The underspecified connective *and* in a parallel TedTalk corpus: functions, translation and implicitation

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

Outline

- Research questions
- Theoretical background (*and,* discourse domains)
- Annotation scheme (Crible & Degand 2017)
- Methodology
- Examples from our corpus
- Findings
  - distribution of domains and functions in English versus in their translations
  - omissions
  - translation equivalents
- Future perspectives

Research questions

(1) What is the functional spectrum of *and* in English TedTalks? What kind of relations does *and* signal?

(2) How is *and* translated in Czech, French, Hungarian and Lithuanian? Are specific functions of *and* associated with specific translations?

(3) Which uses of *and* tend to be omitted in the translations?

Theoretical background on *and*

- *and* encodes very little information in its core meaning
- yet, it is used in a variety of contexts where additional meanings can be identified, such as contrast or consequence

Schiffrin (1987) on the uses of *and*:

- two basic discourse uses: **coordination and continuation**
- it **coordinates ideas and units**, however, it has little semantic meaning
- it can **connect events**: [POSITION EVENT] *and* [SUPPORT EVENT] *and* [EVENT]
- it can also **connect reasons** or two pieces of support at a higher level of idea structure: [POSITION SUPPORT 1] *and* [SUPPORT 2].
- it often **links structurally similar clauses** and it doesn't favor tense switching (unlike temporal connectives).
- it has **contrastive** uses, e.g. *We tried to win.* ***And*** *we lost*

Theoretical background on *and* (Schiffrin 1987)

- it can **preface the outcome** of a reason: e.g. *That's one game I remember because we had a driveway and, like we would hide, and they would walk around the driveway? Y'know? **And** I- I remember it so distinctly*

- it can **connect a general conclusion drawn from a list of specific events** which are asyndetically connected, e.g. *I uh I go on trips with 'em, I bring 'em here, we have supper, or dinner here, **and** I don't see any problem because I'm workin' with college graduates.*

- *And* also has **pragmatic** effect: as a marker of **speaker-continuation** in interaction (which is a consequence of the speaker's situated context-bound use).

- It can be used to (try to) **reopen an interactional unit** whose completion has earlier been interrupted (turn-taking) or to continue/return to a previous question as a request for elaboration (turn-giving): QUESTION 1 ANSWER *and* QUESTION 1.

Theoretical background on coherence relations and discourse domains

- A classical approach to coherence relations (which can be explicitly marked by DMs/connectives) is to classify them into relational or discourse domains:

  - external - internal (Halliday & Hasan 1985)

  - semantic - pragmatic (Sanders 1992)

  - ideational - rhetorical - sequential (Redeker 1991)

  - ideational - interpersonal - textual (Degand 1996)

Annotation scheme

In the annotation of *and* tokens, we apply **Crible & Degand's (2017)** revised taxonomy with cross-domain functions.
DMs can signal a relation in 4 discourse **domains**:

- sequential
- ideational
- rhetorical
- interpersonal

Domains and functions are independent,

in any of the domains, DMs can express any of these **functions**:

[addition] [alternative] [cause] [closing] [concession] [condition] [consequence] [contrast] [enumeration] [opening] [punctuation] [resuming] [temporal] [topic-shift] [specification]

Annotation scheme - domains

- **sequential** relation: structures discourse, signals textual relations

*[But] it wasn't [until] he teamed up with a mathematician, James Murray, that they really started to understand what causes these negativity spirals and how they occur. [**And**] the results that they found [I think] are just incredibly impressively simple and interesting.*

- **ideational** relation**:** expresses objective relations between real-world events; connects factual information; the speaker targets content
  *So let's imagine then, that you start dating when you're 15, [**and**] ideally, you'd like to be married by the time that you're 35.*

- **rhetorical** relation: expresses the speaker's subjectivity and metadiscursive effects; the speaker targets illocutionary value

  *There'd be a huge spread in her scores. [**And**] [actually] it's this spread*

  that counts.

