Observations on the Annotation of Discourse Relational Devices in TED Talk Transcripts in Lithuanian

Giedrė Valūnaitė Oleškevičienė[1], Deniz Zeyrek[2], Viktorija Mažeikienė[1], Murathan Kurfalı[3]
[1]Institute of Humanities, Mykolas Romeris University, Vilnius
[2]Informatics Institute, Middle East Technical University, Ankara
[3]Stockholm University, Stockholm and Middle East Technical University, Ankara
{gvalunaite, vmazeikiene}@mruni.eu
dezeyrek@metu.edu.tr, murathan.kurfali@ling.su.se

Abstract

Lithuanian researchers are working on enriching the existing corpora; they are also looking for ways to make the corpora inter-operable and co-searchable through the annotation of discourse relations. One of the goals of the present research is working on the annotation of discourse relations in TED talks transcripts translated into Lithuanian and expanding the set of available resources in the Lithuanian language. A second goal is to compare cross-linguistically the annotated texts with the view of looking for translation tendencies in rendering discourse relations in the Lithuanian language. This, we believe, will open up a new research path in digital humanities leading to an understanding of translation tendencies in TED talks transcripts across languages. According to our research results, noteworthy translation tendencies embrace explicitation - a tendency to use more explicitly marked discourse relations in Lithuanian than the original transcripts, verbatim translations of discourse connectives, and also a tendency to use fewer alternative lexicalizations (a type of discourse-relational devices).

Keywords: discourse, parallel, multilingual corpus, Lithuanian, annotation

1. Introduction

Lithuanian researchers are working on enriching the existing corpora and are also looking for ways to make the corpora inter-operable and co-searchable through the annotation of discourse relations. One of the aims of the current research is extending the available resources and lexicons of discourse-relational devices in Lithuanian cooperating with the international team of researchers brought together by the European COST Project TextLink (http://www.textlink.ii.metu.edu.tr/). The aim is partially achieved by adding Lithuanian annotated texts to the existing TED Multilingual Discourse Bank, or TED-MDB, a parallel corpus annotated at the discourse level following the goals and principles of Penn Discourse Treebank (Zeyrek et al., 2018). The second aim is to compare discourse-annotated texts with English annotations with a view to understanding translation tendencies. Our ultimate goal is to perform cross-linguistic analysis and transform this information into the domain of digital humanities. In the rest of this paper, we describe the addition of Lithuanian annotations to TED-MDB and discuss our first results to the extent that discourse relations are concerned. This, we believe, will serve as the basis for our ultimate aim.

2. Research background

The section provides some general insights on Lithuanian, describes discourse connectives (DCs), and briefly outlines the PDTB annotation scheme. It also describes the data and presents some observations about the data.

2.1. Lithuanian

Lithuanian is a very old Indo-European language. It is a Baltic language which has conservative morphology, e.g. it has preserved morphological aspects of the proto-language, such as the word declensions. It is spoken by about 2,900,000 native Lithuanian speakers in Lithuania and about 200,000 abroad.

There are two main resources for modern Lithuanian: (a) The 9-million-word Corpus Academicum Lithuanicum – CorALit (http://coralit.lt) compiled by Vilnius University. It contains academic texts from the fields of biomedical sciences, humanities, physical sciences, social sciences, and technological sciences. (b) The 102-million-word online corpus of the Contemporary Lithuanian Language (http://tekstynas.vdu.lt), which is of general character and includes publicist texts, fiction, non-fiction, administrative literature and spoken language. However, parallel corpora involving Lithuanian are still insufficient; currently only one parallel two-directional (English - Lithuanian and Lithuanian - English) corpus exits comprising English - Lithuanian (70,813 parallel sentences) and Lithuanian - English (1,614 parallel sentences) (http://tekstynas.vdu.lt). Furthermore, the corpus is not discourse-annotated. Such scarcity of corpora resources is an obvious barrier for machine translation (Šveikauskienė and Telksnys, 2014). Thus, for example, the English phrase *calling him a liar* is translated into Lithuanian as *skambinti jam melagis* (to phone him a liar) in the google translate application. The improvement of such issues clearly requires corpora development, annotation and research.

