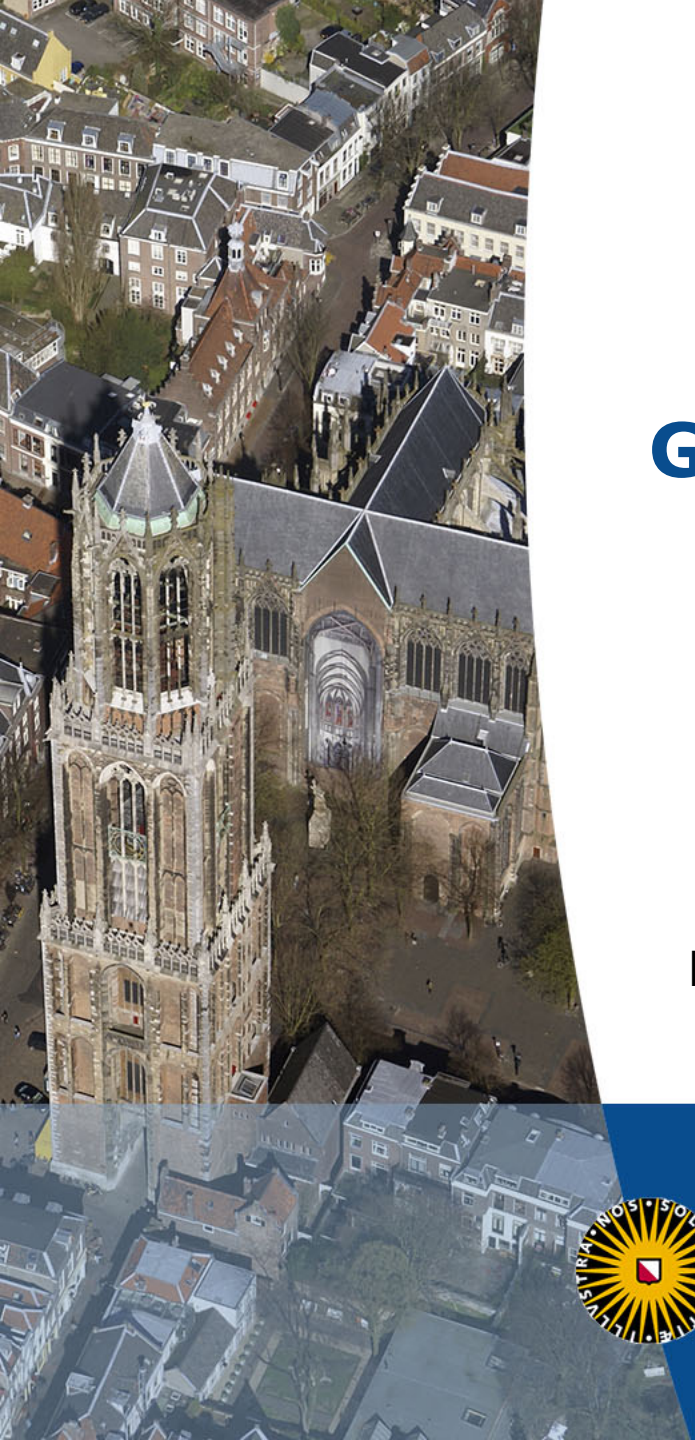


Why we need a typological corpus: Gradience in description & typology

Dr. Sterre Leufkens
Utrecht Institute of Linguistics OTS
December 3, 2021
Descriptive Grammars & Typology, Paris 2021

