Measuring the complexity of grammars: morphosyntactic variation in the Anglophone world

Bernd Kortmann
(Freiburg University)

Structure of the talk

1. Introduction
   1.1 Complexities
   1.2 Complexity debates
   1.3 The present study
2. Measuring complexity in varieties of English
   2.1 Complexity in survey data
   2.2 Complexity in naturalistic corpus data
   2.3 Conclusions
3. Bonus 1: Measuring complexity in different languages (written & spoken)
4. Bonus 2: Measuring analyticity and syntheticity in Pidgins and Creoles: Tok Pisin vs Hawaii Creole

Global vs. local complexity
(Miestamo 2008)

- global linguistic complexity: complexity of a language, dialect, etc. as such
- local linguistic complexity: domain-specific, i.e.
  -> phonological, morphological, syntactic, semantic/lexical, pragmatic (or: ’hidden’) complexity
  -> significant correlations, trade-offs, „balancing effects“?

Absolute vs. relative complexity measures
(e.g. Miestamo 2006a,b, 2008)

- **absolute complexity**
  - theory-oriented
  - objective
  - ’more is more complex’
- **relative complexity**
  - language user-oriented
  - subjective
  - ’cost’, difficulty in processing and learning

Some absolute complexity metrics
(’more is more complex’: structural elaboration)

- number of grammatical categories
  (e.g. Shosted 2006)
- number of phonemic contrasts
  (e.g. McWhorter 2001)
- length of the description of a grammatical/ phonological/ … system
  (Dahl 2004: 21-24)
- token frequencies of grammatical markers
  (Szmrecsanyi and Kortmann 2008, 2009a-c)

Some relative complexity metrics

- L2 acquisition complexity (=difficulty)
  reference point: L2 learners
  (e.g. Trudgill 2001)
- redundancy-induced (-> ornamental rule) complexity
  reference point: language users
  (e.g. Trudgill 1999; McWhorter 2001)
- irregularity-induced complexity
  reference point: language users/processors
  (e.g. Mühlhäusler 1974, Trudgill 2004, McWhorter 2012)
Conditioning factors

• older languages are more complex than younger languages (McWhorter 2001)
• language contact & adult language acquisition \(\Rightarrow\) simplification (Trudgill 2009)
• size of the speaker community, density of social networks (proxies for contact?) (Trudgill 2004)

Complexity/complexification vs. simplicity/simplification: Currently widely debated issues

1.3 The present study

• focus exclusively on structural, „surfacy“ (morphosyntactic) complexity
• large-scale empirical, comparative analyses covering
  – 3 notionally different complexity metrics
  – a number of different dialect variéty types (traditional L1s, high-contact L1s, L2s, PCs)
  – 2 data types (survey data, naturalistic corpus data)

Objectives:

• To what extent are complexity levels sensitive to variety type?
• Are there trade-offs between complexity types?
• Are there trade-offs between syntheticity and analyticity?
• What is the extent to which language contact and/or (adult) language learning might lead to morphosyntactic simplification?
• What is the mileage of our metrics for language-internal variation and cross-linguistic variation?

Types of complexity considered

• quantitative complexity
  – "more is more complex"- complexity (cf. Arends 2001:180)
    • size of marker/rule inventory, number of "ornamental" markers/rules, i.e. those involving more form/code and/or more rules without added communicative bonus (cf. McWhorter 2001, Shustal 2006; Trudgill 1999)
    • verbosity (cf. Dahl 2004) - here: grammaticality
      text frequency of grammatical markers, synthetic or analytic (cf. Greenberg 1960)
  • L2-acquisition complexity
    • number of features in a variety’s inventory known to recur in interlanguage varieties
  • complexity deriving from irregularities and low transparency (cf. McWhorter 2001, Trudgill 2004)
    • text frequency of irregular, lexically conditioned grammatical morphemes

