DE GRUYTER

Raffaela Baechler,
Guido Seiler (Eds.)

COMPLEXITY, ISOLATION, AND VARIATION

FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

LINGUAE & LITTERAE

DE
G

Complexity, Isolation, and Variation

linguae & litterae

Publications of the School of Language & Literature
Freiburg Institute for Advanced Studies

Edited by
Peter Auer, Gesa von Essen, Werner Frick

Volume 57

Complexity, Isolation, and Variation

Edited by
Raffaela Baechler and Guido Seiler

DE GRUYTER

ISBN 978-3-11-034738-8
e-ISBN (PDF) 978-3-11-034896-5
e-ISBN (EPUB) 978-3-11-038645-5
ISSN 1869-7054

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: epline, Kirchheim unter Teck
Printing: Hubert & Co. GmbH & Co. KG, Göttingen
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Contents

Raffaela Baechler and Guido Seiler, University of Munich

Introduction

1 About this volume

Recent years have witnessed a rapidly increasing interest in questions concerning the structural complexity of languages (Sampson, Gil and Trudgill 2009; Miestamo, Sinnemäki and Karlsson 2008; Szmrecsanyi and Kortmann 2012): Are all languages of equal complexity? Does higher complexity of one subsystem (e. g. morphology) correspond to lower complexity of another subsystem (e. g. syntax)? How can complexity be measured? And how can complexity differences between languages be explained?

As the field is growing, specialization is growing, too. In this volume, we put a special focus on three aspects. First, we assume that complexity has to do with variation and change. We pay particular attention to different types of varieties, assuming that the study of language variation and change helps us understand the diachronic pathways leading to both complexification and simplification of languages. Second, we aim at testing hypotheses about the possible relationships between structural complexity and sociolinguistic factors, i. e. the linguistic situation of the speakers of the languages under investigation: Do languages spoken by isolated communities tend to higher complexity, and if so, what exactly does isolation mean: lack of dialect contact, lack of language contact, small community size, lack of L2 learners? Third, the volume addresses the question as to how complexity is to be measured in order to make substantial claims about variation and change in complexity. The contributions will show that appropriate assessments of structural complexity depend on both the selection of linguistic phenomena under investigation and general linguistic assumptions about those phenomena.

2 What is complexity?

In 20th century structural linguistics it was generally assumed that languages do not differ significantly in their overall complexity, an assumption known as the equi-complexity hypothesis. Hockett (1958) puts this view in a nutshell, stating that "[...] impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do [...]" (Hockett 1958: 180). Hockett's statement may also be understood against the background of claims made in the 19th century. Humboldt (1836), for example, assumed a correlation between the structural complexity of a language and the level of culture and mental capacity of the speakers of that language: "Die Sprache ist gleichsam die äusserliche Erscheinung des Geistes der Völker; ihre Sprache ist ihr Geist und ihr Geist ihre Sprache [...]" [Language is the outer appearance of the spirit of a nation; their language is their spirit and their spirit is their language] (Humboldt 1836: 53). For a more detailed discussion about structural complexity in the 20th century see Sampson (2009).

Interestingly, although the equi-complexity hypothesis was widespread in the 20th century, this axiom was implicitly rejected by variationist linguists like Ferguson (1959). Ferguson distinguishes between high (H) and low (L) varieties, and does so with regard to structural complexity, too: "One of the most striking differences between H and L in the defining languages is in the grammatical structure: H has grammatical categories not present in L and has an inflectional system of nouns and verbs which is much reduced or totally absent in L" (Ferguson 1959: 333). At the same time, however, variationist linguists did not propose how complexity might be measured.

Testing the equi-complexity hypothesis not only supposes that there are trade-offs between the different subsystems of language, it also implicates that complexity fulfils a specific function in a language. This means that a correlation is assumed between structural complexity and the complexity of the things a language has to do, i. e. more complex things are expressed by a more complex grammar. This claim has been challenged by more fundamental considerations as well as by several case studies (e. g. Kusters 2003; Miestamo 2006; Miestamo, Sinnemäki and Karlsson 2008; Rescher 1998; Sampson, Gil and Trudgill 2009). A good insight into this discussion is given by Gil (2009) in his paper with the pointed title 'How much grammar does it take to sail a boat?'.

The equi-complexity hypothesis implicitly requires measuring the overall/global complexity of a language. In order to do so, it is necessary to identify all relevant subsystems of a language and measure them all. However, there is no consensus on the taxonomy of the linguistic subsystems. Moreover, measuring different subsystems in detail proves to be an enormous task. Miestamo (2008) refers to these challenges as the problem of representativity and the problem of comparability:

> The problem of representativity means that no metric can pay attention to all aspects of grammar that are relevant for measuring global complexity. Even if this were theoretically possible, it would be beyond the capacities of the mortal linguist to exhaustively count all grammatical details of the languages studied [...]. The problem of comparability is about the difficulty of comparing different aspects of grammar in a meaningful way, and especially

> about the impossibility of quantifying their contributions to global complexity. (Miestamo 2008: 30)

Miestamo's statement explains why most studies survey local complexity, i. e. the complexity of one linguistic subsystem, rather than overall complexity.

This leads to the question of what complexity exactly is. In addition to the distinction between overall and local complexity, a distinction is usually made between absolute and relative complexity: "Absolute complexity is a matter of the number of parts and interrelations in a system, whereas relative complexity is a matter of the cost or difficulty of using or processing a certain grammatical construction [...]" (Sinnemäki 2011: 15).

In the literature on structural complexity there are numerous overviews of what is meant by complexity in linguistics, and what types of complexity may be found in languages (e. g. Szmrecsanyi and Kortmann 2012: 10–12). Rescher (1998), by contrast, in his philosophical analysis of complexity, gives a broader definition and classification of types of complexity. He defines complexity as follows:

> Complexity is first and foremost a matter of the number and variety of an item's constituent elements and of the elaborateness of their interrelational structure, be it organizational or operational. Any sort of system or process – anything that is a structured whole consisting of interrelated parts – will be to some extent complex. Accordingly, all manner of things can be more or less complex: natural objects (plants or river systems), physical artifacts (watches or sailboats), mind-engendered processes (languages or instructions), bodies of knowledge, and so on. (Rescher 1998: 1)

Furthermore, Rescher (1998) discerns different modes of complexity, which have been adapted to linguistic phenomena by Miestamo, Sinnemäki and Karlsson (2008: VIII–IX). Sinnemäki (2011) provides a modified version of that adaption, reproduced below in Table 1.

Epistemic and ontological modes may be categorized as absolute complexity, functional complexity as relative complexity. The types of complexity measured in the contributions of this volume are mainly of the following types: descriptive complexity, generative complexity, constitutional complexity, and taxonomic complexity.

As shown in table 1, the term complexity may refer to a range of different things. Accordingly, depending on which type of complexity is studied, the method used to measure complexity varies; or rather, the type of complexity determines the methodology that is chosen. More generally, the type of complexity and the methodology that is applied are both influenced by and depend on the theory of language structure one adopts.

Finally, one may ask whether there are particular languages or varieties that are more likely to display greater structural complexity, i. e. is structural com-

Table 1: Modes of complexity (Rescher 1998: 9; Sinnemäki 2011: 23)

1. **Epistemic modes**
   A. Formulaic complexity
      a. Descriptive complexity: length of the account that must be given to provide an adequate description of a given system.
      b. Generative complexity: length of the set of instructions that must be given to provide a recipe for producing a given system.
      c. Computational complexity: amount of time and effort involved in resolving a problem.

2. **Ontological modes**
   A. Compositional complexity
      a. Constitutional complexity: number of constituent elements (e. g., in terms of the number of phonemes, morphemes, word, or clauses).
      b. Taxonomic complexity (or heterogeneity): variety of constituent elements, that is, the number of different kinds of components (e. g., tense-aspect distinctions, clause types).
   B. Structural complexity
      a. Organizational complexity: variety of ways of arranging components in different modes of interrelationship (e. g., phonotactic restrictions, variety of distinctive word orders).
      b. Hierarchical complexity: elaborateness of subordination relationships in the modes of inclusion and subsumption (e. g., recursion, intermediate levels in lexical-semantic hierarchies).

3. **Functional complexity**
   A. Operational complexity: variety of modes of operation or types of functioning (e. g., cost-related differences concerning the production and comprehension of utterances).
   B. Nomic complexity: elaborateness and intricacy of the laws governing a phenomenon (e. g., anatomical and neurological constraints on speech production; memory restrictions).

plexity influenced by language external factors? And if so, what kind of speech community favors more or less structural complexity? This discussion was mainly initiated by three contributors to this volume: Johanna Nichols, Peter Trudgill, and Kurt Braunmüller, whose publications were followed by several case studies correlating structural complexity with the type of speech community. Only a few potentially influencing factors are listed here: population size (e. g. Hay and Bauer 2007; Sinnemäki 2011), age, sex, class, region (Sampson 2001), high- and low-contact (isolated) speech communities (e. g. Szmrecsanyi and Kortmann 2009; Trugill 2011) and geography (e. g. Nichols 1992).

3 Complexity, variation, and isolation

The present volume is perhaps the first one to bring together different approaches to a potential correlation between structural complexity and language situation, in particular isolation. Several linguists have already hypothesized that structural complexity may be influenced by geographic and/or social isolation of the speech community. One of the first linguists to discuss this probable correlation may be Roman Jakobson. He observed that central, spreading Ukrainian dialects spoken in a heterogeneous speech community had a smaller vowel inventory than dialects spoken at the periphery of the Ukrainian area:

> Cette différence est due, en premier lieu à la tendance conservatrice qui est caractéristique des parlers de la périphérie, et en second lieu à des différences fonctionnelles. Il n'est pas rare d'observer que la tendance à simplifier le système phonologique croît à mesure que grandit le rayon d'emploi d'un dialecte, avec la plus grande hétérogénéité des sujets parlant la langue généralisée. On n'a pas encore, en linguistique, prêté assez attention à la différence essentielle de structure et d'évolution qui existe entre les parlers gravitant vers le rôle de χοινή ou langue commune, et ceux d'usage purement local. (Jakobson 1929: 73)
>
> [This difference is due, first, to a conservative tendency which is characteristic of the varieties at the periphery, and, second, to functional differences. It is not rare to observe that the tendency to simplify the phonological system increases as the area where the dialect is spoken spreads, too, with a heterogeneous speech community speaking the lingua franca. Linguists have not yet devoted enough attention to the main differences in the structure and evolution existing between those varieties tending to become a lingua franca and those being used only locally.]

Dell Hymnes (1975), too, suggests that languages spoken by a community with a tightly knit network tend to display higher complexity than languages spoken in a larger area: "This latter process may have something to do with the fact that the surface structures of languages spoken in small, cheek-by-jowl communities so often are markedly complex, and the surface structures of languages spoken over wide ranges less so" (Hymnes 1975: 50).

Otmar Werner's (1975) assumption points in a similar direction. He supposes that small and isolated communities preserve complex morphological systems longer than do high-contact languages spoken in a wider area.

> Es ist – vermute ich – ganz allgemein ein Kennzeichen kleinerer, isolierter Sprachgemeinschaften, dass sie lange komplizierte Regelsysteme bewahren; grossräumige Sprachkontakte und damit verbundene Interferenzen fordern dagegen die Analogien, wie sie ja auch von Kindern und Ausländern gerne gemacht werden. (Werner 1975: 791)

> [It is – I assume – a general characteristic of smaller, more isolated speech communities that they preserve complex rule systems for a long period; far-reaching language contact and the resulting interferences, by contrast, require analogies, as they are made my children and foreigners.]

To sum up, there are two types of speech communities: On the one hand, there are low-contact speech communities with a tightly knit network and which are isolated and situated at the periphery. The language of this type of speech community tends to show higher structural complexity. Second, there are larger, high-contact and heterogeneous speech communities which are situated more toward the center. This type of community has a potentially spreading language that displays lower structural complexity.

One of the first linguists who explicitly analyzed the possible correlation between structural complexity and isolation is perhaps Kurt Braunmüller, one of the contributors to this volume. In a paper from 1984, he surveys the complexity of inflectional morphology in Icelandic, Faroese and Northern Frisian. He shows that several phonological rules act at the same time producing opacity in the paradigms and thus morphological complexity. According to Braunmüller (1984), the resulting opacity is compensated only to a small extent by paradigmatic analogy, and this is especially true of languages of small and isolated speech communities (Braunmüller 1984: 48). This is due mainly to two characteristic contexts of those speech communities. First, contacts with others happen in another language, so that the L1 of small and isolated communities is very rarely learnt as an L2 and does not serve as a koiné (Braunmüller 1984: 49). This means that the L1 of small and isolated communities is not or only very slightly influenced by other languages and does not show L2 simplification. Second, small and isolated speech communities are also very homogeneous, as they are not subject to any strong standardizing rules (Braunmüller 1984: 49). By isolated languages, Braunmüller (1984) mainly means low-contact languages. However, the languages of his sample (Icelandic, Faroese and Northern Frisian) are at the same time also geographically isolated, i. e. on islands (all three languages) or, concerning Northern Frisian, at the periphery of the mainland in Schleswig-Holstein (Germany).

In 1992, Johanna Nichols published the book „Linguistic Diversity in Space and Time“. For the first time, at least to our knowledge, a possible correlation between geographical isolation and linguistic complexity was surveyed explicitly, and, also for the first time, tested statistically. Nichols' main goal is to correlate typological features and geography: “In describing the distribution of types and typological features we can often make active use of geography as a predictive factor. This means viewing the languages of a region as a population and demonstrating a correlation between the location or type of the region and the distribution of traits within the population or between populations" (Nichols 1992: 12).

Nichols (1992) distinguishes, among other things, residual and spread zones, whereby the former could be considered as isolated and the latter as non-isolated. Spread zones are characterized as follows:

> (1) Little genetic diversity [...].
> (2) Low structural diversity.
> (3) The language families present in the spread zone are shallow.
> (4) Rapid spread of languages or language families and consequent language succession.
> (5) Classic dialectal-geographical area with innovating center and conservative periphery [...].
> (6) No net long-term increase in diversity. A spread zone is a long-lasting phenomenon, but it preserves little linguistic evidence of its history.
> (7) The spreading language serves as a lingua franca for the entire area or a large part of it. (Nichols 1992: 16–17)

In contrast to spread zones, residual zones show the following typical features:

> (1) High genetic density [...].
> (2) High structural diversity [...].
> (3) The language families [...] are deep.
> (4) No appreciable spread of languages or families. No language succession.
> (5) No clear center of innovation [...]
> (6) Accretion of languages and long-term net increase in diversity. [...].
> (7) No lingua franca [...] for the entire area; local bilingualism or multilingualism is the main means of inter-ethnic communication. (Nichols 1992: 21)

The most important finding of Nichols' survey concerning the possible correlation between isolation and structural complexity is that "[r]esidual zones show relatively high complexity, equal to or greater than that of their respective continents. Spread zones show somewhat lower average complexity, equal to or lower than that of their respective continents" (Nichols 1992: 192). However, Nichols also hypothesizes that there are innovating and conservative areas within spread and residual zones: In a spread zone we can find "[an] innovating center and [a] conservative periphery" (Nichols 1992: 17) and in mountain areas (as a residual zone) we find "innovations at the periphery (lowlands), archaisms in the interior (highlands)" (Nichols 1992: 14). That spread and residual zones indeed have innovating and conservative areas is also shown in Nichols' contribution to this volume.

Peter Trudgill, too, correlates isolated speech communities with higher structural complexity. However, as compared to Braunmüller (1984) and Nichols (1992), he adopts a more detailed definition of what an isolated speech community may be. Just as Nichols' (1992) and Braunmüller's (1984) definitions, Trudgill's definition includes small community size, geographical and mainly social isolation (Trudgill 1992), which means low-contact and few L2 learners. Additionally, however, Trudgill (2011) pays particular attention to the internal characteristics of the speech community: a dense social network, high social stability and a large amount of shared information (Trudgill 2011: 146). So, on one hand, there are small, geographically isolated, low-contact, stable, tightly knit speech communities, which we would here call isolated communities, and on the other hand large, geographically non-isolated, high-contact, fluid, loosely knit speech communities, here called non-isolated communities. Trudgill (2011) assumes that the language of isolated speech communities shows higher structural complexity because in these communities we are "most likely to find not only the preservation of complexity but also an increase in complexity, i. e. irregularity, opacity, syntagmatic redundancy, and non-borrowed morphological categories" (Trudgill 2011: 64). According to Trudgill, both preservation of complexity and increase in complexity can be explained by characteristics of language change influenced by the structure of the speech community. Preservation of complexity is mainly due to three factors. First, for small, low-contact communities with a tightly knit social network it is easier to "enforce and reinforce the learning and use of irregularities" (Trudgill 1992: 204) and thus, more generally, "to enforce and reinforce the learning and use of complexities by children and adolescents" (Trudgill 1996: 13). Second, the rate of change is different depending on the type of speech community: "In small, isolated, stable communities, linguistic change will be slower" (Trudgill 2011: 103). Third, small and isolated communities are in general "more resistant to change as such" (Trudgill 1996: 11), and more specifically "[g]eographically peripheral varieties which have been least subject to dialect contact most strongly resist […] language change leading to simplification" (Trudgill 1996: 6). However, when linguistic change does occur, small, isolated and tightly knit communities are more able "[…] to push through, enforce and sustain changes of a less natural or usual phonological type […]" (Trudgill 1996: 11) and to "[…] promote the spontaneous growth of morphological categories […]" (Trudgill 2009: 109). More generally, in these speech communities "[…] there is a greater chance that [linguistic change] will be of the complexification type […]" (Trudgill 2011: 103). Trudgill (2011) calls this type of complexification "spontaneous, non-additive complexification" (Trudgill 2011: 71). So, it was shown that varieties spoken by small, geographically isolated, low-contact, stable, tightly knit speech communities tend not only to preserve structural complexity, but also to increase complexity. In contrast, high-contact varieties tend to simplify their structure "[…] because high irregularity, low transparency, and high levels of redundancy make for difficulties of learning and remembering for adolescent and adult learner-speakers" (Trudgill 2009: 101). However, another type of contact can be observed, namely "long-term co-territorial contact situations involving child bilingualism" (Trudgill 2011: 34). In this type of contact situation, complexification can occur in the form of additive borrowing, "in which new features derived from neighbouring languages do not replace existing features but are acquired in addition to them" (Trudgill 2011: 27). To sum up, according to Trudgill, we may expect higher structural complexity in isolated varieties and languages due to their tendency to preserve and increase structural complexity.

This volume is a collection of papers which aim at testing these claims: In what sense do isolated languages display higher structural complexity? Is higher complexity only due to the preservation of archaic features or are there also instances of complexification? What is the role of language contact? Are there general diachronic tendencies, and can those tendencies be explained with reference to the linguistic situation of the respective variety? Furthermore, it has been shown above that there are different types of complexity and that these types determine the methodology for measuring structural complexity to a certain degree. This is reflected also in the contributions of this volume.

4 The contributions of this volume

In the following, the contributions of this volume will be summarized according to the following questions: What type of complexity is surveyed, and how is this type of complexity measured? What languages and varieties and what parts of the linguistic system are covered? What external factors are used to explain the differences in complexity?

Raffaela Baechler measures the inflectional complexity of nouns, adjectives and articles in the following German varieties: Old High German, present-day Standard German, and the non-standard Alemannic varieties of Kaiserstuhl, Visperterminen and Issime, whereby the latter two are isolated. The inflectional complexity of the three parts-of-speech is measured by the number of markers. A marker is defined as the distinct pairing of exponent and grammatical feature. Additionally, the degree of complexity of nouns includes the number of inflectional classes. The results show that a) isolated varieties are more complex than non-isolated ones, b) there is a diachronic tendency towards simplification, c) no predictions can be made about the correlation between codification and complexity as well as between contact and complexity.

Kurt Braunmüller argues that complexity emerges whenever a grammatical category is encoded by more than one form or construction and that complexity correlates with isolation and language contact. This is shown by means of several instances in the morphology, syntax and phonology of different Germanic languages. Small and isolated communities tend to preserve inherited linguistic structures and develop higher structural complexity. This is due to functional diglossia preventing simplification, and to the fact that these languages are generally not learned as L2 and thus display no L2-simplification. In addition, they exhibit additive borrowings and more free variation as a result of functional diglossia facilitating code-switching.

Ehret and Szmrecsanyi explore the so-called Kolmogorov complexity, an information-theoretic type which can be approximated by applying modern file compression programmes. The text database contains parallel, semi-parallel and non-parallel texts from ten historical and present-day varieties of English as well as from an array of European languages (Germanic: Dutch, English, German; Romance: French, Italian, Romanian, Spanish; Finno-Ugric: Finnish, Hungarian; Esperanto). Linguistic complexity is measured by obtaining numerical estimates of the relative informativeness of text samples via file compression on the overall, syntactic, and morphological level. It can be shown that (a) the Kolmogorov measure in parallel texts in the observed varieties of English and six European languages yields linguistically meaningful results on both overall and morphosyntactic levels. Concentrating on the historical English dataset, the authors find a statistically significant correlation between morphological and syntactic complexity in all their datasets so that they can trace the development of English from a morphologically complex and syntactically simple to a morphologically simple and syntactically complex language over time. (b) In terms of the morphological and syntactic complexity hierarchies obtained from semi-parallel and non-parallel corpus data, they achieve high to very high correlations between parallel, semi-parallel and non-parallel corpora, whereas overall complexity correlates only on a moderate level.

Jacopo Garzonio measures the syntactic complexity of *yes/no* and *wh* questions in three isolated Northern Italian dialects as well as in the non-isolated standard model of Northern Italian dialects. The syntactic analysis is based on Minimalist and Cartographic approaches. In order to measure the degree of complexity, Garzonio uses two parameters: a) the derivational weight of questions, i. e. the number of Move operations compared to the number of Merge operations, b) the number of synonymous constructions which are freely interchangeable. The results show that isolated dialects have a similar derivational weight, but more freely interchangeable constructions than the non-isolated standard model. Furthermore, *yes/no* and *wh* questions can display different degrees of complexity within the same variety.

JOHANNA NICHOLS compares the complexity of languages at the periphery (isolated) and in the center (many L2-learners and language shift). The language sample contains the Daghestanian branch of the Nakh–Daghestanian language family and several languages of different genetic affiliations spoken in the eastern Eurasian steppe (Turkic, Mongolic, Tungusic, Tibetan, Korean, Japanese, Ainu, Nivkh, Chukchi–Kamchatkan, Eskimo–Aleut, Yukagir, Yeniseian, eastern Uralic). The degree of complexity is measured by taking into account several variables of phonology, morphology, syntax and lexicon (e. g. the inventory size of phonemes, the number of verb inflectional categories, the amount of opacity, etc.). It is shown that small inventory size correlates with center languages functioning as interethnic languages and that opacity correlates with peripheral languages which are generally not learned as L2.

DANIEL SCHREIER critically assesses the claims that high contact equals simplification and low contact equals complexification. Drawing from findings from a number of English varieties around the world, he suggests that such claims have to be revised and modified in order to account for the multifaceted and interwoven phenomena of contact-induced change. Analyzing the phonotactic categories in the languages of the world, Schreier shows that in fact, it is not contact intensity that matters the most, but the systemic and typological differences between the varieties in contact. Moreover, as Schreier demonstrates on the basis of the variation within systems of personal pronouns, low contact may also lead to simplification and high contact to complexification. Another problem for models of high vs. low contact is posed by varieties such as Tristan da Cunha English which display simplification and complexification processes simultaneously. On the basis of these observations, Schreier claims that models have to be diachronically informed and suggests abandoning binary classifications in terms of “high” vs. “low contact” or “simplification” vs. “complexification” and instead adopting a scalar-based approach, viewing contact-induced change as a continuum ranging from minimal contact effects to maximal ones.

PETER TRUDGILL outlines the sociolinguistic conditions that play a role in changes of complexity of a language. The emphasis lies on processes of complexification: Following his argumentation, it is complexification rather than simplification that should be viewed as normal. In a critical review of the Isolation Hypothesis, the emergence of complexity is illustrated using examples from southwestern English western Swedish and North Frisian dialects. They each show patterns of reanalysis, with phonologically based variation being interpreted as grammatical categories. Crucially, these changes require a stable linguistic surrounding providing a number of generations of uninterrupted native-speaker development. On the other side of the spectrum, communities with high contact and loose social networks tend to be unable to provide stability long enough for this kind of change to become likely. In addition to that, the emergent traits of complexity are just the kind of information that is prone to simplification in acquisition processes. Isolation is thus merely a fitting image for a complex interplay of stability, low contact, small size, and tight social networks.

References

Braumüller, Kurt 1984: Morphologische Undurchsichtigkeit – ein Charakteristikum kleiner Sprachen. *Kopenhagener Beiträge zur Germanistischen Linguistik* 22: 48–68.

Gil, David 2009: How much grammar does it take to sail a boat? In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 19–33. Oxford: Oxford University Press.

Ferguson, Charles 1959: Diglossia. *Word* 15: 325–340.

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

Hay, Jennifer and Bauer, Laurie 2007: Phoneme inventory size and population size. *Language* 83 (2): 388–400.

Hockett, Charles Francis 1958: *A Course in Modern Linguistics*. New York: Macmillan.

Humboldt, Wilhelm von 1836: *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts*. Berlin: Dümmler (facsimile 1960).

Hymnes, Dell 1975: Speech and Language: On the origins and foundations of inequality among speakers. In: Morton Bloomfield and Einar Haugen (eds.), *Language as a Human Problem*, 45–71. Guildford/London: Lutterworth Press.

Jakobson, Roman 1929: *Remarques sur l’évolution phonologique du russe comparée à celle des autres langues slaves*. Prague: Travaux du Cercle Linguistique de Prague.

Miestamo, Matti 2008: Grammatical complexity in a cross-linguistic perspective. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*, 23–41. Amsterdam: Benjamins.

Miestamo, Matti, Sinnemäki and Karlsson, Fred (eds.) 2008: *Language Complexity: Typology, Contact, Change*. Amsterdam: Benjamins.

Miestamo, Matti 2006: On the feasibility of complexity metrics. In: Krista Kerge and Maria-Maren Sepper (eds.), *Finest Linguistics*: *Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, 11–26. Tallinn: Tallinn University Press.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Rescher, Nicholas 1998: *Complexity*: *A Philosophical Overview*. New Brunswick/London: Transaction Publishers.

Sampson, Geoffrey 2009: A linguistic axiom challenged. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 1–18. Oxford: Oxford University Press.

Sampson, Geoffrey, Gil, David and Trudgill, Peter (eds.) 2009: *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Sampson, Geoffrey 2001: *Empirical Linguistics*. London/New York: Continuum.

Sinnemäki, Kaius 2011: *Language Universals and Linguistic Complexity*: *Three Case Studies in Core Argument Marking*. Helsinki: University of Helsinki.

Szmrecsanyi, Benedikt and Kortmann, Bernd 2009: Between simplification and complexification: Non-standard varieties of English around the world. In Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 64–79. Oxford: Oxford University Press.

Szmrecsanyi, Benedikt and Kortmann, Bernd 2012: Introduction: Linguistic complexity: Second language acquisition, indigenization, contact. In: Bernd Kortmann and Benedikt Szmrecsanyi (eds.), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, 6–34. Berlin/Boston: de Gruyter.

Trudgill, Peter 1992: Dialect typology and social structure. In: Ernst Håkon Jahr (ed.), *Language Contact: Theoretical and Empirical Studies*, 195–211. Berlin/New York: Mouton de Gruyter.

Trudgill, Peter 1996: Dialect Typology: Isolation, Social Network and Phonological Structure. In: Gregory R. Guy, Crawford Feagin, Deborah Schiffrin and John Baugh (eds.), *Towards a Social Science of Language: Papers in Honor of William Labov, Volume 1: Variation and Change in Language and Society*, 3–22. Amsterdam/Philadelphia: Benjamins.

Trudgill, Peter 2009: Sociolinguistic typology and complexification. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 98–109. Oxford: Oxford University Press.

Trudgill, Peter 2011: *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Werner, Otmar 1975: Flexion und Morphophonemik im Färöischen. In: Karl-Hampus Dahlstedt (ed.), *The Nordic Languages and Modern Linguistics 2: Proceedings of the Second International Conference of Nordic and General Linguistics. University of Umeå, June 14–19, 1973*, 774–791. Stockholm: Almqvist & Wiksell International.

Raffaela Baechler, University of Munich

Inflectional complexity of nouns, adjectives and articles in closely related (non-)isolated varieties

Abstract: The aim of this paper is to introduce a tool to measure morphological complexity in closely related varieties of German. Subsequently, I will discuss whether and to what extent complexity differences between varieties can be explained by factors such as language contact, isolation, codification, and diachronic tendencies. The selected varieties are: Old High German, Standard German (codified), and the Alemannic non-standard varieties of Kaiserstuhl, Visperterminen and Issime. While the latter two are both topographically isolated, only the Issime dialect is subject to intensive language contact.

The results can be summarized as follows: a) there is a diachronic tendency toward simplification, but I also found two instances of complexification; b) isolated varieties are more complex than non-isolated ones; c) no predictions can be made about the relationship between codification and complexity; d) the differences between Visperterminen Alemannic (without contact) and Issime Alemannic (with contact) cannot be explained by language contact.

1 Introduction

In recent typological work, structural complexity of languages has become a center of interest (e. g. Miestamo, Sinnemäki, and Karlsson 2008; Sampson, Gil, and Trudgill 2009). However, from the 20th century until a couple of years ago it was assumed that all languages were equally complex. In particular, if one grammatical area displayed complexity, it would be compensated in another area by simplicity: "[...] impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same" (Hockett 1958: 180). So, whereas structural linguists assumed that there was no difference in overall complexity, another line of research, variationist linguists, assumed that complexity difference did exist. Concerning the differences between High and Low varieties, Ferguson (1959), for instance, claims that "[o]ne of the most striking differences between H[igh] and L[ow] [varieties] [...] is in the grammatical structure: H has grammatical categories not present in L and has an inflectional system of nouns and verbs which is much reduced or totally absent in L" (Ferguson 1959: 241). Such statements about simplified structural characteristics of non-standard varieties as opposed to standard ones are abundant in the dialectological literature and to my knowledge are often intuitive, i. e. no method is proposed to measure the alleged complexity differences.

A third line of research, sociolinguistic typology, tries to detect complexity differences between languages/varieties and to explain these differences on the basis of the structure of the community where the language/variety is spoken. In this context there are indications that languages spoken by small and isolated communities tend to display greater structural complexity (Nichols 1992; Braunmüller 1984, 2003; Trudgill 2004, 2009). Baechler and Seiler (2012) call this the Isolation Hypothesis. If this hypothesis is correct, it predicts something not only about large-scale typological comparison but also about sets of genetically closely related and similar languages or varieties (for details see section 1 in Baechler and Seiler 2012). I will base my analysis of complexity on these sociolinguistic concepts.

The aim of this paper is to introduce a tool to measure morphological complexity, to apply this tool, and to discuss first results. The paper is structured as follows: After presenting a number of research questions (section 2), the language sample (section 3) and previous approaches (section 4), section 5 introduces a tool to measure the complexity of noun, adjective, and article inflection in closely related varieties. Section 6 analyzes the inflectional complexity of nouns, adjectives, articles, and the total inflectional complexity of these parts of speech. Section 7 summarizes the main results and gives an outlook.

2 Research questions

The research questions discussed in this paper can be summarized as follows: Can we explain complexity differences between varieties through language-external factors? And if so, which of these language-external factors contribute to structural complexity? The following questions are based on Baechler and Seiler (2012).

Question 1: Is there an overall diachronic tendency?

There is a certain consensus that languages tend to gradually simplify their grammars over time, especially their morphologies. This may be influenced by "linguist's familiarity with the older Indo-European languages and their intricate inflectional systems" (Baechler and Seiler 2012: 23). However, it has been shown that the sociolinguistic context of isolated languages facilitates increased structural complexity (Trudgill 2011: 73–89). Therefore, I expect a tendency toward simplification in the non-isolated varieties and diachronic complexification only in isolated ones.

Question 2: What are the effects of isolation?
I expect isolated varieties to show a higher degree of complexity than non-isolated ones. It seems that isolated varieties do not only tend to preserve existing complexity because of a slower rate of change (Trudgill 2011: 2–8), but, as mentioned in question 1, they also increase linguistic complexity (Trudgill 2011: 73–89).

Question 3: What are the effects of contact?
High-contact situations can lead to complexification or simplification. Complexification is expected in pre-threshold bilingualism (additive borrowing), i. e. in "long-term co-territorial contact situations involving child bilingualism" (Trudgill 2011: 34). Simplification is expected in post-threshold bilingualism, i. e. in short-term contact situations, where adults learn a second language (Trudgill 2011: 34). As will be shown in Section 3, we are concerned here with pre-threshold bilingualism and therefore I expect complexification.

Question 4: Are there instances of complexification?
If there are instances of complexification, I expect to find them only in isolated dialects. Trudgill (2011) hypothesizes that "it is in low-contact [=isolated] communities that we are most likely to find not only the preservation of complexity but also an increase in complexity, i. e. irregularity, opacity, syntagmatic redundancy, and non-borrowed morphological categories" (Trudgill 2011: 64).

Question 5: What is the role of codification?
Ferguson (1959) claimed that High varieties show greater structural complexity than spoken vernaculars (Ferguson 1959: 241). In addition, there may be conserving effects of codification. Therefore, I assume that codified languages show higher structural complexity than spoken dialects (Baechler and Seiler 2012).

3 Language sample

In order to answer these research questions I selected five varieties of German: Old High German (OHG), the oldest attested variety of German, codified Standard New High German (NHG), and three non-standard varieties. These include the Alemannic dialect of the Kaiserstuhl, an area near Freiburg in south-west Germany, the Alemannic dialect of Visperterminen in the Canton of Valais in Switzerland, and the Alemannic dialect of Issime, a linguistic island in the Aosta

Valley in Italy. The data are based on the following grammatical descriptions: Braune and Reiffenstein (2004) for OHG, Eisenberg (2006) for NHG, Noth (1993) for Kaiserstuhl Alemannic, Wipf (1911) for Visperterminen Alemannic[1], Perinetto (1981) and Zürrer (1999) for Issime Alemannic. I consider OHG, NHG, and Kaiserstuhl Alemannic to be non-isolated, while the two Walser dialects of Visperterminen and Issime are isolated.

OHG is not a single variety, but a label for the German varieties spoken between the sixth and eleventh centuries. The first written documents are from the eighth century. No grammatical description exists of a German variety written in a specific area and at a particular time during the Middle Ages. The only grammar providing an exhaustive description of the complete nominal inflection is Braune's normalized grammar. Even though it is a normalized grammar, it exhibits different variants if there are any between centuries and varieties (Alemannic, Bavarian, Franconian).

There are a few qualitative criteria categorizing the Walser dialects as isolated. Visperterminen is situated in the canton of Valais in Switzerland and has 1373 inhabitants. It is 1378 meters above sea level and located at the dead end of the only road access from Visp. In addition, at the time when Wipf investigated the dialect, Visperterminen was not a tourist destination and people living in Visperterminen only very rarely left their village (Wipf 1911: 1–2).

Issime is one of several Alemannic colonies in northern Italy. In the 13th century people migrated from the canton of Valais to the Aosta Valley in Italy. Since then there has not been any contact with the German speaking language area. Many of the 400 inhabitants in Issime are quinquelingual from early childhood: Alemannic, Franco-Provençal, Piedmontese, Standard French (which is the official language of the région autonome Vallée d'Aoste), and Standard Italian (the official language of Italy) (Zürrer 1999: 96–98). However, they do not speak Standard German (Zürrer 1999: 99). As it is a language community with child multilingualism, we may expect to find additive borrowings.

Although this sample is small, it contains some interesting contrasts: historical (OHG) vs. recent, codified (NHG) vs. vernacular, isolated (Visperterminen, Issime) vs. non-isolated, contact (Issime) vs. monolingual environment.

[1] Unfortunately there are no more recent comprehensive grammars for the dialects in the Valais. This must be kept in mind for the following analysis.

4 Previous approaches to measuring linguistic complexity

The most relevant proposals for measuring complexity come from linguistic typology (McWorther 2001; Shosted 2006; Nichols, Barnes, and Peterson 2006). These adopt quantitative methods, but their tools are too coarse for the micro-comparison of closely related languages and varieties, as they were developed for large-scale comparisons.[2] On the other hand, previous methods of measuring complexity differences between closely-related varieties are too language-specific. For example, Szmrecsanyi and Kortmann (2009) select English-specific features which cannot be used for languages other than English. For these reasons, I tried to develop a tool adapted to the sample under analysis. It should be suitable for the analysis of microvariation and applicable to other inflecting languages. For a more comprehensive discussion of these previous approaches, see section 2 in Baechler and Seiler (2012).

5 Measuring inflectional complexity

5.1 Absolute complexity

In the relevant literature a distinction is made between relative and absolute complexity. Relative complexity is defined as “how difficult a phenomenon is to process (encode/decode) or learn” (Miestamo 2008: 25), i. e. whether a linguistic phenomenon is complex to a speaker, a hearer, an L1 acquirer, or an L2 learner.

Linguists concerned with absolute complexity consider only the language system itself. According to Miestamo (2008) “the [absolute] complexity of a linguistic phenomenon may be measured in terms of the length of the description of that phenomenon […]. A less complex phenomenon can be compressed to a shorter description without losing information” (Miestamo 2008: 24). We can adapt this to the language system and assume that the longer the description of the language system is (i. e. the less it can be compressed), the more complex the language system will be.

Another important point is that I only consider inflectional complexity here, more precisely the inflectional complexity of nouns, adjectives, and articles,

[2] For example, Nichols (2009) suggests counting among other things “[…] the number of elements [a subsystem] […] contains. For instance, the number of consonant phonemes, tones, genders, cases, tenses […]” (Nichols 2009: 111). This would not reveal complexity differences between the varieties under analysis here.

which of course does not mean that phonological or syntactic complexity should be excluded. Rather, they must be included if one wants to calculate the overall complexity of the entire language system. The morphological complexity of pronouns and verbs will be measured at a later stage of my PhD-project. In the following I will introduce my method for measuring inflectional complexity of nouns (5.2) and adjectives/articles (5.3) on the basis of NHG.

5.2 Nouns

I will first give a brief description of the architecture of the inflectional paradigms. The full paradigms are reported in the appendix (tables 14–18). I will then present the procedure for measuring the inflectional complexity of nouns.

OHG distinguishes five cases in the singular (nom, acc, dat, gen, instr) and four cases in the plural (nom, acc, dat, gen). NHG, Issime Alemannic, and Visperterminen Alemannic have four cases (nom, acc, dat, gen) in the singular and in the plural. Kaiserstuhl Alemannic distinguishes morphologically only number, but not case. Case is expressed morphologically in the pronouns, adjectives, and articles only.

For measuring the inflectional complexity of nouns I propose the following method in four steps:

- Step 1: Collect the distinguishable inflectional paradigms of the respective language/variety.
- Step 2: Break each paradigm down into a list of inflectional markers.
- Step 3: Put the markers on a list and remove repeated occurrences of markers. Count the remaining markers.
- Step 4: Multiply the number of markers by the number of marker combinations (= inflectional classes).

Thus complexity of noun inflection is defined as the number of inflectional classes multiplied by the number of markers. Multiplication may result in disproportionately high degrees of complexity. This will be discussed in section 6.4.

Step 1: Each grammatical description reports the paradigms in a different way, even if we are concerned with the same variety: For instance the NHG Duden-Grammatik (1998: 223–224) distinguishes ten declension types, but Eisenberg (2006: 152–154) distinguishes only four types with two subtypes. As I aim to compare the paradigms of different varieties, however, comparable paradigms are needed, i.e. paradigms which are arranged in the same way. Furthermore, to get the shortest description of the noun inflection, each paradigm must be maximally compressed. This allows us to compare the shortest description of variety A with the shortest description of variety B.

Step 2: A marker is defined as a distinct pairing of exponent and grammatical feature. The paradigm of *Tag* ‘day’, for example, consists of three markers (for the full paradigm see table 2):

$$m_1\text{: -es}\quad \begin{pmatrix}\text{NUM} & \text{SG} \\ \text{CASE} & \text{GEN}\end{pmatrix}$$

$$m_2\text{: -e}\quad (\text{NUM} \quad \text{PL})$$

$$m_3\text{: -n}\quad \begin{pmatrix}\text{NUM} & \text{PL} \\ \text{CASE} & \text{DAT}\end{pmatrix}$$

For convenience, the markers are represented without attributes as follows: -*es*:sg.gen, -*e*:pl, -*n*:pl.dat. In cases of multiple exponence, each exponent is counted as a marker. Thus, *Hand–Hände* ‘hand’ is made up of umlaut and the suffix -*e*. Umlaut demonstrates that not only segmentable morphs but also stem alternations can be a marker (unless it is phonologically conditioned). This is noted as a rewriting rule: V → [+front, -low] / [NUM PL]. Again, for convenience, the marker is represented as UL:pl.

Concerning syncretism, I distinguish between “good” syncretism, which does not add to complexity, and “bad” syncretism, which adds to complexity. Table 1 displays the paradigm of *Student* ‘student’, which has homophonous markers -*en*. The plural has the following four markers: -*en*:nom.pl, -*en*:acc.pl, -*en*:dat.pl, -*en*:gen.pl. However, this -*en* can be attributed to one consistent function, namely plural. Therefore the plural has only one and not four markers, i. e. -*en*:pl. This type of syncretism does not add to complexity, thus it is called “good” syncretism. The singular of the paradigm of *Student* shows the following three markers: -*en*:acc.sg, -*en*:dat.sg, -*en*:gen.sg. Since it is impossible to assign one consistent function to this -*en* (the nominative singular is not marked)[3], each of these three suffixes must be counted as a separate marker. Consequently, the paradigm has

Table 1: Paradigm of *Student* (NHG)

	SG	**PL**
nom	Student	Studenten
acc	Studenten	Studenten
dat	Studenten	Studenten
gen	Studenten	Studenten

[3] A consistent function may be non-nominative. However, for the sake of methodological consistency grammatical functions are defined here only positively.

three markers in the singular. This syncretism adds to complexity and I therefore call it “bad” syncretism.

In step 3 the markers are put on a list and repeated occurrences of markers are removed. This is a very important step because varieties of German notoriously re-use the same markers across different paradigms. For instance, if the dative plural is marked in NHG, the marker *-n* is suffixed across the great majority of the inflectional classes (cf. table 15).

Step 4: Inflectional complexity is calculated by multiplying the number of markers by the number of inflectional classes. The inflectional class can be defined as a specific combination of markers. Therefore, both a larger marker inventory and large numbers of inflectional classes add to complexity, but these two factors are independent of each other. Multiplying the number of markers by the number of inflectional classes reflects the intuition that each marker combination is to be counted as one way of making use of the same marker inventory. For instance, if there are five inflectional classes, the morphology uses the marker inventory five times to create a paradigm.

My method adopts underspecification (for German cf. Eisenberg 2006; Thieroff and Vogel 2009) and the Elsewhere Condition (Kiparsky 1973; Anderson 1992). Traditionally, a German nominal paradigm is represented by eight instructions, whereby each instruction contains a full specification of feature content and associated exponent. The paradigm of *Tag* ‘day’ (table 2) contains the following eight instructions: nom.sg → *Tag*, acc.sg → *Tag*, dat.sg → *Tag*, gen.sg → *Tages*, nom.pl → *Tage*, acc.pl → *Tage*, dat.pl → *Tagen*, gen.pl → *Tage*. Assuming underspecification, the paradigm of *Tag* contains only three instructions (cf. table 3): Add *-es* in the genitive singular, *-e* in the plural and *-n* in the dative plural. All the other cells of the paradigm will be filled with the underspecified form *Tag*. However, how does the case-underspecified form *Tag* know that it may not be used as genitive singular? How does the grammar know that it may not generate **wegen des Tag*, but *wegen des Tages* (the preposition *wegen* ‘because’ governs genitive)? Here the Elsewhere Condition comes into play, which says: If there is a more specific instruction you must not follow a less specific one. As *Tages* (*-es*:gen.sg) is more specific for genitive singular than *Tag* (default), *Tages* will be used first, blocking the insertion of *Tag* for the genitive singular.

