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
Two experiences of a typologist & reference grammar user

3.3 Noun morphology
3.3.1 Noun inflexion
3.3.1.1 Cases
3.3.1.1.1 Nominative
3.3.1.1.2 Accusative
3.3.1.1.3 Ergative
3.3.1.1.4 Absolutive
3.3.1.1.5 Genitive
3.3.1.1.6 Dative
3.3.1.1.7 Instrumental
3.3.1.1.8 Locative
3.3.1.1.9 Ablative
3.3.1.1.10 Prolative
3.3.1.1.11 Comitative
3.3.1.1.12 Purposive
3.3.1.2 Number
3.3.1.3 Pertensive
3.3.2 Noun formation
3.3.2.1 Suffixal derivation
3.3.2.2 Conversion
3.3.2.3 Compounding
3.4 Verb morphology
3.4.1 Verb subclasses
3.4.1.1 Action verbs
3.4.1.2 Qualitative verbs
3.4.1.3 Quantitative verbs
3.4.1.4 Denominal verbs
3.4.1.5 The deictic verb
3.4.2 Verb inflexion

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)
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 & Sendroïi 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)
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 \textit{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

<table>
<thead>
<tr>
<th></th>
<th>Pre-planned</th>
<th>(Semi-)spontaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monologue</td>
<td>Folk story</td>
<td>Frog story</td>
</tr>
<tr>
<td>Multilogue</td>
<td>Interview on someone’s past</td>
<td>Conversation on someone’s future</td>
</tr>
<tr>
<td>Written (if available)</td>
<td>Fairy tale</td>
<td>Diary/personal narrative</td>
</tr>
</tbody>
</table>

- Possibly work with conversational questionnaires (Francois 2019) or other elicitation tasks
Typological corpus: annotation

- 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

Comparative Concepts (Haspelmath 2010)

+ detailed
+ cross-lg comparable
+ smaller set
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!

3.3 Noun morphology
  3.3.1 Noun inflexion
    3.3.1.1 Cases
      3.3.1.1.1 Nominative
      3.3.1.1.2 Accusative
      3.3.1.1.3 Ergative
      3.3.1.1.4 Absolutive
      3.3.1.1.5 Genitive
    3.3.1.6 Dative
    3.3.1.7 Instrumental
    3.3.1.8 Locative
    3.3.1.9 Ablative
    3.3.1.10 Prolative
    3.3.1.11 Comitative
    3.3.1.12 Purposive
  3.3.1.2 Number
  3.3.1.3 Pertensive

3.3.2 Noun formation
  3.3.2.1 Suffixal derivation
  3.3.2.2 Conversion
  3.3.2.3 Compounding

3.4 Verb morphology
  3.4.1 Verb subclasses
    3.4.1.1 Action verbs
    3.4.1.2 Qualitative verbs
    3.4.1.3 Quantitative verbs
    3.4.1.4 Denominal verbs
    3.4.1.5 The deictic verb
  3.4.2 Verb inflexion
References


References

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