Data sources

• survey data
  – 46 varieties of English (all spoken)
  – 76 morphosyntactic features (all non-standard)
• naturalistic corpus data
  – transcribed conversational material from 15 corpora sampling spoken varieties of English
    + one written corpus (written British English)
**Method in 2.1**

- classification of features in survey into
  - 'ornamentally complex' features
    i.e. those that complicate the system, vis-à-vis the standard system, without clearly yielding an added communicative bonus
  - simplifying features
    i.e. those that simplify the system, vis-à-vis the standard system
  - L2-simple features
    i.e. those that are known to recur in interlanguage varieties
- establishing corresponding indices for each variety in survey

**Interim summary for 2.1: survey data**

- traditional L1s are most ornamental as expected; cf. McWhorter (2001), Trudgill (2001)
- a puzzle: why don't we find many simplifying and/or L2-simple features in L2s? given the literature we should

**Method in 2.2**

- calculation of 4 Greenberg-inspired (cf. Greenberg 1960) frequency indices:
  - grammaticity indices = total frequency of grammatical markers per sample (quantitative complexity)
  - analyticity indices = total frequency of free grammatical morphemes/function words per sample (quantitative complexity)
  - syntheticity indices = total frequency of bound grammatical morphemes per sample (quantitative complexity)
  - transparency indices = percentage of bound grammatical morphemes in sample which are regular (irregularity/low transparency)

**Interim summary for 2.2: corpus data**

- traditional L1s are most synthetic and least transparent while L2s are least synthetic and most transparent as one would expect – cf. Seuren & Wekker (1986), Klein & Perdue (1997), Trudgill (1999, 2001, 2004), inter alia
- in general: traditional L1s exhibit the highest degree of grammaticity and L2s the lowest degree
- in cross-variety perspective, no trade-off between syntheticity and analyticity, but transparency correlates negatively with grammaticity i.e. the more grammatical markers, the lower the number of transparent grammatical markers; and vice versa
- written E is clearly an outlier while the spoken standard varieties maintain a low profile – one that is akin to high-contact non-standard L1s – in every respect standard dialects are just another type of high-contact varieties (cf. Trudgill 2009)

---

**Table 1. 46 Varieties sampled in the World Atlas**

<table>
<thead>
<tr>
<th>varieties</th>
<th>variety type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orkney and Shetland, North, Southwest and Southeast of England, East Anglia, Isolated Southeast US E, Newfoundland E, Appalachian E</td>
<td>traditional L1 (8)</td>
</tr>
<tr>
<td>Chichewa, Suriname Creoles, Belizean Creole, Tobagonian/Trinidadian Creole, Bahamian E, Jamaican Creole, Bislama, Solomon Islands Pidgin, Tok Pisin, Hawaiian Creole, Aboriginal E, Australian Creoles, Ghanaian Pidgin E, Nigerian Pidgin E, Cameroon Pidgin E</td>
<td>P/C (15)</td>
</tr>
</tbody>
</table>
• thus, based on our English data, language contact very systematically results in a lower degree of complexity
• at the same time, L2s have a strikingly different complexity profile from English-based PCs
• converging evidence, survey & corpus: for L2s in particular, the alternative to L2-difficult syntacticity seems to be no grammatical marking at all, rather than analytic marking or 'overly simple' marking

OUTLOOK - Where to go from here?

• survey data:
  – extending feature catalogue (235 in 74 varieties in WAVES)
  – cooperation with MPI Leipzig (APICS = Atlas of Pidgin and Creole Structures)
• naturalistic data:
  – for Pidgins and Creoles
  – for L1s and L2s of English

2.3 Conclusions – Part II

• advantages over our 2 sets of metrics offer concern
  – absolute holistic complexity measures (cf. Siegel 2004)
  – comparisons across varieties and variety types
  – the trade-off between syntacticity and analyticity
• this kind of large-scale study of complexity in language-internal varieties is, in principle, possible for any language
• large-scale language-internal variation as a testing ground for developing and calibrating complexity metrics which can be used for complexity variation across languages, too

Selected References