Table 2: Paradigm of *Tag* (NHG)

	SG	**PL**
nom	Tag	Tag-e
acc	Tag	Tag-e
dat	Tag	Tag-en
gen	Tag-es	Tag-e

Table 3: Paradigm of *Tag* assuming underspecification

$$\begin{tikzcd}
\text{Tag} \arrow[r] \arrow[d] & \text{-e + PL} \arrow[d] \\
\text{-es + GEN} & \text{-n + DAT}
\end{tikzcd}$$

Given this approach, the following factors add to inflectional complexity:

- number of inflectionally distinguished grammatical features, e. g. the number of cases
- allomorphy created by a number of inflectional classes, e. g. the plural allomorphs (*-e*, *-n*, *-er*, etc.) in NHG
- multiple exponence, e. g. in *Wald–Wälder* 'forest' the plural is expressed by the umlaut and the suffix *-er*
- "bad" syncretism, e. g. the homophonous singular markers as in the paradigm of *Student* (table 1)

The following factors do not add to complexity:

- re-use of markers across inflectional classes, e. g. the suffix *-n* (dative plural) in NHG
- absence of otherwise attested distinctions in particular inflectional or lexical classes, e. g. Kaiserstuhl Alemannic nouns do not distinguish cases, but determiners and pronouns do
- Allomorphy which is predictable on phonological grounds. For instance, in NHG the genitive singular marker is *-es* or *-s* (*Tag-es* 'day', *Brunnen-s* 'fountain'). However, this variation is ignored because *-es* or *-s* is used depending on the phonological structure of the word – specifically its final sound, its represented stress and its number of syllables (Eisenberg et al. 1998: 224–225).

5.3 Adjectives and articles

The inflectional complexity of nouns is measured by multiplying the number of markers by the number of inflectional classes. As there are no inflectional class distinctions within adjectives and articles, the inflectional complexity of articles and adjectives is defined as the number of markers. First I will give a brief description of the paradigms found in the five varieties considered. The complete paradigms are reported in the appendix (tables 19–23 for the adjectives, 24–27 for the articles). In a next step, I will discuss issues concerning the identification of marker inventories which are specific to adjectives and articles.

The present-day varieties exhibit a definite and an indefinite article. The grammaticalization of the demonstrative pronoun into the definite article and of the numeral “one” into the indefinite article begins in the OHG period, but there is no grammaticalized article as such in OHG (Schrodt 2004: 24–26).

Adjectives have a strong and a weak paradigm in each of our five varieties. The two paradigms are in complementary distribution. In OHG weak adjectives express definiteness, strong adjectives express indefiniteness. The weak adjectives can be preceded by a definite determiner (demonstrative and possessive pronouns), the strong adjectives by an indefinite determiner (indefinite and interrogative pronouns) or no determiner (Braune and Reiffenstein 2004: 218; Szczepaniak 2011: 68–69, 105–109).

In the present-day varieties weak adjectives are used if they are preceded by an article or a pronoun and strong adjectives are used if they are not preceded by any article or pronoun. In all the present-day varieties there are some exceptions where in certain cases and genders the strong inflection is used even if the adjective is preceded by an article or a pronoun. For example in NHG: If the adjective is preceded by an article or a pronoun in the masculine nominative singular and in the neuter nominative and accusative singular, the forms of the strong inflection are used (Thieroff and Vogel 2009: 54–55).

I defined the inflectional complexity of nouns as the number of inflectional classes multiplied by the number of markers. As stated above, the inflectional complexity of adjectives and articles is based on the number of markers only: There are no inflectional classes for adjectives and articles, but two different paradigms for the adjectives (strong/weak) and two for the articles (definite/indefinite). To identify the marker inventory of the adjectives and articles, the same procedure is used as for the noun markers. I will not go through every step, but I will illustrate the procedure with the genitive forms.

Table 4 displays the paradigm of the strong inflection of adjectives in NHG; table 5 displays the paradigm of the weak adjectives. Considering the genitive of the two paradigms, we can ascertain that the genitive is mostly formed by adding the ending *-en*. Only the feminine singular and the plural of the strong paradigm form the genitive with *-er*. We can postulate three rules (from the most to the least specific rule): add *-er* in the genitive feminine singular strong, add *-er* in the genitive plural strong, add *-en* in the genitive. The most specific rule (gen.f.sg.st) is applied first, followed by the rule with three features (gen.sg.st) and finally by the least specific rule associated with only one feature (gen). As mentioned above, a more specific rule blocks the application of a less specific one. Thus *-en* can never be used for the genitive feminine singular strong or the genitive plural strong because there are more specific rules for these morphosyntactic properties (*-er*: gen.f.sg.st; *-er*: gen.pl.st). This has the advantage that only three rules are required to describe the genitive. Adopting a traditional description of a paradigm with a full specification of feature contents and associated exponents, we would end up with eight markers for the genitive (m.sg.st, f.sg.st, n.sg.st, pl.st, m.sg.w, f.sg.w, n.sg.w, pl.w). For deriving the rules all the markers of the NHG adjective are listed in table 6. There are markers with one (least specific) to maximally four features (most specific). The complexity of articles is measured using the same procedure.

Table 4: Strong adjectives in NHG

	nom	acc	dat	gen
m.sg	-er	-en	-em	-en
n.sg	-es	-es	-em	-en
f.sg	-e	-e	-er	-er
pl	-e	-e	-en	-er

Table 5: Weak adjectives in NHG

	nom	acc	dat	gen
m.sg	-e	-en	-en	-en
n.sg	-e	-e	-en	-en
f.sg	-e	-e	-en	-en
pl	-en	-en	-en	-en

Table 6: Markers of the NHG strong and weak adjectives

4 features (most specific)	3 features	2 features	1 feature (least specific)
-er:nom.m.sg.st	-em:dat.sg.st	-en:pl.w	-en:gen
-es:nom.n.sg.st	-er:gen.pl.st		-e:nom
-en:akk.m.sg.st			-en:dat
-es:akk.n.sg.st			-e:akk
-er:dat.f.sg.st			
-en:akk.m.sg.w			
-er:gen.f.sg.st			

6 Results and analysis

In this section I will first discuss the complexity of nouns, articles and adjectives separately and then the total complexity with regard to all three parts of speech. The following factors will be considered in each section in the same order: diachronic tendency, codification, isolated (Visperterminen, Issime) vs. non-isolated (Kaiserstuhl), contact (Issime) vs. non-contact (Visperterminen).

6.1 Complexity of noun inflection

First I will discuss the complexity of noun inflection (number of markers multiplied by the number of inflectional classes, figure 1). Then I will compare the number of markers to the number of inflectional classes (figure 2). The number of markers and inflectional classes for each variety as well as the complexity of noun inflection are listed in table 7.

Table 7: Markers – inflectional classes – complexity

varieties	markers	inflectional classes	complexity (markers × inflectional classes)
OHG	40	18	720
Issime	26	19	494
Visperterminen	24	18	432
NHG	11	14	154
Kaiserstuhl	7	7	49

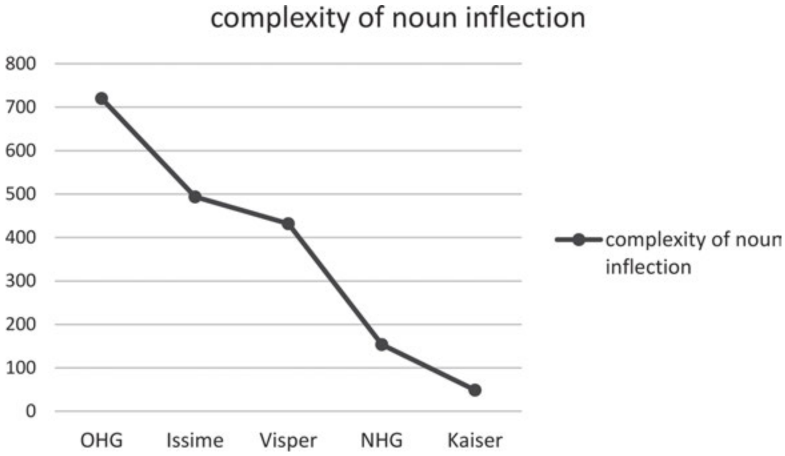

Figure 1: Complexity of noun inflection (number of markers x number of inflectional classes)

First of all, we can see in figure 1 that closely related varieties are not equally complex. Three groups can be identified: The most complex one includes OHG, a second group includes Issime and Visperterminen Alemannic, and a third group includes NHG and Kaiserstuhl Alemannic. Due to the method used (number of markers multiplied by the number of inflectional classes), the complexity differences between the varieties seem disproportionately high. However, if the number of markers and the number of inflectional classes are considered separately, the complexity differences are still clear (cf. figure 2). I will discuss this in more detail later in this section.

Comparing OHG, the oldest attested German variety, to the present-day varieties, we can observe a diachronic simplification process, since OHG is more complex than the present-day varieties.

To answer the question of whether codification leads to complexification or simplification, we compare NHG (codified) with the non-standard varieties. There is a steep decrease in complexity between the Walser dialects (Issime, Visperterminen) and NHG. The noun inflection in Issime and Visperterminen Alemannic is much more complex than in NHG. In contrast, Kaiserstuhl Alemannic is less complex than NHG. However, compared to the Walser dialects the decrease in complexity between NHG and Kaiserstuhl Alemannic is moderate. As NHG is neither more nor less complex than all the non-standard varieties (but lies between these varieties) we can conclude that codification alone does not predict anything about complexity.[4]

We will now turn our attention to the non-standard varieties and especially to the Walser dialects. Figure 1 displays a steep decrease in complexity between the Walser dialects (isolated) and Kaiserstuhl Alemannic (non-isolated). This can be explained by the Isolation Hypothesis: Isolated varieties are more complex than non-isolated ones.[5]

Between Issime and Visperterminen we can observe a moderate decrease in complexity. This is perhaps due to the double isolation of Issime or the language contact situation (which in this case is with Italian and French). First, Issime is not only topographically isolated but also linguistically (it is not part of the West-Germanic dialect continuum). Therefore, if it is correct that complexity correlates with isolation and if we consider Issime as being doubly isolated, Issime's complexity (as compared to Visperterminen) is expected. A second possible explanation is that language contact has a complexifying effect, but (as mentioned in section 2) only in "long-term co-territorial contact situations involving child bilingualism" (Trudgill 2011: 34), which is the case in Issime (Zürrer 1999: 96–99, see section 3). This particular type of complexification, however, is what Trudgill (2011) calls "additive complexification" (Trudgill 2011: 27–33): Morphological categories are borrowed from the contact language/s. However, the noun inflection of Issime Alemannic does not show any additive borrowing from French or Italian (and the

[4] One may also test the possible interactions between codification, contact and population size as suggested by an anonymous reviewer. However, the sample under investigation here is too small for such a statistical analysis. Furthermore, compared to NHG all vernaculars are spoken by language communities with a small population size and the multilingual speakers of Issime Alemannic do not speak NHG.

[5] Nichols (this volume) has very similar findings form the Caucasus.

respective dialects spoken in the Aosta Valley). Hence, the higher degree of complexity in Issime Alemannic is presumably due to the absence of contact with the West-Germanic dialect continuum and supports the Isolation Hypothesis.

Figure 2 shows the number of inflectional classes and the number of markers. In comparison to the overall complexity of noun inflection, the number of markers shows the same order: The variety with the greatest number of markers (40) is OHG, the second group includes Issime (26 markers) and Visperterminen Alemannic (24 markers), and the third group includes NHG (11 markers) and Kaiserstuhl Alemannic (7 markers).

The number of inflectional classes displays a different pattern. It is relatively stable throughout the first three varieties: Issime Alemannic has 19 inflectional classes; OHG and Visperterminen Alemannic have 18. By contrast, we can observe a clear decrease in NHG (14 inflectional classes) and Kaiserstuhl Alemannic (7 inflectional classes). Owing to space constraints it is not possible to show here which inflectional classes are diachronically lost, maintained, or which ones emerged in which variety. Issime Alemannic represents a very interesting case. Concerning the complexity of noun inflection (markers × inflectional classes) and the number of markers, we observed that all the present-day varieties are less complex than OHG, which corresponds to an expected diachronic simplification. However, Issime Alemannic has one inflectional class more than OHG, which we interpret as a case of diachronic complexification. In section 2 it was suggested that instances of complexification could only occur in isolated dialects, so this result supports the Isolation Hypothesis.

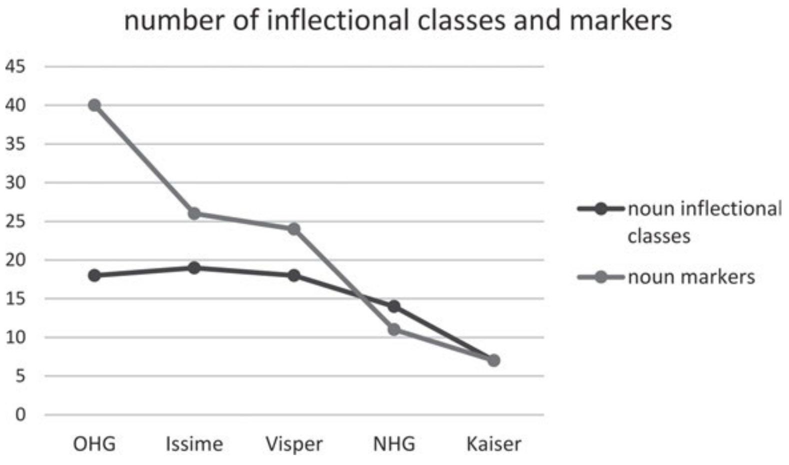

Figure 2: Number of inflectional classes and markers

6.2 Complexity of adjective inflection

The complexity of adjective inflection is defined as the number of rules necessary to generate the strong and the weak paradigms. In table 8 and figure 3 we can see that OHG is more complex than the present-day varieties. Thus we observe diachronic simplification as in the complexity of nouns and the number of noun markers.

Table 8: Complexity of adjectives

	adjective
OHG	38
Visperterminen	17
NHG	14
Issime	12
Kaiserstuhl	9

The comparison of the dialects to NHG shows a pattern for adjective inflection which is reminiscent of the pattern observed for noun inflection: Visperterminen Alemannic is more complex than NHG, Kaiserstuhl Alemannic is less complex. Interestingly, adjectives in NHG – unlike the nouns – display a higher degree of complexity than adjectives in Issime Alemannic. This shows that language-external factors alone cannot explain the complexity differences between varieties. There may be system-internal explanations which should also be considered in future research. Again it seems that codification alone does not predict anything about complexity: NHG is not clearly more or less complex than the non-standard varieties, but is situated between the non-standard varieties.

Like noun inflection, the inflection of adjectives in Issime and Visperterminen Alemannic is more complex than in Kaiserstuhl Alemannic. Again, isolation seems to facilitate complexity: the non-isolated dialect is less complex than the isolated ones.

However, adjective inflection in Visperterminen Alemannic is more complex than in Issime Alemannic – the opposite of what we saw in noun inflection. This is perhaps due to language contact: The four contact languages have less complex adjective morphology (Cerlogne 1971; Brero and Bertodatti 1988; Dardano and Trifone 1997; Riegel, Pellat, and Rioul 2009). It is surprising, however, that contact does not have a similarly simplifying effect on noun inflection. Thus, the low complexity of the adjectives in Issime Alemannic is unexpected given the high complexity of the nouns: Only OHG is more complex than Issime Alemannic with regard to the complexity of nouns and the number of noun markers; additionally

Issime Alemannic shows the highest number of inflectional classes. Therefore, language contact alone cannot explain the variation in the complexity of adjective and noun inflection between Visperterminen and Issime Alemannic. Thus, in future, language-internal mechanisms should also be taken into account. In the following section we will see that articles follow the same pattern as adjectives.

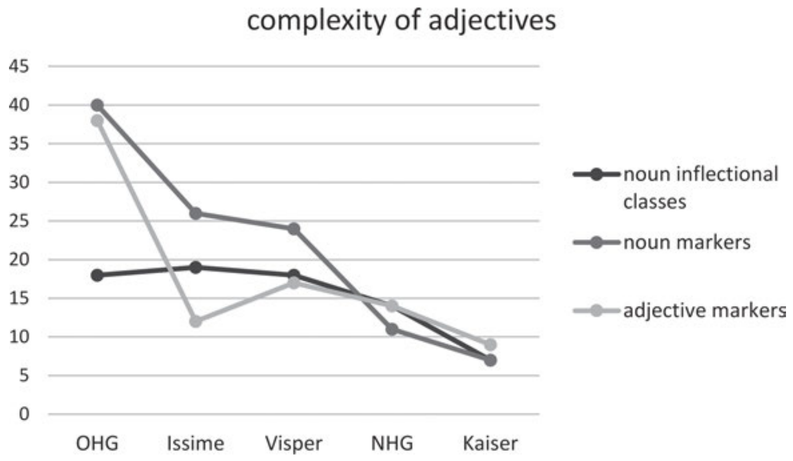

Figure 3: Complexity of adjectives

6.3 Complexity of article inflection

The complexity of article inflection is defined as the number of rules necessary to generate the definite and the indefinite paradigm. Although the differences in complexity are minimal (see table 9), the results are presented in the following.

Table 9: Complexity of articles

	articles
Visperterminen	13
Issime	12
NHG	12
Kaiserstuhl	10
OHG	0

As mentioned in section 5.3, OHG does not have a grammaticalized article and can be regarded as the least complex variety with regard to this part of speech (see figure 4). Thus, this would be a case of diachronic complexification albeit a different kind than the complexification of inflectional classes in Issime Alemannic. Inflectional classes are already present in OHG and expanded in Issime Alemannic. In contrast, there was no grammaticalized article in OHG (Schrodt 2004: 24–26). Expressing (in-)definiteness by means of an article is an innovation which caused a new organization of the language system. One of the consequences of this development is that in many present-day varieties case is encoded by the article rather than the noun. This can be seen in figure 4: The system of adjective and noun markers is much more complex in OHG than in the present-day varieties, which in turn have a grammaticalized article.

The complexity of articles in NHG is situated in between Visperterminen and Kaiserstuhl Alemannic – as in the case of adjectives – and is as high as in Issime Alemannic. Here again, it is difficult to draw a conclusion about the role of codification.

The isolated varieties (Issime and Visperterminen) show a higher degree of complexity than the non-isolated variety (Kaiserstuhl). Since we observed the same pattern in nouns and adjectives, we find further evidence supporting the Isolation Hypothesis.

Articles in Issime Alemannic are slightly less complex than in Visperterminen Alemannic. Adjectives showed the same pattern, nouns the opposite one. I have already mentioned that these observations can hardly be explained by language-external factors alone. The only language-external factor considered here and varying between Issime and Visperterminen Alemannic is contact: As Issime has a long history of multilingualism, we may expect additive borrowings leading to higher complexity. However, the inflectional categories of Issime Alemannic surveyed here, show no additive borrowing. The fact that the inflectional complexity of articles and adjectives is more complex in Visperterminen Alemannic than in Issime Alemannic, but the complexity of nouns is less complex in Visperterminen Alemannic than in Issime Alemannic, suggests that language-internal mechanisms (independent from contact) may also be responsible for the variation in structural complexity.

Moreover, in order to get a better understanding, further non-standard varieties must be considered in future research. First, other non-Walser Highest Alemannic dialects (Issime and Visperterminen make part of the Walser Highest Alemannic group) should be analyzed – for example, dialects in the cantons of Fribourg (Stucki 1917; Henzen 1927) or Uri (Clauss 1929). Second, to test whether contact influences complexity, two other dialects should be included which are not Walser dialects but closely related to each other, for example Kaiserstuhl Alemannic and Alsatian Alemannic. Both dialects are Low Alemannic dialects, i. e. they show clear structural similarities, but only Alsatian Alemannic is subject to language contact (French): All speakers of an Alsatian Alemannic dialect are, at least today, bilingual.

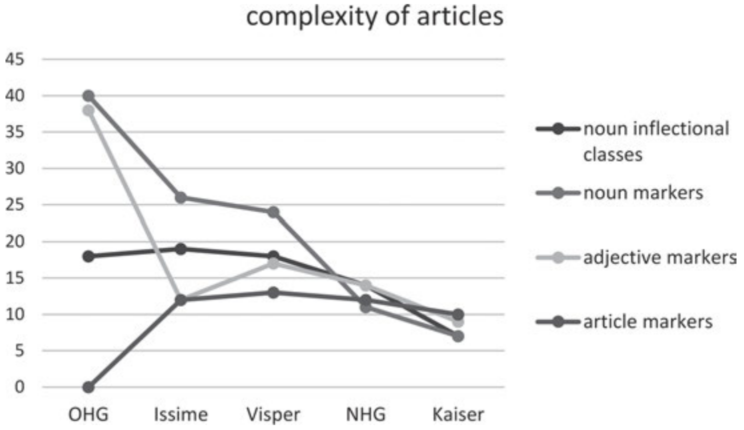

Figure 4: Complexity of articles

6.4 Comparison of inflectional complexity of nouns, adjectives, and articles

In this section, I will first discuss the problems caused by the measurement of noun complexity (the number of markers times the number of inflectional classes) and propose an alternative measurement. Second, I will compare the complexity of the three parts of speech analyzed so far. Third, the total complexity of the three parts of speech under analysis here will be calculated and discussed.

For the complexity of the nouns I took into account the number of markers and the number of inflectional classes because I assume that the more markers and inflectional classes a language has, the more complex it is. However, it is desirable to obtain only one figure for the representation of noun complexity. To achieve this we may either multiply the markers by the inflectional classes or add them together. Originally it seemed more reasonable to multiply because the system combines the marker inventory once, twice, etc., thus giving rise to one, two, etc. inflectional classes. However, the problem is that multiplication causes very high figures for noun complexity and the comparison of the three parts of speech becomes difficult (compare figure 5 with figure 6). We encounter another problem with multiplication when we calculate total complexity. Total complexity of the three parts of speech is defined as the sum of complexity of the three parts of speech: total complexity = complexity of adjectives + complexity of articles + complexity of nouns. As has already been mentioned, figures become very high if noun markers are multiplied by inflectional classes. Consequently, the figures of the adjectives and articles have very little weight in the total complexity: Nouns influence the total complexity much more than adjectives and articles do (compare tables 10 and 11). So, to calculate the complexity of nouns, it may be more suitable to add the number of markers and the number of inflectional classes together.

Comparing the complexity of the three parts of speech (figures 5 and 6), a complexity hierarchy can be identified: noun>adjective>article. Nouns are more complex than adjectives and adjectives are more complex than articles. If noun complexity is calculated as the product of markers and inflectional classes, there are two exceptions to the complexity hierarchy (figure 5/table 10): In Issime Alemannic adjectives and articles are equally complex and in Kaiserstuhl Alemannic adjectives are less complex than articles. If noun complexity is calculated as their sum (figure 6/table 11), we observe the same two exceptions. In addition, in Kaiserstuhl Alemannic nouns and articles show the same degree of complexity. The fact that the complexity hierarchy does not change depending on the method (nouns measured as either the sum or the product of markers and inflectional classes) supports the validity of the method.

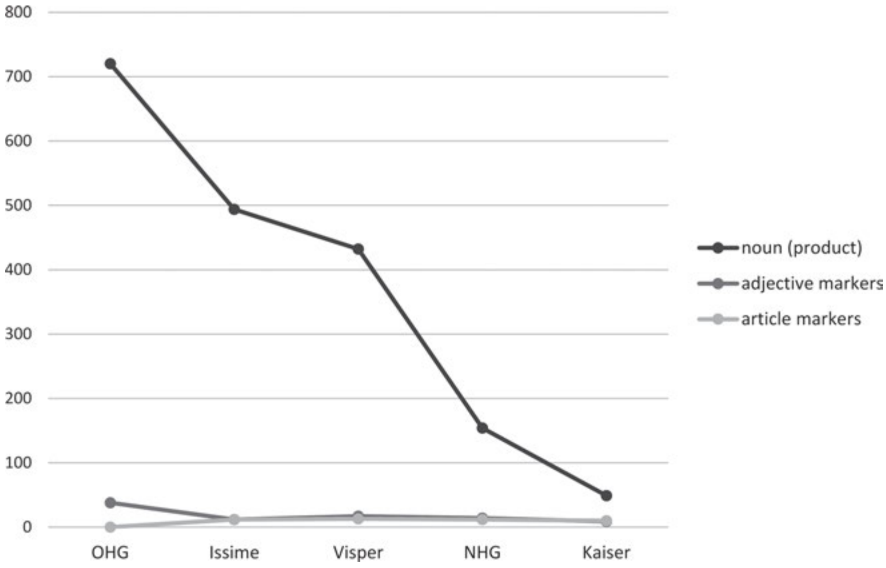

Figure 5: Comparison of parts of speech (noun: markers × inflectional classes)

Table 10: Parts of speech (noun: markers × inflectional classes)

	noun (markers × inflectional classes)	adjective	article
OHG	720	38	0
Issime	494	12	12
Visperterminen	432	17	13
NHG	154	14	12
Kaiserstuhl	49	9	10

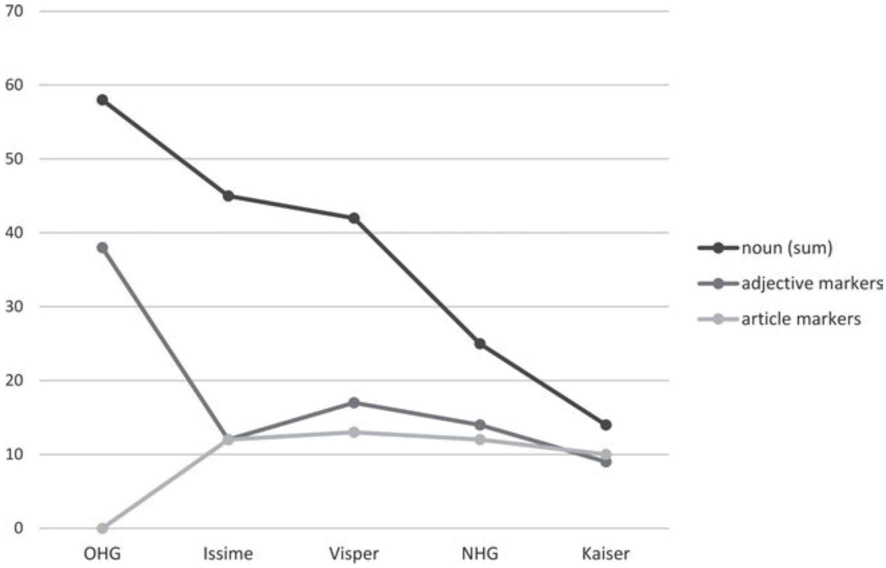

Figure 6: Comparison of parts of speech (noun: markers + inflectional classes)

Table 11: Parts of speech (noun: markers + inflectional classes)

	noun (markers + inflectional classes)	adjective	article
OHG	58	38	0
Issime	45	12	12
Visperterminen	42	17	13
NHG	25	14	12
Kaiserstuhl	10	9	10

Above I defined the total complexity of the three parts of speech under analysis as the sum of the complexity of the three parts of speech: total complexity = complexity of adjectives + complexity of articles + complexity of nouns. Again, I report two figures for total complexity: In figure 7, noun complexity is calculated as the product of markers and inflectional classes, in figure 8 it is calculated as their sum.

In figure 8 we can observe diachronic simplification. NHG is situated in between the Walser dialects and Kaiserstuhl Alemannic, thus we cannot conclude anything about the relationship between standardization and complexity. The

Walser dialects show a higher degree of complexity than Kaiserstuhl Alemannic, which supports the Isolation Hypothesis. Visperterminen Alemannic is slightly more complex than Issime Alemannic.

In figure 7 we can make the same observations as in figure 8, except for Visperterminen and Issime Alemannic: Issime Alemannic now turns out to be slightly more complex than Visperterminen Alemannic. So, the method for measuring noun complexity (sum or product of the number of markers and the number of inflectional classes) influences only the total complexity of Visperterminen and Issime Alemannic. This can be explained by the fact that the number of markers and the number of inflectional classes in Visperterminen and Issime Alemannic are very close to each other.

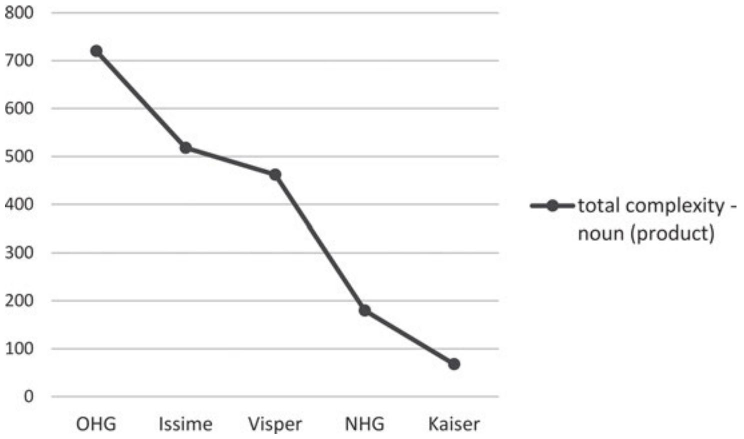

Figure 7: Total complexity (noun: markers × inflectional classes)

Table 12: Total complexity (noun: markers × inflectional classes)

	total complexity – noun (noun: markers × inflectional classes)
OHG	758
Issime	518
Visperterminen	462
NHG	180
Kaiserstuhl	68

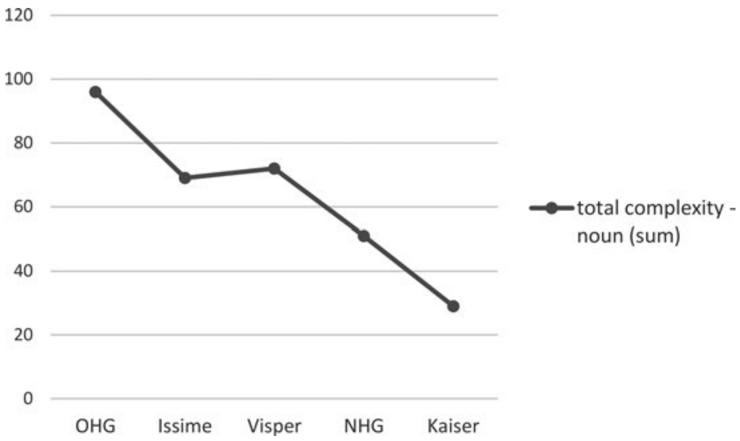

Figure 8: Total complexity (noun: markers + inflectional classes)

Table 13: Total complexity (noun: markers + inflectional classes)

	total complexity – noun (noun: markers + inflectional classes)
OHG	96
Issime	69
Visperterminen	72
NHG	51
Kaiserstuhl	29

7 Outlook

In this paper four main points have been made: First, despite the relatively small sample, substantial evidence was gathered in support of diachronic simplification. Second, isolation correlates with greater complexity. Third, there are two instances of complexification (inflectional classes in Issime Alemannic, development of the articles). Fourth, no effect of codification on complexity could be found.

The differences between the two Walser dialects could not be accounted for in terms of language contact. For a better understanding, larger samples must be considered in future research.

In this paper, only the complexity of nouns, adjectives, and articles were measured. In a next step, demonstrative, possessive, and interrogative pronouns will be included too. In order to get the full picture of morphological complexity, it will be necessary also to include the complexity of verb inflection.

References

Anderson, Stephen Robert 1992: *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Baechler, Raffaela and Seiler, Guido 2012: Simplification, complexification and microvariation: Towards a quantification of inflectional complexity in closely related varieties. In: Angela Ralli, Geert Booij, Sergio Scalise and Athanasios Karasimos (eds.), *Morphology and the Architecture of Grammar. Online-Proceedings of the Eight Mediterranean Morphology Meeting*, 22–40. Patras: University of Patras.

Braune, Wilhelm and Reiffenstein, Ingo 2004: *Althochdeutsche Grammatik I. Laut- und Formenlehre*. Tübingen: Niemeyer.

Braunmüller, Kurt 1984: Morphologische Undurchsichtigkeit – ein Charakteristikum kleiner Sprachen. *Kopenhagener Beiträge zur Germanistischen Linguistik* 22: 48–68.

Braunmüller, Kurt 2003: Language, typology and society: Possible correlations. In: Joel Sherzer and Thomas Stolz (eds.), *Minor Languages. Approaches, Definitions, Controversies*, 89–101. Bochum: Brockmeyer.

Brero, Camillo and Bertodatti, Remo 1988: *Grammatica della lingua piemontese: Parola-vita-letteratura*. Torino: Piemont/Europa.

Cerlogne, Jean-Baptiste 1971: *Dictionnaire du patois valdôtain. Précédé de la Petite Grammaire*. Genève: Slatkine Reprints.

Clauss, Walter 1929: *Die Mundart von Uri: Laut- und Flexionslehre*. Frauenfeld: Huber.

Dammel, Antje and Kürschner, Sebastian 2008: Complexity in nominal plural allomorphy: A contrastive survey of ten germanic languages. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), *Language Complexity. Typology, Contact, Change*, 243–262. Amsterdam: Benjamins.

Dardano, Maurizio and Trifone, Pietro 1997: *La nuova grammatica della lingua italiana*. Bologna: Zanichelli.

Eisenberg, Peter 2006: *Grundriß der deutschen Grammatik: Das Wort*. Weimar: Metzler.

Eisenberg, Peter, Gelhaus, Hermann, Henne, Helmut, Sitta, Horst and Wellmann, Hans (eds.) 1998: *Duden: Grammatik der deutschen Gegenwartssprache*. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Ferguson, Charles 1959: Diglossia. *Word* 15: 325–340.

Finkel, Raphael and Stump, Gregory 2007: Principal parts and morphological typology. *Morphology* 17: 39–75.

Henzen, Walter 1927: *Die deutsche Freiburger Mundart im Sense- und suedoestlichen Seebezirk*. Frauenfeld: Huber.

Hockett, Charles 1958: *A Course in Modern Linguistics*. New York: Macmillan.

Kiparsky, Paul 1973: ‘Elsewhere’ in phonology. In: Stephen Robert Anderson and Paul Kiparsky (eds.), *A Festschrift for Morris Halle*, 93–106. New York: Holt.

McWhorter, John 2001: The world’s simplest grammars are creole grammars. *Linguistic Typology* 5: 125–166.

Miestamo, Matti 2008: Grammatical complexity in a cross-linguistic perspective. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), *Language Complexity. Typology, Contact, Change*, 23–41. Amsterdam: Benjamins.

Miestamo, Matti, Sinnemäki, Kaius and Karlsson, Fred (eds.) 2008: *Language Complexity. Typology, Contact, Change*. Amsterdam: Benjamins.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, Johanna 2009: Linguistic complexity: A comprehensive definition and survey. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 110–125. Oxford: Oxford University Press.

Nichols, Johanna, Barnes, Jonathan and Peterson, David A. 2006: The robust bell curve of morphological complexity. *Linguistic Typology* 10: 98–108.

Noth, Harald 1993: *Alemannisches Dialekthandbuch vom Kaiserstuhl und seiner Umgebung*. Freiburg i.Br.: Schillinger.

Perinetto, Renato 1981: *Eischemer's Büjie*. (typewritten) Issime.

Riegel, Martin, Jean-Christophe Pellat and Rioul, René 2009: *Grammaire méthodique du français*. Paris: Quadrige/PUF.

Sampson, Geoffrey, Gil, David and Trudgill, Peter (eds.) 2009: *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Schrodt, Richard 2004: *Althochdeutsche Grammatik II: Sytnax*. Tübingen: Niemeyer.

Shosted, Ryan Keith 2006: Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.

Stucki, Karl 1917: *Die Mundart von Jaun im Kanton Freiburg: Lautlehre und Flexion*. Frauenfeld: Huber.

Szczepaniak, Renata 2011: *Grammatikalisierung im Deutschen: Eine Einführung*. Tübingen: Narr.

Szmrecsanyi, Benedikt and Kotrmann, Bernd 2009: Between simplification and complexification: Non-standard varieties of English around the world. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 64–79. Oxford: Oxford University Press.

Thieroff, Rolf and Vogel, Petra M. 2009: *Flexion*. Heidelberg: Winter.

Trudgill, Peter 2004: The impact of language contact and social structure on linguistic structure: Focus on the dialects of modern Greek. In: Bernd Kortmann (ed.), *Dialectology Meets Typology. Dialect Grammar from a Cross-Linguistic Perspective*, 435–451. Berlin/New York: Mouton de Gruyter.

Trudgill, Peter 2009: Sociolinguistic Typology and Complexification. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 98–109. Oxford: Oxford University Press.

Trudgill, Peter 2011: *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Wipf, Elisa 1911: *Die Mundart von Visperterminen im Wallis*. Frauenfeld: Huber.

Wurzel, Wolfgang U. 1984: *Flexionsmorphologie und Natürlichkeit: Ein Beitrag zur morphologischen Theoriebildung*. Berlin: Akademie-Verlag.

Zürrer, Peter 1999: *Sprachinseldialekte: Walserdeutsch im Aosta-Tal (Italien)*. Aarau/Frankfurt am Main/Salzburg: Sauerländer.

Abbreviations

OHG:	Old High German
NHG:	New High German
Kaiser:	Kaiserstuhl Alemannic
Visper:	Visperterminen Alemannic
sg:	singular
pl:	plural
m:	masculine
f:	feminine
n:	neuter
nom:	nominative
acc:	accusative
dat:	dative
gen:	genitive
instr:	instrumental
st:	strong adjective inflection
w:	weak adjective inflection

Appendix

Table 14: Noun inflection in OHG (Braune and Reiffenstein 2004)

		SG					PL			
	IC	**nom**	**acc**	**dat**	**gen**	**instr**	**nom**	**acc**	**dat**	**gen**
	1	tag	tag	tag-e	tag-es	tag-o	tag-a	tag-a	tag-on	tag-o
	2	hirt-i	hirt-i	hirt-e	hirt-es	hirt-u	hirt-a	hirt-a	hirt-on	hirt-o
	3	gast	gast	gaste-e	gaste-es	gaste-u	gest-i	gest-i	gest-in	gest-o
	4	win-i	win-i	win-e	win-es		win-i	win-i	win-in	win-o
	5	sit-u	sit-u	sit-e	sit-es	sit-u	sit-o	sit-i	sit-in	sit-i
	6	han-o	han-un	han-in	han-in		han-un	han-un	han-on	han-ono
	7	fater	fater	fater-e	fater-es		fater-a	fater-a	fater-un	fater-o
	8	wort	wort	wort-e	wort-es	wort-o	wort	wort	wort-on	wort-o
	9	lamb	lamb	lamb-e	lamb-es	lamb-o	lemb-ir	lemb-ir	lemb-ir-on	lemb-ir-o
	10	kunn-i	kunn-i	kunn-e	kunn-es	kunn-o	kunn-i	kunn-i	kunn-in	kunn-o
	11	herz-a	herz-a	herz-in	herz-in		herz-un	herz-un	herz-on	herz-ono
	12	geb-a	geb-a	geb-u	geb-a		geb-a	geb-a	geb-on	geb-ono
	13	kuningin	-a	-u	-a		-a	-a	-on	-ono
	14	anst	anst	enst-i	enst-i		enst-i	enst-i	enst-in	enst-o
	15	zung-a	zung-un	zung-un	zung-un		zung-un	zung-un	zung-on	zung-ono
	16	hoh-i	hoh-i	hoh-i	hoh-i		hoh-i	hoh-i	hoh-in	hoh-ino
	17	muoter	muoter	muoter	muoter		muoter	muoter	muoter-un	muoter-o
	18	naht	naht	naht	naht		naht	naht	naht-on	naht-o
wa-stem		hleo	hleo	hlew-e	hlew-es		hlew-a	hlew-a	hlew-on	hlew-o
wa-stem		horo	horo	horaw-e	horaw-es		horo	horo	horaw-on	horaw-o

Table 15: Noun inflection in NHG (Eisenberg 2006)

	SG				PL				
IC	nom	acc	dat	gen	nom	acc	dat	gen	
1	gast	gast	gast	gast-es	gäst-e	gäst-e	gäst-en	gäst-e	
2	tag	tag	tag	tag-es	tag-e	tag-e	tag-en	tag-e	
3	schaden	schaden	schaden	schaden-s	schäden	schäden	schäden	schäden	
4	brunnen	brunnen	brunnen	brunnen-s	brunnen	brunnen	brunnen	brunnen	
5	vater	vater	vater	vater-s	väter	väter	väter-n	väter	
6	lehrer	lehrer	lehrer	lehrer-s	lehrer	lehrer	lehrer-n	lehrer	
7	wald	wald	wald	wald-es	wäld-er	wäld-er	wäld-er-n	wäld-er	bild-er
8	matrose	matrose-n	matrose-n	matrose-n	matrose-n	matrose-n	matrose-n	matrose-n	
9	staat	staat	staat	staat-s	staat-en	staat-en	staat-en	staat-en	konto
10	blume	blume	blume	blume	blume-n	blume-n	blume-n	blume-n	pizza
11	stadt	stadt	stadt	stadt	städt-e	städt-e	städt-e-n	städt-e	
12	mutter	mutter	mutter	mutter	mütter	mütter	mütter-n	mütter	
13	zoo	zoo	zoo	zoo-s	zoo-s	zoo-s	zoo-s	zoo-s	
14	pizza	pizza	pizza	pizza	pizza-s	pizza-s	pizza-s	pizza-s	

Table 16: Noun inflection in Kaiserstuhl Alemannic (Noth 1993)

	SG			PL			
IC	nom	acc	dat	nom	acc	dat	
1	braif	=nom	=nom	briaf	=nom	=nom	
2	gumb	=nom	=nom	gimb	=nom	=nom	
3	schdai	=nom	=nom	schdai-n-er	=nom	=nom	wäld-er
4	grab	=nom	=nom	grab-a	=nom	=nom	
5	ghuch-i	=nom	=nom	ghuch-ana	=nom	=nom	
6	dand-a	=nom	=nom	dand-ana	=nom	=nom	
7	baziand-i	=nom	=nom	baziand-inna	=nom	=nom	

Table 17: Noun inflection in Visperterminen Alemannic (Wipf 1911)

	SG				PL			
IC	nom	acc	dat	gen	nom	acc	dat	gen
1	tag	tag	tag	tag-sch	tag-a	tag-a	tag-u	tag-o
2	chopf	chopf	chopf	chopf-sch	chepf	chepf	chepf-u	chepf-o
3	ar-o	ar-o	ar-u	ar-u	arm-a	arm-a	arm-u	arm-o
4	santim	santim	santim	santim-sch	santim	santim	santim	santim
5	han-o	han-o	han-u	han-u	han-e	han-e	han-u	han-o
6	bog-o	bog-o	bog-u	bog-u	beg-e	beg-e	beg-u	beg-o
7	senn-o	senn-o	senn-u	senn-u	senn-u	senn-u	senn-u	senn-o
8	jar	jar	jar	jar-sch	jar	jar	jar-u	jar-o
9	hor-u	hor-u	hor	hor-sch	hor-u	hor-u	horn-u	hor-o
10	chrut	chrut	chrut	chrut-sch	chrit-er	chrit-er	chrit-er-u	chrit-er-o
11	lamm	lamm	lamm	lamm-sch	lamm-er	lamm-er	lamm-er-u	lamm-er-o
12	redli	redli	redli	redli-sch	redli-n-i	redli-n-i	redli-n-u	redli-n-o
13	öig	öig	öig	öig-sch	öig-u	öig-u	öig-u	öig-o
14	farb	farb	farb	farb	farb-e	farb-e	farb-u	farb-o
15	bon	bon	bon	bon	bon-a	bon-a	bon-u	bon-o
16	sach	sach	sach	sach	sach-u	sach-u	sach-u	sach-o
17	mus	mus	mus	mus	mis	mis	mis-u	mis-o
18	tsun-a	tsun-a	tsun-u	tsun-u	tsun-e	tsun-e	tsun-u	tsun-o

Table 18: Noun inflection in Issime Alemannic (Zürrer 1999)

