Wuppertal, 23.1.2013

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 ⇨ simplification
  (Trudgill 2009)
- size of the speaker community, density of social networks (proxies for contact?)
  (Trudgill 2004)

7

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

Kusters, W. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection.* Utrecht: LOT.

Dahl, Ö. 2004. *The growth and maintenance of linguistic complexity.* Amsterdam/Philadelphia: John Benjamins.

Miestamo, M., K. Sinnemäki and F. Karlsson, eds. to appear 2008. *Language Complexity: Typology, Contact, Change.* Amsterdam/Philadelphia: Benjamins.

Sampson, G., D. Gil, P. Trudgill, eds. 2009. *Language Complexity as an Evolving Variable.* Oxford: Oxford University Press.

McWhorter, J. 2011. *Linguistic Simplicity and Complexity.* Berlin/Boston: De Gruyter.

Trudgill, P. 2011. *Sociolinguistic Typology.* Oxford: Oxford University Press.

Kortmann, B., B. Szmrecsanyi, eds. 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact.* Berlin/Boston: De Gruyter.

8

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/variety types (traditional L1s, high-contact L1s, L2s, PCs)
  - 2 data types (survey data, naturalistic corpus data)

9

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?

10

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; Shosted 2006; Trudgill 1999)
    - *verbosity* (cf. Dahl 2004) – here: *grammaticity*
      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

11

Data sources

- **survey data**
  - *The World Atlas of Morphosyntactic Variation in English*
    (Kortmann & Szmrecsanyi 2004)
  - 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)
  - calculation of Greenberg-inspired indices
    (cf. Greenberg 1960: „A Quantitative Approach to the Morphological Typology of Language“. *International Journal of American Linguistics* **26:** 178-194.)

12

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

varieties	*variety type*
Orkney and Shetland, North, Southwest and Southeast of England, East Anglia, Isolated Southeast US E, Newfoundland E, Appalachian E	traditional L1 (8)
Scottish E, Irish E, Welsh E, Colloquial American E, Ozarks E, Urban African-American Vernacular E, Earlier African-American Vernacular E, Colloquial Australian E, Australian Vernacular E, Norfolk, regional New Zealand E, White South African E	high-contact L1 (12)
Chicano E, Fiji E, Standard Ghanaian E, Cameroon E, East African E, Indian South African E, Black South African E, Butler E, Pakistan E, Singapore E, Malaysian E	L2 (11)
Gullah, 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	P/C (15)

13

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

14

Interim summary for 2.1: survey data

- traditional L1s are most ornamental
  as expected; cf. McWhorter (2001), Trudgill (2001)
- English-based PCs attest a substantial number of L2-simple and of simplifying features
  no surprise either – cf. Seuren & Wekker (1986), Trudgill (2004)
- a puzzle: why don't we find many simplifying and/or L2-simple features in L2s?
  given the literature we should

15

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)

16

Method in 2.2

- for each variety/corpus, extraction of a random set of 1,000 tokens (orthographically transcribed words)
  ⇨ total *N*: 16,000 tokens
- morphological/grammatical analysis of those tokens:
  - does the token carry a grammatical suffix/does it bear a grammatical morpheme? If so, is the suffix/morpheme
    - ✓ regular
    - ✓ irregular (lexically conditioned)?
  - is the token a function word, i.e. does it belong to a closed/grammatical class (determiners, pronouns, modal verbs, conjunctions, etc.)?

17

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)

18

2.3 Conclusions – Part I

- variety type predicts observable complexity levels rather well – much along the lines of McWhorter (2001, 2007) and Trudgill (2001, 2009, 2012)
- 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 syntheticity seems to be no grammatical marking at all, rather than analytic marking or 'overtly simple' marking

2.3 Conclusions – Part II

- advantages our 2 sets of metrics offer concerning
  - absolute holistic complexity measures (cf. Siegel 2004)
  - comparisons across varieties and variety types
  - the trade-off between syntheticity 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

OUTLOOK - Where to go from here?