- **interpersonal** relation: manages the speaker-hearer relationship; the speaker targets intersubjective inferences (this domain is irrelevant in the monologous TED Talks)

Methodology

- TedTalks in the researched languages were aligned and manually annotated
- The gold standard for English annotations was set first
- Then the research teams double-coded each of the working languages
- Mostly domain-level disagreement was observed
- The problem areas were resolved during online inter-annotator discussions

Examples 1 - sequential domain

And is most frequently used (in about 50% of the English occurrences) as a means of expressing sequential addition:

En *[So] these equations, they predict how the wife or husband is going to respond in their next turn of the conversation, how positive or negative they're going to be. [**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.*

Domain: **sequential**
Function: **addition** (*and* establishes a text organizing relation between ideationally loosely connected adjacent segments)

Examples 2 - ideational domain

En *The movie is called "These Birds Walk„ **and** it is about wayward street kids who are struggling to find some semblance of family.*

Cz *Filmas vadinasi „These Birds Walk". Jis yra apie gatvės vaikus, kurie iš visų jėgų stengiasi surasti kažką panašaus į šeimą. (**omission** of and)*

Hu *A címe 'These Birds Walk'. Akaratos kis utcagyerekekről szól, akik valami családféleségre vágynak. (**omission** of and)*

Li *Film má název "Tihle Ptáci Chodí" **a** vypráví o těhle svévolných pouličních dětech kteří doopravdy usilují o alespoň nějaké zdání rodiny.*

Fr *Le film s'appelle « These Birds Walk » **et** parle des enfants de rue qui luttent à trouver un semblant de famille.*

Domain: **ideational** (*and* introduces specification and specifies factual information)

Function: **specification**/addition (example for inter-annotator disagreement)

(+ omissions in Cz and Hu; specification is a less common function of *and;* it is often omitted in translation and substituted by a full stop followed by a new a sentence start)

Examples 3 - rhetorical domain

A striking regularity observed cross-linguistically is the use of *and* (sometimes in DM clusters) for the **topicalization** of the previous focus:

En *There'd be a huge spread in her scores. [**And**] [actually] it's this spread that counts.*

Cz *V jejím hodnocení by byl velký rozptyl. [**A**] [právě] na tomto rozptylu záleží.*

Hu *Nagy lesz a szórása a pontoknak. [**És**] ez az a szórás, ami számít.*

Li *Jos balai būtų visiškai pasiskirstę. [**Ir**] [išties], svarbus būtent tas pasiskirstymas.*

Fr *omission of the discourse connective*

Domain: **rhetorical, function**: **addition** / consequence (Li)

- it is difficult to separate the roles of co-occurring DMs: in this DM cluster *and* expresses addition, while *actually* expresses specification
- translation equivalents also help us separate these roles: when *and* is omitted in certain translations, but *actually* is maintained, we can identify that specification is expressed by *actually*, not by *and*
- *and* tokens with RHET-ADD function are rarely omitted in translation

# Distribution of the domains of *and* in English

- The connective *and* is mainly used in the **sequential** domain, representing the structuring of local and global units of discourse, as well as in the **ideational** domain, connecting factual information.