	SG				PL			
IC	**nom**	**acc**	**dat**	**gen**	**nom**	**acc**	**dat**	**gen**
1	weg	weg	weg	weg-sch	weg-a	weg-a	weg-e	weg-u
2	uav-e	uav-e	uav-e	uav-endsch	uav-n-a	uav-n-a	uav-n-e	uav-n-u
3	noam-e	noam-e	noam-e	noam-endsch	noam-i	noam-i	noam-e	noam-u
4	hoan-u	hoan-u	hoan-e	hoan-ensch	hoan-i	hoan-i	hoan-u	hoan-u
5	vus	vus	vus	vus-sch	vüs	vüs	vüs-e	vüs-u
6	att-u	att-u	att-e	att-e	att-i	att-i	att-e	att-e
7	schu	schu	schu	schu-sch	schu	schu	schun-e	schun-u
8	sia	sia	sia	sia-sch	sia-w-a	sia-w-a	sia-w-e	sia-w-u
9	bet	bet	bet	bet-sch	bet-i	bet-i	bet-u	bet-u
10	berri	berri	berri	berri-sch	berri-n-i	berri-n-i	berri-n-u	berri-n-u
11	lam	lam	lam	lam-sch	lamm-er	lamm-er	lamm-er-e	lamm-er-u
12	lan	lan	lan	lan-sch	lenn-er	lenn-er	lenn-er-e	lenn-er-u
13	matt-u	matt-u	matt-u	matt-u	matt-i	matt-i	matt-u	matt-u
14	mum-a	mum-a	mum-u	mum-u	mum-i	mum-i	mum-u	mum-u
15	chötti	chötti	chötti	chötti	chötti-n-i	chötti-n-i	chötti-n-u	chötti-n-u
16	schuld	schuld	schuld	schuld	schuld-in-i	schuld-in-i	schuld-in-u	schuld-in-u
17	nacht	nacht	nacht	nacht	necht-in-i	necht-in-i	necht-in-u	necht-in-u
18	han	han	han	han	hen	hen	hen-e	hen-u
19	geiss	geiss	geiss	geiss	geiss	geiss	geiss-e	geiss-u

Table 19: Strong and weak inflection of the OHG adjectives (Braune and Reiffenstein 2004)

strong									
	SG					**PL**			
	nom	**acc**	**dat**	**gen**	**instr**	**nom**	**acc**	**dat**	**gen**
m	blint-er/blint	blint-an	blint-emo	blint-es	blint-u	blint-e	blint-e	blint-em	blint-ero
	mar-er/mar-i	mar-an	mar-emo	mar-es	mar-u	mar-e	mar-e	mar-em	mar-ero
	garaw-er/gar-o	garaw-an	garaw-emo	garaw-es	garaw-u	garaw-e	garaw-e	garaw-em	garaw-ero
n	blint-az/blint	blint-az/blint	blint-emo	blint-es	blint-u	blint-u	blint-u	blint-em	blint-ero
	mar-az/mar-i	mar-az/mar-i	mar-emo	mar-es	mar-u	mar-u	mar-u	mar-em	mar-ero
	garaw-az/gar-o	garaw-az/gar-o	garaw-emo	garaw-es	garaw-u	garaw-u	garaw-u	garaw-em	garaw-ero
f	blint-u/blint	blint-a	blint-eru	blint-era		blint-o	blint-o	blint-em	blint-ero
	mar-u/mar-i	mar-a	mar-eru	mar-era		mar-o	mar-o	mar-em	mar-ero
	garaw-u/gar-o	garaw-a	garaw-eru	garaw-era		garaw-o	garaw-o	garaw-em	garaw-ero
weak									
	SG					**PL**			
	nom	**acc**	**dat**	**gen**	**instr**	**nom**	**acc**	**dat**	**gen**
m	blint-o	blint-un	blint-in	blint-in		blint-un	blint-un	blint-om	blint-ono
n	blint-a	blint-a	blint-in	blint-in		blint-un	blint-un	blint-om	blint-ono
f	blint-a	blint-un	blint-un	blint-un		blint-un	blint-un	blint-om	blint-ono

Table 20: Strong and weak inflection of the NHG adjectives (Eisenberg 2006)

strong					**weak**				
	nom	**acc**	**dat**	**gen**		**nom**	**acc**	**dat**	**gen**
m.sg	-er	-en	-em	-en	**m.sg**	-e	-en	-en	-en
n.sg	-es	-es	-em	-en	**n.sg**	-e	-e	-en	-en
f.sg	-e	-e	-er	-er	**f.sg**	-e	-e	-en	-en
pl	-e	-e	-en	-er	**pl**	-en	-en	-en	-en

Table 21: Strong and weak inflection of the Kaiserstuhl Alemannic adjectives (Noth 1993)

strong				**weak**			
	nom	**acc**	**dat**		**nom**	**acc**	**dat**
m.sg	-a	-a	-em	**m.sg**	-ø	-ø	-a
n.sg	-ø/-s	-ø/-s	-em	**n.sg**	-ø	-ø	-a
f.sg	-i	-i	-er	**f.sg**	-ø	-ø	-a
pl	-i	-i	-a	**pl**	-a	-a	-a

Table 22: Strong and weak inflection of the Visperterminen Alemannic adjectives (Wipf 1911)

strong					**weak**				
	nom	**acc**	**dat**	**gen**		**nom**	**acc**	**dat**	**gen**
m.sg	-e	-e	-um	-s	**m.sg**	-o	-o	-u	-u
n.sg	-s	-s	-um	-s	**n.sg**	-a	-a	-u	-u
f.sg	-i	-i	-er	-er	**f.sg**	-a	-a	-u	-u
pl	-i	-i	-e	-er	**pl**	-u	-u	-u	-o

Table 23: Strong and weak inflection of the Issime Alemannic adjectives (Perinetto 1981)

strong					**weak**				
	nom	**acc**	**dat**	**gen**		**nom**	**acc**	**dat**	**gen**
m.sg	-e	-e	-e	-s	**m.sg**	-e	-e	-e	-e
n.sg	-s	-s	-s	-s	**n.sg**	-ø	-ø	-e	-e
f.sg	-ø	-ø	-ø	-er	**f.sg**	-u	-u	-u	-u
m.pl	-ø	-ø	-ø	-er	**pl**	-u	-u	-e	-u
n.pl	-i	-i	-i	-er					
f.pl	-ø	-ø	-ø	-er					

Table 24: Definite and indefinite articles in NHG (Eisenberg 2006)

definite article					**indefinite article**				
	nom	**acc**	**dat**	**gen**		**nom**	**acc**	**dat**	**gen**
m.sg	der	den	dem	des	**m.sg**	ein	einen	einem	eines
n.sg	das	das	dem	des	**n.sg**	ein	ein	einem	eines
f.sg	die	die	der	der	**f.sg**	eine	eine	einer	einer
pl	die	die	den	der					

Table 25: Definite and indefinite articles in Kaiserstuhl Alemannic (Noth 1993)

definite article				**indefinite article**			
	nom	**acc**	**dat**		**nom**	**acc**	**dat**
m.sg	dr	dr	em	**m.sg**	a	a	ma
n.sg	s	s	em	**n.sg**	a	a	ma
f.sg	d	d	dr	**f.sg**	a	a	era
pl	d	d	dr				

Table 26: Definite and indefinite articles in Visperterminen Alemannic (Wipf 1911)

definite article					**indefinite article**				
	nom	**acc**	**dat**	**gen**		**nom**	**acc**	**dat**	**gen**
m.sg	dr	dr	dum	ds	**m.sg**	a	a	amu	as
n.sg	ds	ds	dum	ds	**n.sg**	as	as	amu	as
f.sg	di	di	dr	dr	**f.sg**	a	a	anar	anar
pl	di	di	de	dr					

Table 27: Definite and indefinite articles in Issime Alemannic (Perinetto 1981; Zürrer 1999)

definite article					**indefinite article**				
	nom	**acc**	**dat**	**gen**		**nom**	**acc**	**dat**	**gen**
m.sg	den	den	dem	ds	**m.sg**	a	a	am	as
n.sg	ds	ds	dem	ds	**n.sg**	as	as	am	as
f.sg	di	di	der	der	**f.sg**	a	a	ar	ar
pl	di	di	dan	der					

Kurt Braunmüller, University of Hamburg

On the origins of complexity: evidence from Germanic

Abstract: This paper deals with grammatical complexity, from both a theoretical perspective and with special reference to Germanic languages. The starting point is Rescher's (1998) theoretical approach to complexity. Four *causes* of linguistic complexity are presented in more detail: complexity due to (1) language contact, (2) language cultivation and linguistic conservatism, (3) linguistic variation, and (4) the overgeneralization of linguistic reconstructions.

It is argued that linguistic complexity must be seen as a *dynamic process*, oscillating, for example, between formal complexity and semantic uniformity. When investigating the nature of linguistic complexity it turns out, contrary to Hockett's (1958) claim that all languages are equally complex, that there are indeed simple languages (e. g. pidgins or forms of early child language). Moreover, other variables deserve more attention, such as *selective* grammatical complexity or phenomena of *over-* and *under-specification* (following McWhorter's 2007 approach). The main focus, however, lies on Germanic, its early history, and the divergent structures of the modern Germanic languages. It is shown that newly implemented structures, predominantly caused by grammatical replication, are the main sources of complexity. On top of that, special attention is paid to isolated linguistic communities (morphological complexity in Insular Nordic and North Frisian). Finally, instances of over-specification are discussed in detail on the basis of how 'possession' and 'definiteness' are represented in Germanic languages.

1 Introduction[1]

1.1 One of the much-debated recent issues in language typology and language evolution is whether all languages are equally complex or there are actually simple or at least less complex languages. One of the classic answers to this question denies such a disparity in linguistic complexity: "Objective measurement is difficult, but impressionistically [!] it would seem that the total grammatical

[1] I would like to thank two reviewers for their helpful comments on an earlier version of this paper. All remaining errors and deficiencies are mine, of course.

complexity of any language, counting both morphology and syntax, is about the same as any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically" (Hockett 1958: 180–181). But Hockett's (1958) admittedly "impressionistic" statement ignores for example that utterances made in face-to-face communication draw heavily on the speech situation and other contextual factors. All that is well-known or obvious for the interlocutors therefore need not be verbalized, following the maxim: "Don't say more than is necessary!" Only elementary school teachers often insist that pupils answer in full sentences.

1.2 Rescher (1998) approaches complexity from a more general, philosophical point of view. He distinguishes between (A) *formulaic* complexity, which may be descriptive, generative, or computational but is rather peripheral for linguistic studies, (B) two ontological modes: (1) *compositional* complexity, which may be constitutional or taxonomic, and (2) *structural* complexity, which plays an important role in any linguistic description, especially when organizational and hierarchical issues are the focus of interest, and finally (C) *functional* complexity, which is relevant when analyzing modes of operation and functions and dealing with the elaborateness and intricacy of the laws governing the phenomena at hand (Rescher 1998: 9).

Rescher's (1998) provocative question of whether complexity is a *self-potentiating* phenomenon highlights the far-reaching implications of this category not only for linguistic descriptions but also for the languages themselves. He says: "In nature, certain factors – energy, matter, life, and complexity among them – appear to be self-potentiating: the more of them there is, the more powerful the impetus to the production of yet more." (Rescher 1998: 3) and "[m]ore refined distinction and difficulties never introduce more simplicity than there was before, they can only militate in the direction of greater complexity" (Rescher 1998: 7). Later he concludes, following his principle of *Rational Economy*: "we cannot generally tell in advance of the fact just where and how further complexity will arise" (Rescher 1998: 200) because complexity is limitless in its nature (and not teleologically determined, one might add) – albeit there must be some order in such processes. If Rescher (1998) were completely on the right track, we would have to conclude that any differentiation of linguistic categories would automatically produce more and more complexity, and that this process would be irreversible. Such a general unrestricted supposition is highly questionable, however. There might, for example, be cognitive and other limitations.

Dahl (2004: 289–302) argues in a more differentiated way: Forms of system, structural, and output complexity do not develop in the same way. In his view, so-called *maturation* stages in the development of a language do not *per se* reflect something that may be considered 'unnatural' or 'dispreferred', though some grammatical structures in these languages might be out of balance but remain stable over a long(er) period of time. There will always be a conflict between structure-preserving and structure-changing tendencies on the one hand and the representation of genetic and cultural information on the other.

2 Causes of linguistic complexity

2.1 There are, in our opinion, at least four causes for the emergence of linguistic complexity:

(1) Complexity as a result of *language contact*, seen as (a) *additive* borrowing (see Trudgill 2011: 27): Speakers add and integrate new words and structures into their language while the inherited vocabulary and structures still remain in use. This process is typical of superstratal borrowing. Substratal influence, on the other hand, can be seen as an instance of (b) *receptive* borrowing: L2 speakers add calqued or modified syntactic structures to the system of the superimposed language, which leads to an increase in variation in the imposed language as a whole, where native and non-native structures are then found side by side, i. e. L1 speakers accept at least some of the innovations created by the L2 speakers. In both cases the amount of complexity increases in the new variety.

(2) Complexity due to *language cultivation* and *linguistic conservatism*. Outdated terms and constructions are not generally replaced by new items and are therefore still found in that language, at least in some (archaic) sorts of texts, e. g. biblical texts, folk songs, and fairy tales. But these outdated forms are still part of the receptive linguistic competence of most native speakers. Moreover, when taking Weinreich's (1954) diasystematic approach as the foundation of a comprehensive description of a language, we are faced with different (regional, social, stylistic) varieties. But at the latest when languages split up into partly diverging national varieties, these differences become obvious, as can be seen from these examples taken from Germanic languages: Eng.$_{\text{UK}}$ *dinner jacket* vs. Engl.$_{\text{USA}}$ *tuxedo*, Germ.$_{\text{Germany}}$ *Blumenkohl* vs. Germ.$_{\text{Austria}}$ *Karfiol* 'cauliflower' or Dut.$_{\text{Netherlands}}$ *ham* vs. Dut.$_{\text{Belgium/Flanders}}$ *hesp* 'ham'.

(3) Complexity as a natural and logical consequence of any form of *linguistic variation*. The more a linguistic description tries to account for internal diasystematic variation (based on Weinreich's 1954 approach), the more such a description leads to a higher degree of language-internal complexity. Complexity, described in terms of *inter-individual* variation, increases considerably when a language shows a greater dialectal/sociolectal split, manifold stylistic variation,

or sophisticated pragmatic differentiation – and there is no (gradual) way back to linguistic simplicity again (cf. section 1.2). In contrast, *intra-individual* variation can only reflect a (random) selection of this phenomenon and thus only parts of the total complexity of any linguistic diasystem.

(4) Complexity as a *result of linguistic reconstructions*. As is well known, such reconstructions are hypotheses about (more or less attested) ancient or pre-historical languages. Nonetheless, some grammars tend to treat proto-languages (e. g. Proto-Germanic) as they would describe attested languages. Such descriptions may therefore contain all or at least the most salient structures of *all* daughter languages,[2] though not all of these grammatical categories are found in each of the dialects of the reconstructed language.[3] Therefore grammatical descriptions of that kind mirror linguistic constructs which are far more complex than any of the attested dialects. Moreover, such reconstructions would not be in accordance with Labov's Uniformitarian Principle (cf. Labov 1994: 21–23).

2.2 Generally speaking, it should be borne in mind that *written* languages, especially those with a long written tradition, tend to display more linguistic complexity than only context-dependent *oral* languages such as pidgin languages and, more typically and generally, all early stages in L1-acquisition by children from birth. One-, two- or three-word utterances can only be interpreted as

[2] Cf. Hirt (1931): "Das Urgermanische bestimmen wir also als die Sprachform, die allen germanischen Mundarten zugrunde liegt. [...] Dabei finden wir gewisse Unterschiede, die in dem Verlust alten Erbgutes, in der Neubildung gewisser Formen und sonstigen Veränderungen bestehen" [Proto-Germanic has thus been defined as the linguistic form underlying all Germanic dialects. (...) There are certain differences, resulting from the lost inherited material, the formation of new forms and other changes] Hirt (1931: 15). One might get the impression that Proto-Germanic has apparently been considered a linguistic diasystem where certain differences may occur in some of its varieties ("dialects"). Moreover, Krahe and Meid (1969: 25) speak of a Germanic as "[eine] für das gesamtgermanische Gebiet gültige Sprachform" [(a) linguistic form which is valid for the entire Germanic area].

[3] This is e. g. the case for the category "passive" in early Germanic. Only some few residual forms of this grammatical category are attested in only one dialect: Gothic, whose speakers left the common Germanic territory very early, however: "Hätten wir das gotische Passiv nicht, so würde man behaupten, das Germanische hätte kein Passiv besessen" [If not for the Gothic passive, one could claim that Germanic had no passive] Hirt (1934: 136). But according to the general opinion within runology, there are no passive forms attested in the oldest sources of Germanic, viz. the runic inscriptions in older Futhark (vs. Braunmüller 2004, who argues for the emergence of a new passive {-vowel + k(a)}, replicated from Latin, in the oldest inscriptions; see also section 5.2.2 (c)).

meaningful utterances on the basis of contextual, non-verbal information. These utterances are undoubtedly simple in their structure but heavily dependent on their situational context, complex interactional patterns, and the verbal developing process previously experienced by that child, as interpreted by the parents and others.

Moreover, *context-free* utterances tend to be more complex than *context-dependent* utterances. The reason for this observation is quite simple: The more contextual, situational, or pragmatic factors that are verbalized, the more complex the context-free version of this utterance is on the textual/discourse level.

On top of that, performance[4]and pragmatic factors, too, have an important impact on structural complexity. Left-branching constructions or nesting need more, and especially more complex, syntactic structures than constructions with less structural depth or right-branching constructions. In the latter cases, the linearization on the word order level makes those constructions more accessible for the interpreter than left-branching or nesting constructions. Even the degree of grammaticalization within each linguistic domain may be determined by processing complexity (Hawkins 1994: 14).

3 On defining linguistic complexity

We would like to define linguistic complexity as a *dynamic process* and, in accordance with Rescher's (1998) view, consider complexity a self-potentiating phenomenon. Complexity emerges whenever a grammatical category or structure is represented by more than one category, form, or construction with approximately the same meaning – no matter whether these equivalents are found on the same grammatical or diasystematic level or not.[5]

Two explanatory examples of diverging developments with the Germanic languages follow:

(a) The representation of the category "aspect" in the verbal system of tense-prominent languages gives way to more semantic differentiation and thus to

[4] More on this topic can be retrieved from the profound analyses presented in Hawkins (1994), where the ratios of early immediate constituents to non-terminal (syntactic) constituents of typologically divergent languages are discussed in great detail. He also considers the role of (basic) word orders of meaningful elements in relation to information processing and encoding.

[5] More on the multiple dimensions and diverging aspects of this issue can be retrieved from the anthologies published by Miestamo, Sinnemäki, and Karlsson (2008) and Sampson, Gil, and Trudgill (2009). But in contrast to Dahl (2004), no common denominator takes a shape that could serve as the basis of assessment or a definition, because things still seem to be in a state of flux.

more complexity, which, however, might later become nothing more than morphological variation. English and the Scandinavian languages show greater formal and semantic/aspectual complexity in their tense systems ([±completed action] in the past) than e. g. modern colloquial German, where most speakers consider past tense and present perfect forms to be some sort of free morphological variants, both representing the (unspecified) meaning "past". Though modern German formally seems to be as equally complex as e. g. English or Swedish, it no longer makes any semantic distinction between past tense, formerly also called "imperfect [past] tense", on the one hand and present perfect on the other. In other words, morphological variation (viz. formal complexity) no longer has any counterpart in semantics, a fact that can be regarded as a case of neutralization and a decrease in semantic complexity.

(b) The possibility to rearrange the order of indirect and direct objects can make the use of ditransitive constructions more complex. One can either use the unmarked (or default) order: *Patricia slices Peter a piece of pound cake* [$\text{NP}_{\text{subject}}$ V $\text{NP}_{\text{indirect}}$ – $\text{NP}_{\text{direct}}$] or rephrase it as: *Patricia slices a piece of pound cake for Peter* [$\text{NP}_{\text{subject}}$ V $\text{NP}_{\text{direct}}$ – $\text{PP}_{\text{indirect}}$]. The same phenomenon has become an innovation in modern Faroese due to bilingualism with Danish (see Petersen 2010: 21–26 and passim: *geva* 'to give': ... $\text{NP}_{\text{indirect}} + \text{NP}_{\text{direct}} \rightarrow \text{NP}_{\text{direct}} + \text{PP}$). Both constructions and many other calques from Danish co-occur in colloquial Faroese along with the traditional solely case-marked constructions, which considerably increases the morphosyntactic complexity of modern Faroese (we will come back to this point in section 6.4).

4 On the nature of complexity

4.1 Hockett's (1958) presupposition, as quoted in section 1.1, takes for granted that all languages are equally complex. Taking the ratio between inflectional morphology and syntax/word order into consideration, he assumes that all languages are subject to some sort of (linguistic) Principle of Communicating Tubes: If a language or dialect shows rich morphology, less restrictive word order rules are tolerable. Hockett (1958) refers to English as an instance of the reverse case: a low degree of inflectional morphology but with rather restrictive word order rules (Hockett 1958: 181).

But Hockett's (1958) speculation obviously has no solid empirical foundation: There *are* indeed linguistically simple languages, and all of them are highly context-dependent (see e. g. Gil 2009). Pidgins and (at least young) creole languages, foreigner talk, and early stages of non-native (L2) language learning are clear cases of simplified linguistic systems. This also applies to early (child) language acquisition, where we first observe holophrastic one-word utterances, then two- and three-word utterances and other forms of utterances which are highly context-dependent and/or presuppose much previous (discourse) knowledge by the addressee.

4.2 The next question is whether there are signs of greater *selective complexity*, seen from a cross-linguistic perspective. Selective complexity means that some parts of a grammar show more linguistic differentiations than others – compared to the same domains in other languages. When defining the sources of grammatical complexity, McWhorter (2007: chapter 2) therefore distinguishes between instances and grammatical areas of *over-* and *under-specification*. Both terms must be taken as relational typological concepts, of course. As will be demonstrated in section 7, where we will focus on the concepts of "possession" and "definiteness", some languages show more, some less differentiation when considering certain grammatical categories cross-linguistically. Partial or *selective differentiation*, which according to McWhorter (2007) might be described as instances of *over-specification*, occurs when speakers express their specific communicative needs due to their environmental situation or due to the social structures of the community they live in. Thus languages are not only sign systems but also mirror social structures and other forms of extra-linguistic distinctions, since they emerged out of intense interactional relations between interlocutors in a societal framework. As is well known, some languages may therefore show more lexical differentiation with respect to politeness, gender, or social ranking. (For instances of grammatical differentiation, see section 7.)

4.3 A further question concerns the interplay between linguistic complexity and the number of speakers of a language (in relation to social structures). There may be some degree of mutual dependence, but the chain of cause and effect works in a different way: Small and especially isolated languages are normally *not* subject to foreign (adult) language learning and are therefore not subject to linguistic simplification (viz. incomplete acquisition). Moreover, most speakers of so-called "small" languages must be bi- or even multilingual if they do not want to live in isolation from speakers outside their linguistic community. In contact with foreigners they therefore use another, "bigger" language as a *lingua franca*, the final result of which is an unimpeded development of these "small" and often isolated languages. This gives way to maintaining or even creating linguistic idiosyncrasies, often based on the respective societal structures.

An exception to this case, however, may occur when a large group of immigrants with high social prestige and/or economic or political power enter these societies (e. g. due to migration or usurpation), learn the local variety only to a certain extent and thus restructure (parts of) this language. In such cases, various forms of simplification will evolve.

5 Instances of linguistic complexity within Germanic

In the following, I would like to sketch five instances of grammatical complexity within the Germanic languages (the details are presented in section 5.2.1):

(a) Proto-Germanic as a contact language, which later became expanded (i. e. acquired more complexity) due to intense language contact

(b) the irregularity and complexity of small(er) and isolated languages, such as Icelandic, Faroese or the North Frisian vernaculars

(c) Faroese as a complex contact language with many replicated (or code-copied) linguistic structures from Danish

(d) and (e) a discussion of linguistic complexity with respect to "possession" and "definiteness", respectively. These two grammatical categories will be discussed in detail (in section 7), based on examples from different Germanic languages, since they show an abundance of equivalent linguistic structures and thus (morphosyntactic) complexity.

5.1 How Proto-Germanic acquired complexity – again

Proto-Germanic, as documented in the oldest runic inscriptions, and the oldest testimonies of Ancient Germanic dialects diverge considerably in their typological and grammatical structures from other ancient western Indo-European (IE) languages, such as Latin or Classical Greek. As elaborated in Braunmüller (2008a) in detail, Proto-Germanic shows several features that are typical of *contact languages*, such as the reduction or loss of inherited grammatical categories and word formation elements.[6] Moreover, Proto-Germanic shows the

[6] Hirt (1932): "Das Germanische hat von dem reich entwickelten idg. Verbalsystem in vollem Umfang nur das Präsens und das Präteritum gerettet [!]. In diesem sind aber noch der sogenannte thematische Aorist und vielleicht einige andere Aoristformen aufgegangen. [...] Dadurch entsteht etwas wesentlich neues" [Germanic has to a full extent saved [!] not more than the present tense and the preterit from the richly developed IE verbal system. But into this system merged the

formation of new categories in a way that is characteristic of (modern) contact languages (e. g. periphrastic tense formation or the reinterpretation of inherited categories). This judgement, predominantly based on structural changes, is supported by very strong superstratal lexical influence from other non-IE languages (evidently more than one third, probably much more; for details see Vennemann 2003: 1–20).

5.2 The following observations in favor of "Germanic as a contact language" are all based on grammatical (not lexical) elements:

5.2.1 The loss of grammatical categories

(a) *Tenses and moods.* In sharp contrast to other western IE languages, Germanic has preserved only two tenses of the original IE tense system, namely the present and the preterit (cf. Hirt's hitherto overlooked statement, quoted in footnote 6). There is no future tense (as e. g. in Latin) or subjunctive mood either. Optative forms, however, have been preserved and have taken over functions equivalent to those of the lost subjunctive forms.

(b) *Voices.* No traces of the inherited IE passive voice system are found in the runic inscriptions and early non-translated Germanic texts. Only a few remnants of this grammatical category (mostly occurring in the third person) are attested for Biblical Gothic, which, however, might be due to the influence of the model language, Biblical Greek. It is more common to render this category periphrastically by using *waírþan* (cf. Germ. *werden*) 'to become' or *wisan* (cf. Germ. *sein*) 'to be' (Braune and Ebbinghaus 1973: 98, note 2; 103 note 5), the first of which must clearly be considered a grammatical innovation.

(c) *Tense formation.* We observe a nearly complete loss of the IE reduplicating principle in the verbal system and, more revealing, the complete fossilization of the IE ablaut system in early Germanic verb formation, where the *e/o*- and, to a minor extent, the *a/o*-patterns were generalized more than in any other IE dialect (cf. Mailhammer 2007[7]).

so-called thematic aorist and perhaps some other aorist forms as well. (…) Thereby something fundamentally new emerged] Hirt (1932: 130–131).

[7] Mailhammer (2007: 168–179, 188–210) also points out that almost half of all irregular/ablauting verbs in Germanic have no or only doubtful IE etymologies, a fact which further supports our assumption that Germanic is a contact language.

(d) *Cases*. The loss of the ablative, vocative, and, in almost all ancient dialects the loss of the instrumental, however, cannot be taken as a clear indicator of intense language contact because the loss of these categories may also be due to the more general evolutionary syncretism of cases.

(e) *Person*. The category of the dual has also been lost to a great extent in most ancient varieties of Germanic but this, too, can be explained by normal linguistic change (drift: the disappearance of lesser used grammatical categories).

5.2.2 The evolution of new grammatical structures

Germanic has become widely known for the following two *grammatical innovations*:

(a) The abolishment of the (previously unified) ablauting principle in tense formation led to the creation of a *new preterit* by means of enclitic elements that were derived from the semantically unspecified Proto-IE verb **dhê-* 'to do'. Parallel developments have been described in Bossong (2001) and Muysken (2000: ch. 7: "bilingual complex verbs") for many other languages. These enclitic markers operate like T[ense]-M[ood]-A[spect] particles do in contact languages, which may be roughly illustrated by a periphrastic construction such as *did play* instead of *played* (for details see Braunmüller 2008a: 385–393, 399–401).

(b) Proto-Germanic developed a "new" class of so-called weak adjectives ending in *-an* in order to express *definiteness* (or specificity) in noun phrases. This is not actually an innovation but rather should be regarded as a reanalysis – more precisely, an erroneous interpretation of the IE word formation element *-ant* (by second language learners), which was originally used to express discrimination and individualization, as e.g. in Hittite (cf. Josephson 2007 and especially Braunmüller 2008b).

(c) In our view (Braunmüller 2004: 43–46) there are clear indications that writers of Ancient North-Western Germanic, as documented in the oldest runic inscriptions, tried to create new passive voice forms by directly replicating Latin inflectional endings, based on the frequently occurring Germanic equivalent to Latin *(ego) vocor* '(I) am called', which was rendered and replicated, morpheme by morpheme, as *(ek) haiteka*.[8] The cliticized pronoun *ek(a)* 'I' was later generalized as {-vowel+*k(a)*} as in *rAisidoka* 'was errected' or *fAlAhAk* 'was concealed'.

[8] *Haita* 'to be called' (cf. Germ. *heißen*) is the only medio-passive form found in the oldest Germanic sources. Therefore it had the best (semantic) precondition for this grammatical replication.

(d) The most recently revealed feature in the rebuilding and elaboration process of Ancient Germanic is the replication of Latin *verbal prefixes* as productive patterns of creating new complex verbs (Braunmüller and Höder 2012). Proto-Germanic did not have any productive prefix verb constructions. The *ga-/gi-*prefixes in present perfect participles are semantically bleached and did not represent more than the grammatical category "aspect". The default way to form complex verbs in Ancient Germanic was to modify the core meaning of verbs through post-posed particles or adverbs, as in Modern Swedish *bryta av* 'to tear off' [+concrete]. The younger replicated type is *avbryta* 'to interrupt' [+abstract]. The reason for this grammatical borrowing was certainly the need for appropriate translations of abstract verbs occurring in Latin sources when translating religious or elaborated texts.[9]

5.2.3 Grammatical replication as the main source of complexity in Germanic languages

All Ancient Germanic dialects code-copied/replicated grammatical patterns from the model language Latin, in the case of Gothic also from Biblical Greek, and thus gained considerably more complexity through additive language contact. The main purpose was to achieve *grammatical re-implementation* as a compensation for lost IE categories (viz. the elaborated tense and mood system, including the Consecutio Temporum Principle, the subjunctive forms, and the passive voice in almost all dialects) by means of various techniques, e. g. periphrastic verb formation (the "weak" preterit), grammatical reanalysis (the "weak" adjective inflexion derived from a reinterpreted IE word formation element) or the reinterpretation of lesser used grammatical categories (the original IE optative forms were used as subjunctive forms). *Grammatical duplication* became a new and productive means to create complex verbs, especially those with an abstract meaning, in addition to the pre-existing Gmc. "verb + particle/adverb" constructions. The same function can be attributed to *semantic replication* (making a copy of a Latin inflectional passive form by using a duplicated clitic pronoun), the *re-interpretation* of two existent grammatical patterns, the so-called weak preterit (based on a periphrastic verb construction that is often found in simplified

[9] Cf. Braunmüller and Höder (2012: 163–165) where we have shown that in a puristic language like Icelandic only those new complex verbs were accepted when their prefixes had counterparts in adverbs and other inherited particles, such as *fyrir-* 'for' or *aftur-*'after', but not *bí-/be-*, which was obviously due to Danish or Low German influence and therefore ruled out.

speech), and the definiteness marking in noun phrases containing adjectives as well (a reinterpretation of an IE word formation element).

6 Complexity and isolation

6.1 As is well-known, small and isolated linguistic communities behave differently compared to bigger linguistic societies. These small communities have often been labelled "speech islands" because they are typically found on actual islands or in remote mountain villages. These communities tend to strongly preserve inherited linguistic structures, which may be regarded as a reflex of conservatism in many respects, e. g. an attempt to express their genuine local identity and their personal ties to the region. The speakers of these vernaculars obviously favor divergence from the surrounding varieties in their speech (and in their customs as well) and try to inhibit most outcomes of language contact and convergence with other languages/dialects. Moreover, these languages develop more structural complexity, according to Rescher's (1998) self-potentiating principle, since the speakers strive for more contrasts and more linguistic sophistication (such as the preservation of some otherwise outdated kinship terms or politeness forms).

6.2 The main cause of this kind of divergence is *functional diglossia*: The speakers of small and isolated languages use different languages for internal and external communication. Therefore bi- or trilingualism is quite ubiquitous in these societies. But functional multilingualism or diglossia prevent simplifications of their internal language to a high extent, often accompanied by local immobility, especially in remote rural areas. On top of that, there are no L2-learners of these vernaculars, which has the effect that L2-simplifications (i. e. decline of structural complexity) are unlikely to occur (for more details see Braunmüller 2003). As a direct consequence we observe an unimpeded development of phonological rules, which leads to an increase in complexity, according to Rescher's (1998) principle, reinforced by a strong consciousness of being different from others and the pride of belonging to a special community with a unique language of its own.

6.3 Another important kind of divergence is due to the fact that certain phonological rules operate independently from other grammatical rules and are not balanced by morphological, i. e. paradigmatic, adjustments. To give two examples: Some phonological processes, e. g. contact assimilations, such as the *i*-, *a*- or *u*-umlaut in Germanic, but also other contextually conditioned phonological rules, e. g. the so-called Scandinavian breaking rule (Gmc. *e* > *ja* /__ *a*, or *e* > *jǫ* > *jö* /__ *u*), operate "blindly" whenever certain contextual phonological conditions are met. The result is that morphological transparency vanishes and paradigmatic patterns are obscured. The following paradigm (taken from Icelandic), which is not the most complex one by far, however, is completely rule-governed and regular, seen from a phonological point of view, but highly opaque when considered from a morphological perspective.

(1) Icelandic *fjörður* 'fjord; firth' (< Gmc. * *f**e**rþ-uz*; cf. IE *per- 'through')

a. singular	b. plural
N ***fjö***_rð-ur_	N ***fi***_rð-ir_
G ***fja***_rð-ar_	G ***fja***_rð-a_
D ***fi***_rð-i_	D ***fjö***_rð-um_
A ***fjö***_rð-∅_	A ***fi***_rð-i_

Direct consequences of these phonological processes (here: mostly breaking but also *i*-umlaut) are not only a *decrease* in morphological transparency but also an *increase* in complexity for adult and foreign language learners. Such phonological processes are quite typical of all small (and often isolated) Germanic languages. Their speakers primarily use their vernaculars for internal communication but use other languages for contact with their neighbors and/or foreigners (cf. Braunmüller 2003: 97). Thus no morphological alignments and simplifications due to incomplete language acquisition occur because foreigners hardly ever try (or need) to acquire languages such as Icelandic or a North Frisian variety (cf. also the more general discussion as presented in McWhorter 2007). Compared to the instances mentioned in (1), the following examples from two very small West Germanic vernaculars illustrate an even higher degree of morphological opaqueness and complexity (see Braunmüller [1985] 1995: 63, 75, 59):

(2) Mooring ***d***_äi_ – ***d***_eege_; Öömrang ***d***_ai_ – ***d***_aar_
day-SG day-PL day-SG day-PL

(3) Mooring ***sk***_öif_ (***sk***_ouf_) – ***sch***_uur_; Sölring ***sk***_och_ – ***sk***_uur_
shoe-SG shoe-PL shoe-SG shoe-PL.

Only the onsets are more or less identical (marked in bold). Umlaut and other phonological changes, again, obscure morphological transparency, resulting in some sort of suppletion (caused by phonological change). Similarly complex phenomena can be retrieved from Insular Scandinavian languages as well, demonstrated here by examples from the nominal and verbal inflexion (cf. Braunmüller [1985] 1995: 67):

(4) Faroese ***s**íggja* – *hann **s**ær* – *vit **s**íggja*
see-INF he see-3P.SG we see-1.P.PL

(5) Icelandic ***eig**a* – *hann á* – *við **eig**um*
have$_{\text{[possess]}}$-INF he have-3.P.SG. we have-1.P.PL.

The main point is that all phonological rules operate regularly and according to general rules; only a few of them, as in (5), cannot be motivated by a synchronic analysis (*eiga* is a preterit-present verb). The result is an increase not only in morphological complexity but also in morphological *opacity*. This lack of paradigmatic transparency (i.e. morphological complexity) makes it very difficult for foreign language learners to acquire these languages because they are highly irregular and seemingly not or only marginally rule-governed.

6.4 Structural complexity may also be due to language contact as an instance of *additive borrowing*. Faroese, in contrast to Icelandic, has borrowed many Danish complex verbs which are marked by prefixes, such as *an-*, *be-*, *for-* etc. (cf. *betala* 'to pay'). In doing so, the structural complexity of Faroese increases to a certain extent because the speakers of this language can often choose freely between (in this case) the inherited West Nordic verb *gjalda* 'to pay' and the calqued/code-copied verb from Danish *betala* 'to pay'.

Thus, the doubling of lexical items based on language contact may result either in more stylistic variation, as in English (cf. *freedom – liberty*, *to begin – to commence* ...), or, as in Faroese, facilitate code-switching between two (national) languages that are used in everyday life side by side and in frequent alternation. This leads to a *paradox* for those small and isolated languages where the speakers are generally bilingual and need two national languages in their everyday life (for different purposes or when talking to other residents who do not understand this local vernacular). In this case, linguistic complexity turns out to be an advantage and a sort of short-cut for bilinguals rather than a burden. This kind of additive complexity is not restricted to morphology, however. There is also a considerable amount of (typological) variation in the syntax of (in this case) Faroese, where native and borrowed/calqued constructions may be used side by side and without any semantic/stylistic differentiation:

(6) Faroese *mín hundur* – *hundur mín* (poss.pron. + N vs. N + poss.pron.)
‘my dog’ – ‘dog my’

(7) Faroese a. ***tann*** *lítla* *bók(in)* (definite article$_{\text{T-}}$ + adjective + N)
b. [*hin*] *lítla* *bók(in)* (definite article$_{\text{H-}}$ + adjective + N)
‘the little book(DEF)’

(7′) Danish ***den*** *lille* *bog* (definite article$_{\text{D}}$ + adjective + N)
‘the little book[∅]’

Hin in (7b) represents the predominantly West Nordic form of the definite article, whereas *tann* shows a certain morpho-phonological relationship with the East Nordic, here Danish (7′), definite article system which is based on forms with a word-initial *d-*. In fact, the Faroese ***tann***-article (7a) functions as a translingual diamorph in relation to the Danish (***den***) equivalent, which, again, facilitates frequent code-switching and the parallel use of the two genetically related languages: Faroese and Danish.

Other instances of syntactic variation and complexity are the adnominal genitives (8), the vacillation in word order in subordinate clauses (9), and the facultative use of the expletive pronoun (10). Both constructions occur and there is no semantic/stylistic difference between them. In cases (8) to (10), syntactic complexity is due to *additive borrowing* from Danish.

(8) Far. *Móttakarans undirskrift* – *undirskrift móttakarans*
(Gen + NP vs. NP + Gen)
recipient-DEF-GEN signature – signature recipient-DEF-GEN
‘the recipient’s signature’

(9) Far. a. *Jógvan sigur, at han keypir ikki tað breyð(ið).*
[≈ Icelandic word order]
John says that he buys *not* this bread(DEF)

b. *Jógvan sigur, at han ikki keypir tað breyð(ið).*
[≈ Danish word order; (9′)]
John says that he *not* buys this bread(DEF).

(9′) Danish *Hans siger, at han ikke køber det her brød.*
John says that he not buys this here bread.
‘Johns says that he does not buy this bread.’

(10)	Faroese	*Í*	*gjár*	*regnaði*	*∅ / tað*	*i*	*Tórshavn.*
		In	yester[day]	rained	∅ / it	in	Tórshavn

(10′)	Danish	*I*	*går*	*regnede*	*det* / *∅	*i*	*Tórshavn.*
		In	yester[day]	rained	it / *∅	in	Tórshavn

‘Yesterday it rained in Tórshavn.’

But new constructions also entered the Faroese language as replications from the other national language, Danish, such as (a) the *aktionsart* equivalent to the continuous form in English (11), and (b) the use of adverbal particles as *modal particles* functioning as hedges or downgraders (12) (for more details and more examples, see Petersen 2010: 142–144, 64, 117–119, 157; almost all data quoted are taken form the K8 database, collected in Hamburg; German Research Foundation, Colloborative Research Centre 538 on multilingualism).

(11)	Faroese	*Eg*	*standi*	*og*	*spekuleri*	*uppá*	*Lenu.*
		I	stand	and	speculate	up-on	Lena-OBL

(11′)	Danish	*Jeg*	*står/går/ligger/ …*	*og*	*tænker*	*på*	*Lena.*
		I	stand/go/lie/ …	and	think	on	Lena

‘I am thinking about Lena.’

(12)	Faroese	*Vissi*	*mann*	*lurtar*	*væl*	*eftir*	*orðunum,*
		Some	man	listens	carefully	after	word-DEF
		so	*má*	*tað*	***væ***	*vera,*	
		so	must	it	[I think so]	be,	
		at	*alt*	*skal*	*vera*	*javnt*	
		that	all	shall	be	equal	

‘If one listens to the word carefully, so it must ___ be the case that everything has to be equal.’

Normally English, the Romance languages, and Insular Scandinavian – unlike German and the Mainland Scandinavian languages – have no hedging modal particles. But Faroese acquired them from Danish (and Danish from Low German) and thus became categorically more complex. Again, this surplus of complexity, compared to the sister language Icelandic, has the advantage that the speakers of Faroese can more easily switch over to Danish (which is no longer necessary for the speakers of Icelandic).

To sum up, Faroese not only shows more variation and a richer lexicon but has also taken on two new grammatical patterns, namely *aktionsart* and modal particles, which increased the complexity of the language to a certain extent, though it is not felt as a burden but rather as an asset for successfully mastering two languages simultaneously.

7 On grammatical over-specification: possession and definiteness in Germanic

As already pointed out in section 4.2, over-specification, as defined by McWhorter (2007), may be considered an important source of grammatical complexity. However, over-specification can only be revealed cross-linguistically, viz. by investigating other languages which are organized in a typologically different way. The most elucidating results, in our view, are obtained when comparing genetically closely related varieties that are structured in partially different ways: One branch of a language family (or one specific language) may be specified in a divergent way which can only be revealed by a comparison to other closely related varieties. Applying this approach, we can mark the maximal (and the minimal) borderlines for grammatical constructions: Some morphological phenomena may be represented by only one single grammatical category, others by various and diverging grammatical categories and/or periphrastic constructions (see section 7.1 on possession). Moreover, other instances of over-specification can be described in terms of making more or less use of means taken from one and the same grammatical category (see section 7.2 on definiteness).

7.1 *Possession* in Germanic can be expressed in at least four ways:
(a) with genitives (see ex. in (13))
(b) with various periphrastic prepositional constructions (14)
(c) by means of datives (more precisely, benefactives) (15), and occasionally
(d) as in Icelandic and Faroese with local prepositions ('at'), when talking about body parts or inalienable possessions (16):

(13) *Peter's car* (Swedish *Pers bil, min fars bil* 'my father's car', *husets tak* 'the roof of the house' [literary style in Swedish], *today's morphology is yesterday's syntax* etc.).

(14) *the Book of Kells, the director of X* (Swedish *chefen för avdelingen* 'the head of the department', *en vän till mig* 'a friend of mine', *taket på huset* 'the roof of

the house [colloquial style in Swedish]; cf. also German *X von Y* or Dutch *X van Y* etc.).

(15) Colloquial/non-standard German *Wem ist dieses Buch?* 'Who(m) belongs this book (to)?', cf. Latin *mihi liber est* 'that is my book', possession rendered by dative constructions.

(16) Faroese/Icelandic *fótin(n) hjá mér* 'my foot [lit.: the foot at me]'.

A special case of instance (c) is the so-called Garpe genitive (as represented in 17d) that entered Norwegian and other Mainland Scandinavian varieties during the time of intense contact with speakers of Low German (from the era of the Hanseatic League). Norwegian and its dialects acquired even more structural complexity at that time, compared to e. g. Danish or Swedish (cf. (17b/d)):

(17)	Norwegian	a.	*Olav**s** bok*	[N$_{\text{possessor}}$-s N$_{\text{object}}$]$_{\text{NP}}$ 'Olav's book'
		b.	*boken han**s** Olav*	[N+def.$_{\text{object}}$ poss.pron.$_{\text{non-reflexive}}$] N$_{\text{possessor}}$]
		c.	*boken **til** Olav*	[N+def.$_{\text{object}}$[prep.**to** N$_{\text{benefactive}}$]$_{\text{PP}}$]$_{\text{NP}}$
		d.	*Olav **sin** bok*	[N$_{\text{benefactive}}$ poss.pron.$_{\text{reflexive}}$ N$_{\text{object}}$]$_{\text{NP}}$.

