather than ISQ.

Cross-linguistic scoring After all questions of the four languages have been given individual scores, we assign the final score of each question (across languages) based on the individual scores of its translated instances. For that, we align the questions of the four languages based on the verse number of the Bible, e.g. the Spanish verse *Génesis 3:9* is mapped to the French *Génèse 3:9* and the question contained in each of them is mapped to each other. In the case that one verse contains more than one question, we map the complete verses because the corpus is not sentence-aligned. If one language does not have a question in a verse where the other languages feature a question, we only map the verses of those languages containing a question.

After the questions are aligned across languages, the individual scores of all aligned, translated instances of a question are added up to one final score. With this, every question — across the four languages — receives one overall, final score. The following example should make the cross-linguistic scoring clearer. The four aligned verses with the ID *Genesis 27:38* contain the question 'Did you have only one blessing, father?' in the four languages as shown in Table 2.

Language	Verse text	Score
German	Hattest du denn nur einen einzigen Segen, Vater?	2
French	N' as-tu qu' une seule bénédiction?	2
Spanish	¿No tienes más que una sola bendición, padre mío?	1
Greek	ἡ εὐλογία μία σοί ἐστι, πάτερ·	0
	Cross-linguistic Score	**5**

Table 2: Example of the cross-linguistic scoring.

The individual scoring described in the previous subsection assigns a score to each translated instance of the question, based on the existence of the predefined markers. After the four translated instances have been aligned across the languages, the individual scores of all four language instances are added up to one total, cross-linguistic score. With this the question *Genesis 27:38* gets the final overall score of 5.

3.4. Classification

The final step of classifying and annotating the questions as either ISQs or NISQs is solely based on the scoring; there is no training process involved. Different classifications are possible depending on what score is taken to be the threshold for the classification. In general, the lower the score, the more likely the question is an ISQ. The higher the score, the stronger the tendency that the question is an NISQ. For the current version of the KRoQ corpus we provide, we take 0 to be the threshold and thus classify all questions with scores equal to 0 as ISQ and questions with scores greater than 0 as NISQ because this proved the best setting, as shown in the next section.

4. Evaluation

Data In order to evaluate the appropriateness of our heuristics, we manually created a gold standard of the first 200 aligned questions of the Bible, annotating them as ISQ vs. NISQ. Questions that were only realized as questions in German or Spanish or French, but which were declaratives in all other three languages were considered translation anomalies and were left out from the evaluation. However, if the question was only found in Koine Greek as a question and as a declarative in the other languages, it was still considered for evaluation (see §3.2.). Questions of one language that did not exist in the other languages at all (neither as questions nor as declaratives) were also not included in the evaluation set. We also excluded direct questions containing other questions in direct speech, as in those cases we could not decide if we should annotate the main or the embedded question (e.g. *Warum lacht Sara und sagt : Soll ich wirklich noch Kinder bekommen, obwohl ich so alt bin?* 'Why does Sara laugh and say: am I really going to have children, although I am already so old?').

The 200 question instances of each language (800 in total) were each manually annotated by two expert annotators. We then took all eight manual annotations for each question across languages and went with the majority vote, yielding an inter-annotator agreement of 0.75%. We assume that the same question has the same status (ISQ or NISQ) across languages: the same parallel question has the same context and co-referents in all languages and thus the same status. A question was therefore classified as ISQ or NISQ across languages based on what most annotators had marked it as.

Results To evaluate the quality of our system, we compare the automatic scoring (and thus classification) with the manually-created gold standard. In order to test how the system performs given the score, we employ three different evaluation settings where we set different thresholds and compare the results. In setting 1 – which is also the one applied on the provided corpus – we classify all questions with total scores equal to 0 as ISQ and all others with total

scores greater than 0 as NISQ; this means that all questions where at least one NISQ indicator is found are taken to be NISQ. In setting 2, all questions with a score higher than 1 are treated as NISQs, in setting 3 the threshold is set at 2. The two latter settings classify questions as NISQ only if more than one of the indicators are present in the questions. By comparing the different settings we could investigate how many indicators are necessary in order to correctly annotate the questions.