2.2. Discourse Connectives and an Outline of the Annotation Scheme

Discourse connectives signal the way the writer or speaker would like the reader or listener relate the ideas that are about to be said to the ideas that have been said before. According to Baker (2011), DCs could be used to signal differ-

ent relations and the relations could be expressed in many ways; for example, in English, causality might be expressed through verbs such as *cause, lead to* or through DCs signaling the causality relation. Languages vary in terms of the type of connectives preferred as well as their frequency. Since the DCs signal the relations between pieces of information, they are related to the structuring of information and provide an insight into the whole logic of discourse (Smith and Frawley, 1983).

The literature suggests that some languages tend to express discourse relations (DRs) through complex structures while others prefer to use simpler structures and mark discourse relations explicitly, as for example, the difference between English and Arabic illustrates (Holes, 1984). The author finds that while English prefers to present information in smaller pieces of information and signals the relations between them, Arabic prefers to group information into large discourse chunks. So the question arises how the translators deal with DRs when faced with the multitude of explicit DCs in the source text or conversely, how they render DRs when there is a limited number of connectives in the source text. Given that connectives deal with the logic of the text and they are related to text interpretation, the process of aligning the patterns of DCs with target language specifics and the text type of the target language is a complicated process. Translators could have two choices: for the sake of a smooth and clear translation, they could insert additional DCs even when they are not used in the original text, i.e. resort to explicitation, or they could choose to translate the explicit DCs of the original text verbatim, though the resulting translation might sound foreign in the target language. In practice, translators choose something in between or use a bit of both techniques (Baker, 2011).

The PDTB is a 2-million-word corpus manually annotated for discourse-level information (Prasad et al., 2014). The annotation scheme mainly includes explicit and implicit DCs, alternative lexicalizations, entity relations, no relations, and their binary arguments, called Arg1 and Arg2. Senses are assigned to all DRs except entity relations and no relations. PDTB's annotation approach is theory-neutral and lexically grounded. The theory-neutral approach means that the annotation is not based on a specific discourse theory. Lexically grounded perception implies that annotator judgments are effectively elicited both for explicit DRs and implicit DRs; i.e. even for cases where there are no explicit markers of the relation.

2.3. The Data

Our data comprise Lithuanian TED talks transcripts of the original English texts included in TED-MDB (Table 1). TED-MDB is created on the basis of PDTB 3.0 relation hierarchy (Webber et al., 2016). The PDTB is chosen mainly because it has been used reliably to annotate discourse in other languages, e.g. Turkish (Zeyrek et al., 2013), Arabic (Al-Saif and Markert, 2010), Chinese (Zhou and Xue, 2012), and Hindi (Oza et al., 2009). The corpus already includes transcripts of 6 languages: Turkish, English, Polish, German, Russian and Portuguese. As in the TED-MDB project, Lithuanian transcripts are retrieved from the WIT3 website Cettolo et al. (2012) and annotated for DRs. The annotations are saved into annotation files corresponding to the raw texts. They are simple text files where each token is stored as a series of fields, such as *sense, type, argument spans*, delimited by the pipe symbol (|), as explained in Lee et al. (2016).

Both the TED website and the WIT3 website are open resources, which is attractive to research as they present numerous advantages, e.g. subtitles are available in a substantial number of languages, and the topics cover a wide span of knowledge fields, making the data applicable in multiple domains (Cettolo et al., 2012). However, there are also certain disadvantages of the data. Firstly, the talks are translated by (named) volunteers. This does not necessarily ensure a high-quality translation. The data is also limited concerning the use of parallel transcripts for DC research and for translation. For example, the collection of TED Talks is unidirectional, thus they cannot be used for exemplifying the differences for different translation directions. There are also other issues to deal with, such as subtitling, which is a specific type of translation (Lefer and Grabar, 2015), and the genre of TED talks, which is a mix of spoken and written language. Finally, the variety of TED talks speakers (native and non-native speakers or speakers of various regional varieties of English) might be another issue to consider. Despite such issues, given the scarcity of parallel texts involving Lithuanian and the limited research on Lithuanian DCs, we chose to annotate the TED talks transcripts for DRs and examine the translation issues involved.