7.2 *Definiteness* can be expressed in many ways in Germanic languages. The greatest variation occurs in those Scandinavian varieties where we meet both preposed and enclitic definite article forms, which gives way to many kinds of structural complexity.

(I) SIMPLE definiteness marking **A** (English, German, Dutch, Danish, Jutish dialects):

(18)	a. English	*the*	*big∅*	*car*	
	b. German	*das*	*große*	*Auto*	
	c. Dutch	*de*	*grote*	*auto*	
	d. Danish	*den*	*store*	*bil,*	
	e. Jutish	*æ/de*	*stor∅*	*bil*	
		[DET	ADJ	N]$_{\text{DP}}$	and

(I′) SIMPLE definiteness marking **B** (Swedish [now restricted to names], Norwegian, Icelandic):

(19) Swedish ∅ *Nordiska museet*
∅ Nordic museum-DEF
‘The Nordic Museum’
restricted to names but generally in

(20) Icelandic ∅ *gamli maðurinn*
∅ old man-DEF
‘the old man’
[∅ ADJ NP-DET]$_{DP}$

(II) DOUBLE definiteness marking (Swedish, Norwegian, Faroese):

(21) Swedish *den stora bilen*
(22) Norwegian *den store bilen* (modern Dano-Norwegian)
(23) Faroese *tann stora bil(in)*
[DET ADJ N-DET]$_{DP}$

(21′) Swedish *den här stora bilen*
the here big car-DEF
den stora bilen här
the big car-DEF here

(22′) Danish *den store bil∅ her*
the big car here
‘this big car’
(with compound [= complex] demonstrative pronouns):
den her | *der*
the here | there
‘this | that’

(III) DIVERGENT definiteness marking (in North Frisian and in some Western German dialects):
[DET$_{\mathbf{d\text{-}/a}}$ (ADJ) NP]$_{DP}$ displayed as

- discourse deictic article *d-* [North Frisian: Fering (inherited); see Ebert (1971)] or *dä* [the Rhineland German dialect of Amern; see Heinrichs (1954)] or as
- exophoric/direct referential article *a* [Fering (obviously borrowed from South Jutish)] or *dər* [Amern; a semantic split].

The fact that some Germanic dialects show more than one (preposed) definite article is absolutely unique because their etymologies apparently diverge, which also leads to a functional split in their forms. Moreover, this example shows that Rescher's (1998) theoreme that complexity is a "'self-potentiating phenomenon"' has empirical support in natural languages.

8 Concluding remarks

It has been demonstrated that Hockett's (1958) claim that all languages are "equally complex" is hardly tenable. Rescher's (1998) philosophic and general view on the self-augmenting nature of complexity must be modified and restricted to certain domains when applied to natural languages. His claim is obviously too strong but in principle on the right track, as long as no radical outside forces intervene, typically due to intense language contact.

It has been pointed out that externally and internally motivated linguistic factors such as (a) language contact, (b) language cultivation and linguistic conservatism, and (c) linguistic variation but also (d) metalinguistic factors underlying linguistic reconstructions may cause or increase complexity. Moreover, factors of codification and elaborateness (literacy) on the one hand and contextual dependency (orality) on the other also must be taken into account: The more contextual information is verbalized, the more complex the linguistic form will be. But there will always be an interplay, for example, between elaborated and diversified formal (morphological/syntactic) structures and semantic structures and/or their pragmatic use: Divergent grammatical forms and constructions may be used to express (by and large) the same grammatical content. That is why it makes sense to speak of *selective complexity*: Not all domains of grammar are equally complex and it seems that 1:1 correlations between formal and semantic categories are exceptions rather than the rule in inflectional languages such as the Germanic ones.

Two metalinguistic categories seem to be helpful when distinguishing between instances of selective complexity: over- and under-specification, as described and defined by McWhorter (2007). These two complementary parameters help to measure the degree of divergence in complexity in cross-linguistic comparisons. This has been shown with respect to the grammatical categories "possession" and "definiteness" in Germanic languages. Thus, the issue is not which language (or which branch of a language family) is more complex than others, but rather which *domains* of grammar of a certain language display more differentiation than others (selective complexity in terms of over-specification).

Germanic language history has shown that if morphological complexity is lost due to the impact of foreign learners it can still be acquired again, strictly speaking, restored by *replicating* grammatical structures from a (related) model language (here: the IE sister language, Latin). From a theoretical point of view, this means that complexity (a) can be lost, for whatever reason, but it can (b) also be built up again or restored – *both* developments are triggered by external impact, viz. language contact. Thus, new grammatical structures as well as recovered lost structures can emerge and create new or regain lost complexity.

Complexity may be best preserved in bi- or multilingual societies where languages other than the local vernacular are used for communication with non-native speakers (e. g. in the Faroe Islands or North Frisia). All languages that are not (generally) learned by foreigners have a much greater chance of keeping their original linguistic structures than languages that are learned by many immigrated adults or that are used as a *lingua franca* in trans-regional communication. In other words, incomplete second language acquisition (as is typical of adult language learners) evidently does not cause more complexity, but rather leads to less complexity, i. e. to more simplification and regularity in grammatical structures. Completely bilingual societies, as in the Faroe Islands or among locally restricted minorities, make use of more than one grammatical system, which directly leads to more complexity within *both* languages involved. Moreover, if these languages are genetically closely related, their speakers will sooner or later mix these systems, which will likewise lead to more linguistic complexity (for more details on this special case, see Braunmüller 2009).

References

Bossong, Georg 2001: '¡Si me tomas el pelo, te hago pedazos!': Hacia una tipología de los verbos fraseológicos. *Revista de Investigación Lingüistica* 4 (1): 5–28.

Braune, Wilhelm and Ebbinghaus, Ernst A. 1973: *Gotische Grammatik: Mit Lesestücken und Wörterverzeichnis*. 18th edition. (Sammlung kurzer Grammatiken germanischer Dialekte A 1.) Tübingen: Niemeyer.

Braunmüller, Kurt 1985: Morphologische Undurchsichtigkeit – ein Charakteristikum kleiner Sprachen. *Kopenhagener Beiträge zur Germanistischen Linguistik* 22, 1984 [!]: 48–68. Reprinted in: Kurt Braunmüller 1995, *Beiträge zur skandinavistischen Linguistik*, 53–80. (Studia Nordica 1.) Oslo: Novus.

Braunmüller, Kurt 2003: Language, typology and society: Possible correlations. In: Thomas Stolz and Joel Sherzer (eds.), *Minor Languages: Approaches, Definitions, Controversies. Papers from the Conference on 'Minor Languages: Coming to Grips with a Suitable Definition', Bremen, June 2001*, 89–101. (Diversitas Linguarum 3.) Bochum: Brockmeyer.

Braunmüller, Kurt 2004: Zum Einfluss des Lateinischen auf die ältesten Runeninschriften. In: Oskar Bandle, Jürg Glauser and Stefanie Würth (eds.), *Verschränkung der Kulturen:*

Der Sprach- und Literaturaustausch zwischen Skandinavien und den deutschsprachigen Ländern. Zum 65. Geburtstag von Hans-Peter Naumann, 23–50. (Beiträge zur Nordischen Philologie 37.) Tübingen/Basel: Francke.

Braunmüller, Kurt 2008a: Das älteste Germanische: Offene Fragen und mögliche Antworten. *Sprachwissenschaft* 33: 373–403.

Braunmüller, Kurt 2008b: Observations on the origins of definiteness in ancient Germanic. *Sprachwissenschaft* 33: 351–371.

Braunmüller, Kurt 2009: Converging genetically related languages: *Endstation* code mixing? In: Kurt Braunmüller and Juliane House (eds.), *Convergence and Divergence in Language Contact Situations*, 53–69. (Hamburg Studies on Multilingualism 8.) Amsterdam/ Philadelphia: Benjamins

Braunmüller, Kurt and Höder, Steffen 2012: The history of complex verbs in Scandinavian languages revisited: Only influence due to contact with Low German? In: Ernst Håkon Jahr and Lennart Elmevik (eds.), *Contact Between Low German and Scandinavian in the Late Middle Ages. 25 Years of Research*, 151–169. (Acta Academiae regiae Gustavi Adolphi 121.) Uppsala: Gustav-Adolf's-Academy.

Dahl, Östen 2004: *The Growth and Maintenance of Linguistic Complexity.* (Studies in Language Companion Series 71.) Amsterdam/Philadelphia: Benjamins.

Ebert, Karen 1971: *Referenz, Sprechsituation und die bestimmten Artikel in einem nordfriesischen Dialekt (Fering).* (Studien und Materialien 4.) Bredstedt: Nordfriisk Instituut.

Gil, David 2009: How much grammar does it take to sail a boat? In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 19–33. Oxford etc.: Oxford University Press.

Hawkins, John A. 1994: *A Performance Theory of Order and Constituency.* (Cambridge Studies in Linguistics 73.) Cambridge etc.: Cambridge University Press.

Heinrichs, Heinrich Matthias 1954: *Studien zum bestimmten Artikel in den germanischen Sprachen.* (Beiträge zur deutschen Philologie 1.) Gießen: Schmitz.

Hirt, Hermann 1931/1932/1934: *Handbuch des Urgermanischen:* Vol. I: *Laut- und Akzentlehre* [1931], Vol. II: *Stammbildungs- und Flexionslehre* [1932], Vol. III: *Abriß der Syntax* [1934]. (Indogermanische Bibliothek I, 21.) Heidelberg: Winter.

Hockett, Charles F. 1958: *A Course in Modern Linguistics.* New York etc: Macmillan.

Josephson, Folke 2004: Semantics and typology of Hittite *-ant*. In: James Clarckson and Birgit Anette Olsen (eds.), *Indo-European Word Formation: Proceedings of the Conference Held at the University of Copenhagen, October 20th–22nd 2000*, 93–118. Copenhagen: Tusculanum.

Krahe, Hans and Meid, Wolfgang 1969: *Germanische Sprachwissenschaft.* Vol. I: *Einleitung und Lautlehre*. 7th edition. (Sammlung Göschen 238.) Berlin: Walter de Gruyter.

Labov, William 1994: *Principles of Linguistic Change.* Vol. 1: *Internal Factors.* Oxford/ Cambridge, MA: Blackwell.

Mailhammer, Robert 2007: *The Germanic Strong Verbs: Foundations and Development of a New System.* (Trends in Linguistics. Studies and Monographs 183.) Berlin/New York: Mouton de Gruyter.

McWhorter, John 2007: *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars.* Oxford etc.: Oxford University Press.

Miestamo, Matti, Sinnemäki, Kaius and Karlsson, Fred (eds.) 2008: *Language Complexity: Typology, Contact, Change*. (Studies in Language Companion Series 94.) Amsterdam/ Philadelphia: Benjamins.

Petersen, Hjalmar P. 2010: *The Dynamics of Faroese-Danish Language Contact*. (Germanistische Bibliothek 37.) Heidelberg: Winter.

Rescher, Nicholas 1998: *Complexity: A Philosphical Overview*. (Science and Technology Studies.) New Brunswick, NJ/London: Transaction Publishers.

Sampson, Geoffrey, Gil, David and Trudgill, Peter (eds.) 2009: *Language Complexity as an Evolving Variable*. Oxford etc.: Oxford University Press.

Trudgill, Peter 2011: *Sociolinguistic typology: social determinants of linguistic complexity*. (Oxford Linguistics). Oxford etc. Oxford University Press.

Vennemann, Theo, gen. Nierfeld 2003: *Europa Vasconica – Europa Semitica*. Patrizia Noel Azis Hanna (ed.) (Trends in Linguistics. Studies and Monographs 138.) Berlin/New York: Mouton de Gruyter.

Weinreich, Uriel 1954: Is a structural dialectology possible? *Word* 10: 388–400.

Katharina Ehret, University of Freiburg i. Br.
Benedikt Szmrecsanyi, KU Leuven

An information-theoretic approach to assess linguistic complexity

Abstract: For a long time all languages were considered to be, on the whole, equally complex. This dogma has recently been challenged. Unfortunately, much of the current controversy in regard to complexity (in)variance draws on empirically expensive or subjective evidence.

In this paper, we explore an idea proposed by mathematicians and computer scientists (e. g. Juola 2008) to use an unsupervised, algorithmic, information-theoretic measure for assessing linguistic complexity. Essentially, this measure boils down to the notion of Kolmogorov complexity which can be conveniently approximated by using modern file compression programs. Obtaining numerical estimates of the relative informativeness of text samples via file compression we can assess linguistic surface complexity on the overall, syntactic and morphological level.

To furnish a case study, we analyze both a parallel text database sampling the Gospel of Mark in six languages (Esperanto, Finnish, French, German, Hungarian, and Jamaican Patois) as well as some ten varieties of English, and a non-parallel sample of newspaper texts covering 9 European languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, Spanish).

We demonstrate that Kolmogorov complexity measurements yield linguistically meaningful results, and provide complexity rankings that are in line with what more orthodox complexity notions would lead one to expect. We conclude by considering the advantages and drawbacks of the method, and by sketching directions for future research.

1 Introduction[1]

Linguistic complexity is one of the currently most hotly debated notions in linguistics. The long-standing assumption that all languages are of equal complexity (Hockett 1958; Crystal 1987; Edwards 1994; Bickerton 1995; O'Grady, Dobrovolsky

[1] We would like to thank Annemarie Verkerk and the Max Planck Institute for Psycholinguistics (Nijmegen) for making their Alice database available to us. We are also grateful to the audience of

and Aronoff 1997) had remained unchallenged for much of the twentieth century. Recently, however, this dogma has been questioned and scrutinized (McWhorter 2001; Kusters 2003), and the notion of linguistic complexity has received a considerable amount of interest (Dahl 2004; Miestamo, Sinnemäki and Karlsson 2008; Sampson, Gil and Trudgill 2009; Kortmann and Szmrecsanyi 2012). Two central issues in the linguistic complexity debate are, firstly, the problem of finding a generally applicable definition of what exactly complexity is and secondly, how to measure this complexity.

In this paper, we explore proposals to use an unsupervised, algorithmic, information-theoretic measure (Juola 1998; Moscoso del Prado Martin, Kostic and Baayen 2004; Bane 2008; Juola 2008; Sadeniemi et al. 2008) for assessing linguistic complexity in various languages, based on more or less naturalistic text corpora. We specifically draw on Kolmogorov complexity, which defines the complexity of a string/text as the length of the shortest possible description of that string/text. Kolmogorov complexity is a quantitative, (ir)regularity-based type of complexity, which is completely agnostic about subjective agent-related complexity such as second language acquisition difficulty (Trudgill 2001; Kusters 2003, 2008; Szmrecsanyi and Kortmann 2009).

Kolmogorov complexity can be conveniently approximated by using modern, off-the-shelf file compression programs. Obtaining numerical estimates of the relative informativeness of text samples via file compression, we will assess linguistic complexity on the (i) overall, (ii) syntactic, and (iii) morphological level. To furnish a case study, we tap into three datasets sampling an array of European languages from the three major language families using the Latin alphabet (Germanic, Romance, Finno-Ugric), which are customarily used as test cases in the literature (Bakker 1998; Kettunen et al. 2006; Sadeniemi et al. 2008):

1. a parallel text database sampling the Gospel of Mark in six languages (Esperanto, Finnish, French, German, Hungarian, and Latin) as well as some ten historical varieties of English;
2. a parallel – and, after permutation wizardry, semi-parallel – corpus of *Alice's adventures in Wonderland* in nine languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish);
3. and a non-parallel sample of newspaper texts covering nine European languages (Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, and Spanish).

the 2011 FRIAS workshop on "Complexity and Isolation" for constructive feedback, and to Christoph Wolk for methodical advice. The usual disclaimers apply.

Table 1 and 2 list the number of words per database. Note that corpus size is not a crucial factor in research with parallel text corpora (see Section 4.1 for details on parallel texts).

Table 1: Corpus size in number of words for Alice in Wonderland and the Euro-Congo news corpus

	Alice	Euro-Congo News
Dutch	28,897	12,566
English	26,446	18,002
Finnish	18,572	9,358
French	25,327	20,500
German	26,309	12,175
Hungarian	19,517	15,413
Italian	24,709	18,193
Romanian	23,870	16,022
Spanish	27,128	18,245
Words in total	220,775	140,474

Table 2: Corpus size in number of words for the Gospel of Mark database

Mark	
English versions:	
American Standard	15,049
Anglo Saxon	16,021
Basic English	16,463
Darby	15,189
Douay Rheims	15,036
English Standard Version	14,432
King James	15,201
Websters	15,234
Wycliffe	19,183
Young's Literal	15,682
Other varieties:	
Esperanto	13,045
Finnish	10,507
French	15,712
German	14,120
Hungarian	11,781
Latin	10,545
Words in total	75,710

In this paper, which is methodological in nature, we aim to demonstrate that the compression technique yields linguistically rather meaningful results, because it provides complexity rankings that are in line with what more orthodox complexity notions would lead one to expect. Second, we show that the measurements work on parallel as well as non-parallel corpus data.

This paper is structured as follows. In Section 2, we discuss information theory. Section 3 explains how to measure Kolmogorov complexity. In Section 4, we present our empirical analyses. Section 5 concludes by considering the advantages and drawbacks of the method, and by sketching directions for future research.

2 Information theory

Information theory is "the science which deals with the concept 'information', its measurement and its applications" (van der Lubbe 1997: 1). In his landmark paper "A Mathematical Theory of Communication" Shannon (1948) analyzed the information content along a channel between a message source and a listener, establishing the maximum bounds for the efficiency with which messages can be transmitted. In the framework of this theory, the term 'information' refers to the unpredictability or unexpectedness of a proposition, event or, in terms of communication, a message. Thus, the information content of a message is directly related to its unpredictability, i. e. a message is informative if it is not predictable or expected and conveys something surprising and new. The information content of a message is measured in *Shannon entropy*, a measure of unpredictability or disorder, which calculates the information contained in a given message in relation to the predictable part of the message (which is not informative as it is already given).

A related measure of information is *Kolmogorov complexity* which, in contrast to Shannon entropy, refers to the information content of a given string (not message source). Shannon entropy "is an upper bound on (and asymptotically equal to) Kolmogorov complexity" (Juola 2008: 92), and measures the information content or complexity of a string of symbols as the length of the shortest possible description of it. Mathematically speaking, the complexity of a string is measured by the length of the algorithm which is required to (re)generate the exact string (Juola 2008: 92; Sadeniemi et al. 2008: 191; see also Li et al. 2004). Let us take a look at the two example strings of symbols in (1). Both strings consist of the same number of characters, yet string (1a) can be compressed to the expression *5×cd*, counting four characters, whereas the shortest description of string (1b) is the string itself. Measuring the complexity of string (1a) and string (1b) according to the length of their shortest possible description, string (1a) is obviously less complex than string (1b).

(1) a. cdcdcdcdcd (10 characters) → 5×cd (4 characters)
b. cdgh39aby7 (10 characters) → cdgh39aby7 (10 characters)

Adaptive entropy estimation methods can be used to compute and approximate the upper bounds for Kolmogorov complexity (Ziv and Lempel 1977; Juola 1998). File compression programs (such as gzip) actually use a variant of adaptive entropy estimation that approximates Kolmogorov complexity. More specifically, file compression programs compress text strings by describing new strings on the basis of previously seen and memorized (sub-)strings so that the amount of information and redundancy in a given string can be measured (Juola 2008: 93). The idea, then, is to measure complexity by measuring the information content – and hence, unpredictability – in text samples. In this endeavor, a higher amount of information (unpredictability) is taken to indicate increased linguistic complexity of the linguistic sample under analysis (Juola 1998).

Even though Kolmogorov complexity is not fully compatible with traditional notions of linguistic complexity because compression tools are agnostic of, say, form-meaning relationships, they do capture recurrent (linguistic) patterns and (ir)regularities. Kolmogorov complexity conflates to some extent structural complexity and system complexity (Dahl 2004: 42–44) and also adds a substantial amount of frequency weighting. In sum, Kolmogorov complexity is a quantitative, frequency-based, and corpus-based measure of absolute linguistic complexity; it measures linguistic surface complexity by describing new structures on the basis of previously encountered structures.

3 How to measure Kolmogorov complexity

Methodologically, we utilize an open source compression program, namely gzip (version 1.2.4), to approximate Kolmogorov complexity and thus to assess linguistic complexity on the overall, syntactic and morphological plane. The overall complexity of our text samples is measured by obtaining two measurements for each text file analyzed: the file size (in bytes) before compression, which roughly corresponds to Dahl's (2004) notion of verbosity, and the file size (in bytes) after compression. Subsequently, the two values are subjected to regression analysis in order to eliminate any trivial correlation between the two measurements. The resulting *adjusted overall complexity scores* (regression residuals, in bytes), which measure left-over variance, are taken as indicators of the overall complexity of a given language sample. Overall complexity is equivalent to global complexity in the sense of Miestamo (2008: 29–32) in that it refers to the complexity of a language as a whole. Bigger adjusted complexity scores can be equated with higher informativeness of a given text sample and thus indicate higher levels of Kolmogorov complexity.

Complexity at the morphological and syntactic tier can be addressed by manipulating the text files prior to compression. Largely following Juola (2008), morphological distortion randomly deletes 10 % of the orthographic characters in each file. Through this procedure new word forms are created while at the same time morphological regularity is compromised. Subsequently, the distorted samples are compressed in order to determine how well or badly the compression program deals with the distortion. As morphologically complex languages exhibit overall a relatively large amount of word forms in any case, distortion should not hurt them as much as morphologically simple languages, in which distortion creates proportionally more random noise and thus entropy/complexity. Comparatively worse compression ratios thus signify low morphological complexity.

Distortion at the syntactic level is accomplished by randomly deleting 10 % of all orthographically transcribed word tokens in each sample. This procedure is assumed to have little impact on languages with simple syntax (which is defined here as, essentially, maximal flexibility and free word order) as they lack between-word interdependencies which could be compromised. Syntactically complex languages, however, should be greatly affected as word order regularities are compromised. The auxiliary sequence *would have been* (2a), for instance, which occurs twice in one of our text samples (the Bible in Basic English), could be altered to *would ___ been* (2b) through distortion. In this case the compression algorithm would encounter two hapax legomenon patterns – instead of encountering one pattern twice – which leads to uncompressible entropy. This will hurt compression efficiency. To make a long story short, comparatively bad compression ratios after syntactic distortion indicate high syntactic complexity.

(2) a. *no flesh* ***would have been*** *kept from destruction*
    b. *no flesh* ***would*** _____ ***been*** *kept from destruction*
    (Mark 13:20 [Basic English])

On a more technical note, we calculate two complexity scores: a *morphological complexity* score which is defined as – m/c, where *m* is the compressed file size after morphological distortion and *c* is the compressed file size before distortion. The *syntactic complexity score* is defined as s/c, where *s* is the compressed file size after syntactic distortion and *c* the file size before distortion.

4 Measuring linguistic complexity in corpora

4.1 The Gospel of Mark – complexity in parallel texts

The use of file compression programs for measuring linguistic complexity has to date been limited to parallel text corpora, i. e. translational equivalents of the same text in different languages. Such parallel text databases have become quite popular in typological studies (Cysouw and Wälchli 2007; Dahl 2007) as they facilitate comparability across different languages and language varieties due to the fact that differences in propositional content can be ruled out. For the same reason, relatively small databases can be used without compromising comparability. Furthermore, in complexity research parallel text databases permit generalizations beyond individual texts, i. e. the complexity of a set of languages can be inferred from a parallel sample of these languages. In other words, if all languages were equally complex, a given text, be it *Harry Potter* or *Pride and Prejudice*, should be equally complex in all languages. However, this is not the case as shall be demonstrated. The classic database in parallel text studies is the Holy Bible (see, e. g., Juola 2008), and in precisely this spirit we set the stage by applying the compression technique to the Gospel of Mark in a number of historical varieties of English and seven other languages listed below.

Varieties of English:

- West Saxon (approx. 10th century [from Bright 1905])
- Wycliffe's Bible (14th century [1395])
- The Douay-Rheims Bible (16th century [1582])
- The King James Version (17th century [1611])
- Webster's Revision (19th century [1833])
- Young's Literal Translation (19th century [1862])
- The Darby Bible (19th century [1867])
- The American Standard Version (20th century [1901])
- The Bible in Basic English (20th century [1941]), using mostly 850 Basic English words and simplified grammar (Ogden 1934, 1942)
- The English Standard Version (21st century [2001])

Other languages:

- Esperanto (Esperanto Londona Biblio, 20th century [1926])
- Finnish (Pyhä Raamattu, 20th century [1992])
- French (Ostervald, 20th century [1996 revision])
- German (Schlachter, "Miniaturbibel", 20th century [1951 revision])
- Hungarian (Vizsoly Bible [a. k. a. Károli Bible], 16th century)
- Latin (Vulgata Clementina, 4thcentury)

We proceed as described above and establish the file sizes in bytes before and after compression for each file. We then calculate adjusted overall complexity scores for all language samples and obtain a hierarchy of overall complexity (Figure 1).[2] West Saxon, Hungarian, Finnish, Latin, French and German are (in decreasing order) rather complex whereas Esperanto and all English texts after 1066 are rather simple. These findings tie in neatly with previous, more traditional complexity research (Nichols 1992; Bakker 1998).

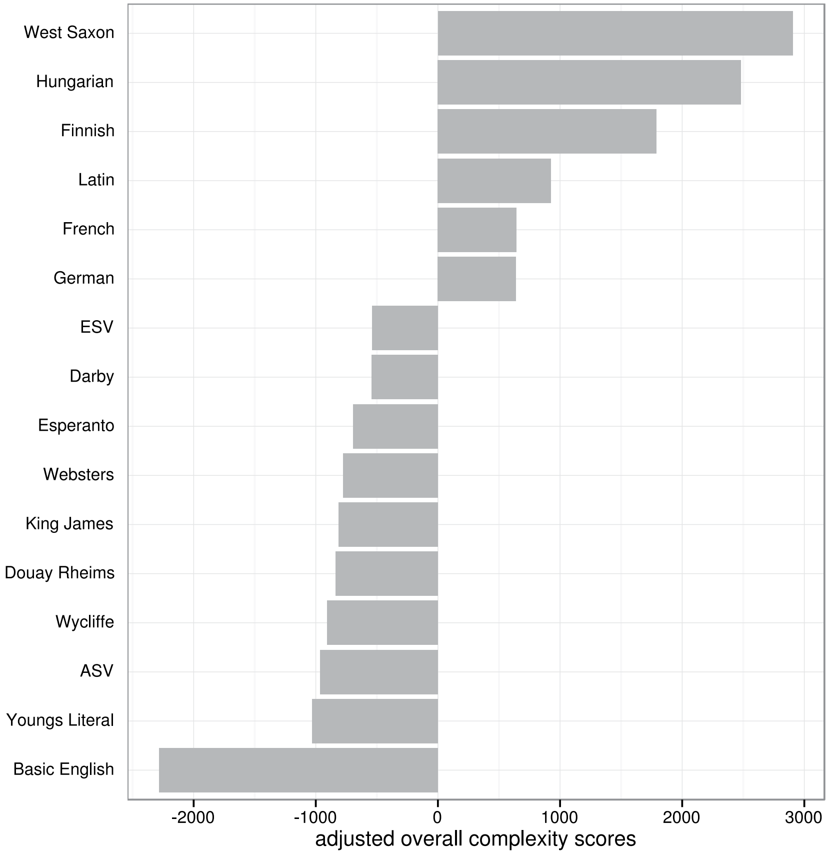

Figure 1: Overall complexity hierarchy. Negative residuals indicate below-average complexity; positive residuals indicate above-average complexity

[2] We note that morphological complexity seems to interact with our overall complexity measure in that morphological complexity is more strongly reflected in the overall measure than syntactic complexity. This is due to the nature of the algorithm which picks up on (structural) irregularities and redundancy in the unmanipulated texts.

The analysis of morphological and syntactic complexity yields equally intuitive results. Thus in Figure 2, languages which are morphologically complex but syntactically simple cluster in the top left quadrant: West-Saxon, Finnish, Latin and Hungarian exhibit the most complex morphology. All the English varieties and French – with the exception of West Saxon – are morphologically simple but syntactically complex and are scattered across the bottom right quadrant with Basic English displaying the lowest morphological complexity. German and, Esperanto cover the middle ground and seem to be balanced in regard to morphological versus syntactic complexity.

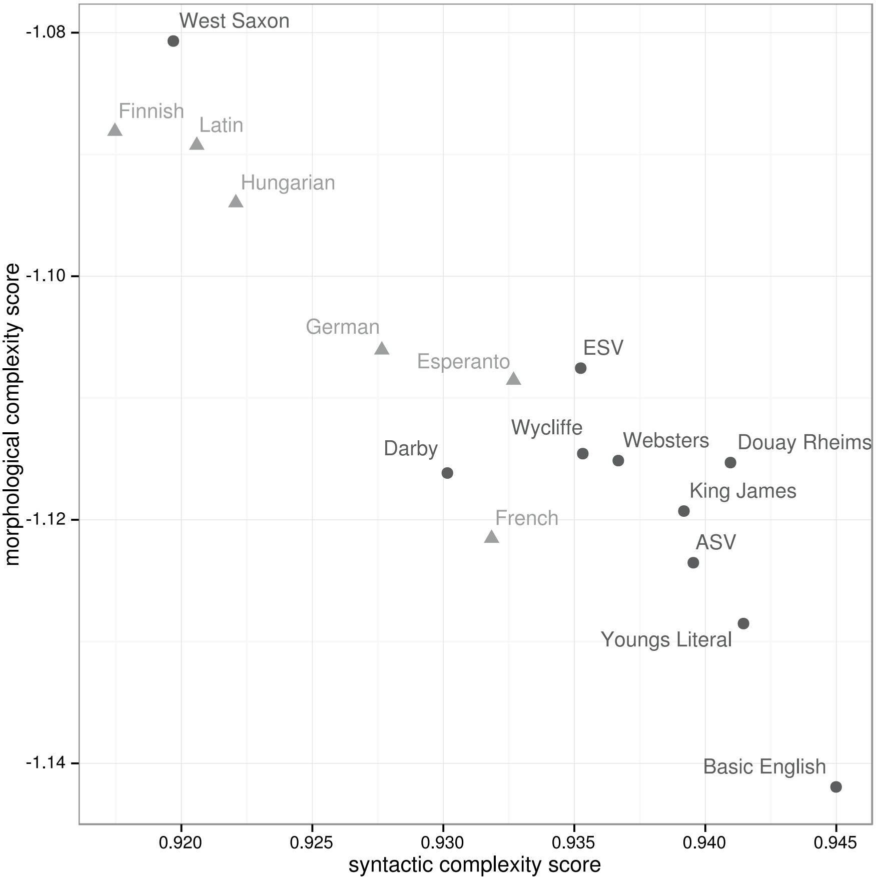

Figure 2: Morphological complexity by syntactic complexity. Abscissa indexes increased syntactic complexity, ordinate indexes increased morphological complexity

In the Bible sample, morphological complexity trades off against syntactic complexity and vice versa. A negative correlation between morphological complexity and syntactic complexity is particularly prominent when focusing on the English varieties; with a Pearson's correlation coefficient of $r = -.92$, $\mathrm{p} = 2.374\mathrm{e}^{-07}$ the correlation between the complexity scores indicates a textbook-style trade-off.

We illustrate the workings of the compression technique with an example passage from Mark 1:8–9 in West Saxon (classified as a morphologically complex but syntactically simple language) and Basic English (classified as a morphologically simple but syntactically complex language). In terms of morphology (see Table 3), we count nine different segmentable inflected word tokens in the West Saxon version whereas we only count three different tokens and two types (*giv-en, day-s*) in the Basic English version.

Table 3: Segmentable inflected word tokens in Mark 1:8–9

West Saxon	Basic English
[8] Ic **fullig-e** eow on **wæter-e**; he eow **full-aþ** on **Halg-um Gast-e**.	[8] I have **giv-en** you baptism with water, but he will give you baptism with the Holy Spirit.
[9] And on ðam **dag-um**, come se Hælend fram Nazareth Galilee, and wæs **ge-full-od** on **Iordan-e** fram **Iohann-e**.	[9] And it came about in those **day-s**, that Jesus came from Nazareth of Galilee, and was **giv-en** baptism by John in the Jordan.

Turning to syntax (Table 4), the West Saxon version features four different word order patterns whereas in the Basic English version word order is relatively rigid (i.e. complex) because the pattern subject-verb dominates throughout the passage. Syntactic complexity is defined in terms of flexibility of word order (cf. Bakker 1998), i.e. less flexibility generates more complexity. Therefore, Basic English is classified as a syntactically complex language – in contrast to West Saxon, it has many word order inter-dependencies which can be violated.

Table 4: Word order patterns in Mark 1:8–9

West Saxon	Basic English
[8] [Ic]$_{\text{subject}}$ [fullige]$_{\text{verb}}$ [eow]$_{\text{object}}$ [on wætere] $_{\text{adverbial}}$; [he]$_{\text{subject}}$ [eow]$_{\text{object}}$ [fullaþ]$_{\text{verb}}$ [on Halgum Gaste] $_{\text{adverbial}}$.	[8] [I]$_{\text{subject}}$ [have given]$_{\text{verb}}$ [you]$_{\text{object}}$ [baptism]$_{\text{object}}$ [with water] $_{\text{adverbial}}$, but [he]$_{\text{subject}}$ [will give]$_{\text{verb}}$ [you]$_{\text{object}}$ [baptism]$_{\text{object}}$ [with the Holy Spirit]$_{\text{adverbial}}$.
[9] And [on ðam dagum] $_{\text{adverbial}}$, [come]$_{\text{verb}}$ [se Hælend]$_{\text{subject}}$ [fram Nazareth Galilee]$_{\text{adverbial}}$, and [wæs gefullod]$_{\text{verb}}$ [on Iordane]$_{\text{adverbial}}$ [fram Iohanne] $_{\text{adverbial}}$.	[9] And [it]$_{\text{subject}}$ [came about]$_{\text{verb}}$ [in those days] $_{\text{adverbial}}$, that [Jesus]$_{\text{subject}}$ [came]$_{\text{verb}}$ [from Nazareth of Galilee] $_{\text{adverbial}}$, and [was given]$_{\text{verb}}$ [baptism]$_{\text{object}}$ [by John in the Jordan] $_{\text{adverbial}}$.

Let us focus now on historical drifts in the English translations of the Bible, as gauged by the compression technique. It is well-known that English has changed from a rather synthetic language – i. e. one that relies heavily on inflections to code grammatical information – in Old English times into a rather analytic language that draws on word order and function words to convey grammatical information. This textbook story is nicely depicted in Figure 3, which plots real time drifts in the history English: our Kolmogorov complexity measurements clearly suggest morphological simplification and syntactic complexification, some outliers notwithstanding.

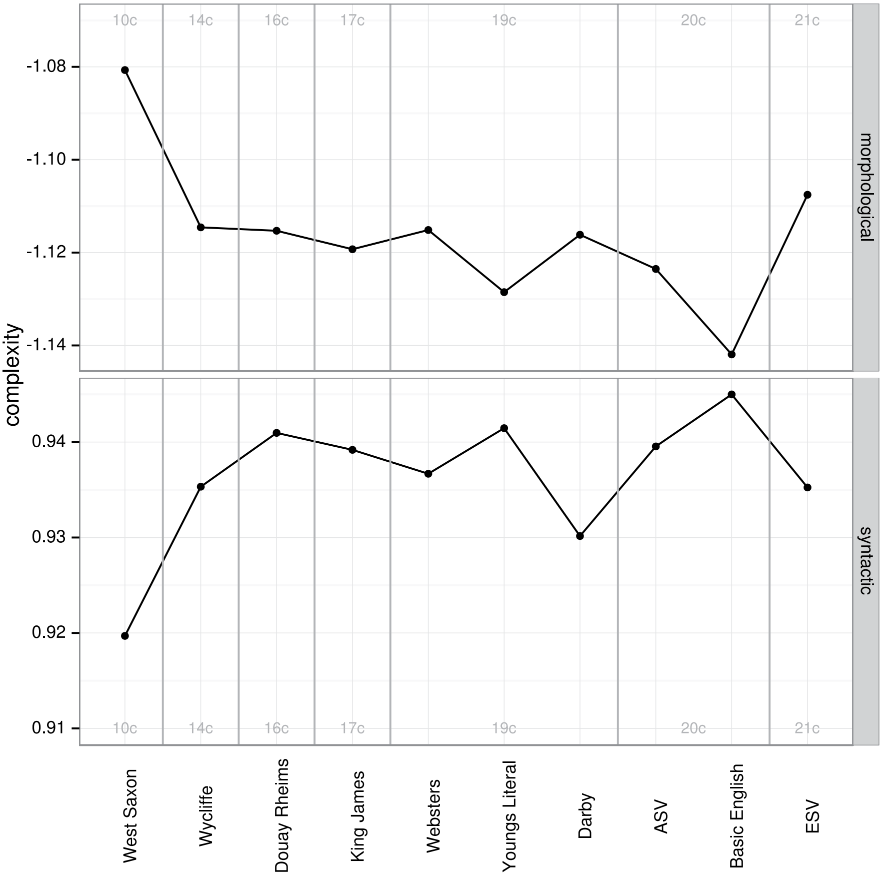

Figure 3: Real time drifts in English: morphological (upper plot) and syntactic complexity (lower plot). Abscissa arranges Bible translations chronologically

4.2 Parallel versus non-parallel texts

In this section we aim to demonstrate that the compression technique need not be limited to parallel texts but can also be applied to non-parallel text databases. We draw on a parallel text database and a non-parallel text database in order to explore the reach and limits of the compression technique in two steps. Firstly, we measure and subsequently compare linguistic complexity in a parallel corpus of *Alice's adventures in Wonderland* and a re-sampled semi-parallel version of the same corpus. Secondly, we measure linguistic complexity in non-parallel newspaper texts and compare our results to the complexity hierarchy obtained from the Alice corpus. In short, we validate our non-parallel corpus results against the parallel Alice corpus.

4.2.1 *Alice's adventures in Wonderland*

In a first step, we sample *Alice's adventures in Wonderland* (by Lewis Carroll) in nine languages chosen from Germanic, Romance and Finno-Ugric languages which use the Latin alphabet and are frequently utilized as test cases in the complexity literature (Bakker 1998; Kettunen et al. 2006; Sadeniemi et al. 2008) – Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian, Spanish – and measure linguistic complexity on the overall, syntactic, and morphological tier.

Establishing the file sizes of each text file before and after compression, we calculate the adjusted complexity scores which indicate the overall complexity of each language sample. Subsequently, we address syntactic and morphological complexity by applying multiple distortion and compression – using an R script[3] which implements the methodology as described in Section 3 but allows for multiple iterations – to the complete Alice corpus. By taking multiple measuring points we ensure that our findings are statistically robust. Notice here that in the process of random deletion, any character or word token of a given text file could be modified. However, the impact of the deletion on complexity may, of course, vary according to the precise character/word which was subject to deletion. Consider example (3), which illustrates syntactic distortion. (3a) is the unaltered sentence. In both (3b) and (3c) two words were deleted, but the impact of the deletion differs greatly: while (3b) is still syntactically intact, (3c) has been rendered incomprehensible because syntax is compromised badly.

[3] R 2.14.0 (R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0).

(3) a. *The Rabbit actually took a watch out of its waistcoat-pocket.*
b. *The Rabbit _______ took a watch out of its waistcoat ______.*
c. *The ______ actually took a _____ out of its waistcoat-pocket.*

In terms of complexity, compression of neither (3b) nor (3c) in isolation would reflect the actual complexity of the sentence. However, taking the average of several measuring points, the actual complexity of the string can be approximated. We therefore apply multiple distortion and compression with $N = 1{,}000$ iterations to each Alice version. Every iteration returns the compressed file sizes for each language sample before and after syntactic/morphological distortion. On the basis of these file sizes we calculate the *average morphological complexity score* and *the average syntactic complexity score*. Intra-sample dispersion turns out to be negligible.[4] The average complexity scores are subsequently obtained by taking the mean of the total number of measuring points ($N = 1{,}000$) for morphological and syntactic complexity respectively.

We also create a semi-parallel Alice corpus by means of permutation. Before every iteration of the multiple distortion and compression script, we randomly sample 10 % of the total number of sentences[5] from the Alice corpus. Even though the original content of this re-sampled corpus is the same across all languages, the concrete content of each sub-sample varies due to the multiple random permutations. We apply the script with $N = 1{,}000$ iterations and thus obtain 1,000 measuring points for the compressed and uncompressed file sizes before and after syntactic/morphological distortion of each permutated language sample. Subsequently, we obtain the *average overall complexity score*[6] by calculating

[4] We measure the dispersion across the individual data points by calculating the *variation coefficient*, which is defined as the ratio of the standard deviation (sd) to the mean: sd(data1)/mean(data1). This measure is more robust than the standard deviation and handles possible outliers in the data much better as it is independent from the mean size (Gries 2009: 203–204). The smaller the value of the variation coefficient is, the smaller is the dispersion in the sample, i. e. the more reliable is the mean. We report the following variation coefficients for morphological complexity ratios in the parallel corpus: Dutch – 0.0012, English – 0.0014, Finnish – 0.0013, French – 0.0014, German – 0.0012, Hungarian – 0.0013, Italian – 0.0012, Romanian – 0.0014, Spanish – 0.0012. The dispersion measures for syntactic complexity ratios in the parallel corpus are: Dutch – 0.0015, English – 0.0017, Finnish – 0.0016, French – 0.0016, German – 0.0015, Hungarian – 0.0017, Italian – 0.008, Romanian – 0.0016, Spanish – 0.0016.

[5] By taking 10 % of the total number of sentences instead, for instance, of words or characters, we ensure that syntactic inter-dependencies remain intact while, at the same time, we keep the sample size across languages constant.

[6] The average overall complexity score is calculated on the basis of compressed and uncompressed file sizes. Variation coefficients for compressed file sizes are as follows: Dutch – 0.076,

regression residuals of the mean compressed file sizes (dependent variable) and the mean uncompressed file sizes (independent variable). The average morphological complexity score and the average syntactic complexity score[7] are subsequently computed as described above. Finally, we compare the overall and morphosyntactic complexity hierarchies of the parallel and semi-parallel Alice corpora.

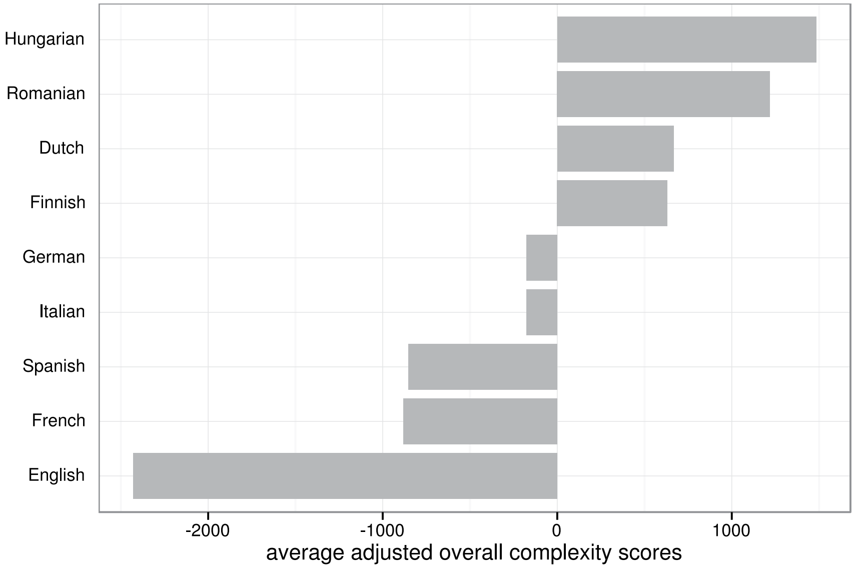

Figure 4: Overall complexity hierarchy in the parallel Alice corpus

The ranking of overall complexity in the parallel corpus (Figure 4) is, in decreasing order of complexity, Hungarian, Romanian, Dutch, Finnish, German, Italian,

English – 0.88, Finnish – 0.074, French – 0.07, German – 0.083, Hungarian – 0.063, Italian – 0.08, Romanian – 0.077, Spanish – 0.07, and uncompressed file sizes: Dutch – 0.074, English – 0.084, Finnish – 0.067, French – 0.069, German – 0.08, Hungarian – 0.063, Italian – 0.079, Romanian – 0.079, Spanish – 0.073.

