DiscAnn 2021

Integrating Perspectives on Discourse Annotation (DiscAnn)

Proceedings of the Workshop

October 4 – 5, 2021
Tübingen, Germany

Preface

Reflecting the considerable interest in analyzing language beyond the sentence level in linguistics, computational linguistics, psycholinguistics, and applied domains, this workshop brought together researchers from various subdisciplines who are working on aspects of discourse annotation.

Advances in formal pragmatics are extending the empirical reach of linguistic analyses. Computational linguistic research on dialogue and discourse structure has produced multi-layer corpus annotation efforts such as NXT Switchboard or the Penn Discourse Treebank. Applications include dialogue systems and argumentation mining. Our DiscAnn workshop was therefore created with the specific goal to foster the interaction and cooperation of researchers working in different frameworks and using different annotations methods and document the current state-of-the-art in the field of discourse annotation. The contributions in the current workshop proceedings thus present on-going research in the areas of

- discourse relations, such as RST, CCR, QUD trees
- creation of new annotation systems, i.e. for communicative functions
- combining the annotation of different frameworks
- different language phenomena, such as metaphorical language and parenthetical discourse markers
- creation of novel data sets
- supporting discourse annotation by compuational approaches such as question generation

The DiscAnn workshop brought together expertise from the above areas to share experiences and brainstorm around the future of the field. All articles in this volume contribute to the ongoing discussion of the sources of evidence that different annotation schemes and annotation methods for discourse phenomena rely on.

Kordula De Kuthy and Detmar Meurers, Tübingen 2021

Organizing Committee

Workshop Chairs

Kordula De Kuthy (Universitiät Tübingen)
Detmar Meurers (Universität Tübingen)

Program Committee

Lisa Brunetti (Université Paris Diderot-Paris 7, France)
Harry Bunt (Tilburg University, Netherlands)
Christian Chiarcos (Goethe University Frankfurt, Germany)
Stefanie Dipper (Ruhr-Universität Bochum, Germany)
Marie-Catherine de Marneffe (The Ohio State University, USA)
Cornelia Ebert (Goethe University Frankfurt, Germany)
Katrin Erk (The University of Texas at Austin, USA)
Eva Hajičová (Charles University, Czech Republic)
Graeme Hirst (University of Toronto, Canada)
Andrew Kehler (UC San Diego, USA)
Alex Lascarides (University of Edinburgh, Scotland)
Anke Lüdeling (Humboldt-Universität zu Berlin, Germany)
Chris Potts (Stanford University, USA)
Ines Rehbein (University of Mannheim, Germany)
Arndt Riester (University of Cologne, Germany)
Hannah Rohde (University of Edinburgh, Scotland)
Merel C. J. Scholman (Saarland University, Germany)
Manfred Stede (University of Potsdam, Germany)
Simone Teufel (University of Cambridge, England)
Judith Tonhauser (University of Stuttgart, Germany)
Klaus von Heusinger (University of Cologne, Germany)
Bonnie Webber (University of Edinburgh, Scotland)
Amir Zeldes (Georgetown University, USA)
Deniz Zeyrek (Middle East Technical University, Turkey)
Šárka Zikánová (Charles University, Czech Republic)

Table of Contents

Jet Hoek, Merel C.J. Scholman & Ted J.M. Sanders:
Is there less annotator agreement when the discourse relation is underspecified? **1**

Regina Zieleke:
German parenthetical discourse markers between perception and cognition - A case in point for an explorative approach to corpus data **7**

Barbara Schuppler & Anneliese Kelterer:
Developing an annotation system for communicative functions for a cross-layer ASR system **14**

Prisca Piccirilli & Sabine Schulte im Walde:
Contextual choice between synonymous pairs of metaphorical and literal expressions: An empirical study and novel dataset to tackle or to address the question **19**

Arndt Riester, Amalia Canes Nápoles & Jet Hoek:
Combined discourse representations: Coherence relations and questions under discussion **26**

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy & Detmar Meurers:
Advancing neural question generation for formal pragmatics: Learning when to generate and when to copy **31**

Is there less annotator agreement when the discourse relation is underspecified?

Jet Hoek
Centre for Language Studies
Radboud University Nijmegen
jet.hoek@ru.nl

Merel C.J. Scholman
Language Science and Technology
Saarland University
m.c.j.scholman@coli.uni-saarland.de

Ted J.M. Sanders
Utrecht Institute of Linguistics OTS
Utrecht University
t.j.m.sanders@uu.nl

Abstract

When annotating coherence relations, inter-annotator agreement tends to be lower on implicit relations than on relations that are explicitly marked by means of a connective or a cue phrase. This paper explores one possible explanation for this: the additional inferencing involved in interpreting implicit relations compared to explicit relations. If this is the main source of disagreements, agreement should be highly related to the specificity of the connective. Using the CCR framework, we annotated relations from TED talks that were marked by a very specific marker, marked by a highly ambiguous connective, or not marked by means of a connective at all. We indeed reached higher inter-annotator agreement on explicit than on implicit relations. However, agreement on underspecified relations was not necessarily in between, which is what would be expected if agreement on implicit relations mainly suffers because annotators have less specific instructions for inferring the relation.

1 Introduction

Discourse-annotated corpora allow coherence researchers to study the distribution and linguistic realization of coherence relations. Such sources of information enable us to take the study of coherence relations an important step forward. However, discourse annotation has proven to be a difficult task, which is reflected in low inter-annotator agreement (IAA) scores (Artstein and Poesio, 2008; Spooren and Degand, 2010). One explanation for this observation is that coherence is a feature of the mental representation that readers form of a text, rather than of the linguistic material itself (e.g., Sanders et al., 1992). Discourse annotation thus relies on annotators' interpretation of a text, which makes it a particularly difficult task.

In order to gain a deeper understanding of the difficulties associated with reaching sufficient inter-annotator agreement on coherence relation annotations, we need more data on the agreement on different types of relations. Unfortunately, many annotation studies report only overall agreement scores (not distinguishing between different connectives or relation types), or only report agreement scores after the annotators have reconciled disagreements.

The few studies that did report separate agreement statistics have shown that annotators tend to agree more when annotating explicit coherence relations, which are signalled by a connective or cue phrase (e.g. *because*, *as a result*; we will use 'connectives' as a shorthand for the combined category), than when annotating implicit coherence relations, which contain no or less explicit linguistic markers on which annotators can base their decision (e.g., Miltsakaki et al., 2004; Prasad et al., 2008).

This can be considered an expected finding, given that connectives provide comprehenders with "processing instructions" on how to connect incoming text inputs to previously read segments (Britton, 1994; Canestrelli et al., 2013; Gernsbacher, 1997; Sanders and Noordman, 2000). However, it does raise concerns about the validity and added value of coherence annotation efforts: if annotators need connectives in order to reach sufficient agreement on the sense of the relation at hand, is the annotation focused on the coherence relation or rather on the connective? If discourse annotation is mainly focused on connectives, one can wonder how valuable the annotated label is? After all, annotations for explicit connectives such as *if* can, depending on the discourse annotation framework, be done largely or completely automatically. The value of manual annotation comes from disambiguating between relational senses when more than one reading could be inferred. This can occur when the connective is ambiguous (underspecified relative to the inferred relation, Spooren, 1997) or when a relation is not explicitly marked with a connective at all.

The current study functions as an initial investigation of agreement on relations with various markers. By annotating relations marked by specific connectives (*because*, *in addition* and *even though*), highly ambiguous connectives (*and* and *but*), or no connective, we aim to investigate to what extent agreement between annotators is dependent on the specificity of the connective that marks the coherence relation. If the amount of inferencing involved in interpreting a coherence relation is the main source of differences in IAA scores between implicit and explicit relations, we expect IAA to decrease as a function of connective specificity: lowest IAA on implicit relations, intermediate IAA on underspecified relations, and highest IAA on relations marked by a specific connective.

2 Method

2.1 Materials

The data set contained 350 relations taken from transcribed English TED talks: 100 implicit relations, 100 relations marked by underspecified connectives (*and/but*), and 150 relations marked by more specific connectives (*because/in addition/even though*). TED talks are highly structured speeches that are minutely prepared and are meant to provide targeted information on various topics.

The 100 implicit coherence relations were randomly selected from the English part of the TED-MDB corpus (Zeyrek et al., 2019), as well as 50 relations marked by *and* and all relations marked by *but* (n=47). We used the Ted Corpus Search Engine (Hasebe, 2015) to randomly select 50 coherence relations each marked by *because*, *in addition*, and *even though*, plus 3 additional *but*-relations. [1] The selected relations were displayed in their original context during annotation.

2.2 Annotation framework

The Cognitive approach to Coherence Relations (CCR) was used to annotate all relations (Sanders et al., 1992, see Hoek et al., 2019 for an up-to-date version). CCR depicts coherence relations in terms of cognitive primitives. Crucial primitives are POLARITY, BASIC OPERATION, SOURCE OF COHERENCE, and ORDER OF THE SEGMENTS.

POLARITY distinguishes between positive and negative relations. A relation is positive if the propositions P and Q (expressed in the discourse segments S_1 and S_2) are linked without a negation of one of these propositions. A relation is negative if the negative counterpart of either P or Q functions in the relation.

BASIC OPERATION distinguishes between causal and additive relations. In causal relations, an implication relation ($P \rightarrow Q$) can be deduced between the two segments. In additive relations, the segments are connected as a conjunction (P & Q). Temporal relations, in which the segments are ordered in time, are considered a subclass of additive relations. Conditional relations are considered a subclass of causal relations.

SOURCE OF COHERENCE distinguishes between objective and subjective relations. Subjective relations express the speaker's opinion, argument, claim, or conclusion. Objective relations, on the other hand, describe situations that occur in the real world. Temporal relations are assumed to always be objective.

ORDER OF THE SEGMENTS applies to causal and conditional relations. In a basic order relation, the antecedent (P) is S_2, followed by the consequent (Q) as S_1. In a non-basic order relation, P maps onto S_2 and Q onto S_1. The ordering of events in temporal relations (chronological, anti-chronological, synchronous) is captured by TEMPORALITY (see Evers-Vermeul et al., 2017).

2.3 Connective choice

Because is a typical, specific marker of causal coherence relations. *In addition* is a typical, specific marker of additive coherence relations. *Even though* is considered a prototypical connective for negative causal relations.

And is considered an underspecified connective: it can mark a variety of relations, including positive additive relations, as in Example (1), and positive causal relations, as in Example (2), but it can also mark negative additive and causal relations (see Crible et al., 2019). It tends to be used most frequently in positive additive relations, however.

(1) I am terrible at playing darts and I don't know how to play pool.

(2) I missed the dart board and someone lost an eye.

[1] The full annotated data set can be accessed at https://tinyurl.com/rgdjear.

But is also considered an underspecified connective: it can mark negative additive relations, as in Example (3), as well as negative causal relations, as in Example (4). Its distribution is different to *and*, in that it has a less strongly associated default interpretation.

(3) I am terrible at playing darts, but I am a champion in pool.

(4) I missed the dart board, but everybody is safe.

2.4 Annotation procedure

The first two authors, both expert coders, annotated discourse relations according to the CCR framework, without specific within-genre training or intermediate discussion. They assigned single values for every primitive. In cases where the two annotators disagreed, the third author provided an additional annotation. The majority vote was then chosen as the true value. This was used to establish ambiguity of connective usage.

2.5 Inter-annotator agreement metrics

In order to evaluate inter-annotator agreement and gain a comprehensive overview of the agreement, we use different metrics and methods.

Regarding the metrics, we report on three different measures: percentage agreement (also known as observed agreement), Cohen's Kappa κ (Cohen, 1960) and AC_1 (Gwet, 2001). Kappa is the most commonly used agreement measure, but it can behave erratically in certain situations; a problem known as Kappa's Paradox (Feinstein and Cicchetti, 1990). Specifically, when data sets are characterized by an uneven distribution of categories, Kappa's values can be relatively low, despite a higher percentage of observed agreement (see also Hoek and Scholman, 2017). AC_1 was introduced to address this issue. Since some types of relations will likely occur more frequently than others in our data set per connective, we consider both Kappa and AC_1 in order to get a full overview of the agreement.

Regarding the method of the inter-annotator agreement, we consider the agreement on the full "label" of the relation (the combination of all values on the dimensions). Full labels give a straightforward impression of a connective's specificity (i.e., the more types of labels, the more ambiguity) and make for better comparison to annotation efforts in other frameworks, which only use end labels (although there is not necessarily a 1:1 correspondence between the full CCR relation labels and relation labels from other approaches, see Sanders et al., 2018).

Connective		%	κ	AC_1
explicit	*because*	84	.68	.68
	in addition	82	.57	.69
	even though	78	.58	.74
underspecified	*and*	74	.58	.71
	but	58	.39	.46
implicit	Ø	66	.58	.64

Table 1: IAA per connective type and connective

3 Results

We exclude the ORDER OF THE SEGMENTS from our analyses. Determining ORDER is largely trivial for specific connectives (indeed, we reached 100% agreement) and the only source of disagreement for the underspecified and implicit relations was the direct result of a disagreement on BASIC OPERATION (i.e., NA order for additive relations versus basic/non-basic order for causal relations).

Connectives and their assigned senses First, we focus on the annotated labels per connective to answer the question of whether underspecified connectives are truly underspecified, when compared to the specific connectives. Moreover, we examine the different senses assigned to implicit relations to determine how "underspecified" such relations are.

Figure 1 shows the distribution of relations per connective. As assumed, *and* and *but* were more ambiguous than *because*, *in addition*, and *even though*. The largest variety of relation labels was used for the implicit relations.

Agreement per connective Next, we compare the inter-annotator agreement of the two coders, to determine whether agreement on underspecified connectives differs from agreement on specific connectives and from agreement on implicit relations.

Table 1 shows the inter-annotator agreement for each connective. In line with IAA statistics from other annotation efforts (e.g., Miltsakaki et al., 2004; Prasad et al., 2008), agreement was lower on the implicit relations than on the explicit relations. Note that this difference is smaller according to

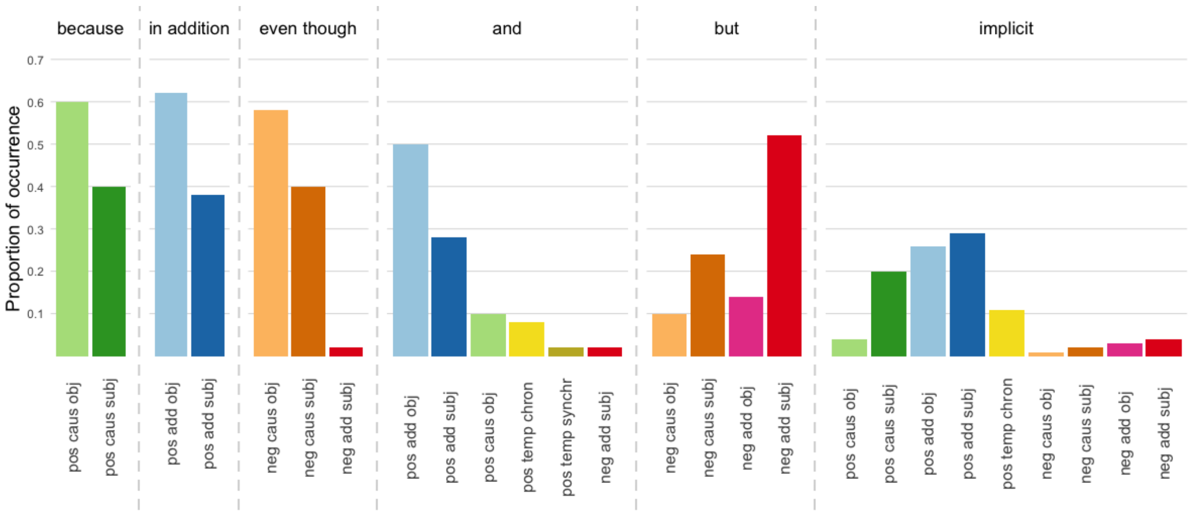

Figure 1: Distribution of relations, per connective.

the two agreement statistics that correct for chance agreement (Kappa and AC_1), but that the difference is still considerable in AC_1, which better takes into account the prevalence of the various categories. However, agreement on the underspecified relations was not necessarily in between. While IAA on relations marked by *and* was comparable with IAA on the explicit relations, agreement was much lower on relations marked by *but*.

4 Discussion

Much like the IAA statistics reported for other annotation efforts, we reached less agreement on implicit relations than on relations that were explicitly marked. However, the level of agreement reached for relations marked by ambiguous connectives *and* and *but* suggests that lower IAA on implicit relations cannot be straightforwardly explained by the specificity of the marker: We did not find an intermediate IAA score on the underspecified relations, and both in the absence of a connective and in the presence of *and*, more types of relations are available than in the presence of *but*, the connective for which we reached the lowest IAA.

Regarding the higher agreement on *and* compared to *but*, we could in part attribute this to the difference in default interpretations. Even though *and* can mark a larger variety of relations than *but*, it is associated more strongly with a default interpretation (78% of *and*-occurrences were positive additive, compared to 66% of *but*-occurrences being negative additive). This stronger default interpretation of *and* likely resulted in more agreement between the annotators. It emphasizes the need for further studies investigating a larger variety of underspecified connectives.

We can further interpret the low agreement on *but* using the primitive-specific annotations: the majority of disagreements on *but*-relations was on BASIC OPERATION, as was for instance the case for example (5).

(5) The US government says it doesn't use torture, and we condemn other countries, like Iran and North Korea, for their use of torture. **But** some people think the so-called worst of the worst deserve it: terrorists, mass murderers, the really "bad" people.

Under the negative causal reading, some people think really bad people deserve to be tortured, *even though* the US government does not support the practice; the fact that your government condemns something might plausibly lead you to condemn it too. Under the negative additive reading, this fragment presents merely two opposite viewpoints: some people condemn torture, *while* others support it (at least in some cases). The distinction between negative additive and negative causal relations corresponds to the distinction between contrast and concessive/denial-of-expectation relations in many other frameworks. Agreement statistics from other annotation efforts indicate that this distinction is a notoriously difficult one to make when coding corpus data (e.g., Robaldo and Miltsakaki, 2014; Degand and Zufferey, 2013).

