# Why we need a typological corpus: Gradience in description and typology

Typological studies often describe morphosyntactic features as having a limited number of discrete values: a language either has pro-drop or it does not, it either displays an alienability split in possessive marking or it does not, and so on. This practice obscures gradience in the actual use of these phenomena. For example, Dutch is not generally considered a pro-drop language, but studies find that first person singular subjects can be omitted in specific contexts (see Example (1), Neeleman & Sendrői 2007). Also, Dutch is not considered to have an alienability split, but under certain conditions, possessive pronouns on kinship terms can be dropped or replaced by definite articles (see Example (2), Leufkens & Van der Meulen, in prep.). This fine-grained information is often acknowledged in descriptive literature, but lost in large-scale comparative studies that would typically classify Dutch as 'non-pro-drop', and 'no alienability split'.

Comparative studies using gradient features do exist, and are often based on another type of data, namely language corpora. Since corpus studies go into much detail and require extensive bodies of intercomparable texts, they usually cannot compare more than 2 or 3 languages at the time. Consequently, typological studies typically compare many languages on limited-value features, or few languages on gradient features, but not many languages on gradient features. Even though the number of languages compared in corpus studies has grown considerably (Hasselgård 2020), samples sizes are extremely small compared to large-scale studies enabled by databases like WALS (Dryer & Haspelmath 2013).

In this talk, I will propose a way to overcome this problem. I will present a plan to create a typological corpus that contains spontaneous language data from an in principle unlimited number of languages of diverse genetic and areal affiliation. The corpus is different from parallel corpora like EuroParl (Koehn 2005), because it does not involve translation of one language into another. Hence, there is no source language imposing linguistic choices on target languages, which makes the corpus uniquely suitable for typological comparison. The corpus also differs from cross-linguistic corpora like Multi-CAST (Haig & Schnell 2021) because texts in the corpus will have pre-determined genres and topics (see Table 1), in order to increase the likelihood of obtaining tokens of crucial linguistic features. Finally, it complements the SCOPIC project (Barth & Evans 2017), which has a similar setup, in containing free narratives rather than picture-description tasks.

**Table 1:** *Genres and topics of texts to be collected in typological corpus*

| | Pre-planned | (Semi-)spontaneous |
|---|---|---|
| Monologue | Folk story | Frog story |
| Multilogue | Interview on someone's past | Conversation on someone's future |
| Written (if available) | Fairy tale | Diary/personal narrative |

First steps in the creation of this corpus will be to develop an annotation and tagging scheme, as well as criteria for the format and minimum length of texts included in the corpora. To get this right, dialogue between descriptive linguists and typologists is absolutely crucial, which is why this conference is the ideal occasion for discussing these plans.

The proposed corpus will not render reference grammars obsolete; rather, it will provide both language describers and typologists with better data on which they can base descriptions and comparisons. Moreover, descriptive linguists will play a crucial role in collecting and annotating the corpus texts.

## Examples

1) Dutch – Neeleman & Sendrői (2007: 674)

a)

| | | | | |
|---|---|---|---|---|
| Ø$_1$ | ken | ik | $t_1$ | niet |
| | know | I | | not |

‘I don’t know (x)’

b)

| | | | | |
|---|---|---|---|---|
| Ø$_1$ | ken | $t_1$ | hem | niet |
| | know | | him | not |

‘(I) don’t know him’

2) Dutch – Leufkens & Van der Meulen (in preparation)

a)

| | | | | | |
|---|---|---|---|---|---|
| De | peuter | vroeg | net | om | ‘sinisap’ |
| the | toddler | asked | just_now | for | ora_juice |

‘The (implied: my) toddler just asked for ora juice’

b)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dochter | heeft | haar | telefoon | met | spelletjes | aan | hem |
| daughter | has | her | phone | with | games | to | him |
| | gegeven... | | | | | | |
| | given | | | | | | |

‘(Implied: My) daughter has given him her phone with games…’

## References

Barth, Danielle & Nicholas Evans. 2017. SCOPIC design and overview. In Danielle Barth & Nicholas Evans (eds.), *Social Cognition Parallax Interview Corpus (SCOPIC)* (Language Documentation & Conservation Special Publication No. 12), 1–21. Honolulu: University of Hawai’i Press.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, accessed on June 22 2021.)

Haig, Geoffrey & Stefan Schnell, Stefan (eds.). 2021. *Multi-CAST: Multilingual corpus of annotated spoken texts*. (Available online at https://multicast.aspra.uni-bamberg.de/, accessed on June 30 2021.)

Hasselgård, Hilde. 2020. Corpus-based contrastive studies. *Languages in Contrast* 20(2), 184–208. https://doi.org/10.1075/lic.00015.has

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*, 79–86.

Leufkens, Sterre & Van der Meulen, Marten. In preparation. Definite articles for close relatives: Lexical variation in marking of kinship terms.

Neeleman, Ad, & Kriszta Sendrői. 2007. Radical Pro Drop and the Morphology of Pronouns. *Linquistic Inquiry* 38(4), 671–714.