- *And* is less often used in the rhetorical domain which is related to the speaker's subjectivity; and in our data it is not used in the interpersonal domain.
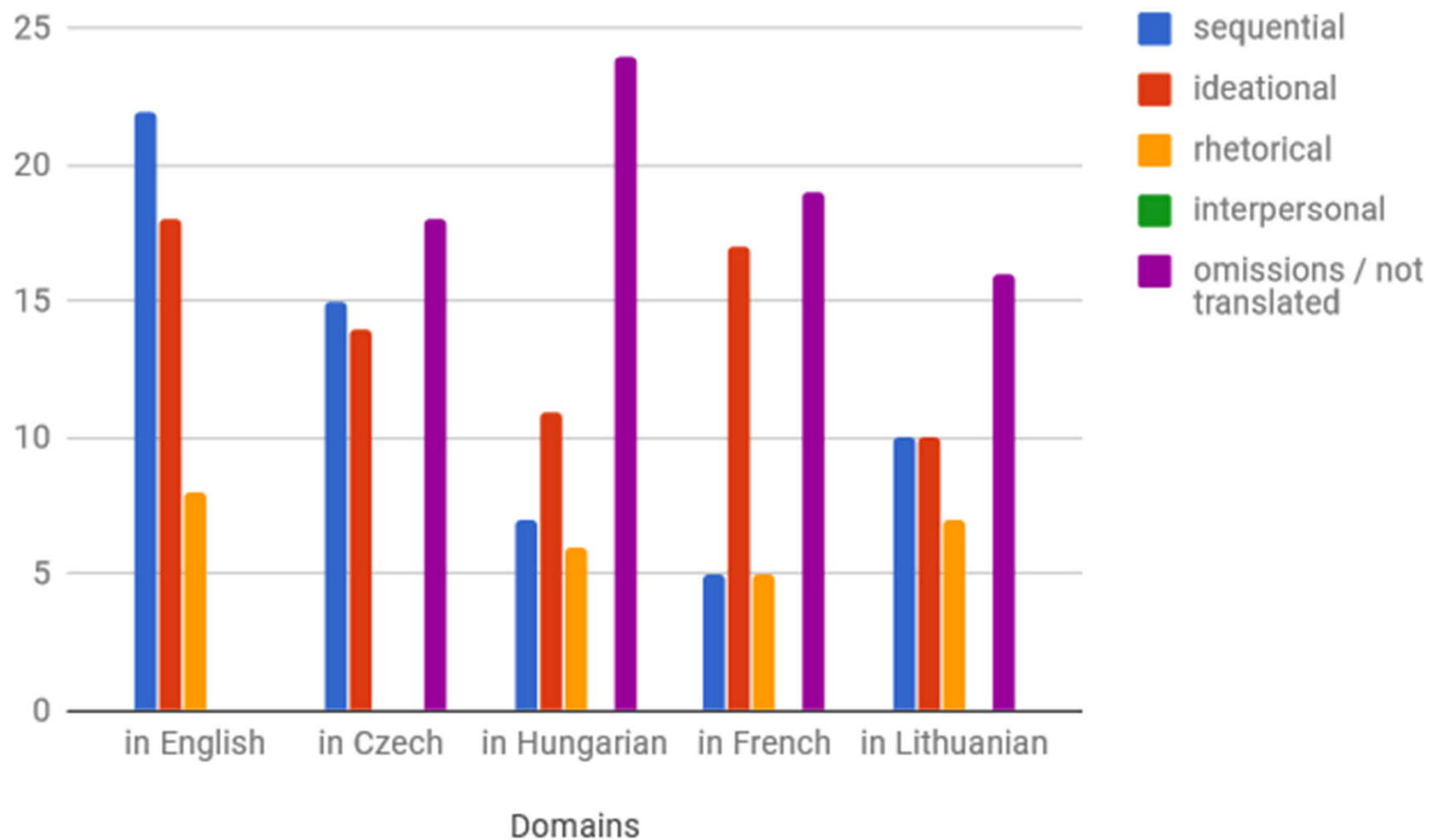
English 'and' – Domains
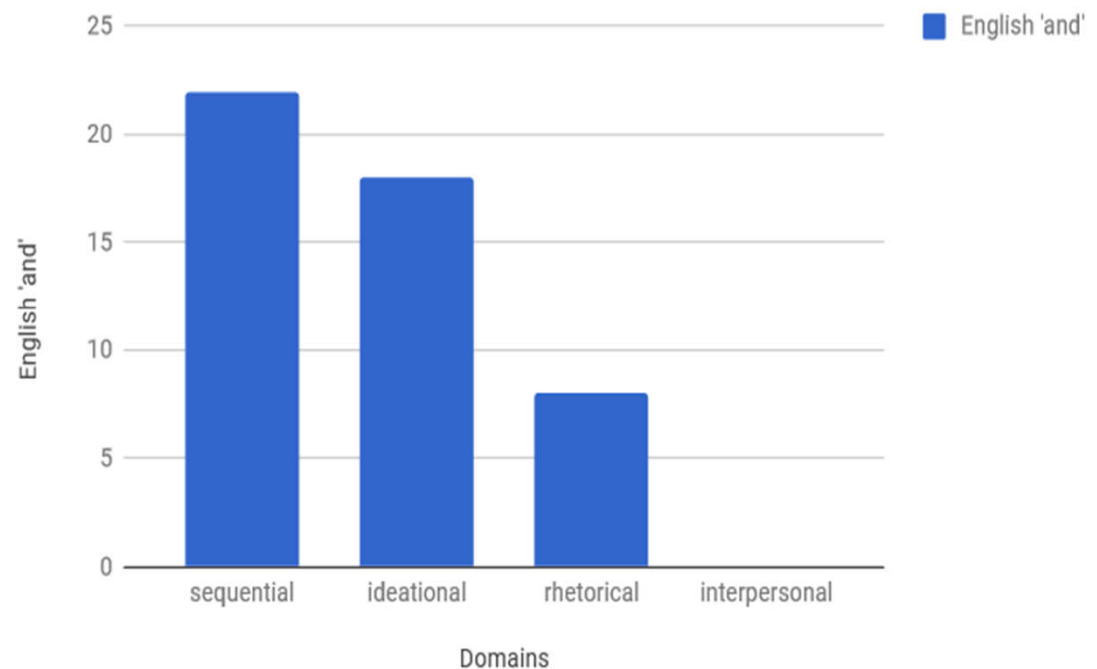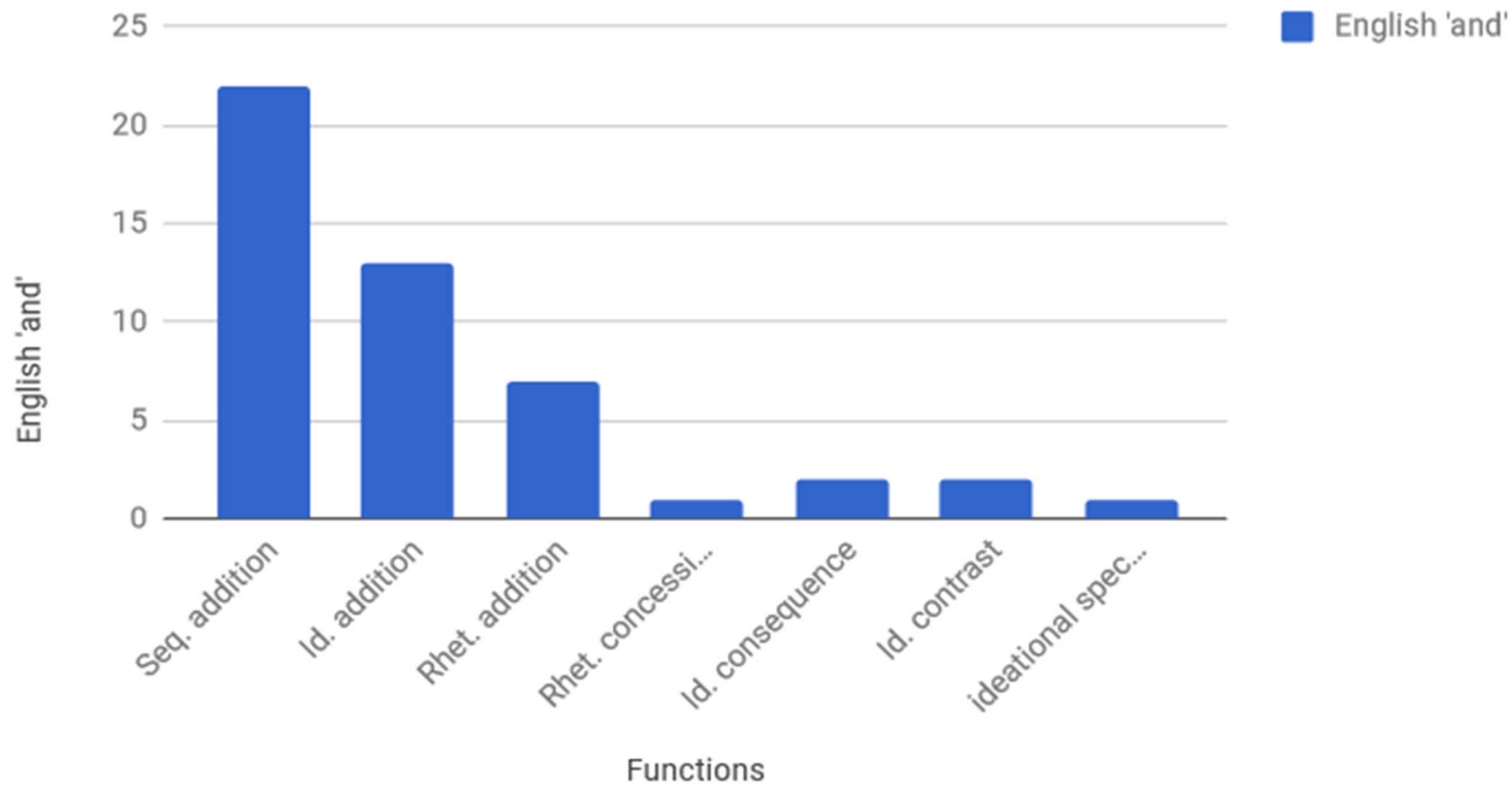
Less common/peripheral functions of *and*