While implicit relations can also express relations with negative polarity (see e.g., Figure 1), the specific interpretation problems with contrastive relations do not seem to have a big effect on the agreement on implicit relations. Negative relations tend to be explicitly marked much more often than positive relations (e.g., Asr and Demberg, 2012; Hoek et al., 2017) and thus only make up a modest percentage of implicit relations. And while negative relations tend to be implicit more often in spoken than in written language, spoken language offers alternative ways to express contrast, such as topicalization and sentence stress (Rehbein et al., 2016).

Although the results suggest that increased inferencing does not necessarily lead to more disagreements, it is likely that the ambiguity of implicit relations does negatively impact the IAA scores. Implicit relation annotation is characterized by the added complexity of it not being clear *which* relation should be annotated, since more than one relation can hold between two segments (e.g., Rohde et al., 2018; Scholman and Demberg, 2017). For example, the originally implicit relation in (6), taken from our data set, can be interpreted in (at least) two ways: the second segment presents a reason for the first segment ('because'), or it supplies an alternative ('instead'). Note that these relations can hold at the same time.

(6) Prudent investing and finance theory aren't subordinate to sustainability. [BECAUSE INSTEAD] They're compatible.

Multiple relations can also hold between segments that are connected by an explicit connective, but in those cases, the connective supplies a clear cue as to which relation should be annotated.

In sum, the current study showed that IAA scores on underspecified relations do not necessarily fall in between the scores of explicit and implicit relations, which is what would be expected if IAA on implicit relations mainly suffers because annotators have less specific instructions for inferring the relation. Hence, our results indicate how implicit and underspecified coherence relations remain a major challenge for the field, both in terms of annotation practice and in terms of theoretical implications: how do humans deal with so many ambiguous relations in everyday communication?

Acknowledgments

The second author was supported by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding".

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684.

Bruce K. Britton. 1994. *Understanding expository text: Building mental structures to induce insights*. Academic Press.

Anneloes R. Canestrelli, Willem M. Mak, and Ted J.M. Sanders. 2013. Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes*, 28(9):1394–1413.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ludivine Crible, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, and Šárka Zikánová. 2019. Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

Liesbeth Degand and Sandrine Zufferey. 2013. Annotating the meaning of discourse connectives in multilingual corpora. *Corpus linguistics and linguistic theory*, pages 1–24.

Jacqueline Evers-Vermeul, Jet Hoek, and Merel C.J. Scholman. 2017. On temporality in discourse annotation: Theoretical and practical considerations. *Dialogue & Discourse*, 8(2):1–20.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Morton Ann Gernsbacher. 1997. Coherence cues mapping during comprehension. *Processing interclausal relationships. Studies in the production and comprehension of text*, pages 3–22.

Kilem Gwet. 2001. *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: STATAXIS Publishing Company.

Yoichiro Hasebe. 2015. Design and implementation of an online corpus of presentation transcripts of ted talks. *Procedia: Social and Behavioral Sciences*, 24:174–182.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2019. Using the cognitive approach to coherence relations for discourse annotation. *Dialogue & Discourse*, 10(2):1–33.

Jet Hoek and Merel C.J. Scholman. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*, pages 1–13, Toulouse, France.

Jet Hoek, Sandrine Zufferey, Jacqueline Evers-Vermeul, and Ted J M Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics*, 121:113–131.

Eleni Miltsakaki, Aravind Joshi, Rashmi Prasad, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 9–16, Boston, MA, USA.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968, Marrakech, Marocco.

Ines Rehbein, Merel C. J. Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 23–28, Portoroz, Slovenia.

Livio Robaldo and Eleni Miltsakaki. 2014. Corpus-driven semantics of concession: Where do expectations come from? *Dialogue & Discourse*, 5(1):1–36.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267.

Ted J. M. Sanders, Vera Demberg, Jet Hoek, Merel C. J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in discourse relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, ahead of print:1–71.

Ted J. M. Sanders and Leo G. M. Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1):37–60.

Ted J.M. Sanders, Wilbert P.M.S. Spooren, and Leo G.M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Merel C. J. Scholman and Vera Demberg. 2017. Examples and specifications that prove a point: Identifying elaborative and argumentative discourse relations. *Dialogue & Discourse*, 8(2):56–83.

Wilbert Spooren. 1997. The processing of underspecified coherence relations. *Discourse Processes*, 24(1):149–168.

Wilbert P.M.S. Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, pages 1–27.

German parenthetical discourse markers between perception and cognition
An explorative approach to parallel corpus data

Regina Zieleke
Eberhard Karls Universität Tübingen
regina.zieleke@uni-tuebingen.de

When perception verbs are employed as parenthetical discourse markers, e.g. English *you see*, French *tu vois*, the concrete visual perceptual meaning of *see* is said to be expanded to a more abstract meaning of general cognition (cf. Brinton, 2008). In this paper, I will show that in German, different cognitive processes map with different parentheticals: visual parentheticals such as *(wie) du siehst* ('(as) you see') are only used in contexts of justification, whereas processes of explanation invoke the use of cognitive parentheticals such as *weißt/verstehst du* ('you know/understand') instead. For this purpose, I will explore data from the parallel corpora Europarl7 and OpenSubtitles2011. The assessment of German equivalents to the English *you see* and French *tu vois* and a paraphrase test aiming at these different cognitive processes provide a pattern linking the latter to German visual vis-à-vis cognitive parentheticals.

1 Parenthetical discourse markers and verbs of perception

Parenthetical discourse markers (also 'pragmatic markers', 'comment clauses', see e.g. Brinton, 2008) such as *you know* or *I mean* are verbal constructions that are "not syntactically connected to the rest of the clause (i.e., [are] parenthetical)" (Brinton, 2008: 1) and are "metacommunicative" in that they "comment on the truth value of a […] group of sentences, on the organization of the text or on the attitude of the speaker" (Peltola, 1982/1983, cited by Brinton, 2008: 5).

On the supposition that 'visual perception is our primary source of information in the outside world' (Bat-Zeev Shyldkrot, 1989, cited by Bolly, 2012: 3[1]), visual perception verbs can be regarded natural candidates for such constructions. However, while English *you see* as in (1) and French *tu vois* as in (2) are frequent, an equivalent construction in German is inacceptable (cf. (3a)) and has to be expressed by a construction involving the cognitive verbs *wissen* ('to know') or *verstehen* ('to understand') instead (cf. (3b)).

(1) I went to three different stores to find the perfect avocado. **You see,** I love guacamole.

(2) J'ai cherché l'avocat parfait dans trois magasins différents. **Tu vois**, j'adore du guacamole.

(3) Ich war in drei Läden, um die perfekte Avocado zu finden.
a. **#Du siehst / #Siehst du**, ich liebe Guacamole.
b. **?Du weißt / Weißt du / Verstehst du**, ich liebe Guacamole.

The link between (verbs of) perception and cognition is well-known (cf. 'I see your point' vs. 'I know what you mean', see e.g. Sweetser, 1990). Viberg (2015: 96), for example, states that "verbs of perception are situated in the middle of a continuum of more raw descriptions of sensations at one end and more abstract reference to thinking and knowledge at the other end". According to Brinton (2008: 159), this path can also be observed in the grammaticalization process of constructions such as English *(as) you see* or *I see*: "the concrete visual perceptual meaning of *see* is bleached or widened to a more abstract meaning of general cognitive perception".

There are two reasons to be nonetheless puzzled by this observation. First, both French and English have parenthetical markers involving an equivalent cognitive verb that are very frequent, i.e. French *tu sais* and English *you know*, but these are ascribed with pragmatic functions distinct from *tu vois* and *you see*. Erman (1987: 117/118), for

[1] Original quote in French: "[…] en tant que 'première source d'information objective et intellectuelle sur le monde ext´erieur' (Bat-Zeev Shyldkrot, 1989: 288)" (ibid.).

example, ascribes English *you see* with an argumentative 'terminating' function making "the addressee accept [the speaker's] ideas and explanations". The cognitive *you know*, on the other hand, is ascribed with an 'introductory' function making "the addressee accept parts of the information conveyed as common ground" (ibid., see also Schiffrin's, 1987 account of *y'know* as appealing to shared knowledge).

Second, there are cases such as in (4)-(6) where German does allow for parenthetical markers with *sehen* ('to see') (cf. (6a)), while cognitive parentheticals are less acceptable (cf. (6b)):

(4) The house isn't cleaned and I didn't go grocery shopping yet. **You see,** I still have a lot to do.

(5) La maison n'est pas propre et je n'ai pas encore fait les courses. **Tu vois**, j'ai du pain sur la planche, là.

(6) Das Haus muss noch geputzt werden und einkaufen war ich auch noch nicht.
a. **Du siehst (also) / Wie du siehst,** ich hab echt viel zu tun.
b. #**Du weißt / #Weißt du / ?Verstehst du,** ich hab echt viel zu tun.

I argue that the two sets of examples involve two different cognitive processes: one of explanation in (1) – (3) and one of justification or provision of evidence in (4) – (6). As (7) and (8) show, the causal explanation marker *because* is acceptable only in the first case, where loving guacamole is the explanation for going through the trouble of visiting three different stores. Such a relation cannot be applied to (8), where the unpleasantness of an uncleaned house and missing groceries are offered as a justification or evidence for the speaker still having a lot to do, instead. This, in turn, correlates with the paraphrase *this is evidence for the fact that.*[2]

(7) I went to three different stores to find the perfect avocado, **because / ? this is evidence for the fact that** I love guacamole.

(8) The house isn't cleaned and I didn't go grocery shopping yet, **#because / this is evidence for the fact that** I still have a lot to do.

It seems, then, that while parenthetical markers involving perception verbs can be used to express both processes in English and French, they are limited to the process of justification/evidence in German. A hint of a different cognitive status of German perception verbs used as discourse markers is provided by Günthner (2017). She studies the German *guck mal* ('look') and *weißt du* ('you know'), and while her verdict for the cognitive *weißt du* resembles Erman's description of *you know* ('projection of a knowledge transfer making the utterance part of the Common Ground' Günthner, 2017: 125), the visual *guck mal* in its discourse marker use is described as merely involving a shift in perception from a purely visual to the 'discourse world and thus the argumentation structure'.

In this paper, I will discuss data from English-German and French-German parallel corpora. The goal of this explorative approach to parenthetical discourse markers with the visual perception verbs *see/voir* is to find out, how the function of these markers is handled in German. In a first step, I will assess the German equivalents in the parallel data – does German make use of parenthetical markers at all, and if so, do they involve verbs of perception (*sehen*) or cognition (*wissen*, *verstehen*)? The second step consists of assessing possible discourse functions relating to the different cognitive processes – is there a pattern linking the different German equivalents to different discourse functions?

2 *You see/tu vois* and their German equivalents

2.1 Data and annotation criteria

The data is taken from two parallel corpora, `Europarl7` and the `OpenSubtitles2011` sub-corpus of `OPUS2`, accessed via SketchEngine. Both corpora consist of aligned transcriptions of spoken language, viz. political discourse data from the proceedings of the European Parliament in the case of Europarl7 and data from subtitles in movies and TV series in the case of OpenSubtitles2011.[3] Both corpora were searched for the two language pairs English–German and French–German each, with

[2] In a way, this is also a kind of an explanation: *I still have a lot to do, because the house isn't cleaned…* . However, the order of explanans and explanandum are inversed resulting in a different relation altogether – as shown by the fact that the paraphrase *this is evidence for the fact that* is not *un*acceptable in (7), but alters the sense in that way.

[3] This choice of corpora comes with two restrictions: first, it is often unclear which language is the source and which the translated language; and second, subtitles tend to involve shortened sentences in order to fit on the screen in the available time (cf. Müller & Volk, 2013: 2).

the English and French parenthetical markers *you see* and *tu vois* as the starting point.

In order to exclude matrix verb uses of these verbal constructions, the search request made use of the observation that parenthetical markers are "marked by "comma intonation" (pauses in speech, or actual commas in writing) that separates [the marker] from its anchor" (Brinton, 2008: 8). The positions of the markers (preposed or postposed) were not restricted in the search request. Since French is a language with strong verb inflection and the discourse in Europarl is of a formal register, the search request in this corpus includes the formal second person equivalents of the parenthetical *tu vois*, i.e. *vous voyez* and *voyez-vous*.)[4].

A random sample of 44 sets of data for each language in the Europarl7 and 48 in the OpenSubtitles2011, respectively, was annotated for simple criteria in line with the explorative nature of the investigation. The main focus lies on the (direct) German equivalent of the parenthetical markers in English or French. This involves the annotation of (i) the specific sequence of words (*verstehen Sie* in (9)), (ii) the lemma of that sequence (*verstehen*), and, most importantly, (iii) a categorization of these lemmas as a) perception verb, b) cognitive verb (as in (9)), c) particle/ connective, or d) no equivalent. Annotation further accounts for the presence of further discourse markers in the parentheticals' environment (e.g. the connective *but* in (9)).

(9) EN Sorry, but **you see**, we've gotta check up on everybody.
DE Tut mir Leid, aber **verstehen Sie**, wir müssen jeden überprüfen.
Ref: OPUS2; #176183352, en/1934/5990/4099372_1of1.xml.gz

2.2 Results: A general pattern

Among the 184 sets of parallel data, there are 37 different German linguistic expressions that can be identified as equivalent to their English and French counterparts *you see/tu vois* – form single word expressions such as *eigentlich* ('actually') to complex tag questions such as *weißt du, was ich meine* ('do you know what I mean'). Figure 1 shows the overall distribution of the German equivalents among the four categories described above:

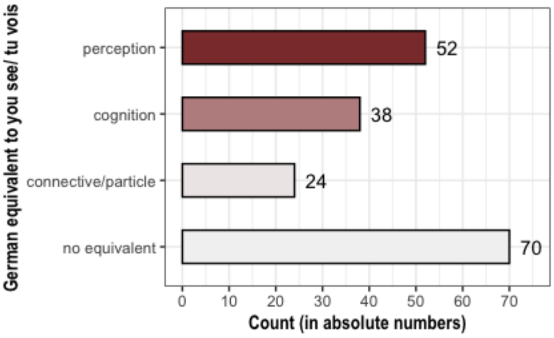

Figure 1: German equivalents to *you see/tu vois*

With 28.3% of the overall data, constructions involving perception verbs such as *siehste* or the phrase *wie Sie sehen* ('as you see') in (10) are more frequent than cognitive verbs such as *weißt du* or *verstehen Sie* in (9) above with only 20.7%. Most frequently, the German data does not contain any equivalent to the French or English parenthetical marker as in (11) (38%), whereas particles/connectives such as *eigentlich* ('actually') or *nämlich* ('namely') in (12) were found least frequently in the corpus data (13%).

(10) EN There is, **you see**, a clear risk that this is just procrastination.
DE […] es besteht, **wie Sie sehen**, eindeutig die Gefahr einer Verschleppung.
Ref: Europarl7; #32016758, /en/ep-08-05-07-014.xml

(11) FR J'étais occupé, **tu vois**.
DE Ich war beschäftigt.
Ref: OPUS2; #228764636, fr/1931/8606/3505132_1of1.xml.gz

(12) FR Ce que nous sommes en train de faire, **voyez-vous**, c'est défendre les secteurs qui ne sont pas compétitifs […]
DE Was wir **nämlich** damit zurzeit erreichen, ist der Schutz und die Verteidigung von nicht wettbewerbsfähigen Wirtschafts-zweigen […]
Ref: Europarl7; #28994717, /fr/ep-06-10-11-016.xml

The comparison between the two languages of origin for the search request reveals similar patterns, cf. **Error! Reference source not found.**. Solely the category connective/particle differs considerably: whereas English *you see* is expressed by a connective or particle in its German equivalent in 19.6% of the time, examples as in (12) only make up 6.5% of French *tu vois*.

[4] See links to the specific CQL concordance search and data: Europarl7_FR: https://ske.li/ikr; https://ske.li/in1; Europarl7_EN: https://ske.li/iks; OpenSubtitles2011_FR: https://ske.li/ikq; OpenSubtitles2011_EN: https://ske.li/ikt

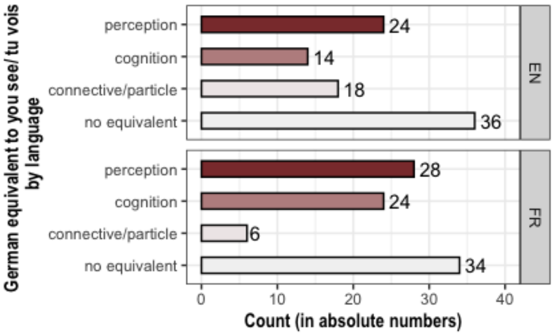

Figure 2: German equivalents by language

The distribution pattern changes completely when comparing the four categories by type of corpus instead, cf. Figure 3. The high number of German equivalents involving perception verbs predominantly relates to the Europarl corpora with 80.7% of the perception-verb-equivalents. The OpenSubtitles corpora seem to be responsible for most of the cognitive-verb-equivalents, instead (94.7%). The majority of German equivalents to *you see* and *tu vois* involving particles/ connectives, in turn, correlates with Europarl again (91.7%).

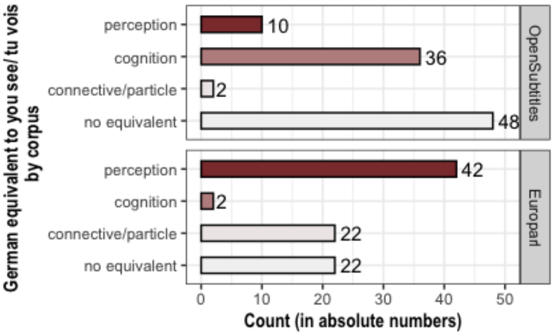

Figure 3: German equivalents by corpus

This considerable change in pattern is particularly interesting if we consider the type of discourse represented in the two kinds of corpora. Since Europarl comprises political discourse data from the proceedings of the European Parliament, the discourse can be said to be more argumentative in nature. This matches well with our assumption that in German visual parentheticals relate to the cognitive process of justification – the speakers use parenthetical markers such as *sehen Sie* or *Sie sehen also* to mark the provision of evidence for their argumentation. OpenSubtitles, on the other hand, comprises discourse that at least tries to imitate everyday interactions. We can thus expect a higher share of the cognitive process of explanation presumably correlating with cognitive parentheticals in German.