The results are shown in Table 3. In setting 1 the automatic system achieves an F-score of 0.83. In settings 2 and 3, precision increases, however at the cost of a significantly lower recall. This means that setting 1, in which we consider all questions where at least one linguistic indicator is found, is the best performing setting. This result tallies with what the theoretical literature has found: First, there is no linguistic cue that consistently marks NISQs. Secondly, one linguistic indicator per question might be sufficient for classifying the question correctly but raising the threshold to two or more indicators improves the precision. This means that the more indicators available and the more languages we can test in parallel, the better and more reliable the classification results become. However, the higher thresholds fail to capture many cases, leading to a low recall. Thirdly, the linguistic indicators that we chose, e.g., particles, are well-motivated and allow for a robust classification.

	Setting 1	Setting 2	Setting 3
Precision	0.85	0.89	0.97
Recall	0.82	0.5	0.31
F-score	**0.83**	0.64	0.46
Accuracy	0.79	0.62	0.53

Table 3: Evaluation results

As was reported in section 2.2., the previous work in the field uses different training and testing data, e.g. Twitter data, the Switchboard Dialogue Act Corpus, etc., which means that no direct comparison of the systems is possible. Nevertheless, the absolute results of our rule-based approach, in particular those of setting 1, are comparable to some of these machine-learning approaches, showing a higher F-score than Ranganath et al. (2016) and Bhattasali et al. (2015) (F-scores of 64,04% and 53,71%, respectively) and a comparable accuracy to Li et al. (2011) and Zhao and Mei (2013) who report an accuracy of 77,5% and 79%, respectively. It is only in comparison to the system of Harper et al. (2009) with 89,7% accuracy, that our system shows a lower score. Note that it is difficult to conduct a uniform comparison of systems as they rely on different approaches and different data. Our rule-based mechanism is designed for parallel, multilingual corpora. As mentioned in the introduction and as discussed in the conclusion, our approach can be seen as a first step for annotating large, parallel datasets so that these can be further used as training data for machine-learning approaches.

5. The KRoQ Corpus

The first version of the KRoQ corpus is made available under `https://github.com/kkalouli/BIBLE-processing`[4] and contains the French, the Spanish and the Greek Bible texts, along with our annotations. The German Bible text cannot be made available because of copyright restrictions. For French and Spanish, the corpus is provided in the CoNLL-U format to facilitate further processing of the corpus. For consistency, the Greek text is also provided in this format, but is not parsed. Every sentence of each of the Bible translations contains its original Bible index (e.g. Génèse 1:1 for Genesis, Chapter 1, Verse 1), the original sentence, its parsed structure and a comment field named "annotation". This field captures the annotations of the questions and is left blank for all non-questions. For the questions, the annotation field contains the final score that was automatically assigned across the four languages and our annotation based on that score and on setting 1, as presented in Section 4.

For better reproduction of the results, we provide the gold standard files we used in addition to the main corpus. The gold standard contains the first 200 aligned questions of the Bible for each of the four languages. For German, French and Spanish, the "gold" questions are provided in the CoNLL-U format. For Greek they are provided as raw text. Each item of the gold standard is numbered according to its occurrence in the Bible and contains the original question, its Bible index (e.g. Génèse 1:1 for Genesis, Chapter 1, Verse 1) and the parsed structure. The annotations are captured in the comment fields of each item: the final score that was automatically assigned across the four languages and the annotation based on that score and on setting 1 of Section 4 (in the comment field *annotation*) and the manual, gold standard annotation of ISQ vs. NISQ (in the comment field *gold_annotation*). In order to see the alignment across languages, we add a spreadsheet where the 200 questions and their instances in the other languages are aligned.

6. Conclusions and Future Work

The performance achieved by our current system (in the first setting) shows that the generated annotated corpus can be used as a reliable resource for further research. Our annotated data can be utilized as training and testing data for machine-learning algorithms, as corpus-data for theoretical linguistic questions and as a resource for further symbolic approaches. Additionally, the quality of the classification gives us confidence that the implemented mechanism can be used to classify further data, which can then be used to augment the existing resource.

In our future work we would like to pursue three courses of action. First, we would like to improve the system by adding more parallel languages because additional languages are bound to give us additional markers which can help us distinguish more reliably between the types of questions (as indicated by the second and third evaluation setting). Second, we would like to investigate a different multilingual corpus to see if our system also delivers satisfactory results for other kinds of corpora. As a third goal, we would like to use the annotated data as training data for a machine-learning approach.

[4] Due to its large size the corpus cannot be made available through the LRE Map.