- *[Now], Peter's not a very greedy man. Of all of the available women in the U.K., all Peter's looking for is somebody who lives near him, somebody in the right age range, somebody with a university degree, somebody he's likely to get on well with, somebody who's likely to be attractive, somebody who's likely to find him attractive. [**And**] comes up with an estimate of 26 women in the whole of the UK.* → **ideational consequence**

- *Of course, we didn't listen to them, [**and**] [instead], we championed the tender gestures of love and headlong flashes of youth.* → **ideational contrast**

- *How ironic that I work in human resources, a profession that works to welcome, connect and encourage the development of employees, 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.* → **rhetorical concession**

Distribution of the domains of 'and'

The distribution of domains

- The functional spectra of the translation equivalents of *and* in CZ, HU, LIT and FR differ in terms of the types of functions/domains as well as their proportions.
- In English and Czech the most prominent domain is the sequential domain; in Hungarian and French the ideational domain is the most common; while in Lithuanian we find an equal number of occurrences of sequential and ideational.

Functions within the rhetorical domain (in each language)


Rhetorical functions
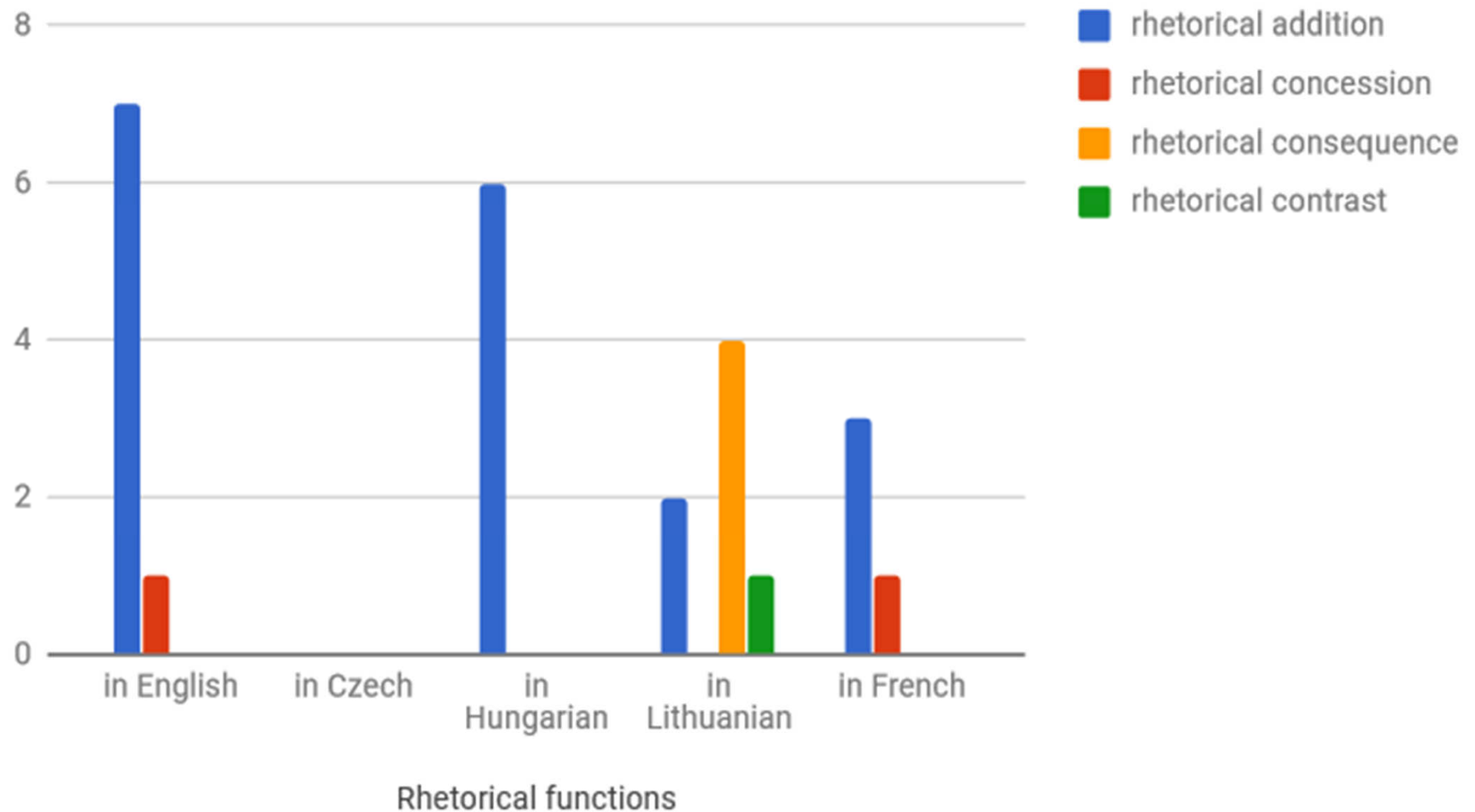
The rhetorical domain