In a second step, we thus have to take a closer look at the possible discourse functions involved with visual and cognitive parentheticals in German.

3 Parentheticals of perception and cognition in German – different discourse functions?

There are three different types of discourse functions that are discussed in the literature on our constructions of departure, English *you see* and French *tu vois* (literature on German *siehst du* is – maybe unsurprisingly, considering the small amount of data and presumably limited discourse functions – as good as non-existent). The first can be entitled as 'interpersonal', i.e. "claim[ing] the addressee's attention (Quirk et al., 1985) or "keep[ing] control over the interaction, maintain[ing] contact with the interlocutor" ('Interpersonal monitoring', Crible & Degand, 2019: 27/35). The second can be summarized under the term 'segmentation', i.e. marking transitions between information units or arguments (Erman, 1987 on English *you see*, see also Bolly, 2012:10/11 on French *tu vois* as a 'ponctuant'). Finally, we have the 'explanation/justification' function as quoted from Erman (1987) above.

As I have argued above, however, I consider explanation and justification to be two different cognitive processes that – at least in German – seem to map with different parenthetical markers. For our purposes, these two should thus be considered as two separate functions that can be distinguished using a paraphrase test along the lines of (7) and (8) above: 'Explanation/Reason' with the paraphrase *because/the explanation for that is* vis-à-vis 'Justification/Evidence' with the paraphrase *this is evidence for the fact that* (or the French equivalent paraphrases for the French part of the data).

The other two functions, 'interpersonal' and 'segmentation' do not involve cognitive processes, but are entirely meta-discursive, relating to the discourse structure and interaction, instead. As such they cannot be identified via paraphrase tests and are more subtle in nature. Examples with question-answer-pairs as in (11) above, for example, seem to be cases where *you see/tu vois* simply marks the transition from question to answer. Other examples, as in (13), seem to mark the beginning of a new, bigger discourse segment, while also "maintain[ing] the contact with the interlocutor":

(13) Let' s see ... where was I? Oh, yes! The master. He was kind, **you see**. He brought me to our mutual acquaintance, Father Karras. Not too well at the time.
Ref: OPUS2; #326900403, en/1990/4253/77639_1of1.xml.gz

Since, at this point, it is unclear how to operationalize a distinction between these two functions (and the focus of the exploration lies on the functions involving cognitive processes), I group them into one category 'Segmentation/ Interpersonal'.

There is one challenge for this part of the exploration of the data, however: The annotation of the presence of discourse markers other than *you see/tu vois* reveals that exactly half of the data provide the combination of *you see/tu vois* with other markers, e.g. English *well*, *and then*, or *but* as in (9) above, cf Figure 4.

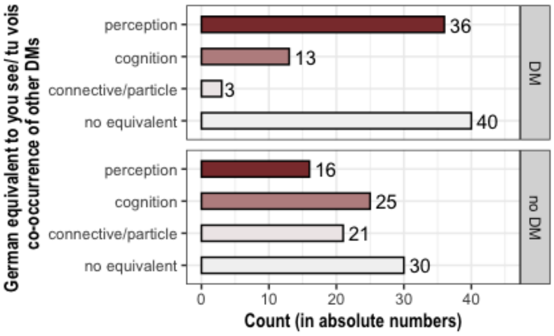

Figure 4: Co-occurrence of other discourse markers

This poses a challenge in so far as the presence of other discourse markers can block the application of the paraphrase test, cf. (14) and the failed attempt to paraphrase *you see* in (14'). Interestingly, omitting *you see* or any paraphrase altogether seems to be closest to the original meaning.

(14) A: Why, Captain John told me I could stay on my place as long as I wanted to. […]
B: Yeah, I know he did, Jeeter ... But **you see**, that land doesn't belong to us anymore.
Ref: OPUS2; #184020757, en/1941/25528/3671553_1of1.xml.gz

(14') I know he told you that you could stay. But **#because / #the explanation for that is that / #this is evidence for the fact that** / ∅ that land doesn't belong to us anymore.

For now, the annotation of discourse functions via the paraphrase test thus has to be limited to the 92 sets of data where *you see/tu vois* is the only discourse marker present. Unfortunately, this leaves us with only 16 instances of German visual parentheticals and 25 cognitive parentheticals. Additionally, the represented languages and types of corpora become slightly skewed with 48 instances from OpenSubtitles compared to 44 from Europarl, and 52 with English *you see* as a point of departure compared to 40 with French *tu vois*.

Nevertheless, the paraphrase test reveals an interesting pattern in terms of cognitive processes and verb types used in German. As Figure 5 shows, German cognitive parentheticals are primarily used to express an explanation process (84%), whereas visual parentheticals primarily occur with the process of Justification/Evidence (62.5%). This latter process is overall least frequent which makes the strong relation with visual parentheticals in German all the more interesting. The observation that German equivalents in form of a connective or particle are used exclusively to express an Explanation/Reason process hardly seems surprising considering that these are mostly causal connectives and particles such as *denn* ('because') and *nämlich* ('namely') as in the French example in (12) above.

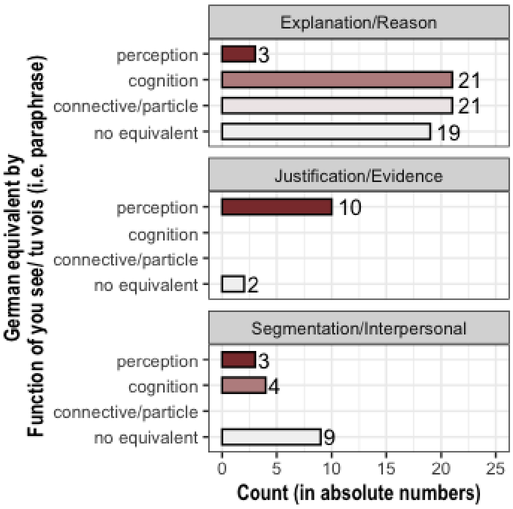

Figure 5: German equivalent by discourse function

4 Discussion

We set out to explore whether the corpus data reveal a pattern considering German parenthetical markers between perception and cognition. The claim was that the use of German visual vis-à-vis cognitive parentheticals as equivalent to *you see/tu vois* depends on the cognitive process involved – an Explanation/Reason that can be paraphrased by *because* or *the explanation for that is* goes in hand with German cognitive parentheticals such as *weißt du* or *verstehen Sie*, and a Justification/Evidence process that can be paraphrased by *this is evidence for the fact that* goes in hand with visual parentheticals such as *siehst du* or *wie Sie sehen*.

A first clue of such a relationship can be observed from the distribution of German equivalents among the different types of corpora. As shown in Figure 3, the Europarl corpora are the origin for 80.7% of the perception-verb-equivalents, while 94.7% of the cognitive-verb-equivalents were found in the OpenSubtitles corpora. The explanation for this distribution was the varying types of discourse represented in the different corpora: the argumentative nature of political discourse represented in Europarl goes in hand with a process of Justification/Evidence, while the everyday discourse in OpenSubtitles would involve more Explanation/Reason processes. Of course, this neither means that there are no explanations in political discourse, nor that everyday conversation lacks justification. However, the high number of connectives and particles used in the German equivalents of the Europarl data (91.7% of this category) and the observation that these are causal in nature suggests that in political discourse the preferred way to express explanations in German are causal connectives and particles, while parenthetical markers are the preferred choice for this process in everyday discourse. A closer look at these different cognitive processes using the paraphrase test to distinguish the three discourse functions Explanation/Reason, Justification/Evidence, and Segmentation/Interpersonal supports these observations, as illustrated in Figure 5.

However, this difference between visual and cognitive parentheticals only concerns the German part of the data. We can thus derive that English and French parenthetical markers involving verbs of perception seem to be situated at different positions in the perception-cognition-continuum described by Viberg (2015, cited above) than their German counterparts: English *you see* and French *tu vois* cover the whole range from 'raw' visual perception over the visually inspired cognitive process of justification all the way to the complex cognitive process of explanation. The German *siehst du/wie du siehst*, on the other hand, only covers the first two functions, or, as Günthner (2017, cited above) put it for the imperative *guck mal* ('look'), merely accomplished the shift from actual visual perception to the abstract perception of argumentation structure (in the sense of 'Look, this is the evidence for my argument!'). The shift to cognitive parentheticals such as *weißt/ verstehst du* in German when expressing the more complex cognitive process of explanation interestingly matches the Common Ground related functions ascribed to both German *weißt du* (cf. Günther 2017, cited above) and English *you know* (cf. Erman, 1987 and Schiffrin, 1987, cited above). In (15), for example, A's explanation that it's a surprise is not exactly presented as unexpected information, but can easily be accommodated (even without further context). This way of making "the addressee accept parts of the information conveyed as common ground" (Erman, 1987, cited above) is perfectly expressed by the German cognitive equivalent *verstehen Sie* ('(do) you understand').

(15) EN A: I don't want them to see me arrive.
B: Oh.
A: It's a surprise, **you see**.
DE A: **Verstehen Sie**, eine Überraschung,
Ref: OPUS2; #220746752, en/1963/1023/4104979_1of1.xml.gz

This raises the question whether, in this use as a marker of Explanation/Reason, English and French *you see/tu vois* and *you know/tu sais* are exchangeable. If we follow Brinton (2008) and many others in the assumption that the original semantics of verbs is bleached on their path towards parenthetical discourse markers, this could be the case. The 'persistence' (i.e. leftover meaning reflected in distributional constraints, cf. Hopper, 1991) in this case, however, might relate to a difference in what kind of information is added to the common ground: English *you see* might involve more objective information, whereas *you know* (in line with Günthner's, 2017 suggestion for German *weißt du*) could be used for (inter)subjective information instead.

Finally, the co-occurrence of parenthetical markers with other discourse markers provides interesting pointers for further research. Since the presence of other markers, especially connectives such as *and (then)* or *but*, impedes the application of the paraphrase test (cf. (14') above), I had to ignore half of the data for this part of the explorative study. As examples such as the following show, however, these cases might be particularly insightful for the as yet somewhat evasive function of Segmentation/Interpersonal, and for the analysis of the multifunctional contribution of discourse markers in general. Example (16) raises the question whether *you see* simply complements the markers it co-occurs with: both *well* and *you see* seem to simply fulfill the same function, i.e. marking the transition from question to complex answer (potentially involving some hesitation as to where to begin). The altered version of (14) shown in (17), on the other hand, seems to provide the opposite case: it seems that the contrastive *but* and the parenthetical *you see* relate two different arguments – *but* marks the contrast between the inferences drawn from the first and second utterances and an implicit argument ('he told you that you could stay' → *you can stay*; 'this land isn't ours' → *you can't*), whereas *you see* marks the second utterance as an explanation for this implicit contrast-argument ('you can't stay here, because this land doesn't belong to us anymore').[5]

(16) A: What made you decide to become a lawyer?
B: **Well, you see**, it's like this, Miss Roy. A white boy, he can take most any kind of job and improve himself. …
Ref: OPUS2; #185696493, en/1942/37804/4037766_1of1.xml.gz

(17) I know he told you that you could stay. **But you see**, that land doesn't belong to us anymore.

To conclude, the explorative approach to parallel corpus data on English *you see* and French *tu vois* and its German equivalents not only provides us with interesting observations on parenthetical markers in the perception-cognition-continuum. It also points us towards important questions for future research: How can we operationalize discourse functions involving processes of different cognitive complexity (such as Justification/Evidence and Explanation/Reason) and those involving meta-discursive functions (such as Segmentation/Interpersonal)? What overlap do visual and cognitive parentheticals provide and what do they add to the Common Ground? Finally, the co-occurrence and interaction of parenthetical markers with other discourse markers prompts an analysis of the multifunctional contribution of discourse markers in general (for example via the two-dimensional model for discourse markers suggested by Crible and Degand, 2019) and the impact of inferences on discourse structure and the interpretation of discourse markers.

References

Bolly, Catherine. 2012. Du verbe de perception visuelle au marqueur parenthétique "tu vois": Grammaticalisation et changement linguistique. In: *Journal of French Language Studies*, 22(02), pages 143-164. (http://hdl.handle.net/2078.1/74646)

Brinton, Laurel J. 2008. *The comment clause in English*. Cambridge: Cambridge University Press.

Crible, Ludivine & Liesbeth Degand. 2019. Domains and functions: A two-dimensional account of discourse markers. *Discours. Revue de linguistique, psycholinguistique et informatique.* 24. (https://doi.org/10.4000/discours.9997)

Erman, Britt. 1987. *Pragmatic expressions in English: A study of 'you know', 'you see' and 'I mean' in face-to-face conversation.* Stockholm: Almqvist & Wiksell.

Günthner, Susanne. 2017. Diskursmarker in der Interkation – Formen und Funktionen univerbierter *guck mal-* und *weißt du*-Konstruktionen. In Hardarik Blühdorn, Arnulf Deppermann, Henrieke Helmer & Thomas Spranz-Fogasy (eds.), *Diskursmarker im Deutschen: Reflexionen und Analysen*, 103–130. Göttingen: Verlag für Gesprächsforschung.

Hopper, Paul J. 1991. On Some Principles of Grammaticalization. In Elisabeth Traugott & Bernd Heine (eds.), *Approaches to Grammaticalization*, 17–35. John Benjamins Publishing.

Müller, Mathias & Martin Volk. 2013. Statistical machine translation of subtitles: From OpenSubtitles to TED. In Iryna Gurevych, Chris Biemann & Torsten Zesch (eds.), *Language processing and knowledge in the Web*, 132–138. Springer. (https://doi.org/10.1007/978-3-642-40722-2_14)

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman Group Ltd.

Schiffrin, Deborah. 1987. *Discourse markers*. No. 5. Cambridge: Cambridge University Press.

Sweetser, Eve. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*, vol. 54. Cambridge University Press.

Viberg, Åke. 2015. Sensation, perception and cognition: Swedish in a typological-contrastive perspective." *Functions of language* 22.1, pages 96-131.

[5] I thank Merel C.J. Scholman for pointing that out in a fruitful discussion of this and similar corpus examples during the DiscAnn workshop.

Developing an Annotation System for Communicative Functions for a Cross-Layer ASR System

Barbara Schuppler
SPSC Laboratory
Graz University of Technology, Austria
b.schuppler@tugraz.at

Anneliese Kelterer
Department of Linguistics,
University of Graz, Austria
anneliese.kelterer@uni-graz.at

Abstract

The investigation of conversational speech requires the close collaboration of linguists and speech technologists to develop new modeling techniques that allow the incorporation of various knowledge sources. This paper presents a progress report on the ongoing interdisciplinary project ”Cross-layer language models for conversational speech” with a focus on the development of an annotation system for communicative functions. We discuss the requirements of such a system for the application in ASR as well as for the use in phonetic studies of talk-in-interaction, and illustrate emerging issues with the example of turn management.

1 Cross-layer language models for conversational speech

In the last decade, conversational speech has received a lot of attention among speech scientists. Accurate automatic speech recognition (ASR) systems are essential for conversational dialogue systems, as these become more interactional and social rather than solely transactional (Baumann et al., 2016). Linguists study natural conversations, as they reveal additional insights to controlled experiments with respect to how speech processing works. Investigating conversational speech, however, does not only require the application of existing methods to new data, but also the development of new categories and modeling techniques, and the inclusion of new knowledge sources.

Here, we present an ongoing interdisciplinary project with two main aims: (1) The project aims at increasing our understanding of how phonetic (and especially prosodic) variation is related to the semantic context and to communicative functions in conversations. For this purpose, we will conduct phonetic corpus studies and perception experiments, both based on data drawn from conversational speech corpora.

(2) Whereas traditional language models (LMs) are trained on text only, we aim at incorporating information on the phonetic variation of words in LMs and at relating this information to the semantic context and to the communicative functions in conversation. This approach to LMs is in line with the theoretical model proposed by Hawkins and Smith (2001), in which the perceptual system accesses meaning from speech by using the most salient sensory information from any combination of levels/layers of formal linguistic analysis. Such a model is reminiscent of the cross-layered optimization principle in wireless communications (Shakkottai et al., 2003). It was introduced as an alternative to the Open Systems Interconnection (OSI) model, where one layer provides services only to its upper layer while exclusively receiving services from the layer below. With the term *cross-layer*, we refer to our view of how humans access meaning and to the system architecture of the envisioned ASR system.

Figure 1 shows the architecture of the ASR system which is currently being developed. Boxes in white show components that have already been developed (Schuppler et al., 2017; Linke et al., 2020; Schuppler and Ludusan, 2020). Those in gray are currently being developed. The LM proposed is aware of the communicative history and dynamics of the conversation (in Figure 1 referred to as ’cache’). Our current ASR experiments show that WERs heavily depend on pronunciation variation, articulation rate, overlapping speech and semantic and syntactic complexity, which in turn strongly correlate with communicative functions. Our knowledge-based approach to LMs is contrary to recent work on end-to-end ASR systems (e.g., Ito et al., 2017), because in addition to improving ASR, we also aim at increasing our knowledge on human speech processing.