[7] This yields the following variation coefficients in the semi-parallel corpus for
(i) morphological complexity scores: Dutch – 0.0056, English – 0.0059, Finnish – 0.0054, French – 0.006, German – 0.0054, Hungarian – 0.0057, Italian – 0.006, Romanian – 0.006, Spanish – 0.0062,
(ii) syntactic complexity scores: Dutch – 0.0048, English – 0.0047, Finnish – 0.0053, French – 0.0051, German – 0.0045, Hungarian – 0.0059, Italian – 0.0049, Romanian – 0.0053, Spanish – 0.0052.

Spanish, French and English. On the whole, these results are as we would expect them to be. In fact, only Dutch and Romanian exhibit surprisingly high overall complexity. Comparing this ranking to the results of the semi-parallel corpus (Figure 5) we find that the Hungarian sample still exhibits the highest overall complexity. Spanish and English are still the least complex languages – even though their order is now reversed, and Spanish is less complex than English. However, the ranking of the other language samples has changed: Italian, German and Dutch as well as French are (in decreasing order) complex whereas Romanian and Finnish have become less complex. The overall correlation of the two complexity hierarchies is moderate but significant (Pearson's $r = 0.59$, $p = 0.05$)[8].

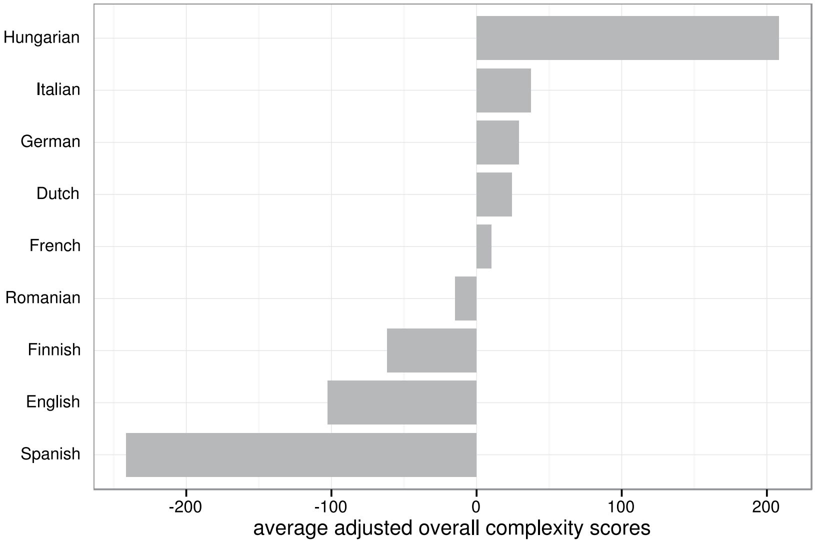

Figure 5: Overall complexity hierarchy in the semi-parallel Alice corpus

Turning to morphological and syntactic complexity, the analysis of the parallel Alice corpus (Figure 6) dovetails with intuitions: Finnish, which according to the Kolmogorov metric is the morphologically most complex but syntactically most simple language in the sample, is located in the extreme top left quadrant whereas morphologically simple but syntactically complex languages (Spanish, French and English) cluster in the bottom right quadrant of the plot. The middle

[8] We use Pearson's correlation coefficient to assess the rankings because it also considers the distance between the points measured in the ranking.

ground is covered by more balanced languages; while Romanian and Hungarian as well as Dutch are more on the morphologically complex side, German and Italian seem to be well balanced. In the semi-parallel corpus (Figure 7) we observe a similar distribution. Finnish and Hungarian are clearly the morphologically most complex and syntactically most simple languages while English, in the bottom right quadrant, is the syntactically most complex but morphologically most simple language. Dutch, Italian, Spanish, Romanian and French cluster in the centre whereas German exhibits medium syntactic complexity and low morphological complexity. Both datasets exhibit a very significant trade-off regarding morphology and syntax: with a negative correlation of Pearson's $r = -0.93$ ($p = 0.0002$) in the parallel and $r = -0.79$ ($p = 0.006$) in the semi-parallel corpus, most languages in our sample trade off morphological for syntactic complexity.

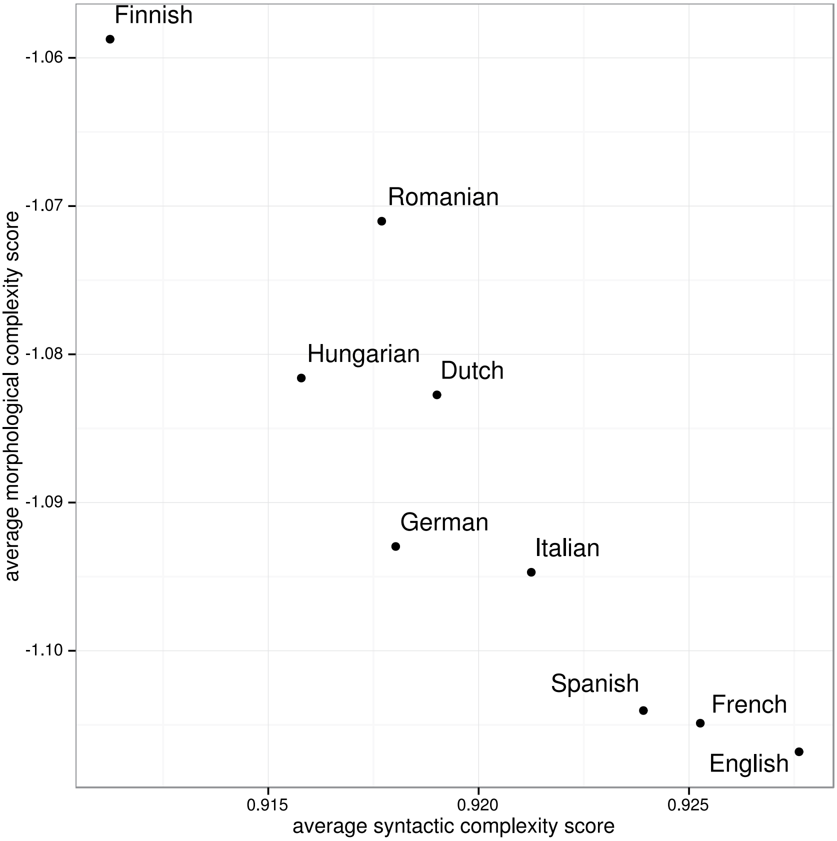

Figure 6: Morphological by syntactic complexity in the parallel Alice corpus

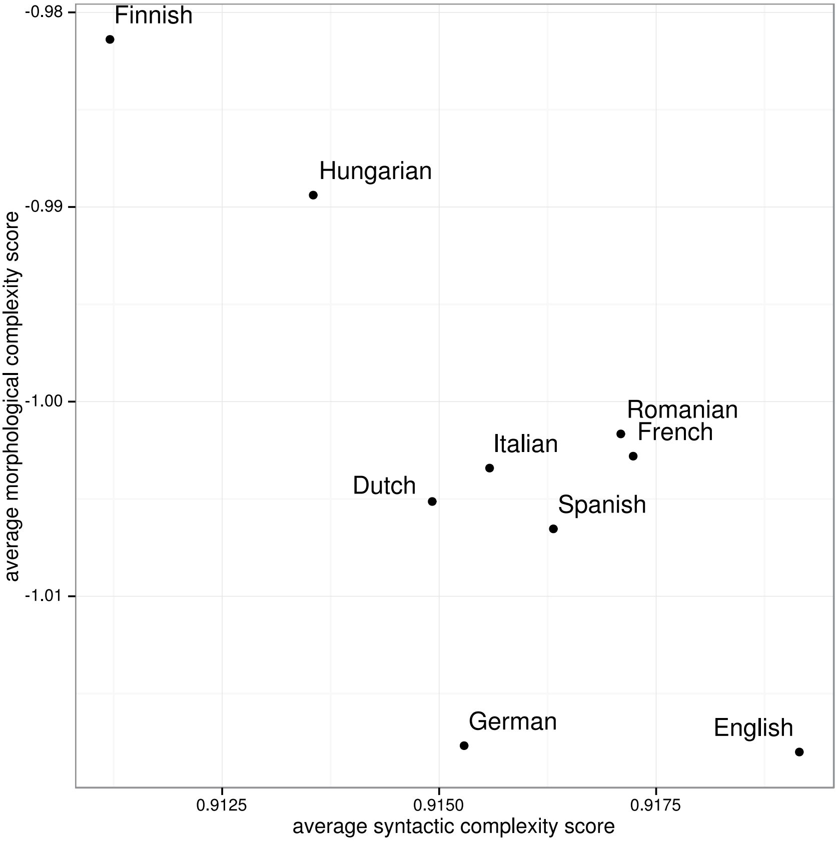

Figure 7: Morphological by syntactic complexity in the semi-parallel Alice corpus

The morphosyntactic complexity analysis of the parallel and semi-parallel Alice corpus would, on the whole, seem to yield similar results. This is another way of saying that the compression technique can be effectively used with semi-parallel texts. In order to back up these findings statistically, we correlate the average syntactic and morphological complexity scores of the two datasets. At the syntactic level the two datasets correlate very highly (Pearson's $r = 0.89$, $p = 0.0006$); at the morphological level the correlation is less strong but still high (Pearson's $r = 0.73$, $p = 0.01$).

In order to see how our findings fare in comparison to more traditional measures, we compare the results obtained via compression to Bakker (1998). Bakker (1998) investigates syntactic complexity on the basis of word-order patterns and their flexibility. He assigns values between 0 and 1 for syntactic flexibility based on twelve word-order variables. Values close to zero indicate less flexibility and thus increased syntactic complexity (Bakker 1998: 387). Table 5 below shows the syntactic complexity rankings of our sample languages in the parallel and non-parallel Alice corpora as well as according to Bakker's (1998) flexibility indices. We compare Bakker's (1998) ranking and the rankings in our datasets by correlating our adjusted complexity scores with Bakker's (1998) flexibility values: the order of syntactic complexity in the parallel Alice corpus very highly correlates (Pearson's $r = 0.74$, $p = 0.02$) with Bakker's (1998) findings. The semi-parallel ranking correlates to a lesser degree but is still significant (Pearson's $r = 0.5$, $p = 0.01$).

Table 5: Syntactic complexity ranking according to Bakker's (1998) flexibility indices and our syntactic complexity scores. Languages are listed in decreasing order of syntactic complexity

			Parallel Alice corpus			Semi-parallel Alice corpus		
rank	**language**	**Bakker's flexibility index**	**rank**	**language**	**syntactic complexity score**	**rank**	**language**	**syntactic complexity score**
1.	French	0.1	1.	English	0.928	1.	English	0.919
2.	Spanish	0.3	2.	French	0.925	2.	French	0.917
3.	Italian	0.3	3.	Spanish	0.924	3.	Romanian	0.917
4.	English	0.4	4.	Italian	0.921	4.	Spanish	0.916
5.	German	0.4	5.	Dutch	0.919	5.	Italian	0.916
6.	Dutch	0.4	6.	German	0.918	6.	German	0.915
7.	Romanian	0.5	7.	Romanian	0.918	7.	Dutch	0.915
8.	Finnish	0.6	8.	Hungarian	0.916	8.	Hungarian	0.914
9.	Hungarian	NA	9.	Finnish	0.911	9.	Finnish	0.911

In sum, the results of the parallel and semi-parallel corpus for overall linguistic complexity are fairly similar, and the complexity rankings we obtain correlate with those reported in previous complexity research. We note minor shifts in the rankings of the parallel and semi-parallel corpus which stem from the loss of alignment of the texts which was achieved by permutation. For some interesting reason – a detailed discussion of which is reserved for another occasion – dislocations are especially pronounced among the languages that are fairly balanced between morphological and syntactic complexity. But in all, the compression technique achieves good results across both corpora. We have thus demonstrated that the use of the compression technique is not in principle limited to parallel corpora, but can successfully be used with semi-parallel corpora.

4.2.2 Newspaper texts

Next, we compile a non-parallel corpus of newspaper[9] texts on several contemporary topics in Dutch, English, Finnish, French, German, Hungarian, Italian, Romanian and Spanish. The topics were chosen according to their availability across the 9 languages. In this paper we analyze articles which were tagged as dealing with the 'Euro crisis' and 'Congo'[10]. For each topic we sample the same number of sentences in order to keep sample size constant across the different language samples. Methodologically, we proceed as described above and calculate adjusted complexity scores as indicator of overall complexity. For the calculation of morphological and syntactic complexity, the news texts are subjected to multiple distortion and compression with N = 1,000 iterations. The average morphological complexity score[11] and the average syntactic complexity score[12] are subsequently calculated as outlined in the previous section.

Figure 8 shows the overall complexity ranking of the non-parallel newspaper texts. The Kolmogorov metric rates Hungarian as the most complex language in the dataset. It is closely followed by Italian and (in decreasing order of complexity) Finnish, German, Romanian, Dutch, Spanish, English and French. Similarly to the semi-parallel Alice corpus, languages which are balanced between morphological and syntactic complexity – such as Italian, German or Dutch – tend to become increasingly complex with increasing lack of 'content control'. Languages on the extreme end of the complexity scale such as Hungarian and English, on the contrary, seem to be hardly affected.

[9] We retrieved articles from the following online newspapers:
Dutch: Volkskrant (http://www.volkskrant.nl/)
English: The Guardian (http://www.guardian.co.uk/)
Finnish: Helsinki Sanomat (http://www.hs.fi/) and Iltasanomat (http://www.iltasanomat.fi)
French: Le Figaro (http://www.lefigaro.fr/)
German: Die Welt (www.welt.de)
Hungarian: HvG (http://hvg.hu/) and Nepszava (http://www.nepszava.hu)
Italian: La repubblica (http://www.repubblica.it/)
Romanian: Adevarul (http://www.adevarul.ro/)
Spanish: ABC (http://www.abc.es)

[10] We also sampled articles on the topics 'Iran', 'Tunisia', 'Kim il Jong' and 'Putin/Russia'. However, these topics yielded less satisfying results. Due to the vast number of articles and the span of languages covered, a manual control of each article's topic was not feasible. For this reason, it is likely that – depending also on the political relations/interests among the respective countries – our sources substantially differ in the topics published under the same topic/tag.

[11] This yields the following variation coefficients for the morphological complexity scores in the non-parallel corpus: Dutch – 0.0017, English – 0.0016, Finnish – 0.0017, French – 0.0016, German – 0.0018, Hungarian – 0.0014, Italian – 0.0014, Romanian – 0.0017, Spanish -0.0016.

[12] This yields the following variation coefficients for the syntactic complexity scores in the non-parallel corpus: Dutch – 0.0022, English – 0.002, Finnish – 0.0021, French – 0.0019, German – 0.0023, Hungarian – 0.0019, Italian – 0.0017, Romanian – 0.0021, Spanish – 0.0021.

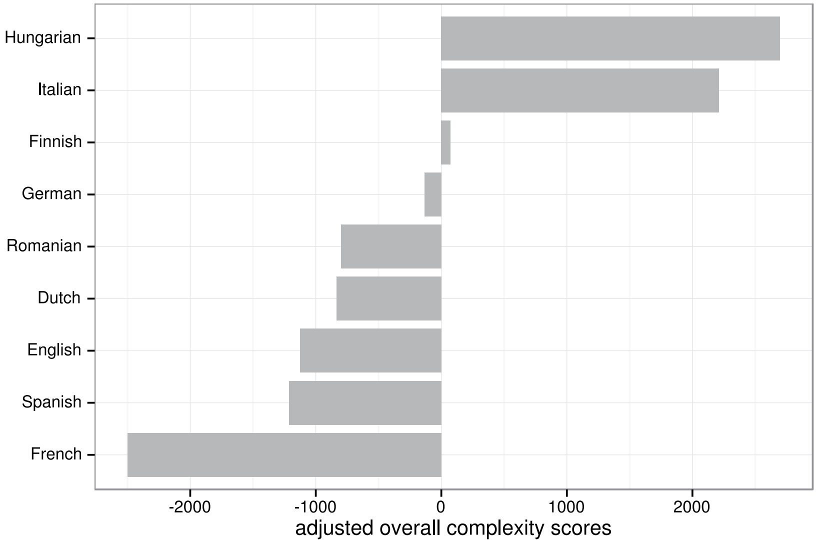

Figure 8: Overall complexity hierarchy in non-parallel newspaper texts

As far as overall complexity scores are concerned, the correlation between the non-parallel news database and the parallel Alice corpus is moderate (Pearson's $r = 0.49$, $p = 0.09$).

The analysis of morphological and syntactic complexity is shown in Figure 9. Morphologically complex but syntactically simple languages are grouped in the top left quadrant: Finnish, Italian, and Hungarian. Dutch, German and Romanian are scattered across the whole middle area of the plot. Syntactically complex but morphologically simple languages, i. e. Spanish, English and French, occupy the bottom right quadrant.

Apart from the Italian data point (which is an as yet inexplicable outlier), these results tie in neatly with our previous findings. In fact, we observe a very high, significant correlation between the parallel and non-parallel complexity scores on the syntactic level (Pearson's $r = 0.84$, $p = 0.002$) and a high, significant correlation on the morphological level (Pearson's $r = 0.63$, $p = 0.03$).

For completeness sake, we compare our findings for syntactic complexity in the non-parallel corpus with Bakker's (1998) flexibility values (Table 6), we still observe a moderate correlation (Pearson's $r = 0.49$, $p = 0.1$) even though the decreased content control in this sample set clearly affects the comparability of our method with other measures.

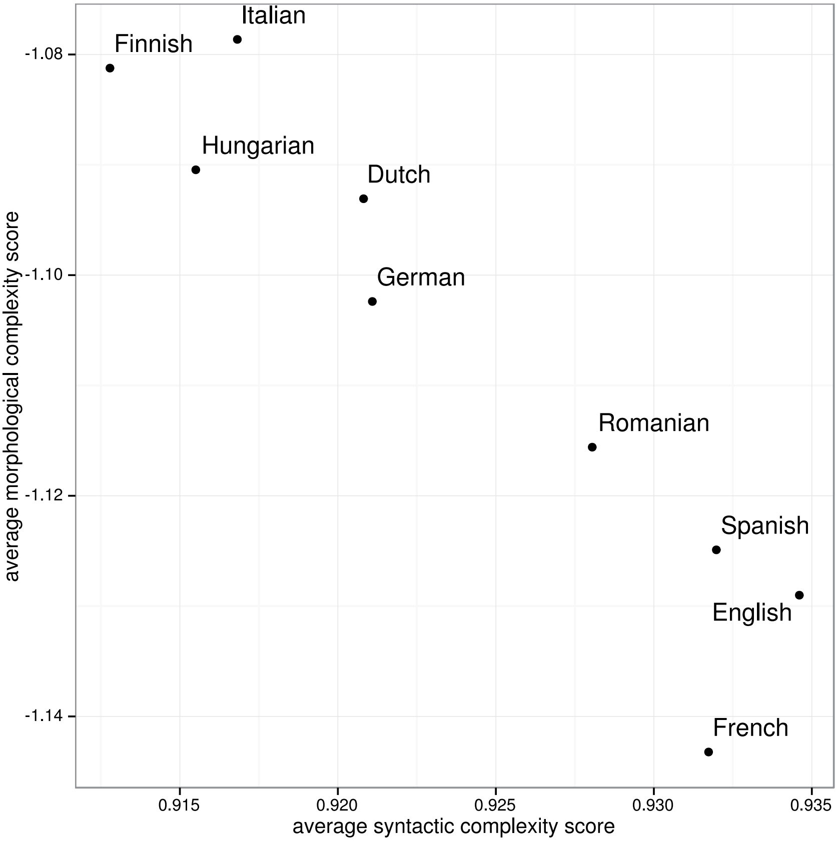

Figure 9: Morphological by syntactic complexity in non-parallel newspaper texts

Table 6: Syntactic complexity ranking according to Bakker's flexibility indices and our syntactic complexity scores in the non-parallel newspaper corpus. Languages are listed in decreasing order of syntactic complexity

rank	language	Bakker's flexibility index	rank	language	syntactic complexity score
1.	French	0.1	1.	English	0.935
2.	Spanish	0.3	2.	Spanish	0.932
3.	Italian	0.3	3.	French	0.932
4.	English	0.4	4.	Romanian	0.928
5.	German	0.4	5.	German	0.921
6.	Dutch	0.4	6.	Dutch	0.920
7.	Romanian	0.5	7.	Italian	0.917
8.	Finnish	0.6	8.	Hungarian	0.916
9.	Hungarian	NA	9.	Finnish	0.912

5 Discussion and outlook

In this chapter, we have explored the reach and limits of information-theoretic methodologies to measure linguistic complexity on the overall, syntactic and morphological level. We set out by measuring complexity in parallel texts using translational equivalents of the Bible in some ten (historical) varieties of English and six European languages. The analysis of both overall and morphosyntactic complexity yields linguistically meaningful results and corroborates findings from previous, more traditional research (e. g. Nichols 1992; Bakker 1998). Focusing on historical English Bible translations, we can trace the development of English from a morphologically complex and syntactically simple to a morphologically simple and syntactically complex language through time. In fact we find a statistically significant, negative correlation between morphological and syntactic complexity in all our datasets. This indicates a trade-off, à la Hockett (1958: 180–181), between morphological and syntactic complexity.

In a second step, we applied the compression technique in a slightly modified version – using multiple instead of simple distortion and compression – to a parallel and re-sampled semi-parallel corpus of *Alice's adventures in Wonderland* in nine European languages. Furthermore, we tested the compression technique on genuinely non-parallel texts using newspaper articles in the same nine languages. The complexity hierarchies obtained from the semi-parallel and non-parallel corpora were compared and correlated to the complexity ranking of the parallel Alice control corpus. In terms of the morphological and syntactic complexity, we achieve high to very high correlations between the parallel, semi-parallel and non-parallel corpora. Overall complexity correlates only moderately. Although control of the topic, if not the precise content, across the different corpus samples seems to be a factor crucial to the successful application of the compression technique, our results show that the compression technique is not limited to parallel texts but can also be successfully applied to semi-parallel and non-parallel texts. In all, using file compression tools to assess linguistic complexity promises an economical and radically objective way to measure linguistic complexity. We also demonstrated that the compression technique reliably works with different text types, i. e. religious, literary and newspaper texts.

The technique has, needless to say, drawbacks. For one thing, the compression technique is completely agnostic about things that are dear to many linguists, such as form-function pairings. The technique, as we use it, is also text based, and so it crucially requires the availability of corpora of text and/or speech. Furthermore, even though we seem to obtain linguistically meaningful results by measuring complexity drawing on off-the-shelf file compression programs, we are not yet able to identify and linguistically interpret the precise patterns and strings which are recognized by these programs to create compression economy. To address this issue, work is under way by the first-named author to develop a custom-programmed compression algorithm which works transparently to aid linguistic interpretation.

In any event, the full potential of the compression technique has not yet been fully explored. Compression algorithms appear to be a useful and quite reliable tool for measuring cross-linguistic complexity variance. Ehret (2014) has also shown how the precise weight of individual grammatical features, such as for instance genitive markers or tense and aspect markers, can be determined via more specific file manipulation.

It has yet to be shown how well the algorithms pick up intra-varietal – (for example, dialectal) complexity variance – say, between more and less isolated dialects of the same language.

References

Bakker, Dik 1998: Flexibility and consistency in word order patterns in the languages of Europe. In: Anna Siewierska (ed.), *Constituent Order in the Languages of Europe*, 384–419. Berlin/ New York: Mouton de Gruyter.

Bane, Max 2008: Quantifying and measuring morphological complexity. *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 67–76.

Bickerton, Derek 1995: *Language and Human Behaviour*. Seattle: University of Washington Press.

Crystal, Bill 1987: *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Cysouw, Michael and Wälchli, Bernhard 2007: Parallel texts: Using translational equivalents in linguistic typology. *Language Typology and Universals* 60 (2): 95–99.

Dahl, Östen 2004: *The Growth and Maintenance of Linguistic Complexity*. Amsterdam/ Philadelphia: Benjamins.

Dahl, Östen 2007: From questionnaires to parallel corpora in typology. *Language Typology and Universals* 60 (2): 172–181.

Edwards, John 1994: *Multilingualism*. London: Penguin.

Ehret, Katharina in preparation: An information-theoretic approach to language complexity: variation in naturalistic corpora. PhD dissertation, University of Freiburg.

Ehret, Katharina 2014: Kolmogorov complexity of morphs and constructions in English. *Language Issues in Linguistic Technology* 11 (2): 43–71.

Gries, Stefan Th. 2009: *Quantitative Corpus Linguistics With R: A Practical Introduction*. New York/London: Routledge.

Hockett, Charles Francis 1958: *A Course in Modern Linguistics*. New York: Macmillan.

Juola, Patrick 1998: Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5 (3): 206–213.

Juola, Patrick 2008: Assessing linguistic complexity. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*. Amsterdam/ Philadelphia: Benjamins.

Kettunen, Kimmo, Sadeniemi, Markus, Lindh-Knuutila, Tiina and Honkela, Timo 2006: Analysis of EU Languages through Text Compression. In: Tapio Salakoski, Filip Ginter, Sampo Pyysalo and Tapio Pahikkala (eds.), *Advances in Natural Language Processing*, Lecture Notes in Artificial Intelligence, 99–109. Heidelberg: Springer.

Kortmann, Bernd and Szmrecsanyi, Benedikt (eds.) 2012: *Linguistic Complexity in Interlanguage Varieties, L2 Varieties, and Contact Languages*. Berlin/Boston: Walter de Gruyter.

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

Kusters, Wouter 2008: Complexity in linguistic theory, language learning and language change. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*, 3–21. Amsterdam/Philadelphia: Benjamins.

Li, Ming, Chen, Xin, Li, Xin, Ma, Bin and Vitányi, Paul M. B. 2004: The similarity metric. *IEEE Transactions on Information Theory* 50 (12): 3250–3264.

van der Lubbe, Jan C. A. 1997: *Information Theory*. Cambridge [England] /New York: Cambridge University Press.

McWhorter, John 2001: The world's simplest grammars are creole grammars. *Linguistic Typology* 6: 125–166.

Miestamo, Matti, Sinnemäki, Kaius and Karlsson, Fred (eds.) 2008: *Language Complexity: Typology, Contact, Change*. Amsterdam/Philadelphia: Benjamins.

Moscoso del Prado Martin, Fermin, Kostic, Aleksandar and Baayen, R. Harald : 2004: Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94 (1): 1–18.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago/London: University of Chicago Press.

O'Grady, William, Dobrovolsky, Michael and Aronoff, Mark 1997: *Contemporary Linguistics: An Introduction*, 3rd ed. New York: St. Martin's Press.

Sadeniemi, Markus, Kettunen, Kimmo, Lindh-Knuutila, Tiina and Honkela, Timo 2008: Complexity of European Union Languages: A Comparative Approach. *Journal of Quantitative Linguistics* 15 (2): 185–211.

Sampson, Geoffrey, Gil, David and Trudgill, Peter (eds.) 2009: *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Shannon, Claude E 1948: A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423.

Szmrecsanyi, Benedikt and Kortmann, Bernd 2009: Between simplification and complexification: Non-standard varieties of English around the world. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 64–79. Oxford: Oxford University Press.

Trudgill, Peter 2001: Contact and simplification: Historical baggage and directionality in linguistic change. *Linguistic Typology* 5 (2/3): 371–374.

Ziv, Jacob and Lempel, Abraham 1977: A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* IT-23 (3): 337–343.

Jacopo Garzonio, University of Venice

On complexity of interrogative syntax in Northern Italian dialects*

Abstract: In this article I discuss the relative syntactic complexity of three different Italo-Romance dialects spoken in the Alpine area. The main focus of the analysis is the syntactic encoding of interrogative force, whose complexity is measured by means of two separate parameters: the derivational weight of both *yes/no* and *wh* questions and the number of synonymous freely interchangeable constructions. Following the current way of proceeding in micro-variation analysis, the systems of the isolated dialects are compared with that generally found in other Northern Italian dialects. The comparison shows that isolation can produce different types of complexity. In some cases syntactic complexity derives from the distributional properties of functional elements (such as *wh* items), while in others it corresponds to more articulated derivational processes and a distinctive encoding of clause sub-types.

1 Introduction

1.1 Goals of this work

In this article I examine the syntax of root interrogative sentences in three different dialects of the northern Italo-Romance domain: Monnese, spoken in Monno (province of Brescia, in northeastern Lombardy); Bellunese, spoken in Belluno (in northern Veneto); and Mendrisiotto, spoken in Mendrisio (in the Ticino canton in Switzerland). All three dialects are from the Alpine area. More specifically, I will show that these varieties clearly have a different grammar compared with related non-Alpine Northern Italian dialects, but these differences should not in all three cases be considered evidence of a more complex syntax. The notion of language complexity is a problematic one, and even more so when discussed in relation to the syntactic component. As I argue in the next subsection, one of the major issues in this field is the individuation of an appropriate

* I wish to thank the audience of the Complexity and Isolation workshop (Freiburg, 8–9 December 2011), Nicola Munaro and Hans-Georg Obenauer for the discussion about the topic of this work. I also thank Cecilia Poletto and Silvia Rossi and two anonymous reviewers, who provided many useful comments on a preliminary version of the paper. All errors are my own.

way of measuring syntactic complexity. In what follows, I develop an analysis of the interrogative constructions of the above-mentioned dialects following some basic assumptions of the current Minimalist and Cartographic approaches to syntactic structures. Hence, the parameters for complexity I propose here have a theoretical grounding. Of course, the complexity which I will be discussing is local, as it is limited to a single clause type, so I will not deal with the complexity of the whole syntactic component or the global complexity of these grammars. However, the proposed parameters can be adopted for clause typing in general, which is one of the main functions of morpho-syntax.

The phenomena I discuss have been discovered and analyzed in some recent works on the syntax of Northern Italian dialects (mainly Munaro 1999; Benincà and Poletto 2004; Poletto and Pollock 2005, 2009). The first phenomenon is a type of *do*-support construction, similar to the well-known *do*-support of English, found in interrogative clauses in Monnese:

(1) a. ***fa**-l majà?*
does-he eat.INF
'Does he eat?'
b. **maja-l*?
eats-he
c. *ke **fa**-l majà*?
what does he eat
'What does he eat?'
d. **ke maja-l*?
what eats he (Monnese; Benincà and Poletto 2004: 52)

The second phenomenon I discuss is *wh* doubling, in which the interrogated variable in *wh* questions is represented by two separate *wh* words in the clause. Both Monnese and Mendrisiotto display this phenomenon (but with relevant differences that I will take into account):

(2) a. ***ngo** fe-t majà **ngont***?
where do-you eat.INF where
'Where do you eat?' (Monnese; Poletto and Pollock 2005: 136)
b. ***sa** ta mangiat **cusè***?
what you eat.2SG what?
'What are you eating?' (Mendrisiotto; Poletto and Pollock 2009: 201)

The third phenomenon is apparent optional *wh* movement in Bellunese, which allows both of the orders in (3):

(3) a. *sié-o 'ndadi* ***andé***?
are-you gone where
'Where have you gone?' (Bellunese; Obenauer 2004: 346)
b. ***andé*** *sié-o 'ndadi*?
where are-you gone
'Where have you gone?'

The article is structured as follows: In the next subsections I outline the approach I adopt for the comparison of syntactic complexity across the three varieties under examination; I also provide some information on the extra-linguistic factors of the three varieties under examination. In sections 2–4 I describe the three systems and compare them with those of non-Alpine dialects. In section 5 I discuss the data and the issue of syntactic complexity, while section 6 contains some concluding remarks.

1.2 Syntax and language complexity

In the "absolute approach" to linguistic complexity the basic assumption is that "the more parts a system has, the more complex it is" (Miestamo 2008: 24). While it is relatively straightforward to apply this idea in phonological or morphological analyses, absolute syntactic complexity is a problematic concept: First, in many cases it is not clear which constructions can be considered members of a paradigm or a natural set (it is possible to count the number of phonemes or phonological features in a language, or the number of plural formations, but it can be a problem to count the number of topic or focus structures, or determine the number and derivational relation of word orders); second, syntax very often displays optionality patterns whose complexity factors do not seem to have a clear status.

Further complications arise also from the fact that, even more than in phonology and morphology, syntactic complexity does not appear to be a theoretically neutral concept, and how complex a construction is seems to depend strongly on the analysis that is adopted. In more general terms, it seems that measuring syntactic complexity requires at least some level of theoretical formalization. This aspect is a potential problem, since the complexity of a linguistic system – especially in recent approaches based on information theory – is considered an objective, mathematically calculable property. Even if we accept a theoretically non-neutral approach to syntactic complexity, it can be hard to determine the relative complexity of different structures or processes. For instance, in a minimalist framework, is a structure involving the operation Move more complex than one involving Agree? Or, in a parametric analysis, does one setting of a certain parameter give rise to more complexity than the opposite setting?

Finally, it is not completely clear to what extent syntactic complexity contributes to the so-called global complexity of a language, since syntax is linked to both morphological encodings (which is a different domain of complexity) and lexical properties, and the measurement of complexity in the lexicon is usually kept separate from the measurement of grammatical complexity (corresponding to what physicist Gell-Mann 1994 calls "effective complexity", which takes into account patterns and regularities of a system).

Given these premises it comes as no surprise that syntacticians usually do not deal with the notion of absolute syntactic complexity. Nevertheless, an intuitive idea of complexity is very often referred to. For all the reasons given in this section, I will not take into consideration the notion of absolute syntactic complexity, but will nonetheless show how an analysis of interrogative patterns can shed light on the link between complexity and isolation in a context of micro-variation.

1.3 Micro-variation and Merge-over-Move: methodological remarks

In this paper I compare the grammars of three Northern Italian dialects. The main advantage of taking into consideration similar and closely related languages (the so-called "micro-variation analysis") is that they have minimally diverging grammars, so that it is possible to focus on a single grammatical property while the rest of the system remains stable, or at least varies in a very slight and/or predictable way. Thus, the study of "micro-varying" languages is very suitable for the measurement of local complexity. It should also be pointed out that I analyze three dialects displaying unique differences in comparison with the majority of the other Northern Italian dialects in the domain of interrogative syntax. This gives us the possibility to compare these rarer (and more isolated) systems with a more general (and widespread) system, which functions as a sort of landmark for comparison.

The theoretical framework I adopt here is the Cartographic approach to clause structure (see among others Rizzi 1997 and Cinque 1999), according to which it is possible to find a determined hierarchical order of functional projections in every syntactic domain or layer (for instance in the clause, usually subdivided into three layers: a complementation layer or Complementizer Phrase, an inflectional layer or Inflectional Phrase, and a lexical layer or Verb Phrase; each of these being formed by a larger set of projections). Each projection is associated with a specific grammatical feature. The number of projections and the relation between the proposed hierarchies and Universal Grammar are under debate. As I will explain in the following sections, the structure of the left periphery of the clause, i. e. the Complementizer Phrase (CP), is our main concern here, since clause type features are encoded in that layer. In recent developments of generative grammar (Minimalism and Phase Theory; see Chomsky 1995, 2001), it is argued that there are two basic operations in the building of syntactic structures: Merge and Move. Merge combines two objects in a phrase (whose properties are identified by its label). Usually, it is assumed that Merge corresponds to the insertion of a lexical object in the derivation (where it forms a phrase with its complement, which is already built up). Move (which has been defined in different ways in the literature) corresponds to the "Move α" operation of the classic Government and Binding Theory: an element is moved from its basic position – the position where it is "merged" – to a different, hierarchically higher position, in order to check or activate some features. The adoption of the Cartographic approach involves the idea that the relative order of Merge operations in a determined sub-layer of the clause structure is fixed.

In the following sections I analyze the structure of questions in these dialects according to the theoretical principles described here. The different structures are the result of different sequences of Merge and Move operations. My proposal is to consider these sequences indicators of the complexity of the syntactic component of the varieties under examination. This idea is based on one of the derivational economy conditions connected to the minimalist principles, the "Merge over Move" preference (MOM). Move is assumed to be a costly operation, as it involves different copies of an element in the structure. Thus Merge is to be preferred in those derivational steps where both operations are possible.[1] Notice that in some cases we will deal with two or more optional constructions. The first task is to determine whether we are observing a case of true optionality or whether some descriptive refinement is required. If the constructions under analysis really are optionally interchangeable, my conclusion is that, at least in the domain of interrogative syntax, the grammar with more options is more complex.

To summarize, two parameters will be taken into account to determine which grammar is more complex in the domain of interrogative syntax: the number of Move operations in comparison to the number of Merge operations, and the number of optionally interchangeable constructions.

[1] In the Minimalism framework, MOM has been proposed as an economy condition that can shed light on the interaction of some problematic but observed derivations and general principles of the theory. It must be pointed out that its introduction has been criticized (for instance, Castillo, Drury, and Grohmann 1999 argue that it is possible to dispense with both MOM and the Extended Projection Principle). My proposal, however, is not based on the MOM as a whole condition, but on one of its formative assumptions, namely that Merge is derivationally cheaper than Move.

1.4 On the syntax of questions in Northern Italian Dialects

In this section I outline the general properties of root questions in Northern Italian dialects and propose a syntactic derivation which will be used as a term of comparison in the analysis of interrogative syntax in the three varieties under examination. As in Italian and dialects of Central and Southern Italy, questions are also characterized by specific intonation patterns (see Grice et al. 2004 among many others). However, I will leave this aspect aside, since the relations and interactions between syntax and prosody in this domain are not yet well understood. In general, root questions are characterized by the presence of subject enclitics, while in corresponding declaratives subject proclitics are found. This phenomenon is usually considered an occurrence of verb-subject inversion, but it must be pointed out that non-clitic subjects are usually non-inverted[2]. In (4) and (5) I provide relevant examples from Paduan.

(4) a. *el vien.*
he comes
'He comes.'
b. *vien-lo*?
comes-he
'Is he coming?'
c. **el vien*?
he comes (Paduan; Poletto 1993: 100–101)

(5) a. *quando ze-lo vegnuo*?
when is he come
'When has he come?'
b. **quando el ze vegnuo*?
when he is come
c. **quando ze Nane vegnuo*?
when is John come (Paduan; Poletto 1993: 100–102)

The sentence in (4a) is a declarative, (4b) is a grammatical yes/no question displaying inversion, (4c) is an ungrammatical yes/no question without inversion.

[2] The exact nature of subject enclitics in Northern Italian dialects (and why they can differ from corresponding proclitics) has been thoroughly discussed in recent years (see Poletto 2000; Manzini and Savoia 2005; Cardinaletti and Repetti 2008). Here, I adopt the generally accepted view that these enclitics are the manifestation of verb movement to a higher position.

Similarly, (5a) is a grammatical *wh* question with inversion, while (5b) is its ungrammatical counterpart without inversion; (5c) shows that only clitics are found in verb-subject inversion constructions. Normally, if a non-clitic subject is present in a question, it is either topicalized (thus it precedes *wh* elements), or it occurs at the end of the clause (a phenomenon usually called “emargination”; see Antinucci and Cinque 1977). However, verb-subject inversion is not attested in all parts of the northern Italo-Romance domain, since in many Ligurian, Emilian, and Lombard varieties, for instance, a root yes/no question normally has the same word order as the corresponding declarative clause. Here I will take into consideration only those varieties such as Paduan in which all root questions have a distinctive overt syntax. The examples in (5) also show that, as in Standard Italian, these dialects have *wh* movement to the left periphery.

In order to account for these two phenomena – namely verb-subject inversion and *wh* movement – I propose a general derivation of root questions as represented in (6) and (7). The IP is merged with a *wh* projection (see Rizzi 1996, 2001), dedicated to the *wh* operator; subsequently the inflected verb and the *wh* item are moved to this projection, to its head and specifier respectively (as represented in (7))[3].

(6) Input: [$_{\text{IP}}$ el vien]
(a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ el vien]]
(b) Move the verb to *wh* and switch the enclitic form →
[$_{wh}$ vien [$_{\text{IP}}$ lo ~~vien~~]]
Merge+Move

(7) Input: [$_{\text{IP}}$ el ze vegnuo [quando]]
(a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ el ze vegnuo [quando]]
(b) Move the auxiliary to *wh* and switch the enclitic form →
[$_{wh}$ ze [$_{\text{IP}}$ lo ~~ze~~ vegnuo [quando]]
(c) Move the *wh* item to the Specifier of *wh* →
[$_{\text{Spec}\,wh}$ [quando] $_{wh}$ ze [$_{\text{IP}}$ lo ~~ze~~ vegnuo ~~[quando]~~]
Merge+Move+Move

[3] I leave aside some complexities related to the form of the clitic and possible instances of Remnant Movement, since they are not relevant for the comparison with the other varieties in the next sections (see Poletto 2000).

According to this analysis, the derivation of root questions involves at least one Merge and one Move operation (two Move operations in *wh* questions) in Northern Italian dialects. With regard to optional constructions, there is much variation among these dialects. However, very often indeed these are not cases of real optionality. For instance, Paduan also displays questions without verb-subject inversion, or with *wh* items *in situ*, or even cases of complementizers in root questions (without verb-subject inversion), but all these constructions are not available in standard questions. In other words, they have additional meanings and interpretative nuances.

1.5 Empirical domain of the investigation

The three varieties under analysis are spoken in the Alpine area. As argued by Pellegrini (1992), the Italo-Romance Alpine area can be considered systematically isolated with respect to the domain of Northern Italian dialects: while there can be cases of contact inside the area, there is strong evidence that external influence is very limited. Monno is a small village in the Camonica valley, in northeast Lombardy. Until 1963 the village was reachable only by a footpath and contacts with people speaking other varieties were rare. Benincà and Poletto (2004) argue that its unique properties are a consequence of this strong isolation.

Bellunese is a Northern Veneto variety, spoken in the main town of the Belluno area, roughly 100 km north of Venice. The dialects of this area share some general properties (and differences with respect to the other dialects of Veneto), but are not very homogeneous. This fragmentation reflects their distribution across different valleys. Munaro (1999) argues that the particuliar *wh* system he describes is found systematically in the Pagotto variety and in Bellunese, but not throughout the whole Northern Veneto domain.

Mendrisio is located in the southernmost part of the Ticino canton in Switzerland. Like other dialects of southern Ticino, Mendrisiotto has been influenced by the Como and Varese dialects in Lombardy. However, it is much more vital than the latter ones.

2 Questions in the variety of Monno

Monnese diverges in a significant way from the “standard” model of Northern Italian dialects presented in section 1.4. First of all, it displays a type of *do*-support similar to the one found in English; second, it has optional *wh* doubling, with more than one *wh* item corresponding to the same variable in the same clause.

The following sentences exemplify the *do*-support construction (see also examples in (1)). In both yes/no and *wh* questions, the subject enclitic is found on an inflected form of the verb *fa* corresponding to 'do', while the lexical verb (bearing infinitive morphology) is left in final or near-final clause position.

(8) a. *fa-l plöer*?
does-it rain.INF
'Is it raining?'
b. *kwata fa-l majà-n*?
how.much does-he eat.INF-of.it
'How much does he eat?' (Monnese; Benincà and Poletto 2004: 52, 67)

As Benincà and Poletto (2004) show, this construction is very similar to the English one: It does not occur in embedded contexts (9a), it cannot apply to auxiliary and lexical 'be' and 'have' (9b–c), it is used even with lexical 'do' (9d) and, finally, it is not used if the *wh* element corresponds to the subject (9e):

(9) a. **i t domandjo ke fe-t majà*
I you ask what do.2SG-you eat.INF
'I ask you what you eat'
b. *ngo è-l na*?
where is-he gone
'Where did he go?'
c. **ngo fa-l esse na*?
where does-he be.INF gone
d. *ke fa-l fa*?
what does-he do.INF
'What does he do?'
e. **ki fa(-l) majà*?
who does(-he) eat.INF
'Who is eating?' (Monnese; Benincà and Poletto 2004: 68–69)

However, unlike English, Monnese does not have *do*-support in negative contexts and lacks both emphatic and pro-sentence *do*, as the following examples show:

(10) a. *l so mìa*.
it know.1SG NEG
'I do not know it.'
b. **fo mìa save-l*.
do.1SG NEG know.INF-it

c. **ma ti te FET kantà bè*!
but you you do.2SG sing.INF well
'But you DO sing well!'

d. **ankö l Mario l maja a l'osteria e an l Carlo l fa*
today the Mario he eats at the restaurant and also the Carlo he does
'Today Mario eats at the restaurant and so does Carlo.'
(Monnese; Benincà and Poletto 2004: 70–71)

According to Benincà and Poletto's (2004) analysis, the differences between Monnese and English derive from the fact that in Monnese the verb raises to a position inside the IP layer and higher than the position targeted by *do*-support in English (where it remains lower than the position where negation and emphatic assertion are encoded). In both languages, however, the verb cannot move to the left periphery of the clause to the position we labeled *wh* in derivations (6) and (7). The derivation I propose for *do*-support in Monnese is represented in (11). Benincà and Poletto (2004) argue that the dummy verb 'do' is not inserted directly in the left periphery, but in the inflectional layer before the merging of *wh* and IP. In (11) the lexical verb remains inside VP, since further movements are not relevant here.