- It is interesting to see and raises a question why the rhetorical domain is absent/inactive in Czech, while it operates in all other languages.
- Among the scrutinized languages, the rhetorical domain is the most productive and most colourful in Lithuanian.
- In Lithuanian the rhetorical domain uncovers various shades of 'and' and reflects a number of differences compared to other languages.
  - We can find the largest number of different types of functions (3 functions) within the rhetorical domain in Lithuanian.
  - We can find rhetorical contrast rendered into the Lithuanian by *o* ('*but and*') which can simultaneously represent both addition and contrast (unlike *ir* '*and*' expressing addition or *bet* '*but*' expressing contrast)
  - Rhetorical consequence is also observed (in the case of the translation of *and* as *taigi* '*so*') which is mostly related to the whole argument.

Functions and languages
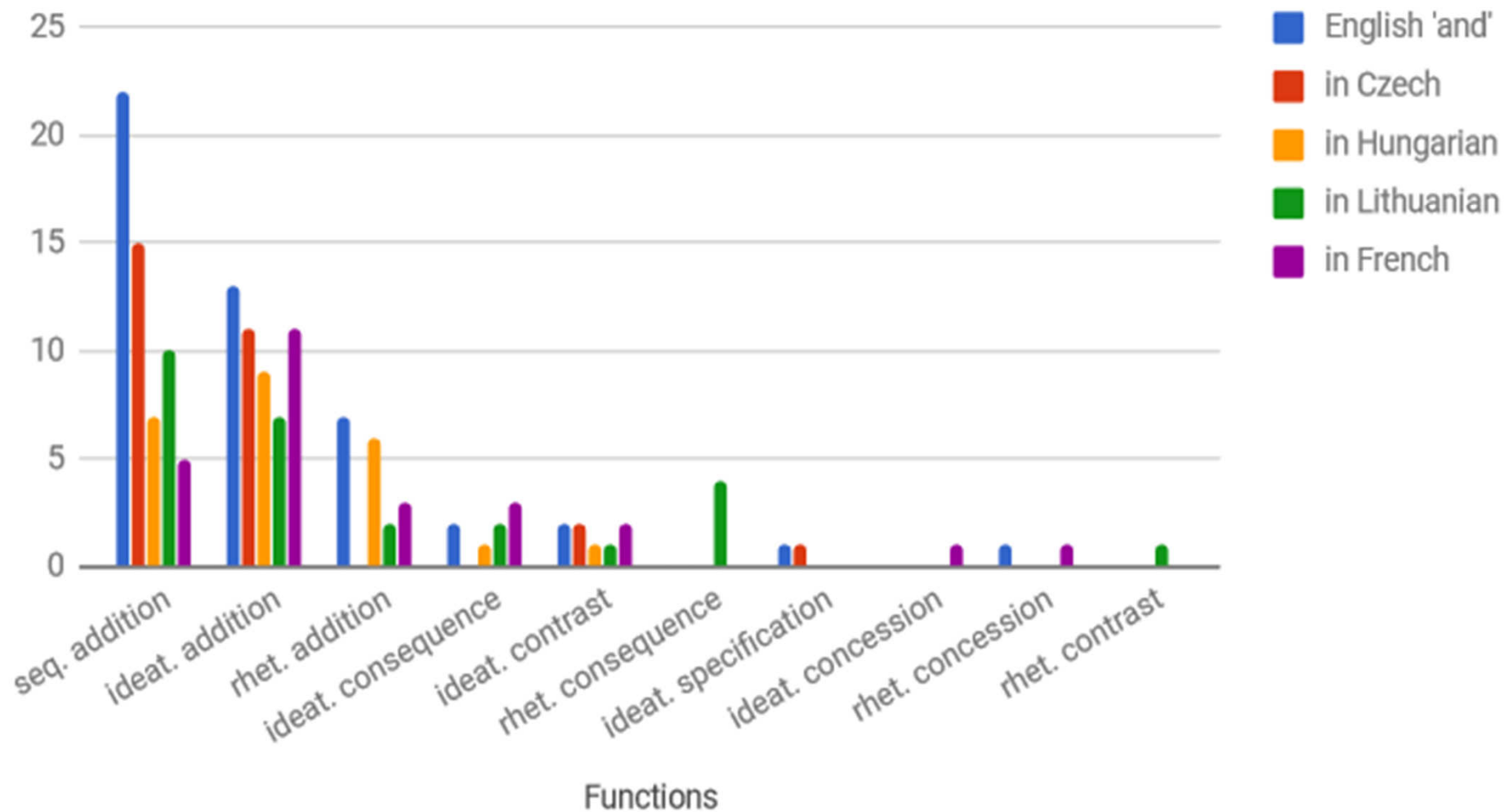
Findings: Functions

- The functions from the original English are not necessarily the same in the translations.
- The functional spectra of the translation equivalents of *and* in CZ, HU, LIT and FR differ in terms of the types of functions/domains as well as their proportions:
  - more translation equivalents and more types of functions in LIT, HU and FR than in CZ.