One important aspect of our work is its interdis-

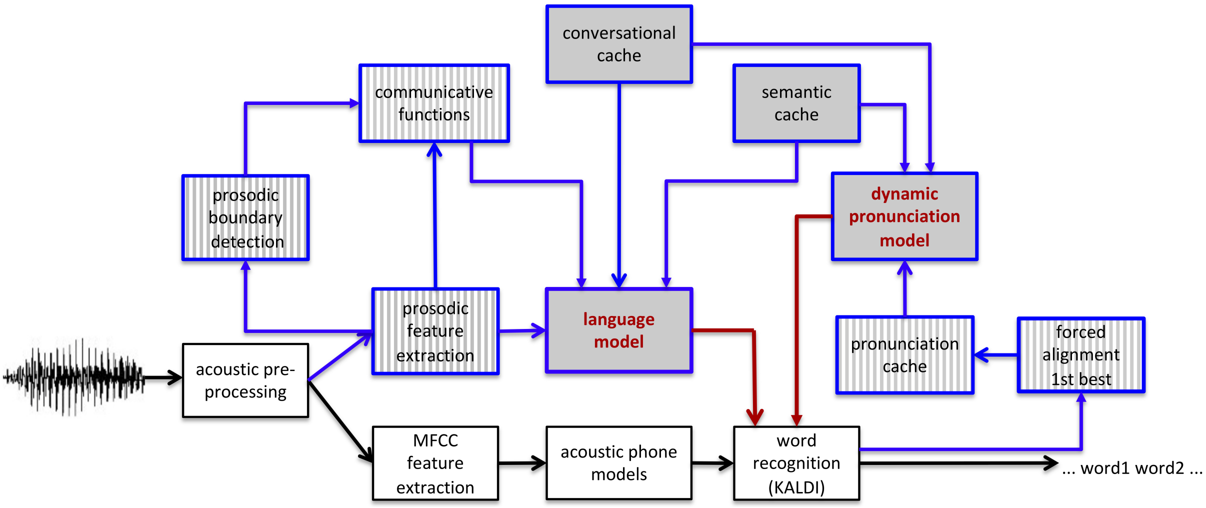

Figure 1: Architecture for an ASR system using a communicative-functions aware language model.

ciplinarity work flow. We create cross-layer LMs which will be tested in ASR systems. In doing so, we will not only investigate which contextual, lexical and acoustic cues work well for speech recognition, but we will also interpret them phonetically. Subsequently, corpus and perception studies will be designed to investigate which of the cues used by the ASR system are also relevant for human speech perception, and which additional cues used by humans might increase ASR performance (e.g., Schuppler et al., 2010). The phonetic studies will be facilitated by ASR technology, i.e., we use tools for the annotation of data, for acoustic feature extraction and we apply advanced statistical methods. Gained phonetic knowledge will again be incorporated into the ASR system. For this interdisciplinary workflow, it is thus necessary to develop an annotation system of communicative functions which is suitable for both phonetic studies and the incorporation into an ASR system.

2 GRASS Corpus

The Graz corpus of Read and Spontaneous Speech (GRASS) contains recordings of spontaneous dialogues of one hour each. They were recorded with 19 pairs of native speakers of (eastern) Austrian German who were friends, couples or family members, resulting in a casual speaking style. The orthographic transcriptions include annotations of disfluencies, breathings and laughter (Schuppler et al., 2014, 2017). Parts of the corpus have been segmented on word and phone-level and were manually annotated prosodically following the KIM system (IPDS, 1997). We have built tools for the detection of prosodic boundaries (Schuppler and Ludusan, 2020) and for the classification of prominence levels (Linke et al., 2020). These tools were created such that they can (1) facilitate the annotation of the not yet annotated parts of GRASS in a semi-automatic procedure, and (2) can be incorporated into the ASR system shown in Figure 1. For the communicative-functions layer of annotations, we also aim to build a tool that serves both mentioned purposes.

3 Annotation of Communicative Functions

For GRASS, we need an annotation scheme that is suitable for speech in naturally occurring conversations. Thus, we take notions from Conversation Analysis (CA), a discipline that focuses on speaker behaviour (rather than, e.g., intentions or intuitions) and stresses the importance of the sequential context for the analysis of speech. Most annotations of communicative functions in the literature are restricted to a limited set of data tailored to a specific investigation (e.g., Ward, 2004; Gravano et al., 2007). One exception are the dialog act categories used to annotate the Switchboard Corpus (Jurafsky et al., 1998; Calhoun et al., 2010). Other corpora that are transcribed in CA terms are searchable for words/lemmata, but not annotated for communicative functions (e.g., *DGD*; Schmidt, 2014). We aim at creating annotations of communicative functions for whole conversations in GRASS. The communicative functions annotations will be used for (1) improving ASR with knowledge about turn-taking, feedback particles with different functions and speaker alignment (e.g., agreement and disagreement), and how they relate to prosody and pronunciation variation; and (2) studying the function-phonetics mapping for various questions in the tradition of Phonetics of Talk-in-Interaction (PTI; Ogden, 2012). Given these two applications, our annotation system has to meet the requirements of (1) annotation consistency, (2) PTI perspective, and (3) ASR application.

Annotation consistency In comparison to PTI studies (e.g., Gorisch et al., 2012; Sikveland, 2012; Zellers, 2016), in which annotations are performed mainly by one or two experts, in our project, large amounts of data are being annotated by a team of approx. 2-4 student assistants. To obtain a high annotation quality and consistency, it is important to keep the annotation task as simple as possible. A way to achieve this is by splitting the annotation into various levels, each of a lower complexity. Another motivation for simplifying labelling tasks for human annotators is that the consistent segmentation and labelling of units are essential to ensure good automatic detection of categories.

PTI perspective For the investigation of prosodic and segmental phonetic variation in an integrated approach such as proposed by Zellers and Post (2011), the annotation of communicative functions has to be methodologically sound following principles of Conversation Analysis (CA). One domain we employ in our annotation scheme is potential transition relevance places (TRP) in terms of *points of potential syntactic completion* (PCOMP). While TRPs are undoubtedly also determined by prosody (e.g., Selting, 1996), it is less clear what constitutes potential phonetic completion. Therefore, even studies within PTI use only syntactic criteria to identify potential TRPs in their investigations of turn management (e.g., Zellers, 2016; Local and Walker, 2012). For the ASR system, the annotation of PCOMPs might pose problems, in particular in cases in which they do not coincide with pauses. Since these domains are not well-defined in terms of prosody, they are harder to detect. In cases in which a pause belongs to the same unit as the stretches of speech around it (e.g., when a speaker makes a pause in the middle of a sentence, cf., Figure 2), units are difficult to recognize automatically.

ASR application For ASR, we want to use communicative functions and prosody features to improve word recognition. Thus, the word level is not available for the identification of PCOMPs in the speech stream. For the application in our ASR system, it is important that boundaries and labels can be detected on the basis of spectral and prosodic features only, as communicative functions are being detected before word-level recognition is done. Moreover, the preference is towards a small number of labels, as a large number of categories (e.g. 42 dialog act categories in Jurafsky et al., 1998, 24 stance type labels in Freeman, 2019) will lead to a high level of confusion in the automatic classification process. From an ASR point of view, the annotation of *Inter-Pausal Units* (IPU) is a viable option, since they are clearly defined and easily detectable in recordings without much background noise. If the minimal pause length is defined, the only misidentification might be extremely long plosive closure durations (e.g., in hesitations). Mismatches between communicative functions and IPUs might cause problems, particularly if one IPU includes several communicative functions, or if a communicative function stretches over more than one IPU.

Annotation labels The set of annotation labels should be suitable for the description of entire conversations without encompassing too many categories in order to reduce potential confusion by the annotator or the ASR system. For turn management, we base our set of labels on four categories used in Zellers (2016), which are defined in terms of CA, i.e., according to the behaviour of participants in the conversation: *Hold* (same speaker continues talking), *Change* (speaker change), *Question* (speaker transfers the turn to another speaker), and *Hearer Response Tokens* (e.g., backchannels; cf., Sikveland (2012)). On the IPU level, three additional labels were necessary to capture incomplete structures before pauses; for incomplete turn-holds (cf., Figure 2b), turn-changes, and in turn competition when one speaker interrupts himself/herself to cede the turn to the other speaker. The annotation of intervals on the PCOMP level is more fine-grained. Thus, we added six labels to the system used in Zellers (2016). We subdivided *Hold* depending on the following context (continuation of syntactic structure vs. new sentence). For the same reason as on the IPU level, we added a label for incomplete turn-changes. A label for incomplete turn-holds is not necessary because no boundaries are set until a PCOMP is reached (cf., Figure 2a). We added labels for collaborative finishes to cap-

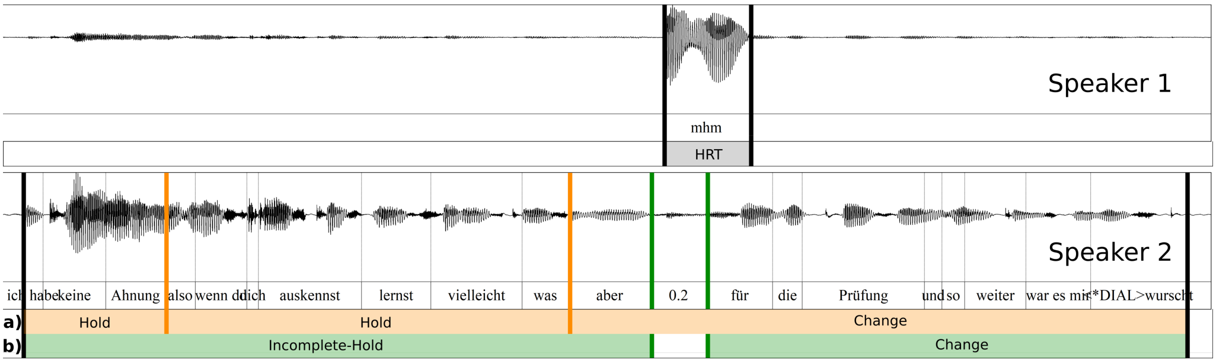

Figure 2: Time-aligned annotation of Speaker 2's turn (engl. 'I have no idea, so if you know your way around maybe you can learn something, but (0.2) for the exam and so on I didn't care.') a) at PCOMPs (orange); and b) of IPUs (green). Speaker 1 aligns his hearer response token with Speaker 2's pause after <aber>.

ture when a syntactic construction stretches over two speakers, and for discourse particles and hesitations that occur at PCOMP boundaries. Finally, we added a label for self-interruptions with subsequent rephrasing. These points in a speaker's turn are not technically PCOMPs, but they are often marked by an abrupt interruption of articulation and the syntactic reset after this point can be relevant for ASR.

Figure 2 shows an example of how PCOMP and IPU annotations are mapped onto each other. In this example, Speaker 2 holds his turn by making a pause at a point of "maximum grammatical control" (Schegloff, 1998: 241; labelled as *Incomplete-Hold* on tier b) after the introduction of a new sentence by <aber>, and completes his turn after the pause. There are two PCOMPs leading up to the pause (labelled as *Hold* on tier a), neither of which give the impression of being complete based on prosody (i.e., slightly rising pitch in <Ahnung> and 'rush-through' in <was>). Even though a pause is produced after <aber>, the next PCOMP is reached only after <wurscht>. Thus, the whole sentence starting with <aber> is grouped into one PCOMP chunk, regardless of any pauses. Speaker 1 times his backchannel (labelled as *Hearer Response Token*) with the pause rather than with the PCOMP just before <aber>. It is predominantly short hearer response tokens that are aligned with pauses at syntactically incomplete positions while participants almost never self-select to produce a longer turn in these positions.

Currently, 90 minutes in 15 conversations have been annotated at the IPU level and the last revision of these labels is in progress. On the PCOMP level, 60 minutes in 12 conversations are being annotated. These annotations are useful for the goals described above, i.e., for application in ASR and for phonetic studies, as well as for the investigation of various hypotheses about the time alignment of hearer response tokens and self-selection.

Outlook

An iterative annotation process while creating manual annotations and developing a classification tool based on acoustics will reveal more fine-grained categories (e.g., a distinction between PCOMPs that are prosodically marked as complete vs. prosodies overarching several PCOMPs). The annotation of more acoustically based categories will, in turn, improve recognition of categories. For instance, we can investigate the prosody at the end of IPUs. In a preliminary study, we performed a Random Forest classification of *Hold*, *Incomplete-Hold* and *Change* on the basis of acoustic features. An analysis of the highest ranked features in the Random Forest with linear mixed effects regression models indicated that *Incomplete-Hold*s (cf., Figure 2b) are characterized by a lower speech rate and a flatter F0 curve at the end. *Hold*s and *Change*s, on the other hand, were not consistently distinguished by prosody. The planned perception experiment of these categories will give us further insights into prosodically different kinds of turn-holds and turn-changes. The developed classifier of communicative functions will aid the annotation process by providing labels for semi-automatic annotations and will also be incorporated into our ASR system to improve word recognition.

Acknowledgments

B. Schuppler was funded by grant V-638-N33, and A. Kelterer by grant P-32700-N, both from FWF.

References

Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. 2016. Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. In *Proc. IWSDS*, pages 1–12.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation Journal*, 44:387–419.

Valerie Freeman. 2019. Prosodic features of stances in conversations. *Laboratory Phonology: The Journal of the Association for Laboratory Phonology*, 10(1):1–20.

Jan Gorisch, Bill Wells, and Guy J. Brown. 2012. Pitch contour matching and interactional alignment across turns: An acoustic investigation. *Language and Speech*, 55(1):57–76.

Augustín Gravano, Stefan Benus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Interspeech*, pages 1613–1616.

Sarah Hawkins and Rachel Smith. 2001. Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistic*, 13(1):99–188.

IPDS. 1997. CD-ROM: The Kiel Corpus of Spontaneous Speech, vol i- vol iii. Available at http://www.ipds.uni-kiel.de/forschung/kielcorpus.de.html (last viewed 25/11/2016).

Hitoshi Ito, Aiko Hagiwara, Manon Ichiki, Takeshi Mishima, Shoei Sato, and Akio Kobayashi. 2017. End-to-end speech recognition for languages with ideographic characters. In *Proc. APSIPA Annual Summit and Conference*.

Dan Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Tailor, and Carol Van Ess-Dykema. 1998. *Switchboard Discourse Language Modeling Project Report*, volume 1. Center for Speech and Language Processing, Johns Hopkins University, Baltimore.

Julian Linke, Anneliese Kelterer, Markus Dabrowsky, Dina El Zarka, and Barbara Schuppler. 2020. Towards automatic annotation of prosodic prominence levels in Austrian German. In *Proceedings of Speech Prosody 2020*, pages 1000 – 1004.

John Local and Gareth Walker. 2012. How phonetic features project more talk. *JIPA*, 42:255–280.

Richard Ogden, editor. 2012. *The Phonetics of Talk in Interaction, Special Issue in Language and Speech 55(1)*.

Emmanuel A. Schegloff. 1998. Reflections on studying prosody in talk-in-interaction. *Language and Speech*, 41(3-4):235–263.

Thomas Schmidt. 2014. Gesprächskorpora und gesprächsdatenbanken am beispiel von folk und dgd. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion*, 15:196–233.

Barbara Schuppler, Mirjam Ernestus, Wim van Dommelen, and Jacques Koreman. 2010. Predicting human perception and ASR classification of word-final [t] by its acoustic sub-segmental properties. In *Proceedings of Interspeech*, pages 2466 – 2469.

Barbara Schuppler, Martin Hagmüller, and Alexander Zahrer. 2017. A corpus of read and conversational Austrian German. *Speech Communication*, 94C:62–74.

Barbara Schuppler, Martin Hagmüller, Juan Cordovilla, and Hannes Pessentheiner. 2014. Grass: The graz corpus of read and spontaneous speech. In *9th edition of the Language Resources and Evaluation Conference*, pages 1465–1470.

Barbara Schuppler and Bogdan Ludusan. 2020. An analysis of prosodic boundary detection in German and Austrian German read speech. In *Proceedings of Speech Prosody 2020*, pages 990 – 994.

Margaret Selting. 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6(3):371–388.

Sanjay Shakkottai, Theodore S. Rappaport, and Peter C. Karlsson. 2003. Cross-layer design for wireless networks. *IEEE Communications Magazine*, 41(10):74–80.

Rein Ove Sikveland. 2012. Negotiating towards a next turn: Phonetic resources for 'doing the same'. *Language and Speech*, 55(1):77–98.

Nigel Ward. 2004. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody*, pages 325–328.

Margaret Zellers. 2016. Prosodic variation and segmental reduction and their roles in cuing turn transition in swedish. *Language and Speech*, 60(3):454–478.

Margaret Zellers and Brechtje Post. 2011. Combining formal and functional approaches to topic structure. *Language and Speech*, 55(1):119–139.

Contextual Choice between Synonymous Pairs of Metaphorical and Literal Expressions: An Empirical Study and Novel Dataset to *tackle* or to *address the question*

Prisca Piccirilli and Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany
{prisca.piccirilli,schulte}@ims.uni-stuttgart.de

Abstract

Research on metaphorical language detection and interpretation has produced a large number of resources mainly focusing on metaphoric vs. literal uses of specific expressions, and on metaphor paraphrases. As to our knowledge, however, no existing NLP resource provides a basis for understanding the choice between a synonymous pair of a literal and a metaphorical expression. E.g., why would one favor the use of *grasp a term* over *understand a term* in a given context, and does the preceding context prime for one or the other usage? We address these questions and provide an empirical study and a novel resource: Based on 50 pairs of English synonymous literal/metaphorical verb–object and subject–verb expressions in discourse, we asked participants in crowdsourcing experiments to (1) rate the degree of metaphoricity of a discourse, and (2) choose the expression that fits best. Our resource contains a total of 1,000 discourses and is ready to be exploited for computational research on discourse conditions for metaphorical vs. literal expression choices.

1 Introduction

Metaphors represent a ”necessary, not just nice” element of everyday thought and communication (Ortony, 1975; Lakoff and Johnson, 1980; van den Broek, 1981; Schäffner, 2004), and frequently manifest themselves in general-domain text corpora (Gedigian et al., 2006; Shutova and Teufel, 2010). Accordingly, metaphors pose a real challenge across NLP applications, and research on metaphorical language detection and interpretation has produced a large number of resources. Up to now, however, there is no empirical study providing a basis for understanding the choice between a synonymous pair of a literal and a metaphorical expression, when they can be used interchangeably in a given context. Consider the following discourse: ”For her, writing is an effective tool to express your viewpoints [...] To write is already to choose, thus, writing should be done along with a critical mind and a caring soul. [...] Reading lets her travel to far-off imagined places and situations.” This discourse might be followed by ”She also learns a lot from *(i) devouring/(ii) reading books*, especially from the socio-political and historical ones.”, where both (i) and (ii) seem equally acceptable.

The underlying choice leads to the following research questions: Why would one favor the metaphorical expression *devour a book* over its literal alternative *read a book*, or vice versa? Is the choice driven by the preceding context? If so, to which extent? These are necessary questions to tackle in order to build a robust NLP system for predicting which choice fits best in a given discourse. Extending the context-induced hypothesis (Kövecses, 2009) to metaphorical vs. literal usage, contextual salience would expect a metaphorical discourse preceding a metaphorical expression, and a literal context preceding a literal expression.

