A pilot study on underspecified discourse connectives in the TED Talk parallel corpus

Šárka Zikánová, Agnes Abuczki, Nijolė Burkšaitienė, Ludivine Crible, Péter Furkó, Anna Nedoluzhko, Giedre V. Oleškevičienė

Motivation and outline

- Multilingual corpus study focusing on most frequent discourse markers in English, and their translations in Czech, French, Hungarian and Lithuanian
- Research questions:
  - underspecification of discourse markers
  - omissions, translations
  - polyfunctionality and domain shifts of DMs
  - monolingual and crosslinguistic approach
  - comparison of typical occurrences of ambiguity in various languages
  - the primary question is whether and where the "weakpoints" coincide in more languages and whether there are some typical language-specific types of underspecification in single languages

Theoretical background

- Discourse markers: “sequentially dependent elements which bracket units of talk” (Schiffrin 1987: 31)

- They come from various syntactic classes:
  - conjunctions (‘and’), adverbials (‘in fact’), VPs (‘I mean’), interjections (‘well’), etc.

- They signal coherence relations between two arguments, such as cause, contrast, specification
- They can also signal new turns or new topics and contribute to the speaker-hearer relationship

Discourse domains

- DMs can work in 4 domains (Crible & Degand 2017)

  – **ideational** : objective relations between real-world events

*we want to contribute to science **but** our links with university are fragile*

  – **rhetorical** : subjective relations and metadiscourse

*I do poetry in 5th grade which may seem traditional **but** well it's how I design the class*

  – **sequential** : hierarchical structure of local and global units

< speaker1> *I like neologisms I like regionalisms but we should be careful*
<speaker2> ***but** about the norm what is it to you?*

  – **interpersonal** : intersubjectivity, contact control

*he will say look uh Jean d'Ormesson again **but** we hear Jean d'Ormesson every year*

Underspecification

- Unbalance between **semantic** encoding and **pragmatic** interpretation : the relation is underspecified
- **Monolingual** underspecification :
  - Spooren (1997) : use of ‘and then’ (Dutch *en dan*) for causal relations, enumerations, etc.
  - Domain shift, polyfunctionality (e.g. ideational ⇒ sequential)
- **Multilingual** underspecification :
  - DM in the original, omission in the translation
  - “strong” DM in the original, “weak” DM in the translation
    - *however* ⇒ Fr. *mais* (‘but’) ; *so* ⇒ Cz. *a* (‘and’)
    - or vice versa

Data

- TED talks: short (pre-planned) spoken lectures on specific topics

  www.ted.com

- multilingual corpus of TED talks: original in English, subtitles in Czech, French, Hungarian and Lithuanian

- 3 texts, 234 sentences, from 5 to 17 minutes
  - Hannah Fry: The mathematics of love
  - Bassam Tariq: The beauty and diversity of Muslim life
  - Morgana Bailey: The danger of hiding who you are

Data

- language families: Czech - Slavic

  French - Roman

  Hungarian - Finno-Ugric

  Lithuanian - Baltic

  (English - Germanic)

- different syntactic structures of the languages:
  - Czech: clauses with finite verbal forms
  - Lithuanian: participle 1 and 2
  - Hungarian: topic-prominent language (emphasis is placed on the verb or phrase preceding the finite verb)

General results

- 261 English tokens, 41 types
- Most frequent English discourse connectives:
  - and, but, so, now, because, when, if, actually, then