Findings: Translations of *and*

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

Examples: Translations of *and*

En   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.

Hu   **S** mikor leugrottam a sarki henteshez 'halal' húsért, fura érzésem támadt.

Li   **A** pokaždé, když jsem šel do místího řeznictví koupit halal maso něco nesedělo. (but and)

Fr   **Mais** lorsque que j'allais chez le boucher acheter de la viande halal, quelque chose coinçait. (but)

Cz   Kiekvieną kartą, kai eidavau pirkti mėsos pas vietinį mėsininką, jaučiau kažką keisto. (**omission**)

Domain: sequential (in FR: rhetorical)
Function: addition (in FR: concession - *and* is translated by a stronger connective, mais '*but*')

The influence of DM clusters on translation

*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' (the meanings of two connectives in English are **merged** and expressed by a single connective in French; as a result, *mais 'but'* is stronger than the original *and yet* in English - cross-lingual shift from weak to strong DM)

Findings: Omissions

- Another striking feature is the high frequency of omissions in the translations of the discourse marker *and.* It results in implicitation in the translation.
- Cross-linguistically, the uses of ***and* in the sequential domain are the most frequently omitted, whereas *and*'s operating in the ideational domain are usually preserved.**
- SEQ-ADD is the basic continuation function and does not bring a lot of information, whereas **IDE-ADD really signals a true semantic addition like a "plus" sign, so we lose less information by removing SEQ-ADD than IDE-ADD.**
- The more omissions there are in a translation, the fewer sequential connectives there are.
- Some *and*'s are maintained (to avoid juxtaposition) and some are removed (partly due to constraints of the translation by subtitles).

Omissions - examples

*En   [So] the math says [then] that what you should do in the first 37 percent of your dating window, you should just reject everybody as serious marriage potential. [**And**] [then], you should pick the next person that comes along that is better than everybody that you've seen before.*

*Cz   [**Ir**] [tuomet] pasirinkite kitą, kuris bus daug geresnis nei kada regėjot.*

*Hu   [Utána] le kéne csapni az első olyanra, akivel összetalálkozunk, és aki jobb bárki korábbinál. (**omission**)*