3. Annotation Procedures in Lithuanian

In Lithuanian, **explicit DCs** include expressions from four grammatical classes: subordinating conjunctions – e.g. *kai, kol, nes, kadangi* (*when, while, because, since*), coordinating conjunctions – *ir, bei, o, tačiau* (*and, but, or, however*), sentential relatives – *tam kad, tuo metu kai* (*so that, at the time when*), and discourse adverbials – *faktiškai, galiausiai* (*actually, eventually*). The main task is to identify if the words and phrases function as explicit DCs as they can have other functions. As in the PDTB, five types of relations are identified and annotated: Explicit relations, implicit relations, alternative lexicalizations, entity relations, and no relations. The argument annotation of explicit DCs and alternative lexicalizations follows the rule that the argument which appears as syntactically bound to the DC is marked as Arg2; the other argument is annotated as Arg1. As in TED-MDB, adverbials called "discourse markers" (Hirschberg and Litman, 1987) are not annotated as they signal the organizational structure of the discourse rather than relating two arguments semantically. For example, Lithuanian *dabar* and its English equivalent (*now*) in the examples below serve to signal discourse organizational structure, so such cases were not annotated.

1. Dabar kaip matote įtampa apie kurią girdėjome San Fransiske apie susirūpinimą dėl būsto kainų ir gyventojų išstūmimo ir technologijų kompanijų, kurios atneša daug turto ir įsikuria, yra tikra.

2. Now you can see, though, that the tensions that we've heard about in San Francisco in terms of people being concerned about gentrification and all the new tech

Talk ID	Title/Speaker	Word count Eng./Lith.
1927	The investment of logic for sustainability (Chris McKnett)	1,614 (1,345)
1978	Embrace the near win (Sarah Lewis)	1,772 (1,362)
2009	A glimpse of life on the road (Kitra Cahana)	694 (512)
2150	Social maps that reveal a city's intersections and separations (Dave Troy)	1,053 (678)
TOTAL		5,133 (3,897)

Table 1: The English and the Lithuanian sections of the corpus included in the study

companies that are bringing new wealth and settlement into the city are real.

According to PDTB annotation guidelines, in annotating **implicit DRs**, the annotator has to insert a DC that best expresses the inferred relation between two adjacent sentences. This procedure is adopted, as in Lithuanian example 3 and its English equivalent in 4. In all the examples, Arg1 is shown in italics, Arg2 is shown in boldface.

3. *Ji tokie sudėtingi ir gali atrodyti mums tolimi, kad galime būti linkę daryti štai ką: slėpti galvą smėlyje ir negalvoti apie tai*. [Implicit=Bet] **Jei tik galite, priešinkitės tam**. (Implicit) (Comparison: Contrast)

4. *...bury our heads in the sand and not think about it*. [Implicit=But] **Resist this, if you can**. (Implicit) (Comparison: Contrast)

Alternative lexalization (AltLex) includes cases of inferred DRs between adjacent clauses, where redundancy appears if an explicit DC is inserted. The reason for this is that the relation is already expressed by some alternatively lexicalized non-connective expression, e.g.

5. *Sėkmė mus motyvuoja, bet beveik pasiekta pergalė skatina mus leistis į nuolatinius ieškojimus*. [Vieną iš ryškiausių to pavyzdžių pastebime], **kai žvelgiame į skirtumą tarp olimpinio sidabro laimėtojų ir bronzos laimėtojų rungtynėms pasibaigus**. (AltLex) (Expansion: Instatiation)

6. *Success motivates us, but a near win can propel us in an ongoing quest*. [One of the most vivid examples of this comes] **when we look at the difference between Olympic silver medalists and bronze medalists after a competition**. (AltLex) (Expansion: Instantiation)

Entity relations (EntRel) are annotated between adjacent sentences when an entity in one argument is described further in the other argument, as in 7 and its English version in 8.