The current paper addresses the above questions by collecting and analyzing judgements on 1,000 instances of 50 pairs of English synonymous literal vs. metaphorical verb–object and subject–verb expressions in corpus-extracted discourses. In Task 1, asking participants to rate the degree of metaphoricity of the discourses sheds light on whether the preceding context plays a role in the choice of metaphorical vs. literal expressions within that discourse. Task 2, in which annotators provide a binary decision that favors one usage over the other, provides insight on the metaphorical vs. literal usage in context. To our knowledge, our work constitutes the first empirical study on conscious discourse-embedded choices about synonymous pairs of metaphoric vs. literal expressions. Our novel dataset constitutes a solid starting point for computational research on salient discourse conditions for contextual metaphorical vs. literal usage.

2 Related Work

Theoretical Background Different metaphor theories were broadly discussed in linguistics and philosophy, first as an attempt to understand what metaphors are. In parallel, researchers looked at what drives people to use metaphors (Glucksberg, 1989; Kövecses, 2010; Ortony, 1975) as well as ”what metaphor actually does” (Hampe, 2017). The cognitive linguistics view of metaphors in Conceptual Metaphor Theory (Lakoff and Johnson, 1980) describes how metaphors are frequently used everyday and by everybody, and moreover in an unintentional way.

A corpus study by Stefanowitsch (2006) provides evidence for this view. He shows that metaphors are used not only as a stylistic choice but also as a cognitive function, since people seem to use them to explicate things or reasonings. However, further studies show clear signs that metaphors can be of stylistic choice, e.g. metaphorical language has a stronger emotional impact than literal language (Mohammad et al., 2016; Köper and Schulte im Walde, 2018). This statement gives support to the idea that there exists a difference in choice between metaphorical versus literal expressions.

In their psycholinguistics study, Thibodeau and Boroditsky (2011) also illustrate that using metaphors over their literal alternatives may influence the way humans conceptualize an act. While on the one hand it seems like people do feel a difference when using one version or the other, on the other hand it also seems that it affects the way people react. It is therefore necessary to find a way for a computational system to capture this difference.

Existing Resources Stefanowitsch (2006) provides a corpus-based study using carefully collected and curated data. He explores whether the use of metaphors is a stylistic choice or a cognitive function, and relies on sentences where both the metaphorical expression and a literal alternative may be used (e.g. *in the heart of* versus *in the center of*). His examples are close to what we aim for in our dataset, but his study is based solely on a handful of metaphorical expressions.

The NLP tasks of figurative language detection and interpretation have led to the creation of several datasets. Mohammad et al. (2016) composed 171 sentences where a verb is used metaphorically, e.g. *abuse* in ”Her husband often abuses alcohol”. For each sentence, the authors of the paper chose a literal synonym of the target verb, such as *drink* in the above example sentence.

Shutova (2010) aimed for metaphor interpretation and collected sentences containing metaphorical verbs from the British National Corpus, e.g. *grasp* in ”Anyone who has introduced speech act theory to students will know that these technical terms [...] are not at all easy to grasp.” She asked annotators to provide an alternative verb with a literal meaning. The dataset consists of a list of metaphorically-used verb–object and subject–verb expressions, with one or more literal verb alternatives. For the verb–object expression *grasp term*, the verb *grasp* was given the literal alternatives *understand* and *comprehend*, for example.

Similarly, the model developed by Bizzoni and Lappin (2018) automatically ranks the best four paraphrases for each metaphorical sentence. The final dataset consists of 200 metaphorical sentences, each with their four automatically generated and ranked paraphrases.

The setup of the latter three datasets is what we were looking for; however, all present only a one-sentence context, which in our opinion is not sufficient when addressing the importance of preceding discourse. Moreover, the dataset from Bizzoni and Lappin (2018) automatically generated the literal alternatives, so they would require additional careful human judgements. Even though we were inspired from all these useful resources, we have not been able to find an existing dataset that can be fully used for our goals.

3 Experimental Setup

3.1 Compiling pairs of expressions

We collect 50 pairs of expressions from Shutova (2010) and Mohammad et al. (2016), 36 of which are verb–object (VO) expressions, and 14 of which are subject–verb (SV) expressions. Our corpus thus consists of 50 expressions where the verb is used metaphorically, and 50 expressions where the verb is a synonymous literal alternative, such as *tackle/address question* for a VO expression and *tension mount/increase* for a SV expression. As the original datasets were created for different purposes, we perform slight changes in some cases. For instance, we exchange *catch contagion* by the more common version *catch disease*. We provide an overview of all pairs in the Appendix B.

(a)	It is true indeed that not a sparrow drops unnoticed by the Mind of THE ALL - that even the hairs on our head are numbered, as the scriptures have said. There is nothing outside of Law; nothing that happens contrary to it. And yet, do not make the mistake of supposing that man is but a blind automaton - far from that. The Hermetic Teachings are that man may use Law to overcome laws, and that the higher will always prevail against the lower, until at last he has reached the stage in which he seeks refuge in the LAW itself, and laughs the phenomenal laws to scorn. Are you able to **grasp** the inner **meaning** of this?
(b)	This wasn't just a play on words, rather it was a demand that they should 'maintain a consistency between their words and their actions'. But I agree, that still does not absolve them from the need to speak truth to power. In our times when people spend so much time with TV and the internet, do they have the interest and time to read poetry? Many people believe that it is difficult to read poetry. Can everyone **understand** the **meaning** of a good poem, or is a skill necessary?

Table 1: Examples of discourses for the synonymous pair *grap/understand meaning*. The metaphorical expression *grasp meaning* is used in (a), its literal paraphrase *understand meaning* is used in (b), and both are applicable in both contexts.

3.2 Collecting discourses

We automatically extract all sentences from the ukWaC (Baroni et al., 2009) containing inflected forms of our compiled expressions, with a maximum of 25 characters in between the verb and its argument in VO/SV. We select 20 instances for each pair of expressions, with 10 instances each for the metaphorical/literal versions. Our dataset thus contains a total of 1,000 discourses. As we are testing the extent to which context plays a role in favoring one expression over the other, we extract four to five sentences preceding the sentence containing the target expression, followed by the actual sentence with the metaphorical/literal expression. The discourses contain 31–216 words, with an average of 98 words. Table 1 shows examples of discourses for a pair of expressions.

3.3 Crowdsourcing experiments

As we are interested in (i) the influence of context in the choice of a target expression and (ii) human preferences for metaphor vs. literal expressions, the annotation process is directed in two tasks.

Task 1: The first task tests for the **degree of metaphoricity vs. literalness** of the expression-preceding discourse, in order to answer the question "Does the discourse influence the choice of a metaphorical/literal expression?" To obtain a minimum of 10 ratings per instance, we present the discourses up to the word preceding the target expression to 15 workers on Amazon Mechanical Turk (AMT)[1] and ask them to indicate on a scale from 1 to 6 where they judge the overall discourse on the range between mostly literal–mostly metaphorical.

Task 2: The second task tests **which expression (metaphorical vs. literal) is favored in a given discourse**, in order to answer the question "Does one favor the use of a metaphorical vs. a literal expression given a specific (metaphorical/literal) preceding discourse?" As in Task 1, we show the discourses to 15 AMT workers, however now including the target sentence but with a blanked spot for the target expression, and ask them to choose which expression fits better (binary choice).

For both tasks, we limit the location of the workers to English-speaking countries, and specify in the instructions that the tasks are only for English native speakers. A total of 183 workers annotated the 1,000 Task-1 instances, on average providing ratings for 81 instances. We disregard 73 workers who completed less than 20 instances, which results in a total of 14,514 judgements by 110 workers on rating the degree of metaphoricity of the discourses. Each instance is rated by at least 11 workers. For Task 2, 238 workers completed 63 instances on average. Similarly to Task 1, we only keep the 136 workers who completed at least 20 instances, which results in a final dataset with 14,378 judgements. Appendix A provides a detailed explanation and examples of the setup of the AMT experiments.

4 Results and data analyses

Task 1: Table 2 shows the workers' ratings on the degree of metaphoricity of the expression-preceding discourses, across all 14,514 judgements, next to the resulting medians for our 1,000 instances. We can see a clear preference of the workers for the middle rather than the extreme categories (1 for mostly literal and 6 for mostly metaphorical), with a slight preference for metaphoricity (also see top part in Table 3, using 3.5 as threshold for literal vs. metaphorical categorisation). For the medians this results in a strong focus on the range 3–4.

Scale	#Ratings	#Median
1	1,905	3
2	2,147	35
3	2,689	340
4	3,496	536
5	3,101	86
6	1,176	0
Total	14,514	1,000

Table 2: Number of ratings and medians across scale.

[1] https://www.mturk.com/

metaphoricity of discourse (annotated)	
metaphorical	622 (62.2%)
literal	378 (37.8%)
metaphoricity of expression (annotated)	
metaphorical	425 (45.6%)
literal	506 (54.4%)
metaphoricity of discourse (annotated–original)	
metaphorical – metaphorical	315 (31.5%)
literal – literal	193 (19.3%)
metaphorical – literal	307 (30.7%)
literal – metaphorical	185 (18.5%)
metaphoricity of discourse–metaphoricity of expression	
metaphorical – metaphorical	260 (27.9%)
literal – literal	187 (20.1%)
metaphorical – literal	319 (34.3%)
literal – metaphorical	165 (17.7%)

Table 3: Summary and comparison of annotations for Tasks 1 and 2 and the original corpus-based discourses+expressions. *Top part:* proportions of annotated metaphorical vs. literal discourses + proportions of annotated metaphorical vs. literal expressions. *Middle part:* metaphoricity of discourses in comparison to metaphoricity of original corpus expression. *Bottom part:* metaphoricity of annotated discourses in relation to metaphoricity of annotated expressions. Threshold for literal/metaphorical categories: median of 3.5.

According to the contextual salience hypothesis, we expect that metaphorically-rated discourses are more likely to be followed by a metaphorical expression, and ditto for literal discourses and expressions. In the middle part of Table 3 we looked at the expressions that were originally collected from the corpus, and compared their categorizations to the discourse ratings. Judging from the discourses where a metaphorical expression was used (*–metaphorical), one may induce that the context-salient hypothesis is valid, since raters mostly judged the respective discourses as metaphorical: 31.5% vs. 18.5%. However, for discourses where a literal expression was used (*–literal), the context-salient hypothesis fails, as raters favored a metaphorical context preceding a literal usage: 30.7% vs. 19.3%.

Task 2: Across 931 binary judgements with an absolute majority (67 discourses are a tie), we note a slight preference for literal expressions (54.4% vs. 45.6%). Only when looking at the individual 50 pairs (see Appendix) we find a more diverse picture. Cases where the literal usages were preferred (e.g., *devour/read book, suck/attract worker, attack/solve problem*) may be explained by the rather strong emotional effect of the metaphorical expressions, cf. Mohammad et al. (2016), if they are not coherent with the context. A preference for the metaphorical expressions, as in *breathe/instill life, painting capture/represent, fate lie/be*, may be explained by the high conventionality of these metaphors.

Combining Tasks 1 and 2: As demonstrated above, the metaphorically- vs. literally-rated contexts were not necessarily in accordance with the original choice of expression in the corpus data. In the bottom part of Table 3 we bring Task-1 and Task-2 results together and relate the binary metaphoricity judgements of the target expressions to the judgements of the respective preceding discourses. As before, we observe tendencies against the context-salient hypothesis: while on the one hand a metaphorically-rated discourse seems to prime for the use of a metaphorical expression (27.9%), it similarly primes for the use of a literal expression (34.3%); on the other hand, only 20.1% of the literal expressions are preceded by literally-judged discourses. Figure 1 shows the average proportions of metaphorical vs. literal expressions across the median ratings for our 1,000 instances. While we observe a slight downward trend for choosing literal expressions – and, in parallel, a slight upward trend for choosing metaphorical phrases – with increasing medians (i.e., when the discourse is rated more metaphorical), we also note that literal expressions are favored over metaphorical ones across all medians (i.e., irrespective of the metaphoricity of the discourse.)

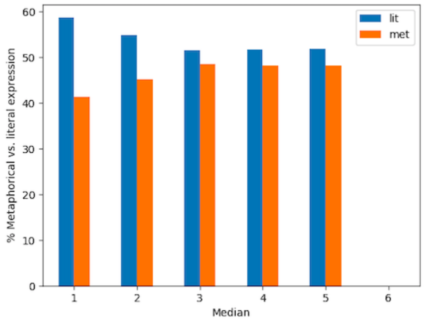

Figure 1: Proportions of metaphorical vs. literal expressions across median ratings for all 1,000 discourse instances.

5 Conclusion

This work offers a new approach and dataset to study metaphor vs. literal language usage in relation to discourse embedding. Our collection counters the theoretical context-salient hypothesis that metaphorical vs. literal usage is expected to be primed by metaphorical vs. literal preceding contexts, respectively. Even more so, it provides a valuable starting point for computational explorations on further discourse conditions for metaphorical vs. literal choices, such as lexical semantic relatedness (Birke and Sarkar, 2006; Sporleder and Li, 2009; Do Dinh, 2013) and contextual abstractness (Turney et al., 2011; Tsvetkov et al., 2014).

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Julia Birke and Anoop Sarkar. 2006. A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, pages 329–336, Trento, Italy.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Erik-Lan Do Dinh. 2013. Automatic Identification of Novel Metaphoric Expressions. Master's thesis, Technische Universität Darmstadt.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching Metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York City, NY.

Sam Glucksberg. 1989. Metaphors in conversation: How are they understood? Why are they used? *Metaphor and Symbolic Activity*, 4(3):125–143.

Beate Hampe, editor. 2017. *Metaphor: Embodied Cognition and Discourse*. Cambridge University Press, Cambridge, UK.

Maximilian Köper and Sabine Schulte im Walde. 2018. Analogies in Complex Verb Meaning Shifts: The Effect of Affect in Semantic Similarity Models. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 150–156, New Orleans, LA, USA.

Zoltan Kövecses. 2009. The Effect of Context on the Use of Metaphor in Discourse. *Iberica*, 17:11–24.

Zoltan Kövecses. 2010. *Metaphor: A Practical Introduction*, 2nd edition. Oxford University Press, New York.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Saif M. Mohammad, Ekaterina Shutova, and Peter D. Turney. 2016. Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany.

Andrew Ortony. 1975. Why metaphors are necessary and not just nice. *Educational Theory*, 25(1):45–53.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding Semantics to Data-Driven Paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1512–1522, Beijing, China.

Christina Schäffner. 2004. Metaphor and Translation: Some Implications of a Cognitive Approach. *Journal of Pragmatics*, 36:1253–1269.

Ekaterina Shutova. 2010. Automatic Metaphor Interpretation as a Paraphrasing Task. In *Proceedings of Human Language Tehcnologies: The Annual Conference of the North American Chapter of the ACL*, pages 1029–1037, Los Angeles, CA, USA.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta. European Language Resources Association.

Caroline Sporleder and Linlin Li. 2009. Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 754–762, Athens, Greece.

Anatol Stefanowitsch. 2006. Words and their Metaphors: A Corpus-based Approach. In Anatol Stefanowitsch and Stefan Th. Gries, editors, *Corpus-Based Approaches to Metaphor and Metonymy*, pages 17–35. de Gruyter, Berlin.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors We Think With: The Role of Metaphor in Reasoning. *PLoS ONE*, 6(2).

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258, Baltimore, MD, USA.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Raymond van den Broek. 1981. The Limits of Translatability Exemplified by Metaphor Translation. *Poetics Today*, 1(4):73–87.

Appendices

A Setup of crowdsourcing experiments

In this section, we provide a detailed explanation of the AMT tasks setup.

We randomly shuffled the 1,000 discourses composing our dataset, and created 50 batches of 20 instances. AMT workers were asked to complete all instances of a batch, and were allowed to complete as many batches as they wished. We discarded workers who completed less than 20 instances.

We provide below an example of what the workers were shown when completing each task.

A.1 Task 1

Workers were asked to rate the degree of metaphoricity they overall perceived, when reading the discourse preceding the target expression. On purpose, we did not give them a definition of what a *metaphor* or a *metaphorical discourse* is, in order to not bias them. Previous cognitive research has shown that metaphors and metaphorical concepts are used without even being aware of using them (Lakoff and Johnson, 1980), and this is what this study has attempted to look at. Below is an example of a discourse that AMT workers were asked to complete:

How metaphorical / literal is this discourse according to you? Please read the following text carefully and rate the degree of literalness or metaphoricity of the discourse from 1 to 6, where 1 means that the discourse is mostly literal and 6 means that the discourse is mostly metaphorical.

> The fact that there's a lunar eclipse that day heightens that need. Indeed, it could be that your focus in the next fortnight (until the solar eclipse on 22nd) will be very strong indeed. True, this could be triggered by unexpected events on Monday which underline the need for change. And true, you've had several sidewinders thrust your way in recent years - and these haven't left you racing for more education. Now though, you may want to [...]

mostly literal 1 2 3 4 5 6 mostly metaphorical

A.2 Task 2

Workers were asked to choose which expression they believe fits best in the given discourse.

Please fill in the blank. Pick the option that fits best in this discourse according to you:

> Why is Saddam Hussein pushing ahead with weapons of mass destruction if at some point he is not going to use them? It's certainly got to be a factor in all of this. Unlike anthrax, the bacteria used in last year's unsolved mail attacks, the highly contagious smallpox virus can be passed from person to person. The virus causes ugly pustules to form both on the skin and inside the mouth and throat. About a third of unvaccinated people who ——————- would die.

Which expression fits best in the blank?
1) caught the disease 2) got the disease

A.3 Inter-annotator agreement and standard deviation

As we obtained over 14,000 judgements from 110 workers (Task 1), we calculated IAA in the same way done by Pavlick et al. (2015), i.e., computing IAA as the average of each rater's correlation with the average of all other workers. We reach Spearman's $\rho = 0.26$, which might seem to be rather low but this is IAA for 110 workers, rather than 7 (as in Pavlick et al. (2015)). Figure 2 gives an idea of annotation reliability as it represents the dispersion from the individual data values to the mean score of each instance. Overall, it seems that raters tend to agree on extreme cases, i.e., agreement is high on rating mostly literal (lower left corner) and mostly metaphorical (upper right corner) discourses. Agreement varies across the in-between average values but stays rather reasonable.