*Li   [**A**] [pak] byste se měli vrhnout do náruče nejbližšímu člověku, kterého potkáte [a] který je lepší než všechno, co jste dosud viděli.*
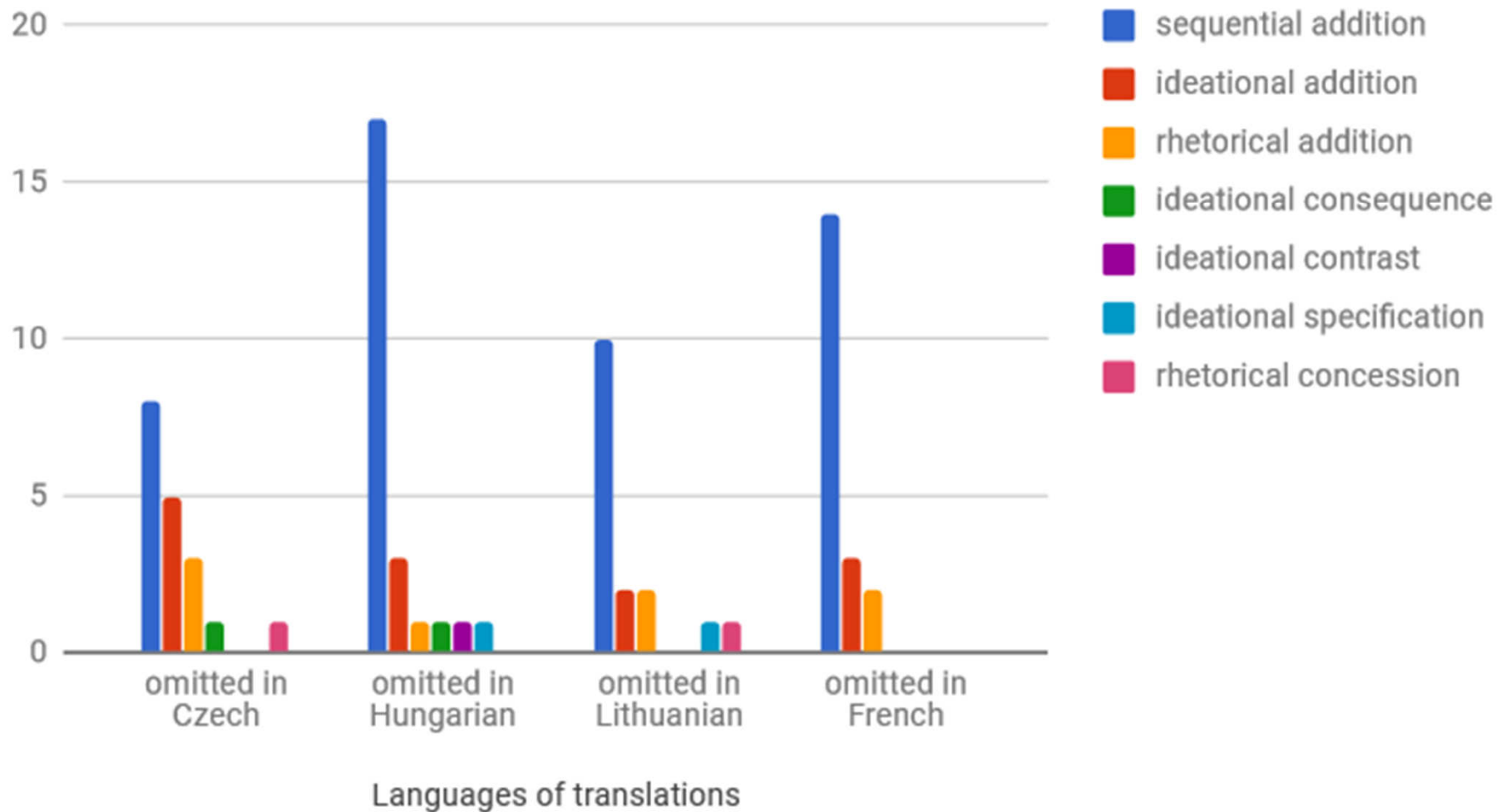
*Fr   [Ensuite], vous devez choisir la première personne meilleure que tous les autres. (**omission**)*

Domain: sequential
Function: addition

Conclusions

- The DM *and* expresses a very wide functional spectrum, much larger than simply addition or continuation.
- Thanks to the distinction of domains in the annotation scheme, we can distinguish various subtypes of additions (sequential, ideational and rhetorical).
- The functional spectra of *and* and its translation equivalents in CZ, HU, LIT and FR differ in terms of the distribution of their domains and functions:
  - There are more translation equivalents and more types of functions in LIT, HU and FR than in CZ.
- The use of domains as well as the use of translation corpora are both beneficial tools for the identification of instances and types of underspecification, both monolingual(ly) and multilingual(ly).

Future perspectives

- annotate more TEDTalks
- analyse *and* in conversation
- translation spotting in parallel corpus
- moving in the other direction: from translated texts back to the original English
- start from French *et*, Lithuanian *ir*, Czech *o*, Hungarian *és*, look at their English counterparts, and identify further instances/types of underspecification and implicitation, e.g.
- **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]
- **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.*
    - → "weak" DM ('et') in the translation (FR), "stronger" DM ('in fact') in the original, in EN

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).
- Degand, L.: A Situation-based Approach in Duth with some Implications for Text Generation. Doctoral dissertation, Université catholique de Louvaine (1996).
- Halliday, M.A.K.; Hasan, R.: Language, Context and Text: Aspects of language in a social semiotic perspective. Geelong, Victoria: Deakin University Press (1985)
- Redeker, G.: Ideational and pragmatic markers of discourse structure. JP 14(3):367-81. (1990)
- Redeker, G.: Review article: Linguistic Markers of Discourse Structure. Linguistics, 29(6):1139-72. (1991)
- Sanders, T. J.M.: Discourse Structure and Coherence: Aspects of a Cognitive Theory of Discourse Representation. Doctoral dissertation, Tilburg University (1992)
- Schiffrin, D.: Discourse Markers. Cambridge University Press, Cambridge (1987).

Acknowledgements