7. *Jie turėtų įvertinti ir tuos efektyvumo rodiklius, kuriuos vadiname ASV: aplinkosauga, socialiniai klausimai ir valdymas*. **Aplikosauga apima energijos vartojimą, prieigą prie vandens, atliekų tvarkymą ir taršą ir ekonomišką išteklių naudojimą**. (EntRel)

8. *Investors should also look at performance metrics in what we call ESG: environment, social and governance*. **Environment includes energy consumption, water availability, waste and pollution, just making efficient uses of resource**. (EntRel)

No relation (NoRel) is annotated when there is no DR inferred by the reader between the adjacent sentences:

9. *Tai 4 milijardai viduriniosios klasės žmonių, kuriems reikia maisto, energijosir vandens*. **Dabar jūs tubūt klausiate savęs: gal tai tik pavieniai atvejai**. (NoRel)

10. *That's four billion middle class people demanding food, energy and water*. **Now, you may be asking yourself, are these just isolated cases**. (NoRel)

TED-MDB adds a new top-level category to the PDTB 3.0 relation hierarchy, called hypophora. This category aims to capture rhetorical question-response pairs, where the question is asked and answered by the speaker. TED-MDB annotates hypophora as a case of AltLex anchored by the question word. Where possible, the additional sense of the Q/R pair may be added.
As in TED-MDB, in Lithuanian, we annotate the question as Arg2, the answer as Arg1. We consider the question as Arg2 because the AltLex is part of the question. The question word (either the wh-word or *ar*, a specific question particle used in Yes/No questions, which can also serve as an explicit DC in Lithuanian) is selected as AltLex since it marks the DR holding between the question and the answer, as in example 11 and its equivalent in 12:

11. *Niekas nepasikeis*, [ar] **mes bandysime pakeisti**, [ar] **tu nieko nebandysi** (Explicit) (Expansion: Disjunction)

12. *Nothing is going to change* [either] **we try to change something** [or] **you don't try anything**. (Explicit) (Expansion: Disjunction)

In the following pairs of examples, we provide more cases of how hypophora is annotated in Lithuanian and English. Lithuanian Q/R pairs are annotated for a primary sense, and tagged as hypophora as the secondary sense.

13. [Ar] **įmonės, atsižvelgiančios į tvarumą, išties finansiškai sėkmingos**? *galintis nustebinti atsakymas yra "taip"* (Explicit) (Altlex: Ar; Expansion: Level-of-detail:Arg1-as-detail; Hypophora).

14. [Do] **companies that take sustainability into account really do well financially?** The answer that may surprise you is yes. (AltLex: Do) (Hypophora)

15. [Kodėl] **kas nors apskritai rinktųsi tokį gyvenimą** - *Atsakymas į šį klausimą gali skirtis, kaip skiriasi ir žmonės sutinkami kelyje, bet keliautojai dažnai atsako vienu žodžiu: laisvė*. (Explicit) (Altlex: Kodėl; Contingency: cause: Reason; Hypophora).

16. [Why] **anyone would choose a life like this, under the thumb of discriminatory laws, eating out of trash cans, sleeping under bridges, picking up seasonal jobs here and there**. *The answer to such a question is as varied as the people that take to the road, but travelers often respond with a single word: freedom*. (AltLex: Why)(Hypophora)

4. Intra- and Inter-Annotator Agreement

The stability of the annotation scheme is evaluated both by intra- and inter-annotator agreement. One transcript (Text ID 1978), which comprises approximately 25% of the Lithuanian section of the data is reannoted by the primary annotator after about 2 months of the first annotation, and it is annotated independently by the secondary annotator (cf. Table 2 for the distribution of the annotated, reannoated and independently annotated DR types).[1] We measured F1 score, which evaluates agreement between the annotators regarding the existence of a DR between the same discourse units. To measure agreement on the types and senses of these DRs, we calculated Cohen's Kappa (Cohen, 1960), which is known to be a robust method to evaluate agreement on categorical items as it takes the chance agreements into account. In this preliminary evaluation exercise, we reached very high scores on both measures: The F1 scores for intra- and inter-annotator agreement are 0.933 and 0.944, respectively. The Kappa values for intra- and inter-annotator type agreement are 0.974 and 0.991, respectively; the Kappa values for intra- and inter-annotator sense agreement are 0.967 and 0.989, respectively.