Universiteit Utrecht
Faculteit Geesteswetenschappen

Two experiences of a typologist & reference grammar user

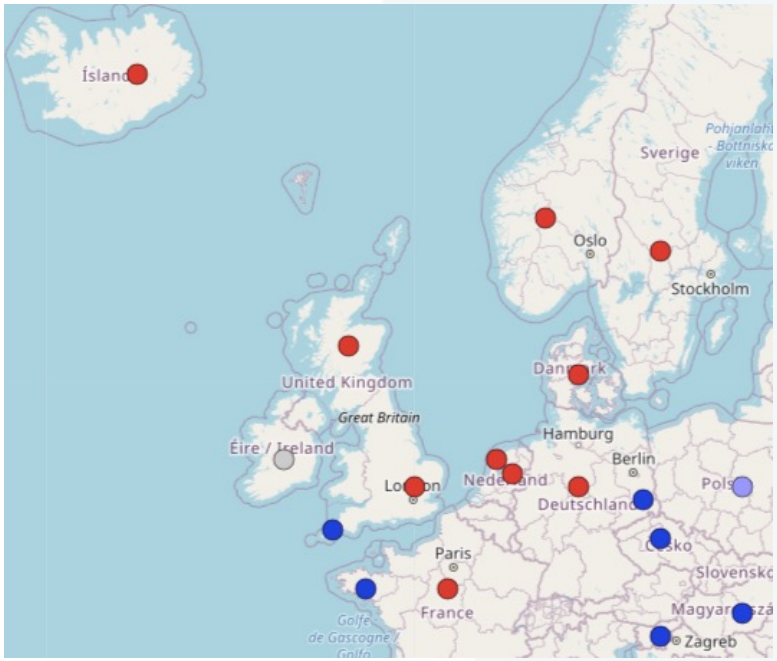
Schmalz 2013: ii

Outline

- Gradience in current typological research, corpus studies & grammars
- The issue
- Proposal for a new type of corpus to complement grammars & existing corpora
- Asking for your comments, thoughts, and collaboration

Example (1): pro-drop

- Large-scale typological study (Dryer 2013)



Values

●	Obligatory pronouns in subject position	82
●	Subject affixes on verb	437
●	Subject clitics on variable host	32
●	Subject pronouns in different position	67
●	Optional pronouns in subject position	61
●	Mixed	32

Example (1): pro-drop

- Reference grammar of Dutch (e-ANS, Haeseryn et al. 2019)

  Most sentences contain a subject, except for
  - Small clauses
  - Contracted clauses
  - Imperatives
  - Incomplete sentences
  - Some idiomatic expressions

Example (1): pro-drop

- Small-scale typological study (Neeleman & Sendrői 2007: 674):

  in Dutch, sentence-initial pro's can be omitted (topic drop)

- Corpus study (Nariyama 2004: 237):

  "in English, subject ellipsis is, in fact, a common phenomenon in conversation and casual letters"

Example (2): alienability split

- Large-scale typological study (Nichols & Bickel 2013)

Values

○	No possessive classification	125
●	Two classes	94
●	Three to five classes	20
●	More than five classes	4

Example (2): alienability split

- Reference grammar on Dutch (Haeseryn et al. 2019)

  Expression of possessive relation does not depend on semantics of possessee noun,

  except for regional use of 'possessive definite article' for body parts ('the dog wagged the tail')

Example (2): alienability split

- Twitter corpus study (Leufkens & Van der Meulen, in prep.):

  *"Dus in joggingbroek met kop koffie op de bank. Terwijl **de** kleuter in pyjama en ongekamde haartjes lekker zit te kleuren. **De** puber is nog niet gesignaleerd."*

  'So in sweat pants with cup of coffee on the couch. While **the** [my] toddler is coloring in their pajamas and with unbrushed hair. **The** [my] adolescent has not been seen yet.'

The issue

- Small-scale typological studies & corpus studies find and describe gradient variation & phenomena that only appear in highly specific contexts, but such studies involve only few lgs

- Reference grammars may or may not describe gradient variation & highly context-specific phenomena

- Large-scale typological studies & databases abstract away from gradience & context-specific phenomena, losing information ("type-based typology", Levshina 2019)

The issue

- Solution: base large-scale cross-linguistic comparison not only on reference grammars, but on multilingual corpora (“token-based typology”, Levshina 2019)

- Problem: existing corpora are often
  - Uni-/bi-/tri-lingual
  - Biased for (Indo-)European languages
  - Poorly cross-linguistically comparable
  - Translations from a source language = not typologically valid