(11) Input: [IP 'l fa [VP majà]]
(a) Merge *wh* and IP →
[*wh* [IP 'l fa [VP majà]]]
(b) Move *fa* 'does' to *wh* →
[*wh* fa [IP l ~~fa~~ [VP majà]]]
Merge+Move

Thus, from a derivational point of view, *do*-support is no more complex than verb-subject inversion. It could be argued that the insertion of 'do' is a further Merge operation. However, *do*-support is not directly involved in the clause typing process, but derives from the fact that the lexical verb is not sufficiently high to be moved to *wh*. In other words, interrogative syntax is no more complex in a question with *do*-support than it is in a question with an auxiliary verb. Notice that this analysis strongly implies that the derivation is not sent to spell-out before the activation of the left periphery. My preliminary conclusion is that from this point of view Monnese is no more complex than the other Northern Italian dialects with verb-subject inversion, but this picture changes as soon as we consider its *wh* system. A *wh* question in Monnese has at least three different forms, all freely interchangeable:

(12) a. *ngo fe-t majà ngont*?
where do.2GS-you eat.INF where
b. *ngo fe-t majà*?
where do.2SG-you eat.INF
c. *fe-t majà ngont*?
do.2SG-you eat.INF where
'Where do you eat?' (Monnese; Poletto and Pollock 2005: 136)

This phenomenon, usually referred to as "*wh* doubling", has the following properties: The variety has more than one form for most of its *wh* elements (usually the one corresponding to 'why' is excluded); of these forms, one has the status of a clitic, the other (following Poletto and Pollock 2005 I will call it the "strong" form) is morphologically richer; each form has a specific linear position in the clause. The clitic is found in the left periphery, the strong one stays in situ. The two forms can be used alone or in combination, each in its position. This last property is exemplified by the following ungrammatical sentences:

(13) a. **ngont fe-t andà ngo*?
where do.2SG-you go.INF where
b. **ngont fe-t andà*?
where do.2SG-you go.INF
c. **fe-t andà ngo*?
do.2SG-you go.INF where
'Where are you going?' (Monnese; Poletto and Pollock 2005: 137)

According to the descriptions available at the moment, the sentences in (12) do not differ in their interpretation, and all of them can be used as standard requests for information. These three constructions are the optional forms that a *wh* question can have in Monnese, a strong difference from the general pattern displayed by Northern Italian dialects, which do not allow for *wh* doubling and *wh in situ* in standard questions. Here I adopt an analysis similar to the one proposed by Benincà and Poletto (2004), which is simpler than the one proposed by Poletto and Pollock (2005). In the latter it is postulated that remnant movement occurs and that there are two separate positions for *wh* items in the left periphery. I also leave aside the problem of the exact position of the clitic *wh* item in the left periphery, simply assuming that it is in the standard position of interrogative *wh* items. The derivation is represented in (14):

(14) Input: [$_{\text{IP}}$ t fe [$_{\text{VP}}$ andà [ngo ngont]]]
a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ t fe [$_{\text{VP}}$ andà [ngo ngont]]]]
b) Move *fe* to *wh* →
[$_{wh}$ fe [$_{\text{IP}}$ t ~~fe~~ [$_{\text{VP}}$ andà [ngo ngont]]]]
c) Move the clitic *wh* item to the Specifier of *wh* →
[$_{wh}$ [ngo] fe [$_{\text{IP}}$ t ~~fe~~ [$_{\text{VP}}$ andà [~~ngo~~ ngont]]]]
Merge+Move+Move

The two *wh* elements are generated together as a single phrase (an idea going back to Kayne's 1991 analysis of clitic doubling in French), and only the clitic is moved to the left periphery, while the strong form remains in situ. The two variants are derived through the deletion of one of the two *wh* items. Since the strong form does not move, it can be assumed either that the clitic form is cancelled after the movement to the left periphery, or that its position is occupied by a silent operator. In any case, Monnese has more options in this syntactic domain and is thus more complex than other dialects. Notice that this complexity is caused only by the presence of these interchangeable variables, since the derivation of *wh* questions in Monnese has the same sequence of Merge and Move operations as it has in the other dialects. I will come back to this issue in section 5.

3 The syntax of *wh* interrogatives in Bellunese

For this work I take into consideration the variety of Belluno as described by Munaro (1999) and subsequent works. The syntactic pattern described here is shared by a group of dialects from the Northern Veneto area (in the Eastern Alps), and we can assume Bellunese to be a typical case, since the whole area can be considered relatively isolated. Like other Northern Italian dialects, and all other Veneto dialects, it displays interrogative verb-subject inversion:

(15) a. *i riva sempre in ritardo.*
they.M arrives always in delay
'They always arrive late.'
b. *magne-li la menestra, i boce*?
eats-they.M the soup the boys
'The boys, do they eat the soup?'
(Bellunese; ASIt database, variety of Civoi)

(15b) is a root yes/no question and the verb precedes the clitic subject *li* (corresponding to the proclitic form *i* in (15a)). Observing these data we can conclude that in the case of yes/no questions this dialect is no more complex than others and that these sentences have a derivational process as proposed in (6): yes/no questions are derived by a Merge+Move sequence.

The *wh* system displays a more articulated pattern. In the case of non-bare *wh* items (and the item *parké* 'why'), questions are formed as in other Northern Italian dialects: by both verb-subject inversion and *wh* movement to the left periphery:

(16) *kwanti libri a-tu comprà*?
how-many books have.2SG you bought
'How many books have you bought?' (Bellunese; Munaro 1999: 64)

However, with bare *wh* items, we observe pairs of sentences as the one exemplified in (17):

(17) a. *sié-o stadi andé*?
are-you been.PL.M where
'Where have you been?'
b. *andé sié-o stadi*?
where are-you been.PL.M (Bellunese; Obenauer 2006: 249)

These examples show that in Bellunese both variants of a *wh* question (with or without *wh* movement) are grammatical. However, much work on this optionality phenomenon (mainly Munaro 2003 and Obenauer 2004) has shown that this is not a case of real optionality, and thus must be analyzed differently from the Monnese system of *wh* syntax. The two sentences in (17) are not exactly synonymous and cannot be freely interchanged: (17a) is a standard question and is used as the canonical form of a request for unknown information; (17b), on the other hand, has a special interpretation. As argued by Obenauer (2004), it is a special question and can receive at least three specific semantic nuances: It can be used to express the speaker's surprise and disapproval towards the place the interlocutors have gone (this Surprise-Disapproval interpretation is similar to that of exclamatory sentences); it can be used as a rhetorical question (corresponding to something like "you haven't been anywhere"); or it can be used with a unique interpretation labeled by Obenauer as "Can't find the value", expressing the fact that the speaker cannot imagine a possible value for the interrogated variable.

For the comparison I adopt the basic derivation for standard *wh* questions in Bellunese in (18), where the *wh* item remains in its low position and the *wh* projection contains a silent operator:

(18) Input: [$_{\text{IP}}$ te va [andé]]
a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ te va [andé]]]
b) Move the inflected verb to *wh* and switch the clitic form →
[$_{wh}$ va [$_{\text{IP}}$ tu ~~va~~ [andé]]] 'Where are you going?'
Merge+Move

Special questions involve the activation of further features. In (19) I give the derivation of a special *wh* question in Bellunese, assuming that its special interpretation is correlated with the activation of some speaker "point of view" feature encoded in the high left periphery (see Giorgi 2010) in a projection I label "Disc(ourse)" following Benincà (2001).

(19) Input: [$_{\text{IP}}$ te va [andé]]
a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ te va [andé]]]
b) Move the inflected verb to *wh* and switch the clitic form →
[$_{wh}$ va [$_{\text{IP}}$ tu ~~va~~ [andé]]]
c) Merge DiscP and *wh* →
[$_{\text{DiscP}}$ [$_{wh}$ va [$_{\text{IP}}$ tu ~~va~~ [andé]]]]
d) Move the *wh* item to the specifier of DiscP →
[$_{\text{DiscP}}$ [andé] [$_{wh}$ va [$_{\text{IP}}$ tu ~~va~~ ~~[andé]~~]]] 'Where are you going?!'
Merge+Move+Merge+Move

Given these derivations, it is difficult to determine the status of Bellunese in the complexity factor: On the one hand, it has canonical verb-subject inversion; on the other hand its *wh* questions display some degree of complexity. Bellunese questions with non-bare *wh* items have the same degree of complexity compared with other dialects, but questions with bare *wh* items are less complex since they are derived by a Merge+Move sequence; however, Bellunese encodes some *wh* special questions as (17b) by means of a distinctive syntax, which makes their derivation more complex (Merge+Move+Merge+Move). This problem will be discussed in section 5, where I will add to the comparison some dialects that encode special questions in a different way.

4 *Wh* doubling in the variety of Mendrisio

The dialect spoken in Mendrisio ("Mendrisiotto") has a very unique interrogative syntax. Like many other dialects of the Italo-Romance area in Switzerland it lacks verb-subject inversion in root interrogatives, where subject clitics precede the verb as in (20):

(20) *ta vet a Milan*?
you go.2SG to Milan
'Are you going to Milan?'
(Mendrisiotto; modified from Poletto and Pollock 2009: 204)

In light of the analysis adopted here, the syntax of yes/no question in Mendrisiotto is simpler than that of dialects with inversion, since there is no difference from declaratives in the linear order of clause elements (of course there is a distinctive intonation). The derivation of a sentence such as (20) can be limited to a single Merge operation between the *wh* projection and the IP, postulating a silent interrogative operator in *wh*.

The syntax of *wh* questions in Mendrisiotto is more problematic. First, it has a doubling phenomenon similar to the one described above for Monnese. Furthermore, most bare *wh* items have two and sometimes three different forms: a clitic one, a strong one, and a "weak" one (so labeled by Poletto and Pollock 2009). The following examples show the distribution of the forms of the *wh* item for 'what'; 'how' has a similar distribution, while the other bare *wh* items lack the clitic form. The clitic form *sa* must appear in the left periphery and can be doubled by the strong form *cusè*, which stays in situ (21a–b). Optionally the strong form can be used alone in the left periphery (21c). If it is used alone and stays in situ, a special Surprise-Disapproval interpretation arises (21d). The weak form *cusa* also must be located in the left periphery and can be optionally doubled by the strong form (22a–b). The clitic and the weak forms cannot be used in doubling configurations (22c).

(21) a. *sa ta mangiat cusè*?
what you eat.2SG what
b. *sa ta mangiat*?
what you eat.2SG
c. *cusè ta mangiat*?
what you eat?
'What are you eating?'
d. *ta mangiat cusé*?
you eat.2SG what
'What are you eating?!' (Surprise-Disapproval interpretation)
(Mendrisiotto; Poletto and Pollock 2009: 203–204)

(22) a. *cusa ta fet cusè*?
what you do.2SG what
b. *cusa ta fet*?
what you do.2SG
'What are you doing?'
c. **sa ta fet cusa*?
what you.2sg do what (Mendrisiotto; Poletto and Pollock 2009: 209, 212)

Leaving the special question (21d) aside for the moment, Mendrisiotto has at least five different optional forms for a given *wh* question with 'what' (or 'how'), and three forms for questions with the other *wh* elements. The derivation I propose for these forms, simplifying some of Poletto and Pollock's considerations, is represented in (23). The *wh* words are generated as a single phrase, as in Monnese; the clitic or the weak form moves to the left periphery; the non-doubling cases are derived by the deletion of the strong form.

(23) Input: [$_{\text{IP}}$ ta mangiat [sa/cusa cusè]]
a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ ta mangiat [sa/cusa cusè]]]
b) Move the clitic or the weak *wh* form to the specifier of *wh* →
[$_{wh}$ [sa/cusa] [$_{\text{IP}}$ ta mangiat [~~sa/cusa~~ cusè]]]
Merge+Move

Alternatively, we can have a case like (21c), with only the strong form in the left periphery. Similar examples indicate that the *wh* phrase generated in the basic position contains only the strong form, which is moved to the left periphery during the derivation:

(24) Input: [$_{\text{IP}}$ ta mangiat [cusè]]
a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ ta mangiat [cusè]]]
b) Move the strong *wh* form to the specifier of *wh* →
[$_{wh}$ [cusè] [$_{\text{IP}}$ ta mangiat [~~cusè~~]]]
Merge+Move

Thus, interestingly, Mendrisiotto is more complex than "canonical" Northern Italian dialects (and Monnese as well) if we consider the number of optional forms of a *wh* question. However, it is less complex if we consider the derivation of these clauses since the linear position of the verb does not change. Before I discuss the three systems from a comparative point of view, I provide a derivation of the Surprise-Disapproval example in (21d). Poletto and Pollock (2009) argue that the strong *wh* form is located in a high position in the left periphery (*Discourse* in the terms of the analysis presented here) and that the whole clause preceding it is moved to a higher position by some type of remnant movement. I propose a different derivation, represented in (25). The clitic *wh* form is first moved to the *wh* projection, then to DiscP to encode the special question, and finally it is deleted[4]:

(25) Input: [$_{\text{IP}}$ ta mangiat [sa cusè]]
a) Merge *wh* and IP →
[$_{wh}$ [$_{\text{IP}}$ ta mangiat [sa cusè]]]
b) Move the clitic *wh* form to the specifier of *wh* →
[$_{wh}$ [sa] [$_{\text{IP}}$ ta mangiat [~~sa~~ cusè]]]
c) Merge DiscP and *wh* →
[$_{\text{DiscP}}$ [$_{wh}$ [sa] [$_{\text{IP}}$ ta mangiat [~~sa~~ cusè]]]]
d) Move the clitic *wh* form to the specifier of DiscP
[$_{\text{DiscP}}$ [sa] [$_{wh}$ [~~sa~~] [$_{\text{IP}}$ ta mangiat [~~sa~~ cusè]]]]
Merge+Move+Merge+Move

Notice that this derivation is no more complex than that of special questions in Bellunese, since it displays the same sequence of Merge and Move.

5 Discussion

The following tables summarize my analysis of the three interrogative systems presented here. Table 1 represents the derivational complexity: The first number in each pair is the number of Move operations, the second one is the number of Merge operations. For the moment I leave aside the problem of special questions. Table 2 represents the number of possible variants that a *wh* question can display in these dialects.

Table 1: Derivational complexity

	yes/no Q	***wh* Q**
NIDs (with clitic inversion)	1/1	2/1
Monnese	1/1	2/1
Bellunese	1/1	1/1
Mendrisiotto	0/1	1/1

[4] At the moment I do not have a definitive explanation for this deletion, but it could be argued that it is motivated by some kind of Blocking condition (Fanselow 1989) in order to avoid linear ambiguity with a standard question with doubling.

Table 2: Optional constructions in *wh* questions

	number of optional constructions
NIDs (with clitic inversion)	1
Monnese	3
Bellunese	1
Mendrisiotto	5 (3)

The tables show that the two parameters lead to different complexity classifications of the dialects under examination: From the point of view of derivational complexity, both Bellunese and Mendrisiotto are less complex than Monnese and "standard" Northern Italian dialects; however, from the point of view of optionality, Mendrisiotto and Monnese are much more complex than Bellunese and the other dialects. So, if we limit ourselves to the syntax of standard questions, varieties with *wh* doubling are more complex than the other dialects, while Bellunese appears to be slightly less complex. Notice, however, that Bellunese encodes special questions by purely syntactic means, as in (17b). In other dialects, also outside the Northern Italian domain, special questions are encoded by other means, for instance discourse particles. In (26) I provide two examples from two different domains. Example (26a) is a Surprise-Disapproval or "Can't find the value" question in Venetian, whereby these interpretations arise due to the presence of the particle *ti*; (26b) is an example from a Southern Italian dialect, Central Calabrian, where a similar interpretation is linked to the particle *ca*.

(26) a. *dove va-lo, ti*?
where goes-he PRT
'Where is he going?!' (Venetian; Munaro and Poletto 2002: 88)
b. *ca cchi sta facennə*?
PRT what stay.2SG doing
'What are you doing?!'
(Central Calabrian; Damonte and Garzonio 2010: 103)

A sentence like (26a) can be analyzed assuming that the particle lexicalizes the appropriate feature for the special question interpretation. If this is correct, Venetian (like Central Calabrian) encodes special questions by Merge and not by Move[5], and its interrogative syntax is slightly less complex than that of Bellunese.

[5] According to Munaro and Poletto (2002) the sentence-final position of the particle is derived

The two parameters adopted here provide us with a general idea about the comparative complexity of the interrogative syntax of these dialects. Leaving aside the obvious problem of further refinements of the method and better descriptions of the grammars, it can be said that this analysis shows us that syntactic complexity in different grammars can have multiple origins. More precisely, it can be argued that there are different levels or types of complexity. Here I propose a tentative classification.

Monnese is a very interesting case. Its *do*-support construction is very uncommon in the Romance domain, but it is not more complex than standard verb-subject inversion (which is in fact typologically rare as an interrogative clause type). In this case, isolation has not produced a more complex syntax, but a different one.[6] I will refer to this phenomenon as "typological complexity", which can be described as typological rarity in a given subset of languages (like, for instance, finding OV in family of languages normally with VO order). The Monnese *wh* system, however, is truly more complex than that of less isolated dialects, since it displays both more optionality and more syntactic operations. Notice, however, that the relatively large quantity of optional *wh* constructions derives from a general property of *wh* words, which are generated as complex phrases with both a clitic and a strong form. Assuming that the "big *wh*" idea is correct, then, the complex syntax of *wh* questions in Monnese is a product of a morpho-syntactic property. To put it simply, more interchangeable forms, each with its specific syntactic behavior, provide more syntactic options and thus more complexity. I will call this "morpho-syntactic complexity".

Bellunese has a slightly less complex interrogative syntax than other Northern Italian dialects. However, its special questions require a more complex syntax, since in order to encode them Bellunese makes use of Move instead of Merge. In this case, this more complex derivation does not derive from non-syntactic properties. My proposal is to consider cases such as Bellunese special questions true cases of pure "syntactic complexity".

Mendrisiotto lacks verb-subject inversion. This property (which is unanimously considered an innovation) is combined with a *wh* system even more complex than the one in Monnese. The case of Mendrisiotto is particularly inter-

through the movement of the whole clause to the specifier of the projection headed by the particle (in our terms, DiscP). Assuming that this is the correct analysis, it must be pointed out that this movement is related to the lexical properties of the particle (which possibly requires this configuration with the clause) and not the clause typing *per se*. Thus, I will not consider it a complexity factor of interrogative syntax.

[6] Compare this with some cases of complexification discussed by Trudgill (2011: chapter 3).

esting because its relative isolation has produced these two properties – the first one simplifying syntax, the other making it more complex. As in Monnese, this is a case of morpho-syntactic complexity.

6 Concluding remarks

In this paper I have proposed to measure syntactic complexity by taking into account the derivational process of a given construction and the number of its synonymous variants. I have used this type of analysis to compare the complexity of the interrogative syntax of three Italian dialects of the Alpine area. Even if the comparison is limited to a single clause type, some interesting general tendencies have emerged: First, isolated dialects do not generally exhibit more derivational complexity than other dialects, but their morpho-syntactic complexity can be very high; second, purely syntactic complexity seems to be very stable among related grammars, the only possible case of greater complexity being the encoding of special questions in Bellunese; finally, it seems that yes/no and *wh* questions can display different degrees of complexity (as in Mendrisiotto, which has less complex yes/no questions and much more complex *wh* questions than other dialects). This fact suggests that further refinements of similar comparative methods should consider separate subsets of the same clause type. In more general terms, even with different degrees of complexity, these dialects still share some abstract properties, suggesting that most parameters of Universal Grammar are set in the same way. Thus, further investigations should consider isolated languages in a greater chronological depth.

References

Antinucci, Francesco and Cinque, Guglielmo1977: Sull'ordine delle parole in italiano: L'emarginazione. *Studi di Grammatica Italiana* 6: 121–146.

Benincà, Paola 2001: The position of Topic and Focus in the left periphery. In: Guglielmo Cinque and Giampaolo Salvi (eds.), *Current Studies in Italian Syntax: Essays Offered to Lorenzo Renzi*, 39–64. Amsterdam: Elsevier.

Benincà, Paola and Poletto, Cecilia 2004: A case of do-support in Romance. *Natural Language & Linguistic Theory* 22: 51–94.

Cardinaletti, Anna and Repetti, Lori 2008: The phonology and syntax of subject clitics in interrogative sentences. *Linguistic Inquiry* 39 (4): 523–563.

Castillo, Juan Carlos, Drury, John and Grohmann, Kleanthes K. 1999: Merge over move and the extended projection principle. *University of Maryland Working Papers in Linguistics* 8: 63–103.

Chomsky, Noam 1995: *The Minimalist Program*. Cambridge, MA: The MIT Press.

Chomsky, Noam 2001: Derivation by phase. In: Michael J. Kenstowicz (ed.), *Ken Hale: A Life in Language*, 1–52. Cambridge, MA: The MIT Press.

Cinque, Guglielmo 1999: *Adverbs and Functional Heads: A Cross-Linguistic Perspective*. Oxford/New York: Oxford University Press.

Damonte, Federico and Garzonio, Jacopo 2010: Per una tipologia delle particelle interrogative nei dialetti italiani. *Rivista Italiana di Dialettologia* 32: 97–111.

Fanselow, Gisbert 1989: Konkurrenzphänomene in der Syntax. *Linguistische Berichte* 123: 385–414.

Gell-Mann, Murray 1994: What is complexity? *Complexity* 1 (1): 16–19.

Giorgi, Alessandra 2010: *About the Speaker: Towards a Syntax of Indexicality*. Oxford/New York: Oxford University Press.

Grice, Martine, Avesani, Cinzia, D'Imperio, Mariapaola and Savino, Michelina 2004: Strategies for intonation labelling across varieties of Italian. In: Sun-Ah Jun (ed.), *Prosodic Typology*, 55–83. Oxford/New York: Oxford University Press.

Kayne, Richard 1991: Romance clitics, verb movement, and PRO. *Linguistic Inquiry* 22: 647–686.

Manzini, M. Rita and Savoia, Leonardo M. 2005: *I dialetti italiani e romanci: Morfosintassi generativa*. Alessandria: Edizioni dell'Orso.

Miestamo, Matti 2008: Grammatical complexity in a cross-linguistic perspective. In: Matti Miestamo, Kaius Sinnemäki and Fred Karlsson (eds.), *Language Complexity: Typology, Contact, Change*, 23–41. (Studies in Language Companion Series 94.) Amsterdam: Benjamins.

Munaro, Nicola 1999: *Sintagmi interrogativi nei dialetti italiani settentrionali.* (Rivista di Grammatica Generativa Monograph.) Padova: Unipress.

Munaro, Nicola 2003: On some differences between interrogative and exclamative *wh*-phrases in Bellunese: Further evidence for a split-CP hypothesis. In: Christina Tortora (ed.), *The Syntax of Italian Dialects*, 137–151. Oxford/New York: Oxford University Press.

Munaro, Nicola and Poletto, Cecilia 2002: Ways of clausal typing. *Rivista di Grammatica Generativa* 27: 87–105.

Obenauer, Hans-Georg 2004: Nonstandard *wh*-questions and alternative checkers in Pagotto. In: Horst Lohnstein and Susanne Trissler (eds.), *Syntax and Semantics of the Left Periphery*, 343–384. (Interface Explorations 9.) Berlin/New York: Mouton de Gruyter.

Obenauer, Hans-Georg 2006: Special interrogatives – left periphery, *wh*-doubling, and (apparently) optional elements. In: Jenny Doetjes and Paz Gonzalvez (eds.), *Romance Languages and Linguistic Theory 2004*, 247–273. Amsterdam: Benjamins.

Pellegrini, Giovan Battista 1992: *Studi storico-linguistici bellunesi e alpini*. Belluno: Fondazione "Giovanni Angelini".

Poletto, Cecilia 1993: Subject clitic / verb inversion in North Eastern Italian dialects. *University of Venice Working Papers in Linguistics* 3 (1): 95–137.

Poletto, Cecilia 2000: *The Higher Functional Field: Evidence from Northern Italian Dialects*. (Oxford Studies in Comparative Syntax.) Oxford/New York: Oxford University Press.

Poletto, Cecilia and Pollock, Jean-Yves 2005: On *wh*-clitics, *wh*-doubling and apparent *wh*-in-situ in French and some North Eastern Italian dialects. *Recherches linguistiques de Vincennes* 33: 135–156.

Poletto, Cecilia and Pollock, Jean-Yves 2009: Another look at *wh*-questions in Romance: The case of mendrisiotto and its consequences for the analysis of French *wh*-in-situ and embedded interrogatives. In: W. Leo Wetzels (ed.), *Romance Languages and Linguistic*

Theory 2006: Selected Papers from 'Going Romance', Amsterdam, 7–9 December 2006, 199–258. Amsterdam: Benjamins.

Rizzi, Luigi 1996: Residual verb second and the *wh*-criterion. In: Adriana Belletti and Luigi Rizzi (eds.), *Parameters and Functional Heads*, 63–90. Oxford/New York: Oxford University Press.

Rizzi, Luigi 1997: The fine structure of the left periphery. In: Liliane Haegeman (ed.), *Elements of Grammar: Handbook of Generative Syntax*, 281–337. Dordrecht: Kluwer.

Rizzi, Luigi 2001: On the position INT(ERROGATIVE) in the left periphery of the clause. In: Guglielmo Cinque and Giampaolo Salvi (eds.), *Current Studies in Italian Syntax*, 287–296. Amsterdam: Elsevier.

Trudgill, Peter 2011: *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford/New York: Oxford University Press.

Johanna Nichols, University of California, Berkeley

Complex edges, transparent frontiers: grammatical complexity and language spreads

Abstract: The Caucasus and the eastern Eurasian steppe are geographically very different but both offer typological and sociolinguistic situations where isolation and complexity can be measured and their correlations with other structural and sociolinguistic factors tested. This paper uses cross-linguistic surveys in both areas to measure two kinds of grammatical complexity: (1) size of inventory or structural complexity (number of phonemes, of genders, of verb inflectional categories, of verb derivational classes, of alignment types, of basic word orders, etc.) and (2) opacity (non-transparency, non-biuniqueness: e. g. allomorphy, suppletion; sandhi, fusion, multiple exponence; synthesis, incorporation; classification; etc.). In each of the two areas complexity proves to be greater at peripheries of language spreads than at centers; opacity is greater in sociolinguistically isolated languages (true whether the sociolinguistic isolation is ascertained directly or through a geographical or demographic proxy). Especially favorable to low opacity and complexity is a variable standing monopoly on inter-ethnic function and spread, where several typologically and/or genealogically close languages have a back-and-forth history of language shift and accommodation now in one direction and now in another. Thus both areas support the correlation of complexity with isolation. The paper uses results of my own fieldwork in the Caucasus as well as published grammars and historical and ethnographic information

1 Introduction

This paper uses language spreads in two areas to test the claims that sociolinguistic isolation entails complexification and that contact and inter-ethnic languages are simpler than others (other things being equal) (Trudgill 2009, 2011). I look at correlations between complexity and sociolinguistic factors in two regions – the eastern Caucasus (section 2) and the eastern part of the Eurasian steppe (section 3) – and the conclusion (section 4) will be yes on both points. (For the languages surveyed see Appendix 1.)

I distinguish two types of complexity: *inventory size* or structural complexity, the number of elements in an inventory (e. g. the number of phonemes, of tones, of declension classes, of genders), and *opacity*, non-transparency or non-biuniqueness between meaning or category and form (factors making for opacity include allomorphy, suppletion, sandhi, multiple exponence, synthesis, incorporation, classification, arbitrary gender assignment, and lexically conditioned alloposition [variously prefix, infix, suffix, etc.] of the same gender agreement marker in the same inflectional paradigm). For the variables surveyed see Appendix 2. Kinds of complexity not considered here include non-canonicality (Corbett 2007; note, however, that it overlaps in large part with opacity), entropy (Ackerman and Malouf 2012), processing effort (e. g. Hawkins 2004), and information-theoretical complexity (Szmrecsanyi, this volume). Most of these strike me as superior measures of complexity, but they would be prohibitively time-costly to survey cross-linguistically.

Language spreads in the two areas investigated proceed in somewhat different ways and for different reasons. In the eastern Caucasus, as in mountain areas more generally, there is overall general uphill spread of languages due to the generally higher economic productivity and inter-society connectivity in the lowlands together with the specific Caucasus factor of transhumance of the male population (Lavrov 1953; Volkova 1967; Wixman 1980; Nichols 2005; Karpov and Kapustina 2011; Nichols 2012c). Uphill spreads are exaggerated in the two local regions investigated here, where a state and/or major inter-ethnic language moves uphill rapidly. Such regions are useful because they create a sharp distinction between the spread area and the languages at the periphery, which in mountain areas is the highlands. Highland villages are tight closed networks, almost entirely endogamous, so there is near-zero L2 learning of those languages, while ascending lowland and foothill languages climb by language spread and thereby take in appreciable populations of adult L2 learners. The situation makes it straightforward to test for correlations between complexity and altitude.

The eastern Eurasian steppe comprises Kazakhstan, Mongolia, and the parts of today's Xinjiang and southern Siberia that connect them; adjacent regions, chiefly Tibet and Manchuria, are also involved in the language spreads discussed here.[1] The area is a spread zone and has been the site of language spreads and

[1] The Eurasian steppe is often segmented into three parts in historical and ethnographic work: the western steppe, in Ukraine and southern Russia to the Urals; the Kazakh or central steppe, east of the Urals and largely coinciding with Kazakhstan; and the Mongol steppe, in northern Mongolia. The latter is sometimes called the eastern steppe, a narrower sense than my use of the term here.

cultural/economic expansions beginning in the third or even late fourth millennium (for the dates, but not specifically for language spreads, see Frachetti 2008, 2012). Large-scale language spreading affecting the entire steppe began with the Indo-Iranian spread of about 4000 years ago (Anthony 2007: Chs. 15, 16),[2] followed by the spreads of Turkic and Mongolic in historical times. The spread of the Tungusic family emanated from Manchuria beginning perhaps 2000 years ago (Janhunen 1996: 216, 218 et pass.). For the general dynamic of spreads in the area see Nichols 1998, 2011, 2012b, 2012c.[3]

The easternmost part of the eastern steppe is a closed spread zone, meaning that entries of new languages to the area are rare and therefore the history of spreads takes the form of back-and-forth shifting and interaction between members and descendants of the same pool of languages. For the last few millennia the main contenders have been Turkic and Mongolic (for their early interaction see Janhunen 1996: 185–193, 240–242; Golden 1998; Schönig 2003), with eastern Uralic, eastern Iranian, Tocharian, Tibetan, and Tungusic less centrally involved. The net result is very strong typological resemblance among the several families found in the area, and a tendency for attractor states to expand (Nichols 2012a, 2012b). Most languages have had more than one episode of acquiring a substantial population of L2 learners and themselves acting as L2 learners, and the result is an accelerated rate of decomplexification. Languages at the periphery of the spread zone, especially those that have spread north of the steppe, lack the most recent episodes of this history (and some may have experienced none of it) and have therefore had more time to undergo natural complexification and fewer pressures to simplify, making it straightforward to test for correlations between complexity and peripherality.

I assume there are two kinds of complexity-related sociolinguistic effects associated with language spreads that can be tested in the geographical situations I have outlined. One is that, since the spreading language absorbs a

[2] The spread of Uralic, probably from northern Kazakhstan or nearby, began earlier than any of these – over 4000–6000 years ago, by various estimates. The historical and late prehistoric distribution of the family arches over the northwest, north, and east of Kazakhstan. Some of the northwestward expansion drew momentum from the later Indo-Iranian expansion from western Kazakhstan c. 2000 BCE, and early Indo-Iranian and Iranian loans into Proto-Finno-Ugric confirm this chronology and directionality of influence, but overall the linguistic geography of Uralic is suggestive of a radial or half-radial expansion from the Kazakh steppe or nearby impelled by the first stages of cultural/economic expansion there. This is my speculation, but if it is right, the Uralic spread is our only surviving linguistic evidence of the initial stage of that expansion.

[3] The eastern steppe homeland proposed for Indo-European in Nichols (1998) is incorrect, but the rest is valid.

sizable population of adult L2 speakers, it undergoes simplification especially of the opacity-reducing (i. e. transparency-increasing) type. Declension and/or conjugation classes become fewer and/or more predictable from lexical semantics; genders likewise; allomorphy and irregularity are reduced; and so on. Therefore the highland languages (in the eastern Caucasus) and the peripheral languages (around the eastern Eurasian steppe) should exhibit measurably greater opacity than the lowland and/or centrally located spreading languages. In addition, the buildup of inventory size that seems to occur naturally over time in L1 transmission is reduced or halted in spread episodes; thus the spreading languages should be simpler in this regard than the highland and peripheral ones.

2 Daghestan

Daghestan is approximately the eastern one-third of the Great Caucasus range. It is a republic of the Russian Federation, a fairly coherent biogeographical area, and the range of the Daghestanian branch of the ancient endemic Nakh-Daghestanian (or East Caucasian) language family. The family is old and much differentiated; the Daghestanian branch alone is probably of nearly Indo-European-like age. (See Figure 1.) The region includes the thickest part of the Great Caucasus range; it is high, rocky, and fairly dry, with deeply eroded river canyons. The highest permanently inhabited places in western Eurasia are found here, as well as some of the smallest stable language communities. Traditionally, towns were basically city-states (Aglarov 1998), each with its own speech variety and often with its own separate language. Each highland town was highly inbred, with most marriages endogamous along both descent lines. There was no traditional lingua franca; most individuals were multilingual, speaking the language of their own town and those of all other towns where they had friends or business connections (including the local market town, usually downhill, and a major lowland market town). Almost never were highland languages learned by lowlanders. At least for the last few centuries until the mid 20th century, Avar, a large lowland language and the court and military language of the Avar khanate until its defeat by Russia in the mid 19th century, was widely used in northern Daghestan for inter-ethnic communication. To the east and south, the major lowland languages were Turkic, chiefly Kumyk and Azeri.

In highland Daghestanian towns, the working-age male population was transhumant, spending the winter half of the year in the lowlands, taking livestock to winter pastures and/or working in business, construction, agriculture, etc. (and often maintaining a second wife and family there). Women and children remained in the highlands all year; men spent only summers there. It might be

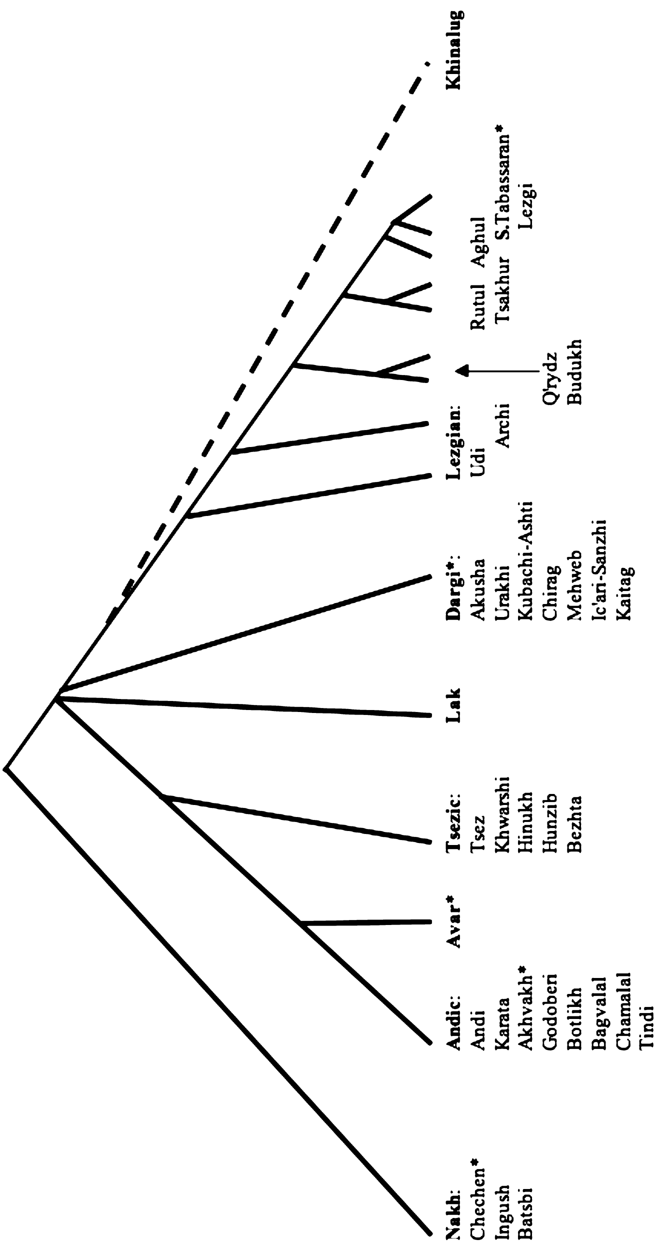

Figure 1: The Nakh-Daghestanian family tree. Individual languages are simply listed under branch nodes, though in fact for most branches there is some further branching structure. Dashed line = possibly earlier split (i. e. a separate branch of Daghestanian).
* = usually treated as one language but actually subsumes two or more mutually unintelligible varieties

said that the female population resided in the highlands and the male population in the lowlands, the two halves forming a single population only during the summer months. Among the working males, often an entire family, clan, or village would work together in the same lowland area, so they – and by extension their upland families – were susceptible to shift to a lowland language. (This probably explains the enclaves of lowland Avar dialects found in highland Daghestan.)

Thus, overall, lowland languages influenced highland languages and could replace them in language shift, so that lowland languages tended to spread uphill while upland languages almost never spread downhill. If an entire family moved downhill, as sometimes happened, it usually joined a co-ethnic enclave which retained the uphill language as long as newcomers continued to move in but whose individual members generally shifted to the lowland language within a generation or two. In view of the endogamy and the asymmetrical bilingualism and multilingualism, highland languages had almost no L2 learners, while lowland languages had many.

Lowland and foothill Daghestan was loosely organized into states and state-like formations for the last few millennia. Of particular sociolinguistic interest is the Avar khanate centered at the Avar capital Khunzakh, which succeeded (at least in name) the Sarir kingdom when Khunzakh converted to Islam. The Sarir kingdom goes back nearly 3000 years, while the Avar spread is unlikely to have occurred more than about 500 years ago (since to my knowledge all dialects of Avar are more or less mutually intelligible). The Avar-Andic branch is unlikely to be much older than Slavic or Romance, i. e. under 2000 years, and Avar-Andic-Tsezic, while more diverse, is still fairly close-knit. Therefore, assuming that language spread from the Avar lowland and foothill region began as the Sarir kingdom formed, it must have been Proto-Avar-Andic-Tsezic that initially spread, followed by successive spreads of subbranch ancestors.

This picture is supported by the linguistic geography. The major upland rivers, the Avar Koisu and the Andi Koisu, converge in the lowlands to form the Sulak. In historical times both banks of the Sulak have been Avar-speaking, with a cluster of Andic place names on the left bank near the confluence. (Andi, the largest Andic language and closest to where those place names are found, was an economic power into the early 20th century and a successful military contender with Avar until the 17th century [Aglarov 1994: 24–25].) All three of Avar, Andic, and Tsezic have branch representatives along both the Avar Koisu and the Andi Koisu: Avar along the Sulak and the entire length of the Avar Koisu with one dialect group on the Andi Koisu; Andic primarily along the Andi Koisu but with its more divergent daughter Akhvakh on the Avar Koisu; and Tsezic in the upper highlands of both rivers (the most divergent daughter, Hunzib, is on a tributary of the Avar Koisu) (it is Tsezic speakers who live at the highest inhabited altitudes of western Eurasia). Thus it seems that the Tsezic languages represent the farthest reaches of the Proto-Avar-Andic-Tsezic uphill spread; the Andic languages represent a second phase, which has progressed well uphill and has absorbed all relicts of the first spread except for the five Tsezic languages; and the Avar spread represents a still later phase, which has absorbed the lowest Andic languages and whatever transitional varieties may have existed between the now-discrete Avar and Andic subbranches.

At the southeastern end of the Great Caucasus, along both north and south slopes, languages of the Lezgian branch of Daghestanian contact and interact (or have interacted) with Azeri, Persian varieties, and Armenian, and before that Urartean. There is a good deal of areality in this contact zone and nearby, saliently including high productivity of light verb constructions (Stilo 2008 and work in progress). There is a long history of states and kingdoms in the vicinity. Their languages are unknown until the early centuries CE, when Caucasian Albanian, the direct ancestor of today's Udi or a close sister of its ancestor, emerged as one of the important inscriptional languages of the Transcaucasus. Udi has since receded to two small enclaves in Azerbaijan (and these were scattered in the early 1990's) plus a Georgian offshoot, its speakers shifting variously to Azeri, Persian, and Armenian. To the east, Lezgi is a very large language extending from lowlands to highlands and still functioning as an inter-ethnic language along its uphill frontier; though its prehistory is not known, it evidently results from a sizable spread from a lowland or foothill center (within the last 500–1000 years to judge from the near-intelligibility of its dialects). Uphill to the north and west lie its closest sisters (some of which it has been absorbing in recent centuries), likely relicts of an earlier spread from the same vicinity. Thus one can infer a history of repeated spreading from the Lezgi lowlands on the north slope and the Alazani-Kura area and historical Caucasian Albania on the south slope.

In both the Avar cultural and political sphere and the Lezgi-Udi area in the southeast, there have been multiple spreads from more or less the same centers and back-and-forth spreading between major lowland languages. That is, there is a standing center of spread, defined geographically and fairly approximately, and whatever language happens to dominate in that center is prone to spread, have inter-ethnic function, and absorb L2 speakers. The center is not a pinpoint but a region accommodating more than one language, so that dominance in spreads can and does shift from one language to another. Hence the language that spread in one episode, absorbing L2 speakers, may be the substratum in the next spread, shifting and contributing grammatical and phonological features to the spreading language. Enough such episodes of back-and-forth spread bring about considerable decomplexification of all languages in and near the center of spreading. Therefore, in both of these parts of Daghestan, and by extension across all of Daghestan, we can expect to find lower complexity (fewer elements in systems, and/or more transparent grammar) in the lowland languages and higher complexity (more elements in systems, more opacity) in the highland languages.

These predictions are borne out. Figure 2 shows inventory size for the 24 Daghestanian languages for which I have data, plus Chechen and Ingush of the Nakh branch. Inventory sizes are summed over the various grammatical structures surveyed, or else lumped into small/medium/large inventories and those counted as 1, 2, and 3 points (see Appendix 2 for how totals were calculated). This ensured that all parts of grammar surveyed had their highest possible totals on the same order (usually 2–5 possible points each). White dots (smallest inventory size totals) are found only in the Avar sphere and the Lezgi-Udi area. Those in the Avar sphere are Avar and the Andic languages Andi, Chamalal, and Godoberi. Those in the southeast are Lezgi, its close sister Aghul, and Udi. These are the languages with relatively recent histories of spreading (Avar, Andi; Udi, Lezgi) and close sisters that are also close neighbors and may have been involved in the spreads. These are the languages that are likely to have absorbed appreciable numbers of adult L2 learners, and their reduced inventories of grammatical elements are consistent with such a history.

It is striking that the simplification associated with such spreads appears to be quite durable. The spread of Udi was in the early to middle first millennium CE, since which time it has been a receding minority language. 1500 years later, it is still notably non-complex. An only slightly later date applies to the lower Andic languages, if I am right that they represent an immediate pre-Avar spread and only Andi proper has been an important economic power since then. Udi and the Andic languages have been sociolinguistically isolated ever since their spreads, and the fact that they remain relatively simple after a millennium or more suggests that the gradual complexification in isolated languages builds up slowly, while the simplification caused by absorbing L2 learners can proceed quickly.