	Primary annotator		Secondary annotator
Relation Type	1st annot	2nd annot	
AltLex	-	2	-
NoRel	15	15	13
Explicit	105	107	101
Implicit	48	53	44
EntRel	28	30	27

Table 2: Frequencies of annotated, reannotated and independently annotated DR types in one Lithuanian transcript

5. Research Findings

In this section, we focus on the whole unit of the annotated texts in English and Lithuanian and present the frequencies of annotated DR types (Table 3) as well as the frequencies of the annotated top-level senses (Table 4). We then discuss the results.

In Table 3, the low frequency of AltLex annotations in Lithuanian could reveal a certain tendency characteristic reflecting the translators' choices while translating the DCs - it appears that the translators tended to render DCs by the variants provided by dictionaries rather than using AltLexs, e.g. *kai* (when), *kol* (while), *nes* (because), *nes* (since), etc. This resonates with Baker´s (Baker, 2011) observations in that translators might choose to align the patterns of DCs with the target language.

[1]The primary and the secondary annotators are the first and the third authors of the study.

Relation Type	English	Lithuanian
AltLex	33	7
NoRel	38	24
Explicit	225	297
Implicit	132	177
EntRel	43	44

Table 3: Frequencies of annotated relation types in 4 transcripts in English and Lithuanian

Top-level Sense	English	Lithuanian
Temporal	24	25
Comparison	57	66
Hypophora	9	13
Expansion	213	262
Contingency	94	127

Table 4: Frequencies of annotated top-level senses of the PDTB scheme including Hypophora in 4 transcripts in English and Lithuanian

Another interesting feature observed is that there are more explicit DRs in the Lithuanian transcripts than in the English versions. This might be explained by the translators' effort to render the implicit DRs in English explicitly. There are also cases where implicit DRs in English texts are translated explicitly to Lithuanian, which goes in tune with explicitation, as observed by Baker (1996). For example:

17. ... *that's okay, right.* [Implicit=But] **We want more.** (Implicit) (Comparison: Concession: Arg2_as_denier)

18. *Nebogai, tiesa.* [Bet] **mes norim daugiau.** (Explicit) (Comparison: Concession: Arg2_as_denier)

However, there are also cases when the explicit DCs are rendered implicitly, which might lead to the loss of the sense annotated in the original text. For example:

19. ... *only looking at race doesn't really contribute to our development of diversity.* [So] **if we're trying to use diversity as a way to tackle some of our more intractable problems, we need to start to think about diversity in a new way.** (Explicit) (Contingency: Cause: Result)

20. ... *žiūrėti tik į rasę nepadeda bandant prisidėti prie įvairumo vystymo.* [Implicit=Taigi] **Bandome įvairumą naudoti sprendžiant kai kurias sudėtingesnes problemas, turime pradėti kitaip galvoti apie įvairumą.** (Implicit) (Contingency: Cause: Result)

21. [If] **we're trying to use diversity as a way to tackle some of our more intractable problems,** *we need to start to think about diversity in a new way.* (Explicit) (Contingency: Condition: Arg2_as_condition)

22. *Bandome įvairumą naudoti sprendžiant kai kurias sudėtingesnes problemas*, [Implicit=todėl] **turime pradėti kitaip galvoti apie įvairumą.** (Implicit) (Contingency: Cause: Result)

Examples 19-20 and 21-22 show that the translator chose not to render the explicit DCs *so* and *if*. However, even though the sense of 'result' could be felt implicitly in 20, in 22, we observe a meaning loss, where the sense of 'condition' is totally lost.