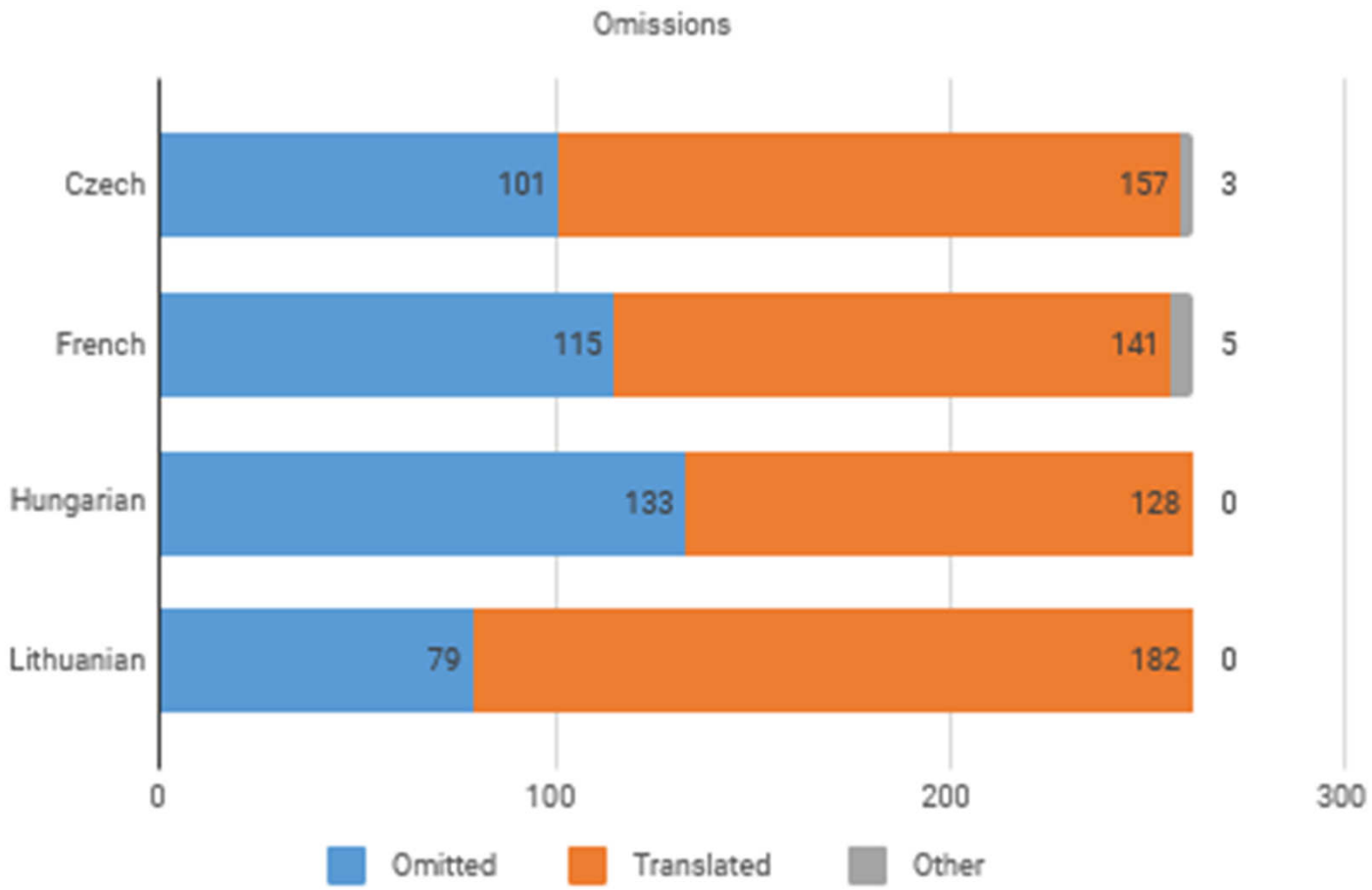
	1st	2nd	3rd	4th	5th
EN	and	but	so	now	because
CZ	a ‘and’	ale ‘but’	když ‘when’ ‘if’	tedy ‘thus’	protože ‘because’
FR	et ‘and’	mais ‘but’	si ‘if’	parce que ‘because’	donc ‘so’
HU	de ‘but’	és ‘and’	ha ‘if’	amikor ‘when’	mert ‘because’
LI	ir ‘and’	bet ‘but’	jeigu ‘if’	taigi ‘so’	kai ‘when’

Implicitation in the translation

DM types frequently lost in translation:

Lithuanian	now (18), and (17), so (12), then (8), actually (5)
Hungarian	and (25), now (23), so (22), but (15), then (10), okay (6), actually (6)
Czech	now (24), and (18), so (17), then (9), but (6), actually (5)
French	now (23), and (20), so (19), but (11), actually (8), then (8)

Proportions of omitted DMs



Underspecification in the translation

Translations of English AND in the different languages :

Lithuanian	*ir [and], o [but/and], ir todėl [and so], taip pat [also], bet [but]*
Hungarian	*és [and], egyébkent [otherwise], ehhez [to this], s [short version of and]*
Czech	*a [and]*
French	*et [and], ensuite [then], alors [so], mais [but], puis [then]*

- Cf. Abuczki's et al. presentation on AND

From weak to strong

- AND can be translated by stronger DMs :

With all the extra time and still no real money, my wife tasked me to cook more for us. ***And*** *whenever I'd go to the local butcher to purchase some halal meat, something felt off.* ⇒ translated by Fr. *mais* 'but'

- Influence of co-occurring DMs :

How ironic that I work in human resources, [...] a profession that advocates that the diversity of society should be reflected in the workplace, ***and yet*** *I have done nothing to advocate for diversity.* [concession]

⇒ translated by Fr. *mais* 'but'

From strong to weak

- Stronger DM (*in fact*) translated by Fr. *et* 'and'

EN *Now if you do this,* [*it can be mathematically proven*]*,* ***in fact****, that this is the best possible way of maximizing your chances of finding the perfect partner.*
FR *Si vous faites cela,* [***et*** *c'est mathématiquement démontrable*]*, c'est la meilleure façon possible de maximiser vos chances de trouver le partenaire idéal.* [comment, aside]

And these equations, they depend on the mood of the person when they're on their own, the mood of the person [when] they're with their partner, ***but*** *most importantly, they depend on how much the husband and wife influence one another.*

⇒ translated by Fr. ***et*** 'and' [contrast or addition?]