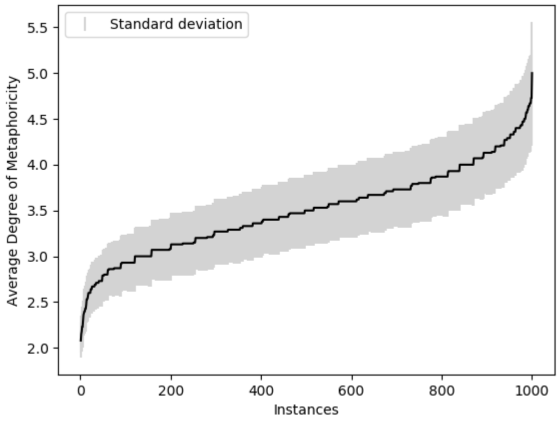

Figure 2: Sorted average values (in black) from all workers across the 1,000 instances. The grey cloud represents the standard deviation values for each average value.

B Summary Tasks 1 and 2

Metaphorical (met)/Literal expressions (lit)	**#metexp** (%)	(a)#met context metexp	(b)#lit context metexp	**#litexp** (%)	(c)#met context litexp	(d)#lit context litexp
subject–verb pairs (SV):						
example illustrate/show	173 (59.45%)	78	66	118 (40.55%)	72	73
fire devour/destroy	127 (44.10%)	90	57	161 (55.90%)	71	72
factor shape/affect	124 (43.66%)	66	84	160 (58.51%)	74	72
painting capture/represent	188 (65.73%)	86	62	98 (34.27%)	78	69
tension mount/increase	146 (51.41%)	83	66	138 (48.59%)	78	70
mess reflect/show	156 (53.98%)	80	67	133 (46.02%)	82	65
moon peep/appear	111 (38.14%)	105	41	180 (61.86%)	86	57
fate lie/be	201 (69.07%)	83	64	90 (30.93%)	75	70
colour harmonise/match	132 (47.31%)	86	59	147 (52.69%)	74	72
story grab/intrigue	125 (44.80%)	91	53	154 (55.20%)	86	58
distinction blur/disappear	143 (50.00%)	72	75	143 (50.00%)	76	67
view reflect/represent	151 (52.80%)	76	68	135 (47.20%)	68	76
result emerge/appear	170 (59.44%)	59	87	116 (40.56%)	75	70
war uproot/displace	113 (38.97%)	78	67	177 (61.03%)	72	72
verb–object pairs (VO):						
mount/organise production	113 (38.57%)	68	78	180 (61.43%)	74	72
recapture/recall feeling	138 (48.08%)	81	64	149 (51.92%)	89	54
grasp/understand meaning	129 (44.48%)	88	55	161 (55.52%)	84	62
drown/forget trouble	147 (50.17%)	96	45	146 (49.83%)	90	56
catch/get disease	162 (55.67%)	83	65	129 (44.33%)	71	74
breathe/instill life	193 (66.55%)	78	67	97 (33.45%)	78	67
flood/saturate market	146 (51.05%)	70	71	140 (48.95%)	74	67
push/urge someone	107 (37.54%)	74	73	178 (62.46%)	73	69
stir/cause excitement	169 (59.09%)	76	68	117 (40.91%)	69	70
cast/cause doubt	191 (65.41%)	66	82	101 (34.59%)	81	67
leak/disclose report	128 (44.60%)	76	73	159 (55.40%)	70	76
devour/read book	98 (34.15%)	81	67	189 (65.85%)	89	56
suck/attract worker	83 (29.23%)	71	76	201 (70.77%)	78	70
dull/decrease appetite	132 (45.52%)	77	69	158 (54.48%)	66	78
frame/pose question	119 (41.18%)	61	80	170 (58.82%)	79	67
abuse/drink alcohol	152 (53.90%)	77	63	130 (46.10%)	75	70
juggle/manage job	158 (55.05%)	79	65	129 (44.95%)	67	77
attack/solve problem	95 (32.31%)	76	71	199.(67.69%)	75	72
disown/reject past	164 (56.94%)	89	53	124 (43.06%)	83	62
pour/invest money	119 (42.20%)	78	70	163 (57.80%)	76	68
follow/practise profession	110 (38.46%)	69	76	176 (61.54%)	63	83
taste/experience freedom	103 (35.27%)	93	54	189 (64.73%)	83	62
break/end agreement	161 (55.71%)	74	69	128 (44.29%)	72	71
sow/instill doubt	138 (47.59%)	73	69	152 (52.41%)	83	64
twist/misinterpret word	151 (53.17%)	92	55	133 (46.83%)	83	60
boost/improve economy	145 (50.88%)	67	81	140 (49.12%)	77	68
throw/make remark	108 (38.03%)	83	63	176 (61.97%)	71	72
tackle/address question	113 (38.18%)	85	62	183 (61.82%)	79	67
buy/believe story	141 (48.62%)	83	62	149 (51.38%)	84	61
fuel/stimulate debate	169 (58.28%)	68	77	121 (41.72%)	76	67
float/discuss idea	120 (41.24%)	67	80	171 (58.76%)	62	78
wear/have smile	175 (61.62%)	89	58	109 (38.38%)	90	55
poison/corrupt mind	142 (48.97%)	71	73	148 (51.03%)	78	67
shape/determine result	118 (41.40%)	82	64	167 (58.60%)	92	54
colour/affect judgement	114 (39.72%)	79	68	173 (60.28%)	71	73

Table 4: Summary of results when combining Tasks 1 and 2. For each pair of metaphorical/literal expression, in order: number of times the metaphorical expression was chosen (%); number of times the preceding context of the metaphorical expression was rated as (a) metaphorical vs. (b) literal; number of times the literal expression was chosen (%); number of times the preceding context of the literal expression was rated as (c) metaphorical vs. (d) literal.

Combined discourse representations: Coherence relations and questions under discussion

Arndt Riester
Department of Linguistics

University of Bielefeld
arndt.riester@uni-bielefeld.de

Amalia Canes Nápoles
Department of Romance Studies
University of Cologne
acanesna@uni-koeln.de

Jet Hoek
Department of Language and Communication
Radboud University Nijmegen
jet.hoek@ru.nl

Abstract

We analyze a text according to three different discourse theories; CCR, RST and QUD trees. We discuss differences with respect to segmentation and show how coherence relations can be mapped onto a discourse representation based on questions under discussion.

1 Introduction

The term *discourse structure* comprises issues relating to the organization and coherence of written or spoken, monologic or multi-speaker discourse. The central, recurring problems related to discourse-structure analysis involve (i) *discourse segmentation*, i.e. the rules that determine which spans of a text form elementary, independent discourse units, (ii) *attachment*, i.e. the question which units are (recursively) grouped together, thereby forming paragraphs and sections, and (iii) the *choice of discourse relations:* how many should be assumed, are they reducible to a set of abstract features?

In this paper, we will address the first and also, partly, the second problem, drawing on three different analyses of the same piece of discourse, within the Cognitive approach to Coherence Relations (CCR), Rhetorical Structure Theory (RST) and Question under Discussion (QUD) trees, briefly introduced in Section 2.[1] Specifically, we will be concerned with the following issues:

1. In Section 3 we discuss whether there are different levels of detail with regard to discourse segmentation. According to what rules are discourse units determined in different frameworks?

2. In Section 4 we compare analyses based on coherence relations (CCR, RST) with an approach that uses questions under discussion. Is the question-answer relation simply a type of coherence relation? Can all relations be represented by means of questions? Can the different tree structures be mapped onto each other?

2 Some frameworks for discourse structure

2.1 RST

Rhetorical Structure Theory (RST: Mann and Thompson, 1988; Taboada and Mann, 2006) is a framework which postulates that discourse can be analyzed in the form of tree structures whose terminal elements are so called *elementary discourse units*. These units are recursively connected by coherence relations. Stede et al. (2017) list 31 RST relations. Most of these relations will subordinate one discourse unit (the satellite) to a second one (the nucleus), indicating that the nucleus is more important. Other relations are multinuclear and therefore coordinating.

2.2 CCR

The Cognitive approach to Coherence Relations (CCR: Sanders et al., 1992, see Hoek et al., 2019 for an up-to-date version) is a taxonomy of discourse relations that uses cognitively relevant primitives to describe the type of relation that holds between discourse segments. It defines discourse relations as the meaning 'surplus' compared to the meaning of the discourse segments in isolation. While CCR is most commonly used to depict relations between individual discourse segments, the approach is compatible with depicting the hierarchical structure of an entire text (e.g., Sanders and Spooren, 2009). Since CCR, unlike RST, does not include a nuclearity principle, an entire relation is related to the rest of the text in the hierarchical

[1] Other important frameworks, not addressed in this short paper, include SDRT (Asher and Lascarides, 2003) and PDTB (Prasad et al., 2008).

discourse structure and attachment points are symmetrically located between two segments (similar to multinuclear relations in RST).

2.3 QUD trees

The QUD tree approach (Reyle and Riester, 2016; Riester et al., 2018; Riester, 2019) allows for a simultaneous analysis of the information structure and discourse structure of a text. The framework is based on the assumption that every assertion of a discourse (more precisely, its focus) is the answer to a typically implicit *question under discussion* (QUD: van Kuppevelt, 1995; Roberts, 2012). Inter-annotator agreement has been studied in De Kuthy et al. (2018). QUDs are hierarchically ordered, mirroring the topical organization of a discourse. As a result, QUDs are the non-terminal nodes of a tree structure, while all non-interrogative discourse segments are interpreted as complete or partial answers to their respective parent QUD node. The reconstruction of QUDs follows common principles of information structure theory (Rooth, 1992; Schwarzschild, 1999; Büring, 2016).

3 Discourse segmentation

3.1 Labeling conventions

In this section we discuss discourse segmentation in the different frameworks.[2] Examples are taken from a section of Barack Obama's famous keynote address at the Democratic National Convention, Boston, July 27, 2004.[3] In order to allow for cross-referencing despite different segmentations, we adopt the following conventions: since both the CCR and RST analyses identified the same 31 segments, we took those as a basis. Whenever the segmentation turned out more granular in the QUD analysis (47 segments in total), sub-labels (1a, 1b, . . .) were assigned, thereby indicating the segmentation differences between the approaches. Discourse relations translated into questions inherit all its immediate children indexes (Q25-29). QUDs as usual only inherit the indexes of their assertion children nodes (Q25a,27). The latter raises an exception in parallel structures where the super-question inherit the indexes of its sub-question children nodes (e.g. Q14,15,16). Since all questions of our sample discourse are implicit (i.e. reconstructed), no labeling conflicts arise.

3.2 Discourse segmentation in RST / CCR

Both RST and CCR take clauses as the basis for identifying discourse segments (see e.g., Stede et al., 2017 for RST and Hoek et al., 2017 for CCR). Exceptions to the clause-as-segment guideline apply, for instance, when a clause connects not to another clause, but to a noun phrase (e.g., some relative clauses), or when a clause does not relate to a complete other clause (e.g., clausal subjects). A discourse relation holds between a segment and another segment (e.g., 2 and 3 in Figure 1) or a group of segments (a complex tree node); for instance, segment 4 with segments 2 and 3. The hierarchical structure of an entire text includes all discourse segments.

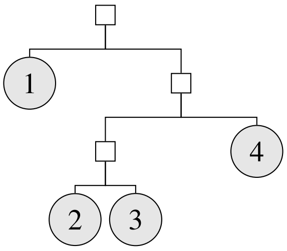

Figure 1: Sample CCR discourse tree

3.3 Discourse segmentation based on QUDs

Discourse segmentation in the QUD tree approach generally follows similar rules as in the aforementioned frameworks; in particular, it shares RST's and CCR's assumption that discourse segments are clauses or sentences, and that adjunct/complement but not argument clauses may form independent segments. However, the problem of segmentation is rephrased in terms of which chunks (not only main and adjunct clauses but also simpler adjuncts) can function as answers to an independent QUD. For instance, the complex sentence in (1) receives the QUD-structural representation in (2),[4] which is homomorphic to the tree in Figure 2.

(1) [They would give me an African name, Barack, or "blessed".]$_{25}$

[2] For the complete RST, CCR and QUD annotations, check http://bit.ly/osf_RST_QUD_CCR.
For an interactive QUD analysis enhanced with discourse relations, see http://bit.ly/ObamaQUD_RST_0-6.

[3] http://bit.ly/Obama_Keynote

[4] Information structure labels: CT: contrastive topic, F: focus, NAI: non-at-issue material, T: (non-contrastive) topic

(2) Q_{25a}: {What would Obama's parents do with him?}
$>$ A_{25a}: [They]$_{\mathrm{T}}$ would [give]$_{\mathrm{F}}$ [me]$_{\mathrm{T}}$ [an African name]$_{\mathrm{F}}$,
$>$ Q_{25b}: {What name would they give to him?}
$>>$ A_{25b}: [Barack,]$_{\mathrm{F}}$
$>>$ Q_{25c}: {What does *Barack* mean?}
$>>>$ A_{25c}: or ["blessed"]$_{\mathrm{F}}$.

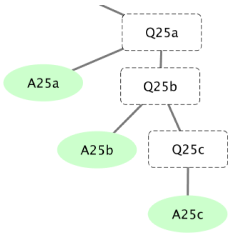

Figure 2: QUD tree for (2), with fine-grained segments

Example (2) shows that sentence [25] is divided into a main (or at-issue) discourse unit A_{25a}, whose denotation provides an answer to Q_{25a}, and two short (appositive, or non-at-issue) units, which do not answer Q_{25a} but instead the subquestions Q_{25b} and Q_{25c}. Another area in which the QUD-tree framework systematically requires a sub-clausal segmentation are NP- or VP-level coordinations, compare (3), analyzed as in (4).[5]

(3) [His father, my grandfather, was a cook, a domestic servant to the British]$_7$

(4) $Q_{7,8}$: {What about Obama's paternal grandfather?}
$>$ A_{7a}: [His father,]$_{\mathrm{T}}$ [my grandfather,]$_{\mathrm{NAI}}$ was [a cook,]$_{\mathrm{F}}$
$>$ A_{7b}: [a domestic servant to the British]$_{\mathrm{F}}$.

In the cases discussed, segmentation is motivated by information structure: every phrase containing a (contrastive) focus counts as a separate information unit, hence a discourse segment. Though these segments are smaller than discourse relation approaches generally allow, the link between these sub-clausal units can be captured by coherence relation labels such as ELABORATION (A_{25a}-A_{25b}), followed by a RESTATEMENT (A_{25b}-A_{25c}) in (2), and a LIST (A_{7a}-A_{7b}) in (4).

Informational backgrounding can occasionally also lead to the situation where adjunct clauses that are separate units according to CCR and RST – and are part of a, respectively, POSITIVE TEMPORAL SYNCHRONOUS and a CIRCUMSTANCE relation in Example (5) – are not separated from their matrix clause in the corresponding QUD analysis in (6). Since in (5), the information that the father studied *here* (i.e. in the US) is given information – thus, a non-informative statement – it is a non-autonomous part of the question background of $Q_{10,11}$.

(5) [While studying here,]$_{10}$ [my father met my mother.]$_{11}$

(6) $Q_{10,11}$: {What happened to Obama's father while he was studying in America?}
$>$ $A_{10,11}$: While studying [here,]$_{\mathrm{T}}$ [my father]$_{\mathrm{T}}$ [met my mother]$_{\mathrm{F}}$.

4 Mapping coherence relations onto QUD tree representations

In this section, we discuss cases that show how discourse relations can be integrated into QUD representations. Because of space limitations, only a few examples are shown. We discuss subordinating and coordinating relations separately. CCR does not make this distinction, so both solutions in Sections 4.1 and 4.2 could be applied to CCR trees.

4.1 Subordinating (hypotactic) relations

Subordinating discourse relations typically correspond to likewise subordinated (and typically anaphoric) QUDs; see the example of a REASON relation in Figure 3, which directly translates into the *why*-question in (7).

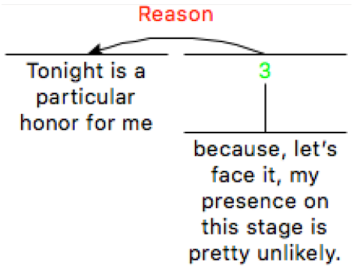
Reason
Tonight is a particular honor for me
3
because, let's face it, my presence on this stage is pretty unlikely.

Figure 3: RST representation of subordinating relation

[5] So far we ignore the likewise independent status of the clause-internal apposition *my grandfather* in A_{7a}, but see related comments in Riester, 2019, 180 ff.

(7) A_2: Tonight is a particular honor for me
Q_3: {Why is it a particular honor for Obama to speak on this stage?}
> A_3: because, [let's face it,]$_{\text{NAI}}$ [my presence on this stage]$_{\text{T}}$ [is pretty unlikely]$_{\text{F}}$.

Since the QUD tree approach is mainly concerned with the identification of information structural patterns, like topical continuity or contrastive parallelism, it can miss certain relations, like the concessive relation between sentences $A_{7a,b}$ and A_8 in Example (8). To capture the relation and represent its subordinating (RST) nature, we introduce an additional subquestion node *Q_8: What did the grandfather want for his son, despite being a servant?*, corresponding to the missing CONCESSION, see A versus B in Figure 4. We represent the relation as a link between the two question nodes.

(8) $Q_{7,8}$: {What about Obama's paternal grandfather?}
> A_{7a}: [His father,]$_{\text{T}}$ [my grandfather,]$_{\text{NAI}}$ was [a cook,]$_{\text{F}}$
> A_{7b}: [a domestic servant to the British]$_{\text{F}}$.
> A_8: But [my grandfather]$_{\text{T}}$ [had larger dreams]$_{\text{F}}$ for [his son]$_{\text{T}}$.