Finally, the annotation of EntRels also revealed some interesting cases. We observed that in some Lithuanian translations, the EntRel is present in two loosely related sentences as in 23, while in the source English text there is just one sentence lacking two separate arguments (see 24):

23. *Tad pasakysiu kai ką, kas gali jus nustebinti: galios balansas, galintis išties paveikti tvarumą, yra institucinių investuotojų rankose*. **Tai tokie didieji investuotojai kaip pensijų fondai, kiti fondai ir labdaros fondai**. (Entrel)

24. And here's something that may surprise you: the balance of power to really influence sustainability rests with institutional investors, the large investors like pension funds, foundations and endowment.

Concerning the frequencies of the top-level senses of DRs, the distribution seems to be approximately equal for both languages as indicated in Table 4.

6. Summary, Conclusions and Outlook

The research findings presented here represent our initial observations and reveal certain tendencies in rendering the discourse of English TED talks in Lithuanian. Our focus has been on how DRs are expressed in Lithuanian transcripts. We observed that there are more explicit DRs in the Lithuanian transcripts, which might be explained by the translators' efforts to render the implicit DRs explicitly - this goes in tune with the observations of Baker (2011). On the other hand, we noticed that the rendering of explicit DCs implicitly might lead to the loss of the sense annotated in the original text. Such choices of the translator could obscure the meaning of the original, and could be explained by the requirements of synchronization during transcript translation. These might be the effect of the issues discussed by Lefer and Grabar (2015) who identify subtitling as a specific type of translation. The annotation of entity relations also reveals interesting cases, such as the translation of a single English sentence into two loosely related arguments in the Lithuanian EntRel version. Finally, it should be kept in mind that there could be some stylistic preferences of the translators, e.g. some translators might want to use more explicit connectives, some less. The investigation of individual translators' choices could be a specific further research topic.

In the future, by annotating more of the Lithuanian transcripts of the English texts in TED-MDB, we hope to reveal and specify more translation tendencies. Also, by exploring the transcripts further, we expect to find out what translation strategies (direct translation, transposition, etc.) are preferably employed by the translators and what this may add to the research field of digital humanities.

7. Acknowledgements

This research is funded by the European Social Fund under the No 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities" measure. For training in annotation and generating ideas for research, we acknowledge the support of the STSM grants by TextLink COST action IS1312.

References

Al-Saif, A. and Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In *LREC*.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *Benjamins Translation Library*, 18:175–186.

Baker, M. (2011). *In Other Words: A Coursebook on Translation*. Routledge.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, volume 261, page 268.

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

Hirschberg, J. and Litman, D. (1987). Now let's talk about now: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, pages 163–171. Association for Computational Linguistics.

Holes, C. (1984). Textual approximation in the teaching of academic writing to Arab students: A contrastive approach. *English for Specific Purposes in the Arab World*, pages 228–242.

Lee, A., Prasad, R., Webber, B. L., and Joshi, A. K. (2016). Annotating discourse relations with the PDTB Annotator. In *COLING (Demos)*, pages 121–125.

Lefer, M.-A. and Grabar, N. (2015). Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures*, 16(2):187–208.

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., and Joshi, A. (2009). The Hindi Discourse Relation Bank. In *Proc. of the 3rd Linguistic Annotation Workshop*, pages 158–161. Association for Computational Linguistics.

Prasad, R., Webber, B., and Joshi, A. (2014). Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Smith, R. N. and Frawley, W. J. (1983). Conjunctive cohesion in four English genres. *Text-Interdisciplinary Journal for the Study of Discourse*, 3(4):347–374.

Šveikauskienė, D. and Telksnys, L. (2014). Accuracy of the parsing of Lithuanian simple sentences. *Information Technology and Control*, 43(4):402–413.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2016). A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., and Çakıcı, R. (2013). Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse*, 4(2):174–184.

Zeyrek, D., Mendes, A., and Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank. In *LREC 2018*.

Zhou, Y. and Xue, N. (2012). PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 69–77. Association for Computational Linguistics.