Monolingual underspecification (1)

- Unbalance between semantics and pragmatics
- Mainly applies to ***and*** and its equivalents

⇒ cf. Crible's talk and Abuczki et al. (next talk)

Monolingual underspecification (2)

- Type of polyfunctionality : domain shift

- Most DMs originally have an ideational sense
- This basic sense can extend to others domains
  - addition of facts ⇒ addition of arguments (‘moreover’) ⇒ addition of topics or enumeration
  - contrast between facts ⇒ contrast between topics (“but let’s come back to…” ⇒ contrast of opinions (disagreeing)

- Qualitative analysis of such examples

Shifts of temporal discourse markers

Then, now: **from temporal meaning to resulting**

Ideational domain, temporal meaning (succession)

(A) *When I was looking through my London journal and scrapbook from my London semester abroad 16 years ago, I came across this modified quote from Toni Morrison's book, "Paradise." "There are more scary things inside than outside." And* ***then*** *I wrote a note to myself at the bottom: "Remember this."*

- (Czech *pak [then]*, French *puis [then],* Lithuanian, Hungarian – omission)

Shift to ideational domain, resulting

(B) *It's this spread that makes you more popular on an online Internet dating website. **So** what that means **then** is that if some people think that you're attractive, you're actually better off having some other people think that you're a massive minger.*

- (Czech *tedy [therefore],* Lithuanian *tuomet [then]*, French, Hungarian – omission)
- Emphasized by *so*

Shifts of temporal discourse markers 2

Shift to sequential domain, (succession or resulting)

(C) *And the important thing to notice is that it's not totally true that the more attractive you are, the more messages you get.*
*But the question arises **then** of what is it about people up here who are so much more popular than people down here, [even though] they have the same score of attractiveness?*
(French *alors [so],* Czech, Hungarian, Lithuanian – omission)

“in the following part of the text”, “a following thought is coming” or resultatively “when we accept the first part of the text, we have to come to the following thought”,

Shifts of temporal discourse markers 3

(D) Now, in my favorite paper on the subject, which is entitled, "Why I Don't Have a Girlfriend" Peter Backus tries to rate his chances of finding love. ***Now,*** *Peter's not a very greedy man.*

(Hungarian *nos [well* – topic elaboration], Czech, French, Lithuanian – omission)

(“at this point of the text”; “at this point which means after a change - e.g. of the topic”)

Shifts of basic consequence markers *(so)*

From consequence in the ideational domain to consequence in the rhetorical domain

Ideational domain, consequence

(E) *Now, if you're following the maths, I'm afraid no one else comes along that's better than anyone you've seen before,* ***so*** *you have to go on rejecting everyone and die alone.*

(Czech *takže [so],* French *donc [so],* Hungarian *így [so],* Lithuanian *taigi [so])*

Rhetorical domain, consequence of intentions

- (F) *Because I believe that mathematics is so powerful that it has the potential to offer us a new way of looking at almost anything. Even something as mysterious as love. And* ***so,*** *to try to persuade you of how totally amazing, excellent and relevant mathematics is, I want to give you my top three mathematically verifiable tips for love.* (French *donc [so],* Lithuanian *na* [particle: *let's, well, to tell you shortly],* Hungarian, Czech – omission)

Shifts of basic consequence markers *(so)* 2

From consequence in the ideational domain to a border marker in the sequential domain

- the following part of the text is resulting from the previous part

Opening border marker:

(G) *And so, to try to persuade you of how totally amazing, excellent and relevant mathematics is, I want to give you my top three mathematically verifiable tips for love. Okay,* ***so*** *Top Tip #1: How to win at online dating.*

(Czech *tedy [so],* Lithuanian taigi *[so],* French, Hungarian – omission)

Closing border marker:

(H) *I think this is conclusive proof, if ever it were needed, that everybody's brains are prewired to be just a little bit mathematical. Okay, so that was Top Tip #2.*

- (Czech *tedy [so],* French *donc [so],* Hungarian, Lithuanian – omission)

Conclusion

- Some processes of underspecification are general and occur in many languages in parallel. This concerns especially systematic shifts between discourse domains (ideational, rhetorical, sequential, interpersonal, cf. Crible and Degand, in press) which are most typical for spoken language.
- Regular tendencies regarding the implicitation, multiple translation equivalents and functional shifts of discourse connectives across languages
- Coherence as a value of communication

References

- Crible, L.; Degand, L.: Reliability vs. Granularity in discourse annotation: What is the trade-off? In: Corpus Linguistics and Linguistic Theory 14(2), 1–29 (2017).
- Crible, L.: Discourse markers, (dis)fluency and the non-linear structure of speech: a contrastive usage-based study in English and French. Université Catholique de Louvain, Louvain-la-Neuve (2017).
- Crible, L.: Identifying and describing discourse markers in spoken corpora. Université Catholique de Louvain, Louvain-la-Neuve (2014).
- Schiffrin, D.: Discourse Markers. Cambridge University Press, Cambridge (1987).
- Zufferey, S., Degand, L.: Annotating the meaning of discourse connectives in multilingual corpora. In: Corpus Linguistics and Linguistic Theory 13(2), 1–24 (2017).

Acknowledgements

This research was supported by the following projects:

TextLink: Structuring Discourse in Multilingual Europe (ISCH COST Action IS1312)

Implicit Relations in Text Coherence (Grant Agency of the Czech Republic, GA 17-03461S)

PhraDiCo (F.R.S.-FNRS project, MIS grant F 4520.16)