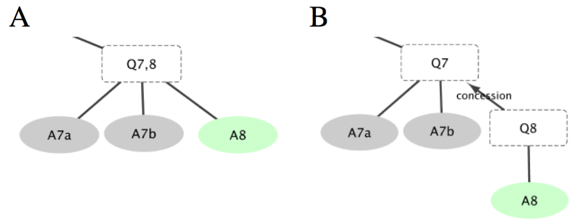

Figure 4: Inserting a subordinating discourse relation into a QUD tree

4.2 Coordinating (paratactic) relations

Coordinating RST relations, such as LIST, JOINT, DISJUNCTION or CONJUNCTION, are also easily translated into QUD structures: a QUD node dominates all coordinated segments, which are, in turn, interpreted as denoting partial answers to the QUD. In order to account for the slightly more complex meaning expressed by (adversative) CONTRAST or (temporal) SEQUENCE, we make use of subquestions and contrastive topics, as proposed by Büring (2003), Riester et al. (2018, 422ff.). For example, the RST SEQUENCE in Figure 5 corresponds to the original QUD analysis in (9) below.

(9) $Q_{14,15,16}$: {What did the grandfather do after the Pearl Harbor attack?}
> Q_{14}: {What did the grandfather do on the (exact) day after Pearl Harbor?}
> > A_{14}: [The day after Pearl Harbor]$_{\text{CT}}$ [my grandfather]$_{\text{T}}$ [signed up for duty,]$_{\text{F}}$
> A_{15}: [joined Patton's army,]$_{\text{F}}$
> A_{16}: [marched across Europe]$_{\text{F}}$

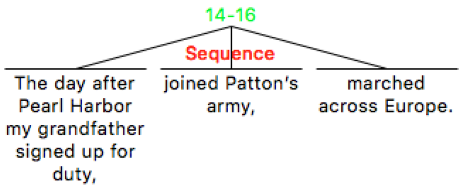

Figure 5: Example of a RST paratactic relation

For economic considerations, the analysis in (9) only contains a subquestion Q_{14} for the segment which contains an explicit temporal contrastive topic *(the day after Pearl Harbor)*. This results in a representation which is not yet entirely parallel. It is, however, permitted to add more subquestions that make the temporal background of each segment explicit. Each event takes place at its own topic time (cf. von Stutterheim and Klein, 1989; Klein, 1992), even if this is not always overtly expressed by an adverbial. The only caveat in this context is that the additional subquestions, in this case questions like *Q_{15}: What did the grandfather do then (at t_{15})?*, should not introduce any more specific information than their respective answers. The augmented representation corresponding to Figure 5 is shown in Figure 6.

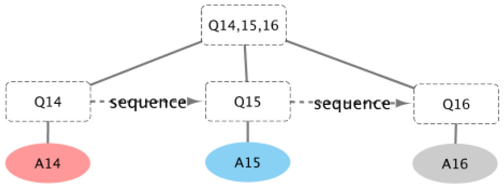

Figure 6: SEQUENCE relation expressed as a QUD tree

5 Conclusions

Analysing a discourse from different theoretical angles can bring many benefits. In our case, CCR and RST analyses, which capture discourse relations, are augmented with QUDs and information structure. The QUD approach also offers a more fine-grained but nevertheless pragmatically motivated discourse segmentation, which we intend to examine more closely in future work. On the other hand, by integrating discourse relations into QUD tree analyses, we can expect to improve and solidify the resulting discourse structures. The QUD tree framework generally allows for the introduction of additional, and potentially more specific, QUDs, which, of course, has an impact on (the representation of) discourse structure itself. By considering coherence relations, we may expect the introduction of these additional questions, as well as their wording, to become more principled.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Daniel Büring. 2003. On D-trees, beans, and B-accents. *Linguistics and Philosophy*, 26(5):511–545.

Daniel Büring. 2016. *Intonation and Meaning*. Oxford University Press.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, JP.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted Sanders. 2017. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14:357–386.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted Sanders. 2019. Using the cognitive approach to coherence relations for discourse annotation. *Dialogue & Discourse*, 10(2):1–33.

Wolfgang Klein. 1992. The present perfect puzzle. *Language*, 68(3):525–552.

William Mann and Sandra Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *The Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech.

Uwe Reyle and Arndt Riester. 2016. Joint information structure and discourse structure analysis in an Underspecified DRT framework. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*, pages 15–24, New Brunswick, NJ, USA.

Arndt Riester. 2019. Constructing QUD trees. In Malte Zimmermann, Klaus von Heusinger, and Edgar Onea, editors, *Questions in Discourse. Vol. 2: Pragmatics*, pages 163–192. Brill, Leiden.

Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for Questions under Discussion and information structure. In Evangelia Adamou, Katharina Haude, and Martine Vanhove, editors, *Information Structure in Lesser-Described Languages: Studies in Prosody and Syntax*, pages 403–443. Benjamins, Amsterdam.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69. Earlier version (1996) in OSU Working Papers in Linguistics, Vol. 49.

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Ted Sanders and Wilbert Spooren. 2009. The cognition of discourse coherence. *Discourse, of course*, pages 197–212.

Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Roger Schwarzschild. 1999. GIVENness, AvoidF, and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. Annotation guidelines for rhetorical structure. University of Potsdam and Simon Fraser University.

Christiane von Stutterheim and Wolfgang Klein. 1989. Referential movement in descriptive and narrative discourse. In Rainer Dietrich and Carl Graumann, editors, *Language Processing in Social Context*, pages 39–76. North Holland, Amsterdam.

Maite Taboada and William Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8:423–459.

Jan van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31:109–147.

Advancing Neural Question Generation for Formal Pragmatics: Learning when to generate and when to copy

Haemanth Santhi Ponnusamy, Madeeswaran Kannan,
Kordula De Kuthy, and Detmar Meurers
SFB 833, Project A4
University of Tübingen

Abstract

Question generation is an interesting challenge for current neural network architectures given that it combines aspects of language meaning and forms in complex ways. The systems have to generate question phrases appropriately linking to the meaning of the envisaged answer phrases, and they have to learn to generate well-formed questions using the source.

Complementing the substantial strand of research on question generation in application contexts, some recent work also highlighted the role that questions and question generation can play conceptually in formal pragmatics for linking the information structure of sentences to the discourse structure of texts in so-called Question-under-Discussion (QuD) approaches (De Kuthy et al., 2020).

In this paper, we show that the sequence to sequence architecture employed in the previous work fails to capture a key property of the task: the required question-answer congruence ensures that the lexical material needed for the question is explicitly given by the answer generated from. Extending the architecture with a pointer component helps overcome this shortcoming. Second, we enrich the model with part-of-speech and semantic role information to improve question phrase generation. The resulting approach quantitatively advances the state of the art in terms of BLEU scores and question well-formedness, and we qualitatively discuss key linguistic characteristics of the generated question.

1 Introduction

Question generation, the task of creating natural questions for a given sentence or paragraph, is a challenging task with potentially many practical applications – from question answering, via dialogue systems, to reading comprehension tasks. Following the first, rule-based QG systems, the recent state-of-the-art approaches are generally based on neural networks. The task of QG is typically formulated as a sequence-to-sequence (seq2seq) learning problem in which a sentence is mapped to a corresponding question (cf., e.g., Pan et al., 2019).

In formal pragmatics, questions also play an increasingly prominent role in so-called Questions-under-Discussion (QuD, Roberts, 2012; Velleman and Beaver, 2016) approaches. Here, questions are used to make explicit the interface between the information structure of a sentence and the particular discourse structure that the sentence can function in. Under this QuD perspective, for every sentence in a text, a question needs to be formulated – and indeed explicit guidelines have been defined to support reliable manual QuD annotation (Riester et al., 2018). In a recent paper, De Kuthy et al. (2020) argue that such question generation should be automated to support the analysis of large corpora, and they propose a seq2seq neural network approach to generate all potential questions for a given sentence. Their approach learned to (often) predict the correct question word for a given answer phrase and generated questions that correctly reflect the word order properties of questions in German.

While this result confirms that neural networks can be successfully trained to generate meaningful, well-formed questions to pursue a formal pragmatic vision, there are clear challenges for such a seq2seq architecture that is supposed to generate questions for any type of large data set. One problem are rare or unknown words that have to be predicted. In most neural generation architectures, words are the basic input and output tokens. Pretrained word embeddings are used to initialize the token embedding matrix and generally a fixed vocabulary is used for both input and output sequences. With a restricted vocabulary, given the Zipfian distribution of words in language use, in any authentic corpus material serving as input there are likely to be rare

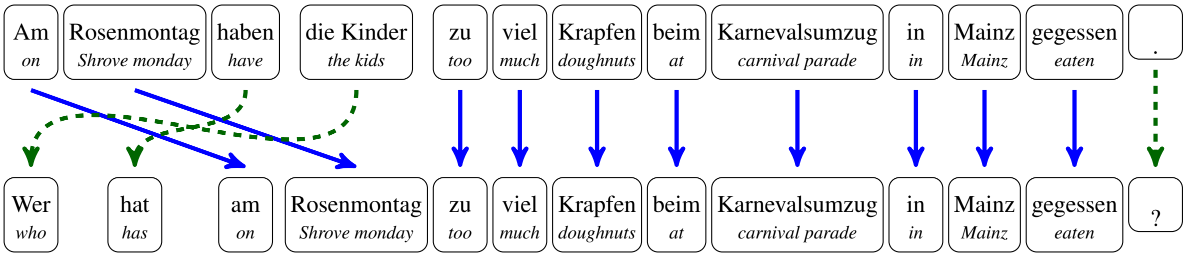

Figure 1: An example showing identical words in source sentence and question (with solid blue links) and the question word and subject-verb agreement requiring changes in the question formulation (dashed green relation)

or unknown words that are not part of the fixed vocabulary and therefore cannot be predicted in the output layer, the generated question. This indeed is a major issue mentioned for the question generation approach of De Kuthy et al. (2020). To overcome this problem, they implemented an ad-hoc post-processing step: All generated questions are checked for markers indicating the places where an out-of-vocabulary token appears. A heuristic then tries to identify that missing word in the source sentence and insert it in the right place of the output.

When we conceptually consider the task of question generation from source sentences, this is a problem that should not arise – after all, the source sentence is explicitly provided and the words in the question to be generated can be selected from that source material, to which the question words, which can be drawn from a fixed set of language expressions for a given language, need to be added. So the task of generating a question based on a given sentence conceptually consists of two subtasks: (i) Identifying the material that is identical between source sentence and question and can simply be copied over, and (ii) predicting the new material appearing in the question, in particular the correct question words. This is illustrated by the sentence-question pair in Figure 1. In that example, the specialized carnival terminology, *Karnevalsumzug* and *Rosenmontag*, are typical rare words, and the use of the city name *Mainz* illustrates the occurrence of named entities.

A related problem has been discussed in recent work for question generation (Zhao et al., 2018) and for text summarization (See et al., 2017). (Zhao et al., 2018) propose to extend a seq2seq attention model with a pointer mechanism in order to improve their task of paragraph-level QG. They show that their model with the copy mechanism can learn to generate some words in a question and to copy others from the input text. But they do not provide any details about which parts of the question are copied and which are generated and in which way the copy mechanism improves the generated questions. (See et al., 2017) observe that baseline seq2seq models for summarization often replace an uncommon (but in-vocabulary) word with a more common alternative and fail to reproduce out-of-vocabulary words. They show that the copy mechanism of their pointer-generator model overcomes both these problems, but again they do not provide any details based on which information the model chooses to copy or generate the different parts of the sentences in the summarization.

In this paper, we adopt a pointer-based architecture for the generation of questions in pursuit of the formal pragmatic vision of generating QuDs for every sentence in a text. Such an architecture turns out to be more successful than the seq2seq based model, replacing the ad-hoc heuristic post-processing step used in previous work into a design feature of the neural network architecture. In architecturally separating the copying from the generation component, it also supports the integration of further linguistic information needed to successfully determine which parts of the sentence can be copied over to the question and which parts have to generated, as for examples the question phrase.

For the mentioned example, Figure 2 identifies the minimal case, i.e. the rare or unknown words that should be copied using the pointer component, whereas other words can or need to be generated to fit the output context, such as the question word *wer (who)* and the subject-verb agreement that needs to be adjusted from plural *haben (have)* to singular *hat (has)*. We will show in a detailed analysis of the attention scores that our model learns to generate material in the question only in four cases: question word, question mark, lower-cased first word and verb form. All other material from the source sentence are copied over to the question.

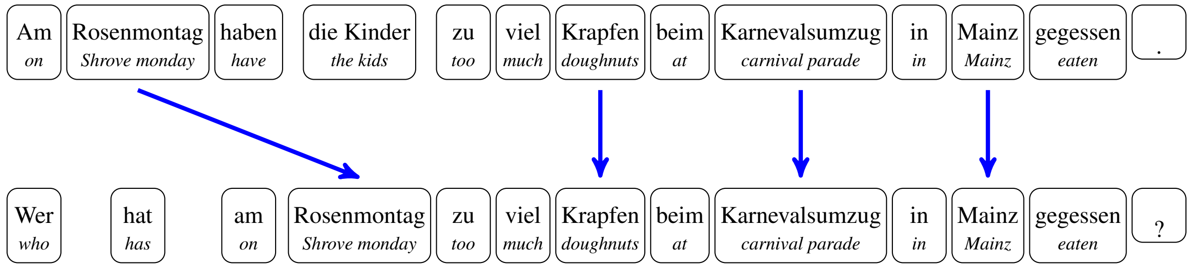

Figure 2: Example illustrating minimal identification of rare words and named entities in support of QG

The paper is structured as follows: After providing some background on the different architectures used for neural QG, in section 3 we spell out the specifics of the question-answer data we use to train the different QG models. In section 4, we present our neural QG models. As a baseline, we train a seq2seq model, similar to the one presented in (De Kuthy et al., 2020) and compare this to two versions of a pointer-based neural model, a baseline model and a model extended by two word-level features, POS tags and semantic role labels. The methods are all evaluated in quantitative terms using BLEU scores. For a qualitative analysis, we verify for how many of a random set of 500 sentences the different models produce how many well-formed questions and provide some examples illustrating the pros and cons of the different models.

2 Related Work

Generating questions is a challenging task regardless of the language. Prior to the advent of deep neural networks, question generation was largely restricted to transformation-based methods that leverage linguistic characteristics and syntactic structures (Liu et al., 2010; Curto et al., 2012; Heilman, 2011). These methods are inherently limited in that they are designed to generate questions based on pre-programmed rules that manipulate parse trees and scale very poorly with linguistic complexity. Deep neural methods, on the other hand, learn - from large sets of data - latent representations of syntactic and semantic language characteristics, amongst others.

Framing question generation as a sequence learning task enables one to exploit sequence-to-sequence architectures (Sutskever et al., 2014) that have seen significant success in neural machine translation. Sequence-to-sequence (*seq2seq*) architectures consist of two networks: 1) an encoder network that learns a representation of the source sequence, and 2) a decoder network that generates target words one at a time. This architecture was improved upon by incorporating local and global attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015) that modulate the contribution of each token (in the source sequence) to the tokens in the target sequence.

A recent survey of neural question generation research (Pan et al., 2019) shows that the above-mentioned architectures form the basis of many NQG models. Du et al. (2017) condition a generative model on target answers by encoding the position of the answer in the context as an input feature. Sun et al. (2018) split the QG task into first determining the question type and then generating the question using a template-based approach with two *seq2seq* models. Kumar et al. (2018) leverage linguistic features such as POS and NER tags and deep reinforcement learning techniques such as policy gradient methods to add additional task-specific rewards to the training objective. Rare words present a challenge to generative NLP models, and NQG models are no exception. Gulcehre et al. (2016) propose a neural model for machine translation that uses a MLP in tandem with dual softmax layers to determine when to predict a word from a fixed vocabulary and when to point to one in the source sentence. Gu et al. (2016) and See et al. (2017) showed the efficacy of pointer-generator networks at the task of abstractive text summarization. Zhao et al. (2018) adapt the same network by augmenting the encoder with gated self-attention and the decoder with a maxout pointer mechanism to deal with larger contexts. While all these implementations of pointer-based mixture models exemplify their ability to solve the unknown word problem and the advantage of copying words from the context, it is still unclear how exactly the model adapts to the task at hand, i.e., how the competing generator and pointer networks contribute to the final score and under what circumstances one is preferred over the other.

We employ the task proposed by De Kuthy et al. (2020), in which question generation is defined in the context of the formal pragmatic QuD approach where a question is generated from every sentence in a given text. We replace their sequence-to-sequence model and post-process copy step with a unified pointer-generator network that significantly outperforms the former. We also provide detailed insight into the generation and copying characteristics of the latter.

3 Data

Since question generation has primarily been approached as a sub-task of question answering, a large majority of the relevant corpora are generally tailored as QA datasets first and foremost. While datasets such as SQuAD (Rajpurkar et al., 2016), Coqa (Reddy et al., 2019), Quac (Choi et al., 2018) can nevertheless be used to train and evaluate pure question generation models, they unfortunately come up short in the context of our task.

The most obvious downside to datasets such as the above is that nearly all of them are exclusively available in English. The few that are multilingual such as XQUAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2019) are too limited in size to be used for training neural models since they are originally intended to be used as evaluation datasets for question generation/question answering models. This limitation, however, is not insurmountable; one could leverage machine translation[1] to automatically translate one of the above corpora to the target language to create a potentially usable dataset. An alternative, more active approach could involve developing a neural model that is able to jointly translate, align and generate questions (Carrino et al., 2019). Unfortunately, both approaches have a significant disadvantage in that their outputs can be expected to be of significantly lower quality then human-generated output. This can in turn increase the potential of affecting the model's performance in the actual downstream task of question generation due to the increased error propagation at the translation stage.

The second downside to using corpora such as SQUAD is that they are designed to provide paragraph-level contexts for questions. Each question can potentially have multiple ground-truth answers that can be spans of any sentence in the context. This fundamentally changes, i.e., decreases the Q-A-Congruence of the question-answer pair, making them unsuitable for the generation of assertion-level questions, as is our goal here, following the approach proposed in (De Kuthy et al., 2020) for the generation of sentence-level QUDs.