Good examples of what I mean by opacity come from gender systems. Daghestanian gender systems are semantically transparent in Avar and the Andic languages (in the Avar sphere), where all nouns referring to human males belong to the gender marked by *w-*, all nouns referring to human females belong to the gender marked by *j-*, and all other nouns (names of animals, plants, structures, tools, substances, abstractions, etc.) belong to the gender marked by *b-*. Also very transparent is the system of northern Tabasaran (Lezgian branch), where human nouns take *w-* and all others take *r-*. Maximally transparent are the four languages that have lost gender entirely: Lezgi, southern Tabasaran, Aghul, Udi (all in the Lezgi-Udi area). Semantically very opaque is the gender system of Ingush (Nakh branch), shown in (1). Words referring to humans – both nouns and pronouns – take *v-* (masculine) or *j-* (feminine) depending on the sex of the referent; this much is transparent. But for all other nouns the system is opaque. There are four

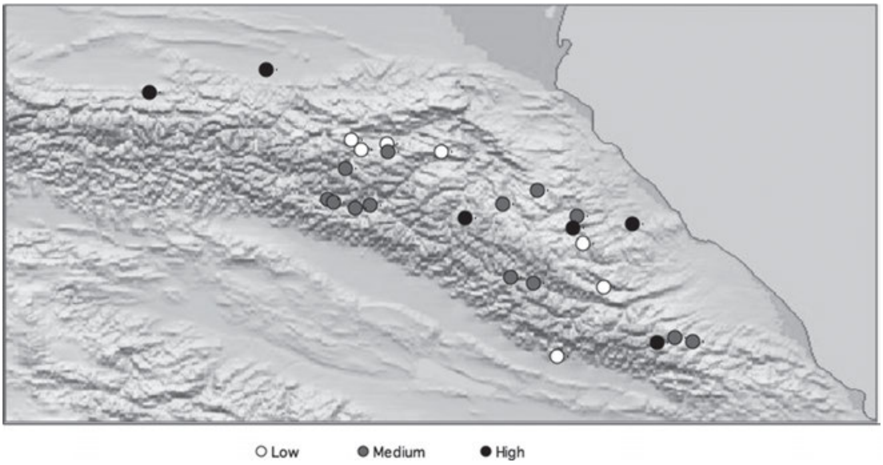

Figure 2: Inventory size for 26 Nakh-Daghestanian languages. (The northernmost and westernmost two are Chechen and Ingush of the Nakh branch; the others are in the Daghestanian branch.) Here and below, darker plot symbol indicates higher complexity. Low = approximately 1 s. d. below the mean (curved to accommodate data clusters); high = approximately 1 s. d. above the mean

controller genders (in the terminology of Corbett 1991), and all contain a mix of animate and inanimate nouns of various semantic types.

(1) Ingush gender classes. M = masculine, F = feminine (human).

		singular	*plural*	*examples*
1st, 2nd person pronouns	M	v	d	me, you, us
	F	j	d	me, you, us
3rd person pronouns (human)	M	v	b	him, her, them
	F	j	b	him, her, them
male human nouns		v	b	man, Ahmed
female human nouns		j	b	woman, Hava
some animals, inanimates		b	d	ox, head
some plants, inanimates		b	b	apple, family
inanimates, some animals		j	j	wolf, fence
inanimates, some animals		d	d	dog, house

Another example of opacity and transparency comes from the formal treatment of the causative alternation. In the list of nine verb pairs used by Nichols, Peterson, and Barnes (2004) to typologize the morphology of the causative alternation, most Nakh-Daghestanian languages use a variety of different formal pairings:

causativization, decausativization, ambitransitive verbs, suppletion, different light verbs, etc. Such systems are fairly opaque in that there is no predicting, across the board, how a verb and its semantic causative will be formally related. In addition, pairs using morphological causativization, where the semantic causative is marked by overt affixal material, are transparent in the further sense that there is a straightforward correspondence between semantics and morphology, shown in (2).

(2) Avar. (*-ize* is the infinitive suffix.)

	Semantics	*Form*
'fear':	'fear'	hˤinq'-ize
'scare'	'fear' + Causative	hˤinq'-**ab**-ize

Only in the Avar sphere is the set of verbs highly regular and highly transparent: nearly all pairs use causativization (Creissels 2010), especially in Avar and Andic.[4]

Figure 3 is a map of overall opacity levels, summed over several different areas of grammar. These are the Nakh-Daghestanian factors listed in Appendix 2. White dots (the lowest opacity, i. e. highest transparency) are found only in the Lezgi-Udi area. The next lowest opacity (pale blue dots) is found across the northern lowlands and foothills. Black dots (highest opacity) are found in the highlands, and especially among the sociolinguistically isolated languages of the highlands.[5] From northwest to southeast they are: Khwarshi, Bezhta, and Hunzib (Tsezic branch); Archi (a geographically and sociolinguistically isolated and divergent Lezgian language); Lak; Tsakhur (Lezgian); and three Lezgian languages cut off from the rest of the family and surrounded by Azeri: Kryz, Budukh, and Khinalug.

For this survey, opacity levels were lumped as low, medium, and high (1, 2, 3 points) in all parts of grammar surveyed, and the attested totals range from 2 to 11. Figure 4 plots these totals against altitude and shows that there is a moderately strong correlation. This is because it is the highland languages that are sociolinguistically isolated, so altitude is a reasonably good proxy for sociolinguistic isolation.

[4] Languages in the Udi-Lezgi sphere make extensive use of light verb constructions in these pairs. These are not necessarily transparent, as the light verb used with any given verb is usually not predictable; but they have the advantage, in a contact situation, of making it very easy to incorporate loan vocabulary into the language.

[5] Baechler (this volume) has very similar findings from the Alps.

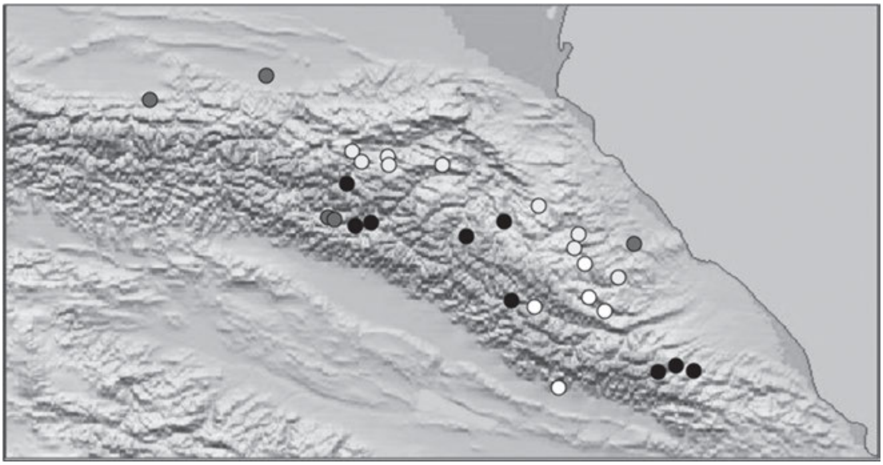

Figure 3: Opacity in 28 Nakh-Daghestanian languages. Four-way breakdown. Darker plot symbol = more opacity

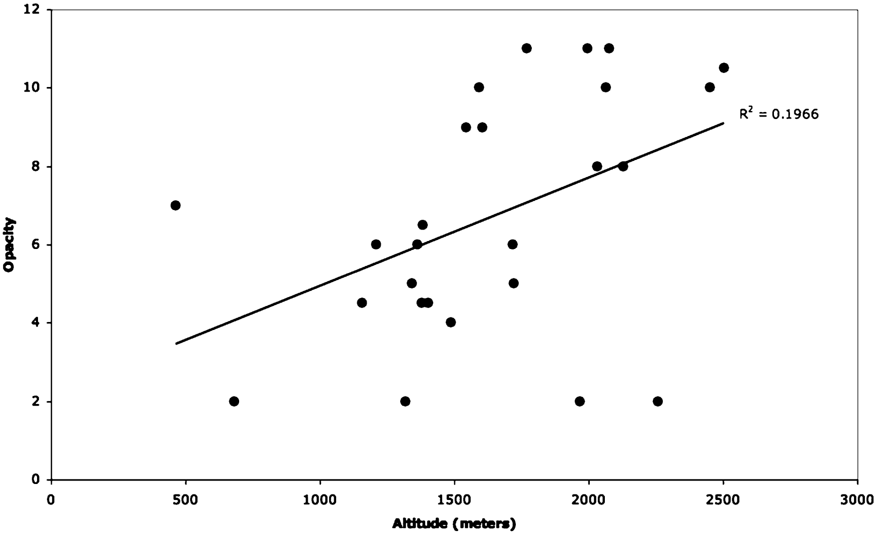

Figure 4: Opacity plotted against altitude, for 26 languages of Daghestan

Only opacity correlates well with altitude; inventory size does not. Somewhat similarly, opacity is negatively correlated with the number of verb pairs (out of the nine surveyed, as discussed above) that use morphological causativization to implement the causative alternation; inventory size is not. See Figures 5 and 6. These differences in patterning show that the two kinds of complexity,

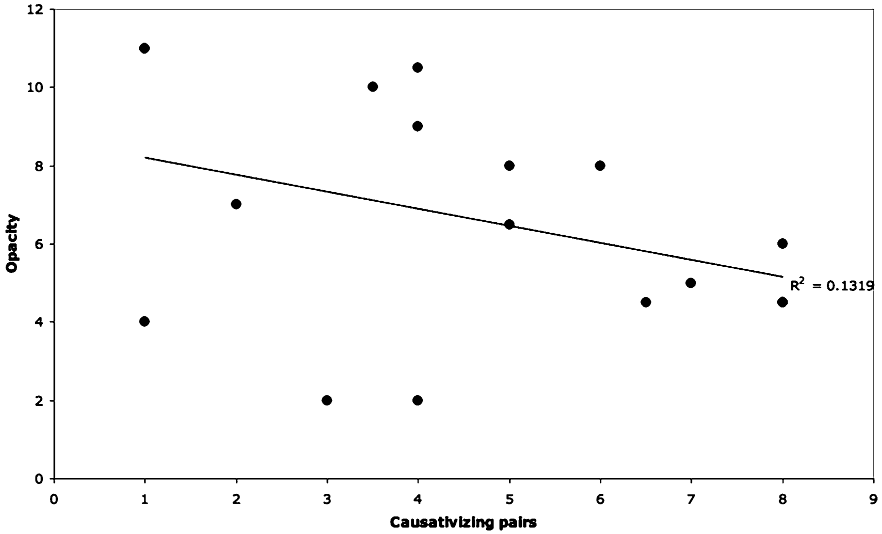

Figure 5: Overall opacity level plotted against the number of verb pairs (out of the 9-pair wordlist of Nichols et al. 2004) that use causativization. 17 Daghestanian languages

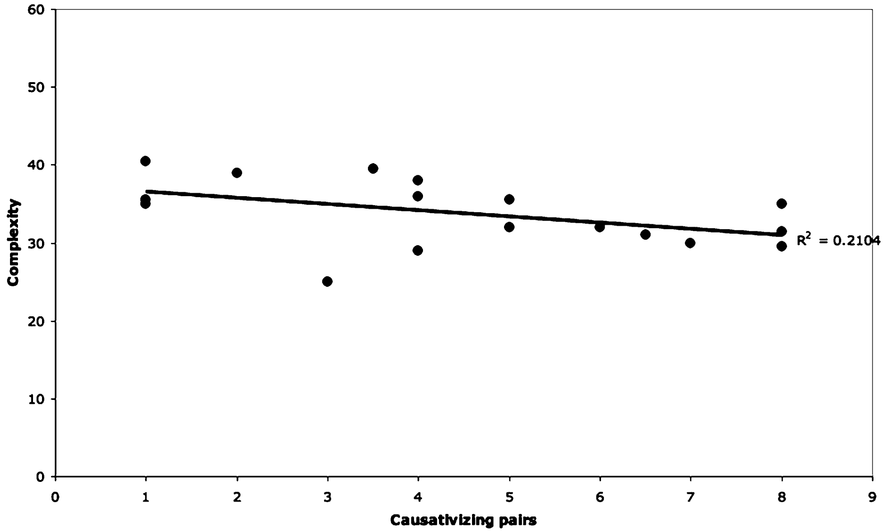

Figure 6: Overall inventory size plotted against the number of verb pairs (out of the 9-pair wordlist of Nichols et al. 2004) that use causativization. 17 Daghestanian languages

inventory size and opacity, are different things. From the Daghestanian evidence it appears that sociological isolation (and its proxy, altitude) is associated with higher overall opacity, while functioning (or having functioned) as inter-ethnic language is associated with both transparency and a smaller overall inventory of elements (here there is no correlation with altitude). Whether these are general distributions or just an accidental pattern found in this small sample remains to be determined. If they are not accidental, perhaps they mean that transparency builds up faster than the inventory of elements expands.

3 The eastern Eurasian steppe

I surveyed inventory size and opacity in 22 languages representing the language families and isolates of the eastern steppe zone and nearby: Turkic (Tuva, Yakut. Uighur), Mongolic (standard modern Mongolian), Tungusic (Manchu, Nanai, Even, Evenki), Tibetan (spoken modern standard), Korean, Japanese, Ainu, Nivkh, Chukchi-Kamchatkan (Chukchi, Itelman), Eskimo-Aleut (Siberian Yup'ik, Aleut), Yukagir, Yeniseian (Ket), and eastern Uralic (northern Khanty, Mansi, Tundra Nenets). The hypotheses are that the languages farthest from the steppe will have the highest inventory sizes and opacity levels; those at the steppe periphery will be intermediate and those at the center or involved in the latest large spreads will be simplest.

The languages conform rather well to these predictions. Figures 7 and 8 show inventory size levels on two breakdowns: two-way (above vs. below mean) and three-way (high/medium/low, curved to break points); Figures 9 and 10 show opacity levels on the same two breakdowns. Table 1 shows the actual levels. The lowest inventory sizes are found in languages that have undergone, or been affected by, major spreads: on the three-way breakdown, Uighur, Mongolian, and Manchu, plus Nanai (which was at the edge of the Manchu expansion), and Khanty, one of the languages peripheral to the Indo-Iranian expansion and the later Turkic spreads (I cannot explain why only Khanty of the three eastern Uralic languages shows up here; this may be random). The two-way breakdown brings in all three eastern Uralic languages, of which Khanty and Mansi have similar prehistories and Nenets results from a relatively recent expansion of its own; and Ainu and Japanese, both with relatively recent expansions (see Janhunen 2002). High opacity levels ring the far eastern and northwestern peripheries of the steppe zone, on either breakdown. Calculated mean opacity levels shown on Table 1 do not reflect this last claim well, for three reasons: numeral classifiers and grammaticalization of honorifics, etc. give Japanese and Korean higher levels; perhaps Khanty and Mansi should be counted as Edge rather than Outer as noted in the legend; and if word formation, synthesis, etc. were surveyed, Eskimo-Aleut, Chukchi-Kamchatkan, and Nivkh would rank as much more complex.

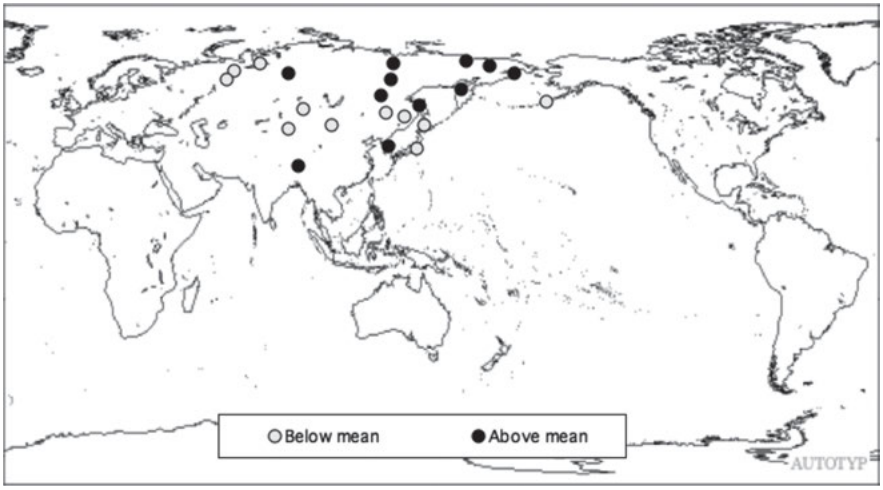

Figure 7: Inventory size levels for languages of eastern Eurasia. ($N = 22$)

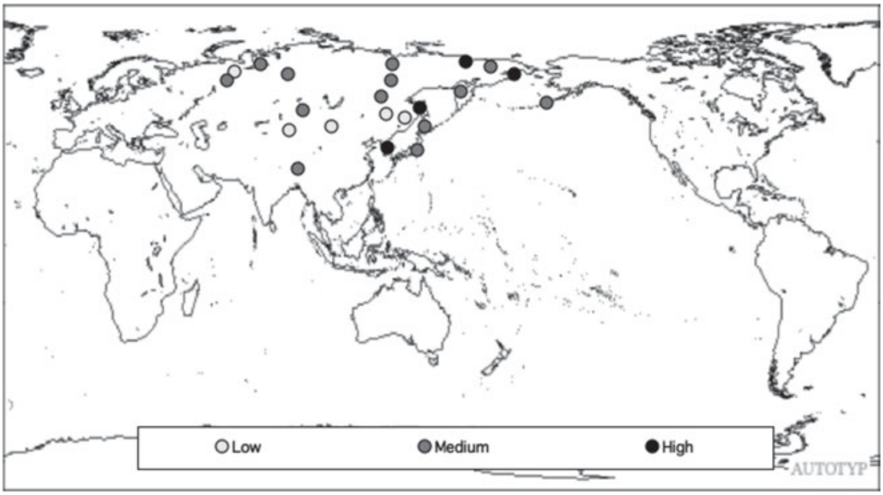

Figure 8: Inventory size levels for languages of eastern Eurasia, three-way breakdown. ($N = 22$)

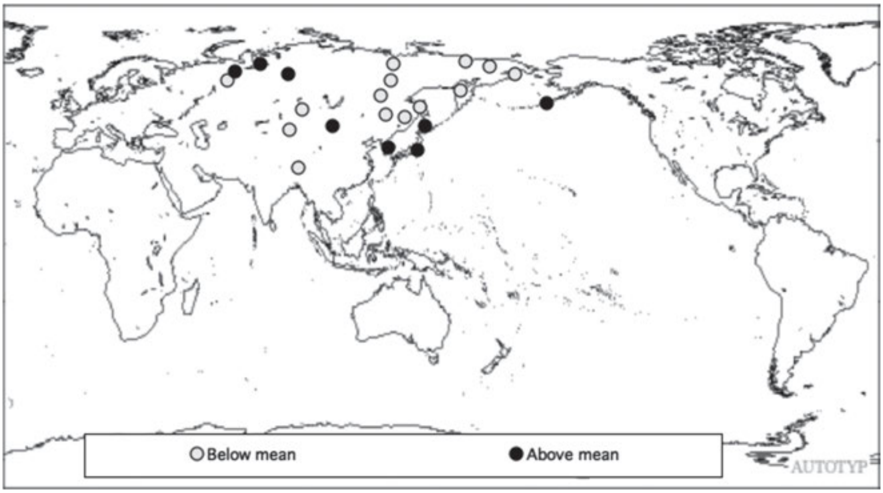

Figure 9: Opacity levels for languages of eastern Eurasia. (*N* = 22)

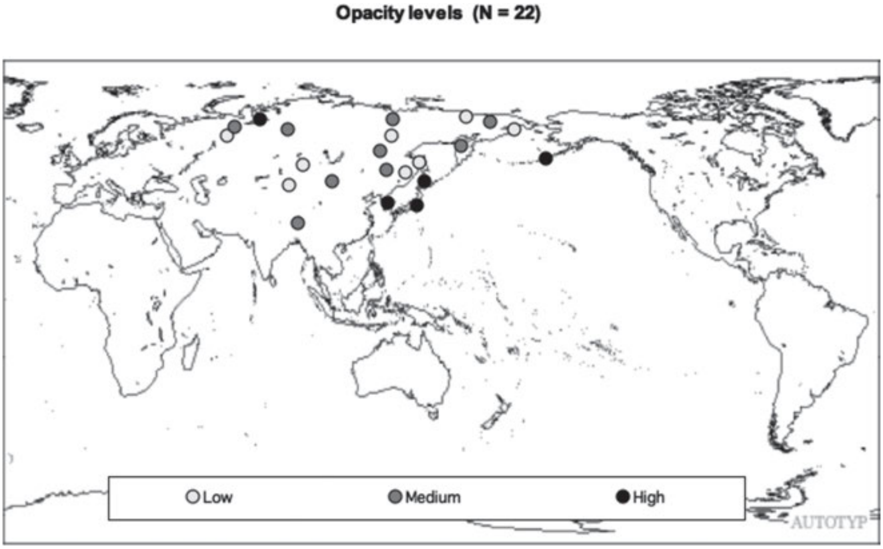

Figure 10: Opacity levels for languages of eastern Eurasia, three-way breakdown. (*N* = 22)

Table 1: Inventory sizes and opacity levels for the 22 languages, geographically arranged. Inner = at centers of spreading; edge = peripheral but not distant; outer = not adjacent (usually distant). * = outer languages that were adjacent to the Indo-Iranian spread and perhaps the early phases of the Turkic spreads. (If they are counted as edge, the means for inventory size on the two-way breakdown are polarized: 15.3 and 18.2.) Mean = mean for the set of languages beginning with that row and continuing to the next entry (e. g. 15.13 = mean inventory size for Mongolian, Manchu, Uighur, and Tibetan; also the next 5 languages on the 2-way breakdown).

Language	Geography	Inventory	Mean		Opacity	Mean	
			3-way	2-way		3-way	2-way
Mongolian	Inner	14	15.13	15.50	5	4.00	4.89
Manchu	Inner	14			4		
Uighur	Inner	14			3		
Tibetan	Inner	18.5			4		
Ainu	Edge	15	15.80		6	5.60	
Japanese	Edge	15			8		
Korean	Edge	19			8		
Nanai	Edge	14			3		
Tuva	Edge	16			3		
Khanty (Northern)	Outer*	14	17.62	17.62	5	3.96	3.96
Mansi	Outer*	15			3		
Nenets (Tundra)	Outer	16			6		
Ket	Outer	18			5		
Yakut	Outer	17			3		
Evenki	Outer	18			4		
Even	Outer	17			4		
Yukagir	Outer	19			3		
Nivkh	Outer	23			2		
Chukchi	Outer	18			4		
Itelmen	Outer	18			4		
Aleut	Outer	16			5.5		
Yup'ik (Siberian)	Outer	20			3		

This support by eastern steppe and Siberian languages of the complexity-isolation correlation is gratifying, given that sociolinguistic isolation in this region is less clear-cut than in Daghestan; in Siberia there was traditionally intermarriage and second-language learning both in and out of most societies.

4 Conclusions

In both areas surveyed here the languages at the periphery emerge as more complex and those at spread centers as simpler. Inventory size and opacity pattern somewhat differently, small inventory size characterizing the languages that have functioned as inter-ethnic communication vehicles and high opacity characterizing those with sociolinguistic isolation and not generally learned as L2. In the course of this study it has become clear that it is difficult to capture differences within areas and within families using variables developed for cross-linguistic work, and difficult to find morphological variables that generalize well across families and types. The variables for Daghestan include some more or less Nakh-Daghestanian-specific variables as semantic transparency of genders, prefixal vs. infixal location of gender agreement markers, the number of oblique and plural stem extensions, and others. As noted just above, consideration of word formation and synthesis would change complexity figures for Siberia. Despite these obstacles, the expectation of higher complexity associated with sociolinguistic isolation does appear to hold up in both areas.

References

Ackerman, Farrell, Blevins, James P. and Malouf, Robert 2009: Parts and wholes: Implicative patterns in inflectional paradigms. In: James Blevins and Juliette Blevins (eds.), *Analogy in Grammar: Form and Acquisition,* 54–82. Oxford: Oxford University Press.

Ackerman, Farrell and Malouf, Robert 2012: Information-theoretic morphology: The low entropy conjecture. MS, UCSD.

Aglarov, Mamajxan A. 1994: Andis. In: Paul Friedrich and Norma Diamond (eds.), *Encyclopedia of World Cultures, VI: Russia and Eurasia/China*, Boston: Hall.

Aglarov, Mamajxan A. 1988: *Sel'skaja obschina v Nagornom Dagestane v XVII – nachale XIX v*. [The village *communitas* in highland Daghestan, 17th – early 19th centuries]. Moscow: Nauka.

Corbett, Greville 2007: Canonical typology, suppletion, and possible words. *Language* 83 (1): 8–42.

Corbett, Greville 1991: *Gender*. Cambridge: Cambridge University Press.

Creissels, Denis 2014: P-lability and radical P-alignment. *Linguistics* 52 (4): 911–944.

Golden, Peter B. 1998: The Turkic peoples: A historical sketch. In: Lars Johanson and Éva Ágnes Csató (eds.), *The Turkic Languages*, 16–29. London: Routledge.

Hawkins, John A. 2004: *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Janhunen, Juha 1996: *Manchuria: An Ethnic History*. (Mémoires de la Société finno-ougrienne, 222.) Helsinki: Suomalais-ugrilainen Seura.

Janhunen, Juha 2002: On the chronology of the Ainu ethnic complex. *Bulletin of the Hokkaido Museum of Northern Peoples* 11: 1–20.

Karpov, Jurij Ju. and Kapustina, Ekaterina L. 2011: *Gorcy posle gor: Migracionnye processy v Dagestane v XX – nachale XXI vv.: ix social'nye i ètnokul'turnye posledstvija i perspektivy*. [Mountaineers after the mountains: Migratory processes in Daghestan in the 20th and early 21st centuries: their social and ethnocultural consequences and perspectives.] St. Petersburg: Rossijskaja AN, Muzej antropologii i ètnografii.

Lavrov, L. I. 1953. Nekotorye itogi raboty Dagestanskoj èkspedicii 1950–52 gg. *Kratkie soobschenija Instituta Ètnografii* 193–7. Moscow: AN.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Nichols, Johanna 1998: The Eurasian spread zone and the Indo-European dispersal. In: Roger Blench and Matthew Spriggs (eds.), *Archaeology and Language II: Archaeological data and linguistic hypotheses,* 220–266. London: Routledge.

Nichols, Johanna 2009: Linguistic complexity: A comprehensive definition and survey. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 110–125. Oxford: Oxford University Press.

Nichols, Johanna 2011a: Causativization and contact in Nakh-Daghestanian. *Berkeley Linguistics Society* 37: 68–80.

Nichols, Johanna 2011b: Forerunners to globalization: The Eurasian steppe and its periphery. In: Cornelius Hasselblatt, Peter Houtzagers and Remco van Pareren (eds.), *Language Contact in Times of Globalization,* 177–195. Amsterdam: Rodopi.

Nichols, Johanna 2012a: The history of an attractor state: Adventitious *m* in Nakh-Daghestanian. In: Tiina Hyytäinen, Lotta Jalava, Janne Saarikivi and Erika Sandman (eds.), *Per Urales ad Orientem: Iter Polyphonicum Multilingue* (Festschrift for Juha Janhunen), 261–278. Helsinki: SUST.

Nichols, Johanna 2012b: Selection for *m: T* pronominals in Eurasia. In: Lars Johanson and Martine Robbeets (eds.), *Copies versus Cognates in Bound Morphology*, 47–69. Leiden: Brill.

Nichols, Johanna 2012c: The vertical archipelago: Adding the third dimension to linguistic geography. In: Peter Auer, Martin Hilpert, Anja Stukenbrock and Benedikt Szmrecsanyi (eds.), *Space in Language and Linguistics*, 38–60. Berlin/Boston: de Gruyter.

Nichols, Johanna, Peterson, David A. and Barnes, Jonathan 2004: Transitivizing and detransitivizing languages. *Linguistic Typology* 8 (2): 149–211.

Schönig, Claus 2003: Turko-Mongolic relations. In: Juha Janhunen (ed.), *The Mongolic Languages*, 403–419. London: Routledge.

Stilo, Don 2008: An introduction to the Araxes-Iran linguistic area. Presented at SOAS, December 2008. http://www.soas.ac.uk/linguistics/events/deptseminars/02dec2008-an-introduction-to-the-araxes-iran-linguistic-area.html

Trudgill, Peter 2009: Sociolinguistic typology and complexification. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 98–109. Oxford: Oxford University Press.

Trudgill, Peter 2011: *Sociolinguistic Typology: Social Determinants of Linguistic Structure and Complexity*. Oxford: Oxford University Press.

Volkova, Natalja G. 1967: Voprosy dvujazychija na Severnom Kavkaze. [Issues of bilingualism in the North Caucasus]. *Sovetskaja ètnografija* 1: 27–40.

Appendix 1: Languages surveyed

Nakh-Daghestanian family (eastern Caucasus):

Nakh branch			Ingush
			Chechen
Daghestanian branch:			
	Avar-Andic-Tsezic		
		Avar	standard Avar
		Andic	Akhvakh (northern)
			Godoberi
			Karata
			Bagwalal
		Tsezic	Tsez
			Khwarshi
			Hinuq
			Hunzib
			Bezhta
	Lak		Lak
	Dargi		Akusha/standard
			Kubachi
			Ic’ari
	Lezgian		Lezgi (standard)
			Tabasaran (northern)
			Aghul (Burschag dial.)
			Rutul
			Tsakhur
			Kryz
			Budukh
			Archi
			Udi
			Khinalug

Eastern steppe region:

Uralic	
Ugric branch	Mansi
	Northern Khanty
Samoyedic	Tundra Nenets
Yeniseian	Ket
Turkic	Uighur
	Tuva
	Yakut
Mongolic	Khalkha (modern standard)
Sino-Tibetan	Tibetan
Tungusic	Manchu
	Nanai
	Evenki
	Even
Yukagir	Kolyma Yukagir
Ainu (isolate)	Ainu
Nivkh (isolate)	Nivkh
Chukchi-Kamchatkan	Chukchi
	Itelmen
Eskimo-Aleut	
Eskimoan	Siberian Yup’ik
Aleut	Aleut
Japonic	Japanese
Korean (isolate)	Korean

Appendix 2: Linguistic variables surveyed

Those calculated for all languages (the first group) are described briefly in Nichols (2009). The Nakh-Daghestanian set is new here. Inventory size is counted as the sum of entries for all variables (for yes/no variables, yes = 1 and no = 0). Gender (under Nakh-Daghestanian) is counted as one point for each of these that is present in the gender agreement system: allomorphy, suppletion, sandhi, multiple exponence, synthesis, incorporation, classification, arbitrary gender assignment, and lexically conditioned alloposition (see § 1).* = variable surveyed for opacity (the others contribute only to inventory size).

All languages:

Phonology:	Number of contrastive obstruent manners of articulation
	Vowel qualities (high/medium/low)
	Tones (yes/no)
Synthesis:	Index of verb inflectional synthesis (apart from agreement)
	Verb polyagreement (2 arguments or more)
	Noun number categories: plural (yes/no)
	Noun number categories: dual (yes/no)
Classification:	Numeral classifiers (yes/no)
	Overt possessive classification (0/1/2 classes)
	Gender agreement (yes/no)
	Overt inherent gender (non-agreement, marked on the noun itself)
Syntax	Number of major clause alignments (other than neutral)
	Number of basic clause word orders
	Noun incorporation (yes/no)
Lexicon	Inclusive/exclusive opposition
	Number of different formal types of causative alternation in 9-verb list
	Number of suppletive pairs in that list

Additional Nakh-Daghestanian variables:

Gender*	Total points of non-transparent gender agreement
	% underived verbs taking gender agreement
Declension*	Number of singular stem extensions (forming oblique stem)
	Number of plural stem extensions (forming plural stem)
	Number of suppletive roots in 1st–2nd singular personal pronouns and 'who', 'what'
Verb stem	Number of different types of prefixes (deictic, spatial, bipartite, mutating initial)

Daniel Schreier, University of Zurich

A true *split*? Typological and sociolinguistic considerations on contact intensity effects

Abstract: In this chapter, I argue that contact between linguistic systems (no matter whether typologically similar or not) may give rise to *both* simplification and complexification processes, in one and the same variety and at one and the same moment in time. These processes do not depend on the variety's contact history. I show that simplification under contact conditions is evident in regularization (for instance of verbal paradigms, particularly in the case of inflectional morphology), whereas complexification may simultaneously operate in tense, mood and aspect systems or pronoun systems. The sociolinguistic factors involved are: the properties of the systems in contact, the social grounding of the contact scenario (including population demographics and peopling), as well as the intensity of contact, particularly the strength of substrate effects. I suggest that general claims such as "low contact entails complexification, high contact simplification" have to be revised and modified so as to do justice to the multifaceted and interwoven (social and linguistic) phenomena of contact-induced change. My claims are supported by findings from a number of English varieties around the world. I will specifically focus on the variety of Tristan da Cunha English (TdCE), where these processes manifested themselves in extreme isolation, i. e. prime low-contact conditions, thus substantially challenging current taxonomic models.

1 What contact has to do with it

In a recent (2009) article, Peter Trudgill argued that current typological approaches to World Englishes were misguided in that they focused, as had been tradition for over a century, on standard vs. non-standard/vernacular varieties[1]. For instance,

[1] This paper is the result of sharing ideas with a number of colleagues over the years (not all of whom would agree with me, of course). Most importantly, I wish to acknowledge here discussions with Jack Chambers, Ray Hickey, Bernd Kortmann, Salikoko S. Mufwene, Edgar W. Schneider, Guido Seiler, Benedikt Szmrecsanyi, and Peter Trudgill, as well as the feedback received from participants when I presented the paper at the workshop on complexity held at the FRIAS, University of Freiburg i. Brsg., in December 2011. I also wish to thank Balthasar Bickel, Brook Bolander, Nicole Eberle and Danae Perez-Inofuentes for their most valuable comments on a draft version of this chapter.

the standard vs. non-standard dichotomy is one of the driving forces behind Kachru's (1985, 1986) concentric circles model, where inner-circle L1 varieties such as British and American are "norm-providing" for the varieties of English as a Second (ESL) or Foreign Language (EFL). McArthur (1987) and Görlach (1990) developed alternative models, both taking for granted that English could be classified into first- (or native), second- and foreign-speaker varieties and that standard varieties with high social recognition were of particular importance for the diversification of English or the emergence of new varieties around the world (see also Hundt 2012). Trudgill (2009), in contrast, suggests that these models should focus on contact instead. Diversification should be looked at from another angle since there exists no English (or any other) variety that has not undergone at least some extent of language or dialect contact.

The main dividing line between English varieties is therefore not one of standardization (perhaps also degree of nativization, if we follow Kachru 1985) but of contact history (classified as "high" vs. "low"): high contact entails processes of simplification, whereas low contact gives rise to complexification (Trudgill 1986, 2009, 2011), and this has resulted in a "true typological split" in English (Trudgill 2009: 304). In the former category, we find 1) transplanted L1 Englishes or colonial standards (e. g. New Zealand English, White South African English, but also Bahamian English, Falkland Islands English, Channel Island English or Maltese English); 2) language-shift Englishes (shift or shifted varieties, e. g., Irish and Welsh English); and 3) Standard L1 varieties (e. g. colloquial British and American English). Varieties in the latter category include traditional, regional non-standard varieties with a long tradition of mother-tongue speakers (and natural transmission), e. g., the English Southwest, Southeast (as well as East Anglia) and North, Newfoundland English, Appalachian English and Ozark English in North America. Importantly, all varieties classified by Trudgill (2009) as "low contact" have considerable time depth and are mostly spoken in geographically isolated regions, often at a country's periphery and near national borders. Varieties such as Appalachian English or Newfoundland English are now fully indigenized varieties that had native speakers from early on; they were formed within the last 400 years, typically in former settlement colonies, and shaped by settlers with diverse dialect backgrounds, often in adstrate relationships and with negligible substrate influence from other languages.

Based on these claims, Trudgill (2010) outlines a "sociolinguistic typology": "the distribution of linguistic structures and features over languages is sociolinguistically not entirely random [...] there may be a tendency for different types of social environment and social structure to give rise to, or at least be accompanied by, different types of linguistic structure" (Trudgill 2010: 16). Following Trudgill (2009), first language acquisition vs. second language learning and the strategies applied in these processes are most important; simplification vs. complexification is consequently not so much a function of contact intensity but of whether or not languages are naturally transmitted from one generation to the next – or learnt as L2 by adults or post-adolescents. Trudgill (2009) claims that

> just as complexity increases through time, and survives as the result of the amazing language learning abilities of the human child, so complexity disappears as a result of *the lousy language learning abilities of the human adult*. Adult language contact means adult language learning; and adult language learning means simplification. (Trudgill 2009: 372–373) [emphasis mine]

As is well-known, children are crucial agents in new-dialect formation (Trudgill 1986; Kerswill 1996, 2001; Schreier 2012). They face the task of constructing new grammars in dynamic social circumstances of first dialect acquisition, characterized by diffuse heterogeneity and heterogeneous inputs (Trudgill, Gordon, Lewis, and Maclagan 2000; Trudgill 2004). Moreover, children participate in "pick'n'mix" processes (Schreier 2014), so that they are mainly in charge of selecting features from various feature pools and combining them in new systems. Adults, on the other hand, have, according to Trudgill (2009) "lousy language learning abilities" (Trudgill 2009: 372) and thus fail to master complex rules, preferring regular (easily learnable) paradigms without redundancy instead. These, so the story goes, are the conditions that favor simplicity.

One problem, however, is that all contact situations are to some extent messy. Learning and acquisition occur simultaneously, population demographics vary within the speech community, and social inequalities are reflected by typical outcomes of social stratification (e. g., the difference between basilectal and acrolectal varieties in creole-speaking communities). Following Schreier (2008), all explanatory approaches or models suggest

> clarity (as models usually do) and [give] the impression that this is a nice and tidy process, and a chronologically ordered one on top. Nothing could be further from the truth, of course. Koinéisation, like all processes of contact-induced language change, is messy and far from straightforward. The complicating factors are numerous: Processes may co-occur and influence each other, interdependent developments are impossible to predict, social factors may intervene at all times, changing the outcome, etc. (Schreier 2008: 31)

Consequently, then, I will critically assess arguments such as "high contact equals simplification, low contact equals complexification" with examples of English around the world.

2 Testing the claims

2.1 High contact = simplification

In fact, it can easily be demonstrated that high contact between systems may give rise to simplification, as a look at the *World Atlas of Linguistic Structures* (WALS) shows. Maddieson (2011) suggests that the world's languages fall into three phonotactic categories in regard to their syllable types, namely in how consonants (C) and vowels (V) combine into larger units (syllables). These are: 1) *simple*: consisting of (C)V types exclusively; 2) *moderately complex* ("modest expansions of the simple CV syllable type"): CV(C), (C)CV, (C)CV(C) ("most elaborate syllable permitted"); or 3) *complex*: containing up to three Cs in onset or coda positions ((C)(C)(C)V(C)(C)(C)(C)). A search for the distribution of these systems in the world's languages shows that while CV(C) is the universal syllable type, found in all the world's languages, languages that allow moderately complex and complex types are less frequent. Simple (C)V types are by and large restricted to the equatorial belt (West and Central Africa, Papua New Guinea, the area in or to the north of the Amazon basin; Figure 1; note there is a typologically interesting correlation with small overall consonant inventories).

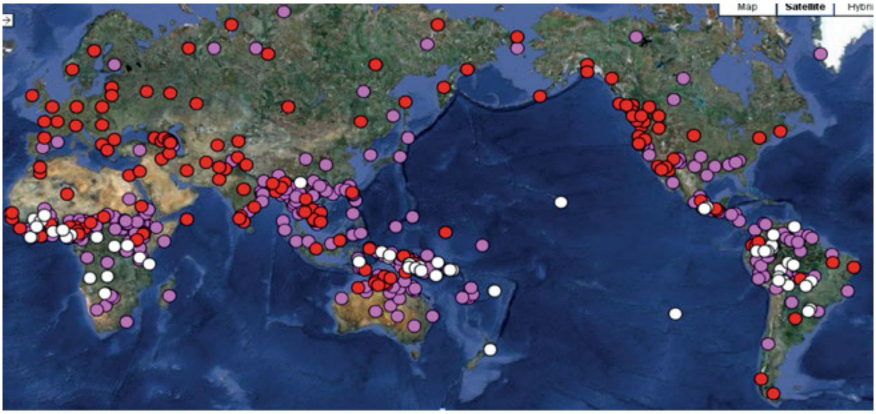

Figure 1: The distribution of phonotactically simple/complex systems around the world (simple [white] – 64, moderately complex [pink] – 275, complex [red] – 151)

English belongs to the *complex* category, which means that its phonotactic system admits clusters in both syllable onsets and codas: (C)(C)(C)V(C)(C)(C)(C), e. g. /spr-/ in *spring* or /skr-/ in *scratch*, /-mps/ in *glimpse*. These clusters are highly variable and subject to a contextually sensitive deletion process: consonant cluster reduction (CCR), also referred to as "word-final stop deletion" (Guy

1980). CCR typically affects cluster-final plosives (/t/ in *fast*, or /k/ in *flask*) and operates in English universally ((CC) > (C), (CCC) > (CC); (Schreier 2005); CCR is found in all speakers (regardless of whether they are mono-, bi- or multilingual), in all regional, social and ethnic varieties and all speech styles (casual, careful, formal). It has been particularly well researched in varieties of American English, for instance in Tejano English (Bayley 1995; Santa Ana 1996), Appalachian English (Wolfram and Christian 1976), Philadelphia English (Guy 1980), Lumbee English (Torbert 2001), but also in York/UK English (Tagliamonte and Temple 2005), Indian English (Khan 1991), mesolectal Jamaican Creole English (Patrick 1991) and New Zealand English (Holmes and Bell 1994; Schreier 2003).

While all varieties of English variably reduce syllable-coda clusters to at least some extent (see Schreier 2005), they differ in frequency and the factors that condition CCR application. Looking at total reduction rates in varieties of English around the world (Table 1), we note that varieties that have undergone recent dialect contact (such as New Zealand English) differ very little from long-term established varieties that have merged via koinéization (thus, in Trudgill's view, traditional, regional non-standard varieties with a long tradition of mother-tongue speakers). It is striking indeed that York English, White American and White New Zealand English have almost identical overall reduction rates (between 24 and 29 per cent).

At the same time, the varieties as a whole differ extensively in how often CCR applies, which, however, is not a function of contact histories and definitely not the consequence of high vs. low contact. Table 1 suggests that varieties fall into several categories. The first group is made up of those that have moderately low rates: they have either long-standing historical continuity of speakers and natural transmission with little contact (Yorkshire, eastern USA) or else undergone extensive dialect contact in colonial settings (New Zealand). Language shift varieties with a high percentage of English as L2 speakers form a separate group (Chicano English), whereas those that underwent heavy language contact and/or creolization have extremely high CCR rates (Vietnamese English, St Helenian English, Black Bahamian English, mesolectal Jamaican Creole English).

One variety that merits special attention is African American English (AfrAmE). Its origins and evolution have been debated prominently (Winford 1997, 1998; Poplack and Tagliamonte 2001; Wolfram and Thomas 2002). All studies consulted show that AfrAmE has *higher* CCR rates than white cohort varieties, which has been interpreted as earlier language contact and a subsequent hold-over effect of a substratum feature (or more precisely: the adoption of CVC syllable types from African languages). Consulting the WALS, one finds that West Africa, from where the vast majority of the slaves originated, has an unusually high percentage of languages with simple syllable types (Figure 2).

Table 1: CCR rates in 14 varieties of English (adapted from Schreier 2005)

Typology	Varieties and studies	Total CCR (100 % means reduction of all clusters)
Group 1 long-standing historical continuity of speakers and natural transmission with little contact, dialect contact in colonial settings	York English (Tagliamonte and Temple 2005)	24 %
	Pakeha New Zealand English (Schreier 2003)	27.8 %
	Philadelphia English (Neu 1980)	28.2 %
	White Hyde County NC English (Wolfram and Thomas 2002)	28.8 %
Group 2 language shift varieties with a high percentage of English as L2 speakers	African American English, Washington DC (Fasold 1972)	40.2 %
	Texas Tejano English (Bayley 1995)	48 %
	Los Angeles Chicano English (Santa Ana 1996)	52 %
Group 3 heavy language contact and/or creolization	Early Maori New Zealand English (Schreier 2003)	66.5 %
	African American English, Hyde County NC (Wolfram and Thomas 2002)	67.2 %
	Mesolectal Jamaican Creole English (Patrick 1991, 1999)	72.3 %
	St Helenian English (Schreier 2008)	86.5 %
	Black Bahamian English (Childs, Reaser, and Wolfram 2003)	87.6 %
	Tristan da Cunha English (Schreier 2005)	87.8 %
	Early Vietnamese English (Wolfram et al. 1986)	ca. 92 %

Most West African languages have either simple or moderately complex syllable structures, and languages with large speaker groups, such as Yoruba and Igbo, are exclusively simple (Maddieson 2011). Several of these are spoken along the coast, the slaves' areas of origin.