Existing typological corpora

- There are exceptions, such as (among others):
  - Multi-CAST (Haig & Schnell, eds.)
    - Annotation scheme Grammatical Relations and Animacy in Discourse (GRAID; Haig & Schnell 2014)
    - Spoken monologues
  - SCOPIC (Barth & Evans 2017)
    - Semantic domain: social cognition
    - Spoken language elicited by means of picture task

Proposal

- Create typological corpus
  - without restrictions on semantic or grammatical domain
  - with (some) control on content, to enable comparison

- Benefits
  - Allows for detail & gradience, no *a priori* abstraction
  - Enables cross-linguistic comparison of any phenomenon a typologist could be interested in
  - Phenomena can be found regardless of terminology
  - Capture intralinguistic variation of any type

Typological corpus: content

spontaneous ——— controlled

+ easily available

+ real lg data

+ cross-lg comparable

	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

- Possibly work with conversational questionnaires (Francois 2019) or other elicitation tasks

Typological corpus: annotation

Lg-specific ——— Comparative Concepts (Haspelmath 2010)

+ detailed

+ cross-lg comparable

+ smaller set

- Universal Dependencies scheme (de Marneffe et al. 2021), possibly combined with more specific schemes
- Interface + (online) recording & annotation training such that any fieldworker/expert could add content

Typological corpus

- Reference grammars remain essential
  - for description & analysis of items/rules > basis for annotation
  - for cross-linguistic generalisation & categorisation, type-based typology

- Typology would benefit from a wider range of descriptive sources
  - Reference grammar + small-scale corpus work + large-scale typological corpus

Discussion Qs

- What would fieldworkers/descriptive linguists need to be able to contribute to a typological corpus?

- What would typologists need for a typological corpus to be useful and usable?

Thank you!

References

- Barth, D., & Evans, N. (2017). The social cognition parallax corpus (SCOPIC): Design and overview. In D. Barth & N. Evans (Eds.), *Social Cognition Parallax Corpus (SCOPIC),* 1–21. Honolulu: University of Hawai'i Press. http:/hdl.handle.net/10125/24742
- Dryer, Matthew S. (2013). Expression of Pronominal Subjects. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/101, Accessed on 2021-11-28.)
- François, A. (2019). A proposal for conversational questionnaires. In A. Lahaussois & M. Vuillermet (Eds.), *Methodological Tools for Linguistic Description and Typology* (Vol. 16), 155–196. Honolulu: University of Hawai'i Press. http://nflrc.hawaii.edu/ldc
- Haeseryn, W., Romijn, K., Geerts, G., Rooij, J. & Toorn, M. (2019). *20.2.1 Inleiding. Algemene Nederlandse Spraakkunst*. https://e-ans.ivdnt.org/topics/pid/ans200201lingtopic (geraagdpleegd 28 November 2021).

References

- Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts.* (multicast.aspra.uni-bamberg.de/) (accessed 2021)
- Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, *86*(3), 663–687. https://doi.org/10.1353/lan.2010.0021
- Leufkens, Sterre & Marten Van der Meulen, in prep. Definite articles for close relatives: Lexical variation in marking of kinship terms.
- Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, *23*(3), 533–572. https://doi.org/10.1515/lingty-2019-0025
- de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402
- Nariyama, S. (2004). Subject ellipsis in English. *Journal of Pragmatics*, *36*(2), 237–264. https://doi.org/10.1016/S0378-2166(03)00099-7

References

- Neeleman, A., & Szendro, K. (2007). Radical Pro Drop and the Morphology of Pronouns. *Linquistic Inquiry*, *38*(4), 671–714.
- Nichols, Johanna, Balthasar Bickel (2013). Possessive Classification. In Dryer, Matthew S. & Haspelmath, Martin (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/59, Accessed on 2021-11-28.)
- Schmalz, Mark (2013). *Aspects of the grammar of Tundra Yukaghir*. (Doctoral dissertation, Universiteit van Amsterdam).