Given the above-mentioned limitations of using pre-existing QA corpora, we obtained the German QA answer corpus descrined in (De Kuthy et al., 2020). This corpus of 5.24 million sentence-question-answer triples is based on sentences from the German newspaper *Die Tageszeitung (taz)* [2] and questions were generated using the only available comprehensive transformation-based question generation system (Kolditz, 2015) for German.

Due to inherent limitations of transformation-based approaches to question generation, such systems are not always capable of producing a question for a given sentence. Furthermore, the system in question only contains a limited domain of transformation rules that mainly selects NPs and PPs as answer phrases and transforms sentences into wh-questions asking about subject and object NPs and several types of PP adjuncts and adverbial phrases. The example in (1) illustrates some types of questions and answer phrases that are produced by the transformation rules and that are part of the QA corpus created by (De Kuthy et al., 2020).

(1) a. Die Kinder essen am Sonntag Kuchen im Garten.
The children eat cake in the garden on Sunday.
b. Wer isst am Sonntag Kuchen im Garten. - Die Kinder
Who eats cake in the garden on Sunday - the children
c. Was essen die Kinder am Sonntag im Garten? - Kuchen
What do the children eat in the garden on Sunday? - cake
d. Wann essen die Kinder Kuchen im Garten? - am Sonntag
When do the children eat cake in the garden? - on Sunday
e. Wo essen die Kinder am Sonntag Kuchen? - im Garten
Where do the children eat cake on Sunday? - in the garden

4 Neural Question Generation Architectures

The task of question generation is formulated as a sequence learning problem where given a source sentence or context as the input sequence $x_1, ..., x_n$ and a target answer phrase a, the model learns the conditional probability $p(y|x, a)$ of generating the

[1] https://cloud.google.com/translate

[2] https://taz.de/

target question $y_1, ..., y_m$:

$$log\, p(y \,|\, x, a) = \sum_{j=1}^{m} log\, p(y_j \,|\, y_{<j}, x, a)$$

spaCy's `de_core_news_sm` model was used for parsing and tagging the input sentences for both models. Answer phrase spans were encoded in IOB format. *fastText* embeddings (Bojanowski et al., 2017) were used as pre-trained token embeddings. Input and target vocabulary sizes were fixed to 100K most frequent words in the corpus.

4.1 Sequence-to-sequence Model

The baseline *seq2seq* model is identical to the one used by De Kuthy et al. (2020). The input sequences to the model are the source sentence's word tokens, their part-of-speech tags, and the answer phrase span. Since this architecture does not implement an explicit mechanism to handle out-of-vocabulary words, an ad-hoc post-processing pass is performed on the model's predictions to automatically resolve OOV tokens by locally aligning the parses of the source sentence and the predicted question.

4.2 Pointer Model

Our pointer model is an extension of the work done by Zhao et al. (2018), who implement a Maxout pointer mechanism with gated self-attention.[3]. We experimented with two variants of input sequences. In the first variant, the input sequences were restricted to surface form tokens of the source sentence and the span of the answer phrase. Then we added the parts of speech (POS) and semantic role labels (SRL) in the next variant. Canonical representations of the encoder variants are shown below:

$$u_t = RNN(u_{t-1}, [e_t, a_t]) \tag{1}$$

$$u_t = RNN(u_{t-1}, [e_t, a_t, p_t, s_t]) \tag{2}$$

e_t is the embedded word tokens of the source sentence, a_t answer tagging embedding, p_t represents the POS embedding and s_t indicates the embedded semantic role labels. Now the encoder hidden state u_t is computed through the function of previous encoder hidden state u_{t-1} and the concatenated feature embeddings $[e_t, m_t]$ or $[e_t, a_t, p_t, s_t]$. Further, the hidden states $\{\widehat{u}_t\}_{t=1}^{M}$ are refined using the self-attention. The raw attention scores (Luong et al., 2015) computed between the encoder hidden state U and the decoder hidden state d_{t-1} are used to compute the copy scores. The general approach of the copy mechanisms is to treat each word in the source sentence to be a unique target to point to and to compute the scores separately. In the end, the scores of the words that occur repeatedly in the source sentence are added to get a final copy score. This leads to an overshoot of the copy scores for the words that are repeated in the source, resulting in repeated predictions of the same in the target sequence. To overcome this issue, only the maximum copy score of each word is used (Goodfellow et al., 2013; Zhao et al., 2018). An expression of the scoring mechanism is shown below:

$$sc^{copy}(y_t) = \begin{cases} \max_k r_{t,k}, & y_t \in \chi; x_k = y_t \\ -inf, & otherwise \end{cases}$$

x_k is the k^{th} word in the source sequence and y_t is the t^{th} word in the output sequence. χ is the vocabulary of all words in the input sequence, and $r_{t,k}$ is the raw attention score between x_k and r_t.

The scores from the copy mechanism and the decoder are softmaxed and combined to get the final probability distribution over the extended vocabulary containing the OOV tokens. Since the raw copy and generation scores are added together as a single vector, the copy module and the generation module essentially compete with each other for the final prediction at each timestep.

Hyperparameter	Value
Batch size	64
Epochs	10
Encoder RNN Unit	Bi-LSTM
Decoder RNN Unit	LSTM
Encoder/Decoder Hidden Size	300
Encoder/Decoder Dropout	0.3
Word Embedding Dim	300
Answer Span Embedding Dim	3
POS Embedding Dim	25
SRL Embedding Dim	25
Min Decode Step	8
Max Decode Step	100

Table 1: Pointer Model Hyperparameters

5 Evaluation

The *seq2seq* model and the pointer network were trained on the same 400K training samples. Validation and test sets were set to 15K samples each. For quantitative evaluation, questions predicted by the

[3] Unofficial implementation: https://github.com/seanie12/neural-question-generation

models were compared to the ground-truth questions from our QA corpus and their corresponding BLEU (Papineni et al., 2002) scores were calculated (Table 2).

Even though the *seq2seq* models lack a copy mechanism in their architecture, they adequately learn to mimic the behaviour by positively biasing the generative probabilities of (in-vocabulary) words that appear in the source sequence. The post-processing copy operation, though error-prone, extends this to out-of-vocabulary words, improving performance even further. In contrast, the pointer models unequivocally show that implementing copying directly in the neural architecture improves performance even in the absence of additional linguistic features such as part-of-speech tags and semantic role labels.

5.1 Model Comparison

The high BLEU scores for all of our models indicate that the models are all capable to learn the task of generating questions. To investigate where the differences and particular strengths of the different models are, we provide a more in-depth qualitative analysis of the three models. We performed a manual evaluation of a random set of questions produced by all our models for the same set of sentence - answer phrase pairs. The sample set was obtained by randomly sampling 500 sentences from the original TAZ corpus. For the comparison of our three question generation models, the 500 sentences plus the answer phrases from the rule-based output described in section 3 were used. Based on this set of 500 sentences plus answer phrases, the three neural QG models generated 500 questions each, i.e., one question per sentence - answer phrase pair. Next, the quality of the generated questions was manually evaluated by two human annotators with good annotation agreement ($\kappa = 0.74$), i.e., whether a question is well-formed and whether there is question-answer congruence between the question and the source sentence.

For the 500 questions generated by each model, the baseline seq2seq model shows the worst performance with only 31% well-formed questions out of 500. Adding the post-processing step of replacing OOV words by a word from the source sentence increased the number of well-formed questions to 52%. The two pointer architectures produced well-formed question with improved accuracy: The baseline pointer model produced 55% well-formed questions, while the pointer model with POS and SRL features produced 61% well-formed questions, the best performance for this sample set. The table in 3 sums up the results of this evaluation.

Table 4 shows a systematic analysis of the most frequent errors in the 500 sample questions made by the three models. One can, for example see, that while the questions from seq2seq model still contained unknown words in 47 cases (even after the post-processing), the questions of both pointer models did not have this problem anymore.

5.2 Copying vs Generation

The pointer model with attention and a copy mechanism successfully learned to point to the OOV tokens from the input string and copy them over to the predicted question. The generated questions thus do not contain any OOV tokens anymore. What is not obvious right away is whether the pointer architecture also learned to point and copy over other parts of the sentence and to generate only where really necessary in order to produce a new form. An investigation of the raw attention scores used to compute the copy scores that determine whether a token can be copied over between input and output or needs to be generated showed that indeed the model learned to simply copy over many parts of the source sentence into the question. Figure 3 shows a typical sentence-question pair from our 500 sample, containing 4 instances of generated tokens: *Wer* question word, *hält* Infinite verb to match the subject, *deshalb* lower case transformation to the first word and *?* question mark.

Figure 4 shows the softmaxed scores of the attention between the previous decoder hidden state at every timestep to all the encoder hidden states. Each column here shows the distribution of weights corresponding to hidden representations of each word in the input sequence towards the computation of the context vector. The output token at that time step is produced as the result of the function of this context vector and the previous decoder hidden state. The tokens with the highest attention scores in each column indicate the primary focus of the model before generating the respective output. Since the words *Wer, hält, deshalb and ?* are generated in the output and not copied, we can infer that the hidden states corresponding to the highest scores in each column have a direct influence in generating these words. The higher attention on the word *Professor* in the input sequence to gener-

Model	Training Size	Features	BLEU-1/2/3/4	Cumulative
seq2seq	500k	Word, Ans, POS	84.9/75.0/67.1/60.3	71.25
seq2seq + Copy	500k	Word, Ans, POS	93.8/86.5/81.0/76.5	84.24
Pointer	500k	Word, Ans	97.0/91.0/86.7/83.4	89.40
Pointer	500k	Word, Ans, POS, SRL	98.0/92.9/89.1/86.3	**91.45**

Table 2: Evaluation results

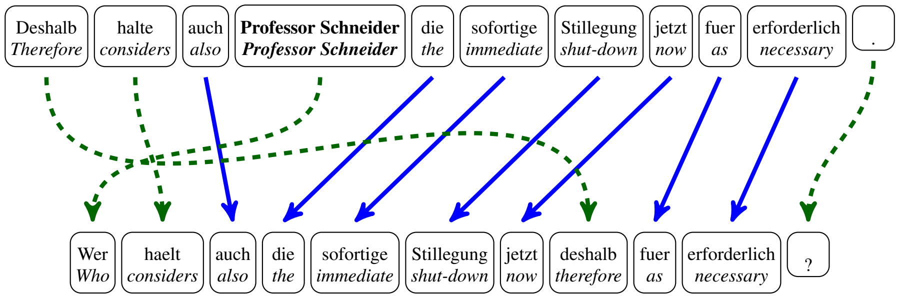

Figure 3: A question generation example, highlighting copy and generate decisions

Model	Well-formed Questions
Baseline seq2seq	31%
seq2seq + Copy	52%
Baseline Pointer	55%
Pointer + Ling. Features	**61%**

Table 3: Results for random sample of 500 sentences

Error Type	Seq2Seq	Ptr1	Ptr2
Question word	88	105	88
Unknown Word	47	0	0
Word Order	40	29	24
Different Word	18	31	15
Missing Word	6	7	6
Verb Form	6	7	5

Table 4: Distribution of error types in the 500 samples

ate the appropriate question word *Wer* shows that the model has learned the relationship between the nature of answer phrase *Professor Scheider* and the type of the question phrase.

To illustrate how the model uses information from the attention scores in the decoding step and also to compute copy scores, Figure 5 shows the raw attention scores between the previous decoding hidden state at every timestep to each of the encoding hidden states corresponding to the input tokens. The maximum scores in each column directly correspond to the score used by the copy module to compete with the generated scores. The streaks of high scores as diagonals shows that a chunk of the source sentence is copied with high support from the attention. This behaviour of replicating most of the information from the source sentence instead of generating new tokens shows that the model has

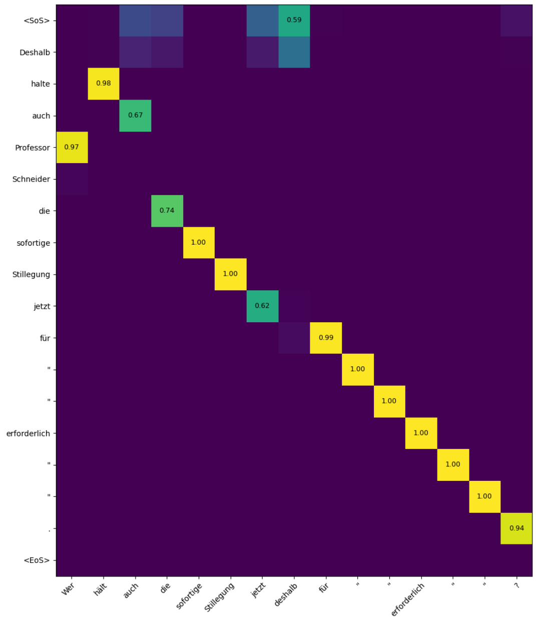

Figure 4: Softmaxed attention weight used for computing the context vector as input to each decoding step

adapted to the nature of the task including the right decision between copying or generating based on linguistic features.

We can now also precisely determine how often the pointer model is generating and copying and what type of tokens are being generated. As shown above, the decision for predicting each word in the output sequence is influenced by their intermediate scores. We can thus categorize the decisions into four categories: *Copy* - Only the copy module has suggested the final prediction with high confidence, *Generate* - Only the generate module has

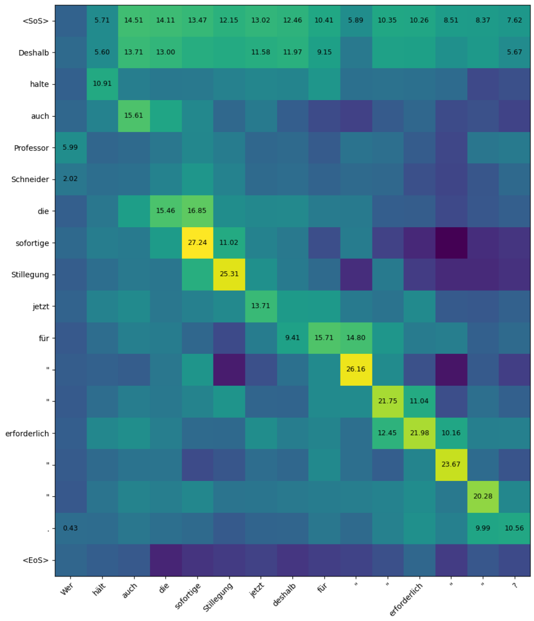

Figure 5: Raw decoder attention scores used directly as the copy scores

suggested the final prediction with high confidence, *Both* - Both the modules has suggested with high confidence and *Neither* - Neither of the modules suggested the final prediction with high confidence but jointly achieved the final prediction.

Category	Avg. % of a question
Copy	79.32%
Generate	17.57%
Neither	2.12%
Both	0.48%

Table 5: Direct influence of the modules on the final prediction of each question

Table 5 shows that around 79% parts of the questions are being copied and only around $17 - 18\%$ being generated. 2% parts of the question are being jointly predicted by both the generation and the copy modules and just less than 1% are mutually agreed by both the modules. We also determined that the model only generates tokens in four cases: Question word, question mark, lower-cased first word and verb form.

5.3 Greedy vs Beam search

We here briefly discuss the effect of different sequence search strategies like beam search vs the greedy approach to achieve a balance between the quality of the generated output and the computational cost. It has been observed that the beam search strategy might not be advantageous in all the cases of sequence generation. (Cohen and Beck, 2019) highlighted the effect of degrading performance of the sequence generation models with the increase in beam width. In our model, we face a similar scenario, where the increase in beam width during the decoding stage harms the model's performance both quantitatively and qualitatively.

Beam width	BLEU (1/2/3/4)	(Cummulative)
1 (Greedy)	98.0/92.9/89.1/86.3	91.45
3	96.7/91.2/87.4/84.5	89.83
5	96.0/90.4/86.4/83.5	88.94
7	95.7/90.0/85.9/83.1	88.54
15	95.3/89.5/85.4/82.5	88.05

Table 6: Degrading effect of beam width on the pointer model's performance

In Table 6, the model version with the greedy search (beam width=1) approach performs much better than the other versions with the increased beam width. This behaviour is due to the nature of our task, which requires predominantly the exact words to be copied from the source sentence into the output sequence. As we have shown above, the model prefers to copy around 79% of the question with very high confidence. So choosing an alternative for the exact words that are supposed to be copied and choosing words that maximize the overall probability in subsequent steps lead to a mispredicted sequence. This error mainly happens when the copy module suggests the words with relatively lesser confidence.

6 Conclusion and Outlook

Given the task of question generation in a formal pragmatics context, we successfully trained and tested two different neural network architectures on a dataset of natural question-answer pairs from a German newspaper corpus. We showed that a pointer-based architecture is advantageous for this task since it can employ task specifics to overcome problems with unknown or rare words, learning to copy those words from the input. We extended the approach by integrating information designed to improve those aspects that need to be generated, especially the appropriate question words. The quantitative evaluation using BLEU scores and an in-depth qualitative evaluation showed that indeed the pointer-based model with additional linguistic features is the best performing system for this task.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Casimiro Pio Carrino, Marta R Costa-jussà, and José AR Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *International Conference on Machine Learning*, pages 1290–1299.

Sérgio Curto, Ana Cristina Mendes, and Luísa Coheur. 2012. Question generation based on lexico-syntactic patterns learned from the web. *Dialogue & Discourse*, 3(2):147–175.

Kordula De Kuthy, Madeeswaran Kannan, Haemanth Santhi Ponnusamy, and Detmar Meurers. 2020. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. 2013. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. *arXiv preprint arXiv:1603.08148*.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. thesis, Carnegie Mellon University.

Tobias Kolditz. 2015. Generating questions for German text. Master thesis in computational linguistics, Department of Linguistics, University of Tübingen.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018. A framework for automatic question generation from text using deep reinforcement learning. *arXiv preprint arXiv:1808.04961*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Ming Liu, Rafael A Calvo, and Vasile Rus. 2010. Automatic question generation for literature review writing support. In *International Conference on Intelligent Tutoring Systems*, pages 45–54. Springer.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for questions under discussion and information structure. In Evangelia Adamou, Katharina Haude, and Martine Vanhove, editors, *Information structure in lesser-described languages: Studies in prosody and syntax*, Studies in Language Companion Series. John Benjamins.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Leah Velleman and David Beaver. 2016. Question-based models of information structure. In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*, pages 86–107. Oxford University Press.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910.