7. Acknowledgements

We thank the German Research Foundation (DFG) for the financial support within the projects P8 "Questions Visualized" and P2 "Word Order Variation in Wh-questions: Evidence from Romance" of the Research Group FOR 2111 "Questions at the Interfaces" at the University of Konstanz.

8. Bibliographic References

Bhatt, R. (1998). Argument-adjunct asymmetries in rhetorical questions. In *North East Linguistic Society (NELS 29)*, Dalware.

Bhattasali, S., Cytryn, J., Feldman, E., and Park, J. (2015). Automatic Identification of Rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, page 743–749.

Bohnet, B. and Kuhn, J. (2012). The Best of Both Worlds – A Graph-based Completion Model for Transition-based Parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 77–87.

Bohnet, B. and Nivre, J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1455–1465, Jeju Island, Korea.

Bonifazi, A., Drummen, A., and de Kreij, M. (2016). *Particles in Ancient Greek Discourse: Five Volumes Exploring Particle Use across Genres*. Hellenic Studies Series 74. Washington, DC: Center for Hellenic Studies, Washington, DC.

Cattell, R. (1973). Negative Transportation and Tag Questions. 49(3):612–639.

Dayal, V. (2016). *Questions*. Oxford University Press, Oxford.

de Vries, L. (2007). Some remarks on the use of Bible translations as parallel texts in linguistic research. 60:148–157.

Efron, M. and Winget, M. (2010). Questions are content: a taxonomy of questions in a microblogging environment. In *Proceedings of ASIST '10*.

Escandell, V. V. (1999). Los enunciados interrogativos. Aspectos semánticos y pragmáticos. In I. Bosque et al., editors, *Gramática descriptiva de la lengua española. Vol. 3: Entre la oración y el discurso. Morfología*, pages 3929–3991. Espasa Calpe, S.A, Madrid.

Gale, A., Church, K., and Jarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *Computer and Humanities*, 26(5/6):415–439.

Ginzburg, J., Fernandez, R., and Schlangen, D. (2013). Self-addressed questions in disfluencies. In *Proceedings of Disfluency in Spontaneous Speech (DiSS) 2013*.

Han, C. (2002). Interpreting interrogatives as rhetorical questions. *Lingua*, 112:201–229.

Harper, F., Moy, D., and Konstan, J. A. (2009). Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2009)*, pages 759–768.

Kaiser, G. A. (2015). Zur Verwendung von Bibelübersetzungen in der (romanistischen) Sprachwissenschaft. In *Biblicum Jassyense*, pages 5–18.

Kübler, S., Baucom, E., and Scheutz, M. (2012). Parallel syntactic annotation in CReST. *Linguistic Issues in Language Technology (LiLT)*, 7(4).

Li, B., Si, X., Lyu, M. R., King, I., and Chang, E. Y. (2011). Question Identification on Twitter. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM'11)*.

Maibauer, J. (1986). *Rhetorische Fragen*. Max Niemeyer Verlag, Tübingen, 1 edition.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Paul, S. A., Hong, L., and Chi, E. H. (2011). What is a question? Crowdsourcing tweet categorization. In *CHI 2011, Workshop on Crowdsourcing and Human Computation*.

Ranganath, S., Hu, X., Tang, J., SuhangWang, and Liu, H. (2016). Identifying Rhetorical Questions in Social Media. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM 2016)*.

Sadock, J. (1971). Queclaratives. In D. Adams, et al., editors, *Papers from the 7th Regional Meeting of the Chicago Linguistic Society*, pages 223–232, April.

Tov, E. (2011). Early Bible translations into Greek and renderings derived from them. In H. Kittel, et al., editors, *Ein internationales Handbuch zur Übersetzungsforschung. An International Encyclopedia of Translation Studies. Encyclopédie internatioonale de la recherche sur la traducion.*, volume 3, pages 2351–2354. Berlin: de Gruyter.

Wang, K. and Chua, T.-S. (2010). Exploiting salient patterns for question detection and question retrieval in community based question answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING10)*, page 1155–1163.

Wheatley, J. M. O. (1955). Deliberative questions. *Analysis*, 15(3):49–60.

Zhao, Z. and Mei, Q. (2013). Questions about Questions: An Empirical Analysis of Information Needs on Twitter. In *Proceedings of the International World Wide Web Conference Committee (IW3C2)*, pages 1545–1555.

Zymla, M.-M. (2014). Extraction and Analysis of non-canonical Questions from a Twitter-Corpus. Master thesis, University of Konstanz.