- survey data:
  - extending feature catalogue (235 in 74 varieties in WAVE)
    → Kortmann/Lunkenheimer 2011 and 2012
  - cooperation with MPI Leipzig (APiCS = *Atlas of Pidgin and Creole Structures*)
- naturalistic data:
  - for Pidgins and Creoles
    → Siegel/Szmrecsanyi/Kortmann in press
  - corpus analyses for more L1 and L2-varieties of English
- learner varieties of English (ILCE, Louvain-la-Neuve)
  → Szmrecsanyi/Kortmann 2011
- stylistic genre analyses (BNC) → Szmrecsanyi 2009
- history of English (e.g. ARCHER) → Szmrecsanyi 2012
- language comparison → Szmrecsanyi in press

Publications on complexity by Kortmann and/or Szmrecsanyi

Kortmann, B. and B. Szmrecsanyi (2009). World Englishes between simplification and complexification. In: Siebers, L. and T. Hoffmann, eds. *World Englishes - Problems, Properties and Prospects: Selected Papers From The 13th IAWE Conference*. Amsterdam: Benjamins, 265-285.

Kortmann, B. and B. Szmrecsanyi (2011). Parameters of morphosyntactic variation in World Englishes: prospects and limitations of searching for universals. In: Siemund, P., ed. *Linguistic Universals and Language Variation*. Berlin/New York: De Gruyter Mouton. 264-290.

Kortmann, B. and B. Szmrecsanyi (eds.) (2012). *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin/Boston: De Gruyter.

Siegel, J., B. Szmrecsanyi and B. Kortmann (in press). Measuring analyticity and syntheticity in creoles. *Journal of Pidgin and Creole Languages*.

Szmrecsanyi, B. and B. Kortmann. (2009a). Vernacular universals and angloversals in a typological perspective In: Filppula, M., J. Klemola and H. Paulasto, eds. *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. London/New York: Routledge. 33-53.

Szmrecsanyi, B. and B. Kortmann. (2009b). Between simplification and complexification: Non-standard varieties of English around the world. In: Sampson; G., D. Gil and P. Trudgill, eds. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 64-79.

Szmrecsanyi, B. and B. Kortmann (2009c). The morphosyntax of varieties of English worldwide: A quantitative perspective. Special issue of *Lingua* 119/11: 1643-1663.

Szmrecsanyi, B. and B. Kortmann (2011). Typological profiling: learner Englishes versus indigenized L2 varieties of English. In: Mukherjee, J. and M. Hundt, eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: Benjamins. 167-187.

Szmrecsanyi, B. and B. Kortmann (2012). Introduction: Linguistic complexity – Second Language Acquisition, indigenization, contact. In: Kortmann, B. and B. Szmrecsanyi, eds. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: de Gruyter. 6-34.

Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21(3): 319–353.

Szmrecsanyi, B. (2012). Analyticity and syntheticity in the history of English. In: Nevalainen, T. and E. Closs Traugott, eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press. 654-665.

Szmrecsanyi, B. (to appear). Forests, trees, corpora, and dialect grammars. In: Szmrecsanyi, B. and B. Wälchli, eds. *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation and Text and Speech* (working title). Berlin: de Gruyter.

Selected References

Ansaldo, U., S. Matthews and L. Lim, eds. (2007). *Deconstructing Creole.* Amsterdam/Philadelphia: John Benjamins.

Bakker, P., A. Daval-Markussen, M. Parkvall and I. Plag. (2011). Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages* 26: 5-42.

Chambers, J.K. (2004). Dynamic typology and vernacular universals. In: Kortmann, B. ed. *Dialectology meets Typology.* Berlin/New York: Mouton de Gruyter. 127–145.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity.* Amsterdam/Philadelphia: John Benjamins.

Dixon, R.M.W. (1997). *The rise and fall of languages.* Cambridge: Cambridge University Press.

Greenbaum, S. (1996). *Comparing English worldwide: the International Corpus of English.* Oxford/New York: Clarendon Press/Oxford University Press.

Greenberg, J.H. (1960). A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics* 26: 178-194.

Hockett C.F. (1958). *A course in modern linguistics.* New York: Macmillan.

Hyltenstam, K. (1984). The use of typological markedness conditions as predictors in second language acquisition: The case of pronominal copies in relative clauses. In Andersen, R. (ed.) *Second Languages.* Rowley, MA: Newbury. 39-58.

Klein, W. and C. Perdue (1997). The basic variety (or: Couldn't natural languages be much simpler?). *Second Language Research* 13: 301-347.

Kortmann, B./K. Lunkenheimer (eds.) (2011).*The Electronic World Atlas of Varieties of English.* [eWAVE]. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ewave-atlas.org>

Kortmann, B. and K. Lunkenheimer, eds. (2012). *The Mouton World Atlas of Variation in English*. Berlin/New York: Mouton de Gruyter.

Kortmann, B. and B. Szmrecsanyi (2004). Global synopsis: morphological and syntactic variation in English. In Kortmann, B., E. Schneider, K. Burridge, R. Mesthrie and C. Upton (eds.) *A Handbook of Varieties of English.* Berlin/New York: Mouton de Gruyter. 1142-1202.

Kusters, W. (2003). *Linguistic Complexity: The Influence of Social Change on Verbal Inflection.* Utrecht: LOT.

Littlewood, W. (2006). Second Language Learning. In Davies, A. and C. Elder (eds.) *The Handbook of Applied Linguistics.* Malden, MA: Blackwell. 501-524.

McWhorter, J. (2001a). The world's simplest grammars are creole grammars. *Linguistic Typology* 6: 125-166.

McWhorter, J. (2001b). What people ask David Gil and why: Rejoinder to the replies. *Linguistic Typology* 6: 388-413.

McWhorter, J. (2007). *Language Interrupted: Signs of Non-native Acquisition In Standard Language Grammars*. Oxford: Oxford University Press.

McWhorter, J. (2011). *Linguistic Simplicity and Complexity.* Berlin/Boston: De Gruyter.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In: Miestamo, M., K. Sinnemäki and F. Karlsson (eds.) *Language Complexity: Typology, Contact, Change.* Amsterdam/ Philadelphia: Benjamins. 23-41.

Miller, J., R. Weinert. (1998). *Spontaneous Spoken Language. Syntax and Discourse*. Oxford: Clarendon Press.

Seuren, P., H. Wekker (1986). Semantic transparency as a factor in creole genesis. In: Muysken, P. and N. Smith (eds.) *Substrata versus Universals in Creole Genesis.* Amsterdam, Philadelphia: Benjamins. 57-70.

Shosted, R.K. (2006). Correlating complexity: A typological approach. *Journal of Linguistic Typology* 10: 1-40.

Siegel, J. (2004). Morphological simplicity in Pidgins and Creoles. *Journal of Pidgin and Creole Languages* 19: 139-162.

Siegel, J. (2008).*The emergence of pidgin and creole languages*. Oxford: Oxford University Press.

Towell, R., R. Hawkins (1994). *Approaches to second language acquisition.* Clevedon, Philadelphia: Multilingual Matters.

Trudgill, P. (2001). Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5: 371-374.

Trudgill, P. (2004). Linguistic and Social Typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8**:** 305-20.

Trudgill, P. (2009). Vernacular universals and the sociolinguistic typology of English dialects. In Filppula, M., J. Klemola and H. Paulasto (eds.) *Vernacular Universals and Language Contacts: Evidence from Varieties of English and Beyond*. London/New York: Routledge.

Trudgill, P. (2011). *Sociolinguistic Typology.* Oxford: Oxford University Press.

Wekker, H. (1996). *Creole languages and language acquisition.* Berlin, New York: Mouton de Gruyter.