The analysis of CCR demonstrates that high contact entails large-scale simplification, as a result of which high-contact varieties such as creoles or shift-varieties have high CCR rates. However, and this is the flip side of the coin, there is evidence that contact intensity itself is of secondary importance when compared with the phonotactic properties of the systems in contact. It is the similarities/differences of syllable types in the languages (or dialects) that is a more important criterion. This is in line with Schreier's (2005) second principle of CCR: "Contact between systems with similar or identical phonotactic systems does not lead to phonotactic simplification. CCR remains stable in dialect-contact situations and is not modified during koinéisation" (Schreier 2005: 200).

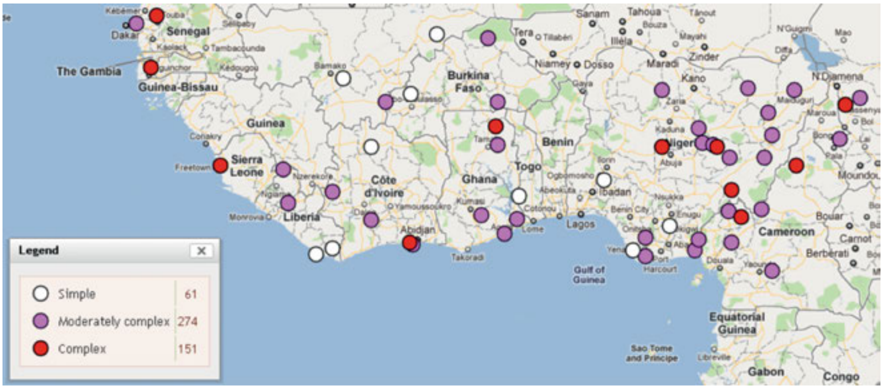

Figure 2: Simple syllable types in the West African input varieties to African American English

Consequently, it is not the intensity of contact that matters the most, but the systemic and typological differences between the varieties in contact. High contact between (C)(C)(C)V(C)(C)(C) varieties may have little effect (if any), as witnessed in White New Zealand English, whereas less intense contact with a (C)V language is likely to have a major effect on the phonotactic system of the newly developing variety. Similarity is thus of primary importance, contact intensity is secondary overall.

Therefore, a claim of the type "high contact equals simplification" is correct for cases of contact between *distinct* phonotactic systems (evident in creolization, for instance); however, I would argue that it is only correct in such circumstances. The equation is wrong in cases of high contact between *similar* phonotactic systems, for instance when there is dialect contact and koinéization. Whether or not such an assessment is accurate primarily depends on the nature of the systems in contact and not for how long and how intensely they are in contact (the same may be true in language shift conditions, where the degree of typological similarity is an influential factor).

2.2 Low contact = complexification

Turning to Trudgill's alternative hypothesis, 'low contact leads to complexification', we already saw above that it can to some extent be supported by varieties of English around the world. However, it is not generally valid and there is counterevidence, for instance in variation of personal pronoun systems. Table 2 shows that the standard system, reported in practically all the major varieties, has person and number concord (singular vs. plural, 1st, 2nd and 3rd person). The one noteworthy exception is the 2nd person, where *you* is used both for singular and plural. The referent of a question such as "where are you?" is therefore not morphologically coded (needless to say, the context of the utterance nearly always clarifies this so that genuine misunderstandings occur rarely; grammatically, however, the uncertainty exists). Consequently then, the fact that sing./pl. forms are identical means that they are regular and that they can easily be learnt. *You* has a natural advantage in language learning situations and one would expect to find it all over the English-speaking world, particularly in high-contact varieties.

Table 2: Personal pronouns in standard varieties of English

Person	Singular	Plural
1st	I	We
2nd	You	You
3rd	He/she/it	They

In the Caribbean, one of the hot spots of creolization in English, there are frequent reports of variation in 2nd pers. pronouns. In fact, local varieties differ from the standard system in that more than half a dozen alternative forms are in use. Table 3 lists some of the more common ones.

Table 3: 2nd plural forms in the Caribbean (Allsopp 1996; Mühleisen 2005; Patrick and Holm 2008)

Region	2nd p. plural form reported
Guyana	All a yu
Anguilla and Barbuda, Antigua, Barbados, Guyana, St Kitts, Trinidad	Aa-yu/all yuh
Guyana	A(ll)-yo-dis
Carriacou, Grenada	Among-you
Jamaica	Ono
Belize, Cayman Islands, Jamaica, Tobago	Unu
Barbados	Unna
Jamaica	Unno
Nevis	Yu aal
Bahamas	Yinnah

At the same time, one notes that there is sociolinguistic variation within the varieties, so that in Jamaican Creole English we find *ono* (Baker and Huber 2001), *unno*/*onnoo* (Baker and Huber 2001), *unu*, *uno*, *una* (Cassidy and Le Page 1980).

On Barbados, we find *wunna* (Allsopp 1996), *unna* (Baker and Huber 2001) and *all you* (Baker and Huber 2001), *all you* being most recent historically. These forms are subject to social stratification, placed at acrolectal/basilectal ends of the continuum, thus attesting to the "messy face" of contact linguistics as discussed above.

Similar findings are reported in Pacific creoles. Varieties such as Tok Pisin and Bislama distinguish between *yumi* (inclusive 1st plural) and *mipela* (exclusive 1st pl.), or *yumitupela* vs. *mitupela* (Verhaar 1995: 19). Here too we find that a regular paradigm has become more complex due to the inclusion of inclusive vs. exclusive information and more number distinctions. This is nothing new and it is well-known why a more complex and irregular pattern was adopted in high-contact scenarios. The distinction is made in many Polynesian languages: there is a separate morphological encoding of singular vs. plural, inclusive vs. exclusive, and also two vs. three or more addressees. These patterns are then adopted in the newly emerging English creole varieties as a result of persistent substratum influence (resembling the long-term effects of phonotactic simplification in AfrAmE, as discussed above). It goes without saying that this poses a real challenge to the simplification claims above.

Of course, one might seek alternative explanations and look at variation within personal pronouns systems in low-contact varieties that may have served as inputs in the respective scenarios. It is well-known that dialects of British English have separate plural forms. *Yous* (Northern Ireland), *all you*, *you together* (East Anglia; Hughes, Trudgill, and Watts 2005), for instance, so one might argue that standard forms are not relevant since the majority of inputs had alternative (non-standard) features. However, the form *yous*, one of the better known and more frequently used pronouns, is hardly used outside the British Isles, let alone in English-based creoles. It would be a long stretch indeed to claim that it may have triggered a variable pronoun system with more than six different forms in the Caribbean or throughout the Pacific.

To sum up: by and large, low-contact varieties have comparatively regular personal pronoun systems, which is indicative (and which one would typically expect in cases) of simplification. On the other hand, high-contact varieties have considerably more complex personal pronouns systems (encoding alternative information: exclusive vs. inclusive, duality). As a result of substratum influence, they display complexification and thus behave more like low-contact varieties (and are in fact more complex). We thus find evidence against both the "low contact equals complexification" and the "high contact equals simplification" claims, if we wish, and they are not generally valid. The case becomes even more complicated when looking at a variety that is at the same time simplified and complexified: Tristan da Cunha English.

3 Tristan da Cunha English: when simplification and complexification co-occur

Tristan da Cunha is one of the most unusual offspring varieties of English around the world. Its social history and sociolinguistic evolution have been described in great detail elsewhere (Schreier 2003, 2010), so I will focus only on the most important characteristics of the community. The island was settled in 1816 under *tabula rasa* conditions, so Tristan da Cunha English is a very young variety in the canon of World Englishes. There was no pre-existing population. The origins of the founding members of the population (Mufwene 2012) are well-known: Kelso (Scottish Lowlands), London, Hull, Hastings, Table Bay area, St Helena, and there were less influential short-term members as well (from Plymouth and Denmark). Later on, settlers arrived from the Netherlands, Connecticut and Italy, though not all of them were equally relevant sociolinguistically (date of arrival, length of stay, etc.). The community grew gradually with abrupt growth following the arrival of the St Helenian group in 1827, only to fall into a state of near-total isolation from 1850 to WWII. Hardly any new settlers arrived in this period – population growth was due to a high birth rate – and the contact with the outside world diminished (Tristan is geographically hyper-isolated, 3'000 km distant from Cape Town, the closest major harbor). Isolation ended when a British garrison was installed on the island in 1942, and the community opened up in the second half of the 20th century (Schreier 2003).

Given these unusual sociolinguistic characteristics, it is by no means easy to assess the contact status of Tristan da Cunha English. Cooperating on the *World Atlas of Variation in English* (WAVE) project, the question was raised repeatedly whether it was high or low. In my opinion, such a distinction was not possible and I pointed out that, with reference to Trudgill's (2009) taxonomy (low contact found in "traditional, regional non-standard varieties with a long tradition of mother-tongue speakers (and natural transmission)" and high contact in "transplanted L1 Englishes or colonial standards"), one could make a case in point that Tristan da Cunha English was in fact both, depending on how we argue. The initial 1816–40 phase saw extensive contact of different varieties of British and Connecticut (American) English, perhaps also early South African English (Afrikaans substratum?), and at least two L2 varieties. Adults arrived, dozens of children were born, the entire community lived in one settlement with frequent interaction – a clear case of high contact, no doubt. However, if we focus on the period after 1840, when there was practically no immigration and very little contact with the outside world, then conditions were ideal for koinéization and norm-crystallization, and TdCE can only be classified as "low contact", and Chambers (2004) argues in favor of such an interpretation:

> [S]ocial conditions [on Tristan da Cunha] were decidedly eccentric, with total absence of mobility, almost no contact with outside groups and no immigration for several generations. Because there was little formal education and limited literacy, there was not even a codified standard language to measure the vernacular against. (Chambers 2004: 135)

This is as close to a low-contact variety as one can possibly get.

There is thus good reason to classify Tristan da Cunha English either alongside the English Southwest, East Anglia, Newfoundland English, Appalachian English and Ozark English (and thus as "low contact"), or with New Zealand English, White South African English, Bahamian English and Falkland Islands English ("high contact") so that Tristan da Cunha English meets in fact both criteria. Table 4 summarizes some of the major criteria that have to be taken into consideration, illustrating the dilemma at hand.

Table 4: Contact in Tristan da Cunha English: high or low?

Criterion	**High or low contact?**
Contact in formation phase (1816–1840)	High
Contact in consolidation phase (1840–1940)	(extremely) Low
Dense and multiplex networks	High
Local mobility	(extremely) Low
Contact with outside varieties	Low
Mutual intelligibility of inputs	(probably) High
Low speaker numbers	?

The distinction is further blurred when looking at processes of complexification and simplification in the variety. For one, there are various prime candidates of *simplification*. Perhaps the best example here is the unusual process of categorical leveling in the present and past *be* paradigms (documented by Schreier 2003). Leveling mechanisms of this kind have been researched in varieties around the world and particularly the pivot *was* (in *we was*, *the girls was*, *there was two girls*, etc.) has been identified as a "vernacular root", "one of the primitives of vernacular dialects in the sense that they recur ubiquitously all over the world" (Chambers 2009: 258). TdCE has undergone a most radical change, induced by high contact *nota bene* (Schreier 2003), as a result of which both paradigms are perfectly regular, *is* and *was* being used with all grammatical persons and in all environments. According to Chambers (2004), this is

> a breakthrough because the default singulars in their pure basilectal form there are palpable […] they came into being as a natural regression to primitive linguistic instincts, a conclusion that follows from an apparent-time inference of very shallow time depth and therefore seems virtually incontrovertible. (Chambers 2004: 139)

Other examples of simplification brought about by high contact are the (very) high percentage of CCR (87.8 %; see above), which is the legacy of St Helenian English brought to the island in the 1820s, and we might also add to this category the vowel merger between LOT and THOUGHT which can be traced to Eastern New England and Scotland, thus matching the founders' origins (cf. also Schreier and Trudgill 2006). As a result, TdCE shows a clear tendency towards simplification in morphosyntax and in phonology; it thus resembles high-contact varieties.

However, opposite complexification trends have been documented as well. For one, there is a partial redistribution and overlap of the LOT, THOUGHT and NORTH lexical sets, which "clearly demonstrates the complexity of dialect contact" (Schreier and Trudgill 2006: 138). Moreover, TdCE may be the only dialect of English in the world that has four possible realizations for Standard English dental fricatives /θ/ and /ð/. We do find dental fricatives, though they do not occur often, labiodental fricatives, as in Southeastern English English (TdCE usually has them at the end of stressed syllables), labiodental plosives, as in AfrAmE and English-derived creoles (TdCE has them in all positions) and sibilants often found at (stressed) syllable onsets, so-called "TH sibilization" (Schreier 2003: 211). The interplay of /θ ~ ð/, /t ~ d/, /f ~ v/ and /s/ is arguably the most complex set known (though it is possible, as Ray Hickey, p. c. 2012, reminds me, that TH sibilization is lexicalized and only found in a certain number of lexical items). The constraints of variation will only be understood when further quantitative research is carried out in a representative sample of the TdCE-speaking community, so this will certainly be subject to further research. For the time being, however, it suffices to state that where many varieties of English have one pair of phonemes (voiced ~ voiceless), TdCE has four different realizations.

In our analysis of TdCE segmental phonology, Schreier and Trudgill (2006) came to the conclusion that

> no single dialect served as a model per se for the first generations of native Tristanians. Even though TdCE adopted a number of features that resemble English-based Creoles, we nevertheless find features that are undoubtedly of British English origin, i. e. structures that originally stem from the dialects of English spoken by the British founders of the community. It is thus the mixture of features that accounts for the uniquely distinctive blend of a local phonological system. These selection and retention mechanisms were complemented by additional developments (substratal, interdialectal, etc.) which further shaped the status of TdCE as an independent variety of English. The system of segmental phonology found in TdCE is thus the result of linguistic contact, feature selection and retention from several inputs as well as of internal linguistic changes. (Schreier and Trudgill 2006: 138)

One should also add that, due to its geographic hyper-isolation, TdCE is both a conservative and a dynamic dialect. It has retained some very conservative features of British English (THOUGHT in words like *off*, *across*, *froth*, etc., genuine monophthongs in FLEECE and GOOSE, no long mid diphthonging in FACE, no diphthong shift in GOAT, merger of /w/ and /v/, STRUT in *first*, to name a few) while at the same time undergoing some dynamic changes (e. g. an overlap of the LOT, THOUGHT and NORTH lexical sets, LOT in *rope* and *fork*, THOUGHT in *bottle* etc., STRUT in *fur*, VVV > VCV in English English /ɑuə/ triphthongs, e. g. *our* /hʌbə/, *flower* /flʌbə/, which "presumably comes from */ɑuwə/ and is perhaps best described as a case of a language contact feature which was later subjected to independent development" (Schreier and Trudgill 2006: 138). The retention of older features in an archaic, conservative system is characteristic of low contact conditions, the creation and development of new features is typically expected in high contact conditions.

To surmise, quantitative and qualitative analyses (both in morphosyntax and phonology) allow the conclusion that TdCE has undergone both far-reaching simplification and considerable complexification. Moreover, it is at the same time conservative, displaying colonial lag tendencies, and dynamic in that it has undergone processes of local innovation. How are we to integrate this into a coherent model of language or dialect contact?

4 Conclusion

The above analysis has concentrated on a number of varieties around the world, varieties that differ in terms of settlement patterns and founder effects, contact histories or time depth. I am happy to concede that one would have to analyze (quantitatively, if possible) a much larger set of features in more varieties (*Atlas of Varieties of English around the World*, Kortmann et al. 2004, or the more recent *World Atlas of Variation in English*) than was possible here. Still, the findings presented and discussed here allow me to draw some general conclusions and also to critically evaluate commonplace claims on complexity and simplification.

First of all, as should be clear by now, I am rather pessimistic in what regards a simple binary distinction between "high contact" and "low contact" varieties. As is well-known, attempts have been made in the past to correlate contact intensity effects with the strength of social and cultural exposure. Most notably, Thomason and Kaufman (1988: 76–94) suggest a five-step model ranging from (1) casual contact; to (2) slightly more intensive contact; (3) more intense contact; (4) strong cultural pressure; and finally (5) very strong cultural pressure. For stage 1, they claim that "[w]ith a minimum of cultural pressure we expect only lexical borrowing, and then only nonbasic vocabulary" (Thomason and Kaufman 1988: 77). Scenario (5), on the other hand, involves major structural changes and typological disruption, strongly affecting phonology (morphophonemic rules and phonetic changes, loss of phonemic contrasts/morphophonemic rules), morphology (changes in word structure rules, e. g., the adoption of prefixes in suffixal languages, inflectional over agglutinative tendencies), categorical, ordering changes in morphosyntax (added concord rules, bound pronominal elements; Thomason and Kaufman (1988: 88)). In other words, the degree of complexity in these circumstances is a consequence of factors such as input properties, weak vs. strong contact patterns, local settlement patterns, sociocultural pressures, feature pools, identity and prestige, etc. (see also Schneider 2007). I am not convinced at all that a classification into two categories can do justice to this.

Second, such an approach suffers from the fact that it does not take time depth into consideration. Particularly when interested in long-term effects of contact situations, it is imperative to include a diachronic perspective (see the *electronic World Atlas of Variation in English* (Kortmann and Lunkenheimer 2013), where time depth finally became one of the criteria for language-shift varieties [high-contact varieties "over the last 400 years", thus roughly since the foundation phase of American English]). The study of isolated varieties, TdCE being but one example here, shows that isolation effects may override contact effects over time. Imagine the following scenario: a contact variety is formed under high-contact conditions in the 18th century, with extreme diffuseness and perhaps also a random mix of features from various inputs (as attested in New Zealand; Gordon et al. 2004). After the first two generations, a phase of sociolinguistic stability sets in, there is consistent birth growth, children acquire the dialect natively (see above) and leveling and focusing operate (under low-contact conditions). Consequently, after a few generations of stabilization in isolation, it is certainly plausible that complexification trends set in, which in turn supersede initial ones.

A good example here is East Anglian English. Trudgill, who has studied it for more than 40 years now (Trudgill 1974, 1998, 2010), classifies it as a "low contact variety". Indeed, EAE is regionally isolated and mostly rural, an area with a long settlement history and sociolinguistic stability.

On the other hand, we know that EAE has not always been among the traditional, regional non-standard varieties with a long tradition of mother-tongue speakers (and natural transmission) in Trudgill's view. Trudgill himself has documented and studied the effects of immigration of Flemish weavers and workers in the 16th century (Trudgill 1998, 2010), who fled from religious prosecution in the Low Countries only to resettle across the Channel in Norwich and its surroundings. This led to extensive contact between Flemish and EAE, and Trudgill (1998) has argued that one of the most striking features of EAE morphosyntax, 3rd person singular present tense zero, e. g. in "she singØ really loud", originated at least in part as contact-derived simplification of an irregular paradigm. The immigrants from the Low Counties, as non-native speakers, accelerated an ongoing change (see Trudgill 2010 for a detailed discussion). EAE underwent phases of high contact (with Dutch) with lasting sociolinguistic effects, and this raises several questions: if EAE resembled a high-contact variety in the 17th century, is it still justified to classify it as a low-contact variety today? Is it not much more likely that what is truly crucial here is time depth, which means that the transmission over 12–15 generations has in a way superseded (or even eradicated) earlier effects? Other models do recognize the factor of time, Schneider's (2007) influential model for instance, so would one not integrate this with benefit here also? While a distinction between "high" and "low" may have some synchronic validity (if we can find and agree on generally binding characteristics for these two characteristics, that is), there is strong evidence that they change across time and are subject to diachronic flexibility. Contact is not equal across time, it is more intense at times only to decrease again, and models must include historical information and be diachronically informed.

Finally, the discussion of Tristan da Cunha English has shown that it is certainly possible for varieties to simultaneously have both tendencies. These emerge more clearly here, due to the well documented social history, and can be backed up by extensive information on settlement patterns and founders' origins (Schreier 2008). The parallel adoption and/or development of complex and simple features (discussion in Schreier 2003) is problematic for the model; a simple classification of "high" vs. "low" or "complexification" vs. "simplification" (Figure 3) can not be upheld when varieties have both at the same time.

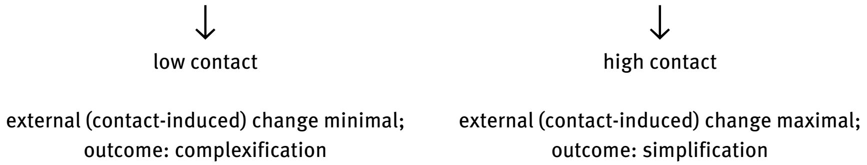
low contact
high contact
external (contact-induced) change minimal;
outcome: complexification
external (contact-induced) change maximal;
outcome: simplification

Figure 3: A model of high vs. low contact (following Trudgill 2009)

As an alternative (and influenced by Thomason and Kaufman's (1988) scalar approach discussed above), I would suggest that it is more promising to view – and attempt to model – contact-induced change as a continuum, ranging from minimal contact effects (borrowing of lexical items) to maximal ones (syntactic reordering, addition of morphological elements; e. g., Russian verb inflections in Copper Island Aleut, Altaic case suffixes in Xinjiang Chinese, Thomason and Kaufman 1988). Modeling and placing them on a continuum (Figure 4) would help us portray the complex character of contact-induced change more clearly and put linguists in a position to attribute features at various points along the continuum, thus accounting for simultaneous occurrence of features.

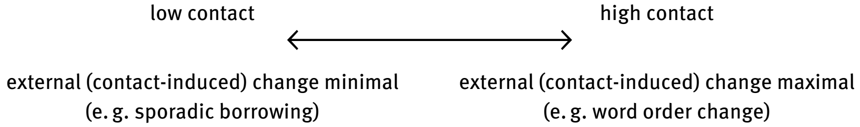
low contact
high contact
external (contact-induced) change minimal
(e. g. sporadic borrowing)
external (contact-induced) change maximal
(e. g. word order change)

Figure 4: A revised model: contact as a continuum

The major advantage of this model is that it is more open and flexible. Since no general distinction is made between the two (separate) categories, it is more suitable to account for the fact that varieties are simple in some ways (consonant cluster reduction) and complex in others (pronouns), such as Tristan da Cunha English (but one could also include Bahamian English, Bislama and no doubt other varieties).

At the same time, viewing contact as non-discrete allows for a more detailed feature-based analysis across separate systems (simplification in morphosyntax but not in phonology, for instance). Third, it is in line with continua-based approaches that have been applied with success elsewhere in contact linguistics, such as lectal continua in creolistics or stylistic continua in variationist sociolinguistics. As a result, we may resolve some current controversies that seem irreconcilable, for instance the ultimate outcome of language contact, as evidenced by the following two quotes:

> [C]ommunities involved in large amounts of language contact, to the extent that this is contact between adolescents and adults who are beyond the critical threshold for language acquisition, are likely to demonstrate linguistic pidginisation, including simplification, as a result of imperfect language learning. (Trudgill 2004: 306)

> [A]lmost all of these high-complexity languages are in areas of considerable linguistic diversity and contact. It can be concluded that contact among languages fosters complexity, or, put differently, diversity among neighbouring languages fosters complexity in each of the languages. (Nichols 1992: 192)

To conclude, I suggest that, as a consequence of the spread of English around the world and the embedding of the language in a myriad of diverse contact settings, there has been a *typological splitting*, not a typological split. I fully agree with Hurford (2011) in that “a good generalization about whether contact induces simplification or complexification is elusive. Just mentioning instances, one way or the other, won’t settle the general case (if there even is one)” (Hurford 2011: 479), which is strong evidence against a “true typological split” indeed. Contact linguistics, no doubt, is too complex for simple, binary classifications.

References

Allsopp, Richard 1996: *Dictionary of Caribbean English Usage*. Oxford: Oxford University Press.

Bayley, Robert 1995: Consonant cluster reduction in Tejano English. *Language Variation and Change* 6: 303–327.

Baker, Philip and Huber, Magnus 2001: Atlantic, Pacific, and world-wide features in English lexicon contact languages. *English World-Wide* 22 (2): 157–208.

Cassidy, Frederic Gomes and Le Page, Robert B. 1980: *Dictionary of Jamaican English*. Cambridge: Cambridge University Press.

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

Chambers, J. K. 2009: *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Chichester, U. K./Malden, MA: Wiley-Blackwell.

Childs, Becky, Reaser, Jeffrey and Wolfram, Walt 2003: Defining ethnic varieties in the Bahamas: Phonological accommodation in black and white enclave communities. In: Michael Aceto and Jeffrey P. Williams (eds.), *Contact Englishes of the Eastern Caribbean*, 1–28. Amsterdam/Philadelphia: Benjamins.

Fasold, Ralph 1972: *Tense Marking in Black English: A Linguistic and Social Analysis*. Arlington, VA: Center for Applied Linguistics.

Gordon, Elizabeth, Campbell, Lyle, Hay, Jennifer, Maclagan, Margaret, Sudbury, Andrea and Trudgill, Peter 2004: *New Zealand English: Its Origins and Evolution*. Cambridge: Cambridge University Press.

Görlach, Manfred 1990: *Studies in the History of the English Language*. Heidelberg: Winter.

Guy, Gregory 1980: Variation in the group and in the individual: The case of final stop deletion. In: William Labov (ed.), *Locating Language in Time and Space*, 1–36. New York: Academic Press.

Holmes, Janet and Bell, Allan 1994: Consonant cluster reduction in New Zealand English. *Wellington Working Papers in Linguistics* 6: 56–82.

Hughes, Arthur, Trudgill, Peter and Watt, Dominic 2005: *English Accents and Dialects: An Introduction to Social and Regional Varieties of English in the British Isles*, 4th edition. London: Hodder Arnold/New York: Oxford University Press.

Hundt, Marianne 2012: The diversification of English: old, new and emerging epicentres. In: Daniel Schreier and Marianne Hundt (eds.), *English as a Contact Language*, 182–203. Cambridge: Cambridge University Press.

Hurford, James R. 2011: *The Origins of Grammar: Language in the Light of Evolution*. Oxford: Oxford University Press.

Kachru, Braj B. 1985: Standards, codification, and sociolinguistic realism: The English language in the outer circle. In: Randolph Quirk and Henry G. Widdowson (eds.), *English in the World*, 11–30. Cambridge: Cambridge University Press.

Kachru, Braj B. 1986: *The Alchemy of English: The Spread, Functions and Models of Non-native English*. Oxford: Pergamon Press.

Kerswill, Paul 1996: Children, adolescents and language change. *Language Variation and Change* 8: 177–202.

Kerswill, Paul 2001: Koineization and accommodation. In: Jack K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds.), *A Handbook of Language Variation and Change*, 669–702. Oxford: Blackwell.

Khan, Farhat 1991: Final consonant cluster simplification in a variety of Indian English. In Jenny Cheshire (ed.), *English around the World: Sociolinguistic Perspectives*, 288–298. Cambridge: Cambridge University Press.

Kortmann, Bernd, Burridge, Kate, Mesthrie, Rajend, Schneider, Edgar W. and Upton, Clive (eds.) 2004: *A Handbook of Varieties of English*. Berlin/New York: Mouton de Gruyter.

Kortmann, Bernd and Lunkenheimer, Kerstin (eds.) 2013: *The electronic World Atlas of Varieties of English [eWAVE]*, <http://www.ewave-atlas.org/> (18 September 2013).

Maddieson, Ian 2011: Feature 12A: Syllable structure. In: *The World Atlas of Linguistic Structures* (*WALS*, http://wals.info/feature/12A; accessed September 26 2012).

McArthur, Tom 1987: The English Languages? *English Today* 12: 21–24.

Mufwene, Salikoko S. 2012: Driving forces in English contact linguistics. In: Daniel Schreier and Marianne Hundt (eds.), *English as a Contact Language*, 204–222. Cambridge: Cambridge University Press.

Mühleisen, Susanne 2005: Forms of address in English-lexicon Creoles: The presentation of selves and others in the Caribbean context. In: Susanne Mühleisen and Bettina Migge (eds.), *Politeness and Face in Caribbean Creoles*, 195–223. VEAW G 34. Amsterdam/Philadelphia: Benjamins.

Neu, Helene 1980: Ranking of constraints on /t, d/ deletion in American English: A statistical analysis. In: William Labov (ed.), *Locating Language in Time and Space*, 37–54. New York: Academic Press.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Patrick, Peter L. 1991: Creoles at the intersection of variable processes: -t, d deletion and past marking in the Jamaican mesolect. *Language Variation and Change* 3: 171–189.

Patrick, Peter L. 1999: *Urban Jamaican Creole: Variation in the Mesolect* (Varieties of English around the World G17). Amsterdam/Philadelphia: Benjamins.

Patrick, Peter L. and Holm, John A. (eds.) 2008: *Comparative Creole Syntax: Parallel Outlines of 18 Creole Grammars*. London: Battlebridge Press.

Poplack, Shana and Tagliamonte, Sali 2001: *African American English in the Diaspora*. Oxford: Blackwell.

Santa Ana, Otto 1996: Sonority and syllable structure in Chicano English. *Language Variation and Change* 8: 63–91.

Schneider, Edgar W. 2007: *Postcolonial English: Varieties Around the World*. Cambridge, UK: Cambridge University Press.

Schreier, Daniel 2003: *Isolation and Language Change: Contemporary and Sociohistorical Evidence from Tristan da Cunha English* (Palgrave Studies in Language Variation 1). Houndmills, Basingstoke/New York: Palgrave Macmillan.

Schreier, Daniel 2005: *Consonant Change in English Worldwide: Synchrony meets Diachrony*. (Palgrave Studies in Language History and Language Change 3) Houndmills, Basingstoke/New York: Palgrave Macmillan.

Schreier, Daniel 2008: *St Helenian English: Origins, Evolution and Variation*. (Series Varieties of English Around the World G37) Amsterdam/Philadelphia: Benjamins.

Schreier, Daniel 2009: Language in isolation, and its implications for variation and change. *Blackwell Language and Linguistics Compass* 3: 682–699.

Schreier, Daniel 2010: The consequences of migration and colonialism II: Overseas varieties. In: Peter Auer and Jürgen E. Schmidt (eds.), *Language and Space: An International Handbook of Linguistic Variation*, 451–467. Berlin/New York: Walter de Gruyter.

Schreier, Daniel 2012: English as a contact language: Lesser-known varieties. *English as a Contact Language*. In: Daniel Schreier and Marianne Hundt (eds.), *English as a Contact Language*, 149–164. Cambridge: Cambridge University Press.

Schreier, Daniel 2014: On cafeterias and new dialects: the role of primary transmitters. In: Sarah Buschfeld, Thomas Hoffmann, Magnus Huber and Alexander Kautzsch (eds.), *The Evolution of Englishes: The Dynamic Model and beyond,* 231–48. Amsterdam: Benjamins.

Schreier, Daniel and Trudgill, Peter 2006: The segmental phonology of 19th century Tristan da Cunha English: Convergence and local innovation. *English Language and Linguistics* 10: 119–141.

Tagliamonte, Sali and Temple, Rosalind 2005: New perspectives on an ol' variable: (t, d) in British English. *Language Variation and Change* 17: 281–302.

Thomason, Sarah G. and Kaufman, Terrence 1988: *Language Contact, Creolization, and Genetic Linguistics*. Los Angeles: University of California Press.

Torbert, Benjamin 2001: Tracing Native American language history through consonant cluster reduction: The case of Lumbee English. *American Speech* 76: 361–387.

Trudgill, Peter 1974: *The Social Differentiation of English in Norwich*. Cambridge, UK: Cambridge University Press.

Trudgill, Peter 1986: *Dialects in Contact*. Oxford: Blackwell.

Trudgill, Peter 1998: Third person singular zero: African American Vernacular English, East Anglian dialects and Spanish persecution in the Low Countries. *Folia Linguistica Historica* 18: 139–148.

Trudgill, Peter 2004: *New Dialect Formation: The Inevitability of Colonial Englishes*. Edinburgh: Edinburgh University Press.

Trudgill, Peter 2009: Vernacular universals and the sociolinguistic typology of English dialects. In: Markku Filppula, Juhani Klemola and Heli Paulasto (eds.), *Vernacular Universals and Language Contacts: Evidence From Varieties of English and Beyond*, 304–322. New York: Routledge.

Trudgill, Peter 2010: *Investigations in Sociohistorical Linguistics: Stories of Colonisation and Contact*. Cambridge: Cambridge University Press.

Trudgill, Peter 2011: *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Trudgill, Peter, Gordon, Elizabeth, Lewis, Gillian and Maclagan, Margaret 2000: Determinism in new-dialect formation and the genesis of New Zealand English. *Journal of Linguistics* 36: 299–318.

Verhaar, John W. M. 1995: Towards a reference grammar of Tok Pisin: An experiment in corpus linguistics. In: *Oceanic Linguistics Special Publication* 26. Honolulu: University of Hawaii Press.

Winford, Donald 1997: On the origins of African American Vernacular English – a creolist perspective. Part I: The sociohistorical background. *Diachronica* 14 (1), 304–344.

Winford, Donald 1998: On the origins of African American Vernacular English – a creolist perspective. Part II: Linguistic features. *Diachronica* 15 (1): 99–154.

Wolfram, Walt and Christian, Donna 1976: *Appalachian Speech*. Arlington, VA: Center for Applied Linguistics.

Wolfram, Walt, Christian, Donna and Hatfield, Deborah1986: The English of adolescent and young adult Vietnamese refugees in the United States. *World Englishes* 5: 47–60.

Wolfram, Walt and Thomas, Erik R. 2002: *The Development of African American English*. Oxford: Blackwell.

Peter Trudgill, Agder University

The sociolinguistics of non-equicomplexity

Abstract: If some languages are more complex than others, then the question arises as to why this is. How does it come about that complexity is not evenly distributed across human languages? How is it that some languages are, in some important sense, simpler than others? Sociolinguists have for a long time had an answer to this question: if language contact leads to simplification, then the more contact a language has experienced, the less complex it will be. But from work in sociolinguistic typology, it has also become clear that we are not just talking here about just any old form of contact. There are also forms of contact which lead to additive borrowing and thus additional complexity. And what about spontaneous complexification? This paper examines the nature of the sociolinguistic conditions which are most favorable to internal linguistic complexity development.

1 Introduction

The notion that all languages have an equivalent degree of complexity was at one time very much part of the conventional wisdom of the linguistics community. And that view was, understandably, particularly strongly maintained in the face of a nonspecialist public who still held to the view that some languages really were more "primitive" than others.

Faarlund (2010), it is true, has disputed the pervasiveness of this view amongst linguists; he writes of Sampson, Gil, and Trudgill (2009) – a book which is derived from a 2007 conference at which this issue of differential complexity was extensively discussed – that it is "probably [...] pushing at an open door" (Faarlund 2010: 749). And indeed it is difficult to find many published expressions of that conventional wisdom in the literature. But the truth is most likely to be, as I personally recall, that this wisdom was so strongly entrenched in the linguistics community – and still is in some circles – that it was taken for granted and actually overtly expressed only very infrequently. One notable and often cited exception was Hockett (1958), who did actually spell it out: he wrote that "the total grammatical complexity of any language, counting both morphology and syntax, is about the same as any other" (Hockett 1958: 180–181). But he was just saying out loud what very many other linguists believed. The idea was that simplification at the level of morphology would be compensated for by complexification at the level of syntax, and vice versa.

However, this *invariance of linguistic complexity hypothesis* or *equicomplexity hypothesis* (Sampson, Gil and Trudgill 2009) – that a change in level of complexity in one part of a language is compensated for by a complexifying trade-off elsewhere – has always been implicitly rejected by certain sorts of linguists, notably those sociolinguists, creolists and dialectologists who had learnt that language contact of certain sorts led to simplification – as in the development of creoles. It had always been obvious to them that, if the same language could be more or less simple at different points in time, then different languages could be more or less simple at the same point in time. More recently the thesis has come under renewed examination, and the equicomplexity hypothesis has been quantitatively demonstrated to be unfounded by a number of writers such as Kusters (2003), Shosted (2006), and Sinnemäki (2008).

2 Simplicity

If, then, some languages are more complex than others, then the question arises as to why this is. How does it come about that complexity is not evenly distributed across human languages? How is it that some languages are, in some important sense, simpler than others? Sociolinguists, as just mentioned, have for a long time had an answer to this question. If language contact leads to simplification, then the more contact a language has experienced, then the less complex it will be.

But from work in sociolinguistic typology it has become clear that we are not just talking here about just any old form of contact. As I have pointed out (Trudgill 2001, 2002, 2009, 2010, 2011), the crucial form of contact for producing simplification is the form which leads to foreign language/dialect learning by adults and adolescents who have passed the "critical threshold" (Lenneberg 1967), beyond which the majority of human beings are incapable of learning a new variety perfectly. As Dahl (2004) says, when it comes to language learning, especially informal untutored language learning, "human children indeed seem to have an advantage compared to [...] adult members of their own species" (Dahl 2004: 294). Where such learning occurs, then, there will be a tendency for phenomena that are "L2 difficult" (Dahl 2004) to disappear. And if the demographic factors are right (Trudgill 2011), this can even have consequences for the language as a whole as spoken by native speakers – consider the simplification of Dutch in southern Africa which produced Afrikaans. But the demographic factors do have to be right – in normal situations, native speech is not influenced by non-native varieties.

Simplification, as illustrated by comparing creoles with their source languages, takes the form of a number of processes, most notably:

1. The *regularization* of irregularities: In regularization, obviously, irregularity diminishes, so that, for example, irregular verbs and irregular plurals become regular, as in the development in English of *helped* rather than *holp* as the preterit of *help*; and the replacement of *kine* by *cows* as the plural of *cow*. Regular forms are certainly less L2-difficult – less difficult to remember – than irregular ones.
2. An increase in lexical and morphological *transparency*: for example, forms such as *twice* and *seldom* are less transparent than *two times* and *not often*, and any (partial or complete) replacement of the former by the latter would represent simplification. These two factors are often linked – forms such as *cows* are more transparent or analytic, and iconic, than forms like *kine*. Any reduction in opacity also represents a reduction in L2 difficulty.
3. The loss of syntagmatic redundancy, as in the case grammatical agreement, where information is repeated. Here, reduction in redundancy will take the form of reduction of the number of repetitions, as in the loss of agreement. For L2 learners, this represents a reduction in the number of operations a speaker has to remember to carry out.
4. The *loss of morphological categories:* The loss of the morphological expression of grammatical categories may be compensated for by the use of more analytical structures, as in usage in Modern English of prepositions instead of the dative case of Old English. Or categories may be lost altogether, as in the loss of grammatical gender from Afrikaans, in which case there is simply less to learn.

Varieties which have undergone any of these processes, 1–4, are less complex than they were before they underwent those processes. It is perhaps still necessary, however, to stress that there is no evaluative dimension at all implied by the use of the term "simple" – simplification in this technical sense has no consequences whatsoever for cognition, expression, or adequacy.

3 Complexity

An important concomitant of the fact that adult language-contact leads to simplification is that absence of this type of contact will tend to lead to the maintenance of existing complexity. But we cannot suppose that differential complexity across languages is due simply to different degrees of simplification. It is clear that it is also due to different degrees of complexification. Indeed, contact between languages, if it is of the appropriate type, can actually lead to greater complexity. This is due to *additive borrowing*, in which grammatical categories are transferred from one language to another and added to the features that the language already possesses without replacing any existing features. A language which acquires an additional grammatical category from a neighboring language is clearly more complex than it was before.

But what type of contact is it that tends to lead to additive complexification? The answer is that we can expect to see the addition of complexity to languages in longterm, co-territorial contact situations which involve childhood, and therefore prethreshold and proficient, bilingualism. It is this kind of development which gives rise to the phenomenon of the Sprachbund, where languages acquire features and structures from each other. Mithun (1999) says, "strong linguistic areas are typically characterized by large numbers of small linguistic communities on good social terms. Their members are in frequent contact and often become multilingual" (Mithun 1999: 314). The length of time that may be involved in this kind of co-territorial contact is very considerable, sometimes stretching back thousands of years (Mithun 2007: 146).

But this is only additive complexity, where morphological categories are acquired from another language. What about other aspects of complexification? If we assume that complexification consists of the reverses of process 1–4 above, then where does irregularity "come from"? What are the "origins" of opacity? Under what sociolinguistic conditions does grammatical agreement develop? And what about the *language-internal* creation of morphological categories? Where, that is, does complexity come from "in the first place"?

My suggestion is in fact that linguistic complexification is most likely to develop in languages spoken in communities with certain social characteristics (see Trudgill 2011, for extended argumentation). The most favorable environment of all for complexity-development is in communities with the following constellation of social characteristics:

- low amounts of adult language contact,
- high social stability,
- small size,
- dense social networks,
- large amounts of communally-shared information.

I have to stress, though, that I am not suggesting that this constellation of features *necessarily* leads to complexification; rather, it is simply the case that these features represent something resembling a precondition for complexity development. These sociolinguistic features are *hospitable* to complexity development, but do not inevitably produce it. We cannot therefore necessarily expect to find, amongst the world's languages, any straightforward correlation between community size and complexity.

4 "Isolation"

But – if it is the case that languages spoken against a certain kind of sociolinguistic background are more likely to develop greater complexification – *why would that be?* This is a very important question, even if it is one which few linguistic-typologists have been very interested in investigating.

In attempting to answer this question, let us consider the five societal features that I have outlined. Why are they so important? Let us think about this by looking at just one type of low-contact-induced complexification, the spontaneous addition of morphological categories. What mechanisms can be involved in this kind of development? And what social contexts would favor the operation of such mechanisms, and how would they do so?

One rather remarkable example of spontaneous complexification (Trudgill 2011) occurred in certain of the traditional-dialects of a small area of the southwest of England. This was the development of a new and fascinating morphological category: the morphological marking of the difference between transitive and intransitive infinitives. Transitive infinitives were unmarked, while intransitive infinitives in these dialects were marked by the word-final morpheme *-y*, e. g. *to hit* vs. *to runny*. So in Dorset we find (Gachelin 1991: 220):

(1) *Can you zew up thease zeam?*
'Can you sew up this seam?' (Gachelin 1991: 220)
vs.
There idden many can sheary now.
'There aren't many who can shear now.' (Gachelin 1991: 220)

This rather extraordinary morphological marking of intransitivity is unparalleled anywhere else in the English-speaking world, and quite possibly unparalleled anywhere else at all – it is at the very least extremely rare in the world's languages (Trudgill 2011). My suggestion is that it is not at all a coincidence that in English this very rare development has occurred in a traditional-dialect and not in any of the larger mainstream varieties. As to why it is not a coincidence, this has to do with the nature of the mechanisms that were involved in the historical development of this additional category.

As far as these mechanisms are concerned, it does seem to be reasonably clear how the English southwestern dialect two-infinitive system came about diachronically. Phonological change was crucially implicated in what eventually became a grammatical development. The intransitive *-y* ending is considered to be a relic of the Middle English infinitive ending *-en*, later [ə] or presumably, in the southwestern dialects in question, [-i̇] (Ihalainen 1991). Originally, all infinitives would have carried this vowel. Later, as is well known, the final unstressed vowel was eventually lost in most dialects. Presumably, while this loss was occurring, there was a period of variability of the type that normally accompanies change, with alternation between older forms and newer forms with and without the final vowel, respectively. A useful hypothesis would be, then, that during this variable period the vowel was originally lost less often in utterance-final position than when another word, e. g. an object noun, followed.

Support for this proposal comes from observations of similar developments in Scandinavian dialects. Torp (2003: 249) describes how in certain Swedish dialects from Bohuslän, unstressed final vowels of infinitives preceding a direct object have been subject to reduction to [ə], while in certain other contexts, such as before an adverbial phrase, the original unreduced vowel *-a* is retained:

(2) *Ve feck järe de vi kunne*
We got to-do that we could
'We had to do what we could.' (Torp 2003: 249)
vs.
Va feck du jära där på Strândräng?
What got you to-do there at S.
'What did you have to do there at S.?' (Torp 2003: 249)

It then seems likely that it was at this stage, where variability had set in, that the southwestern English dialects in question reinterpreted the contrast between forms with and without the vocalic ending in such a way that a difference without a distinction became a difference with a distinction. The difference, which was originally phonologically conditioned and variable, came to be reinterpreted as grammatico-semantic – transitive vs. intransitive – and categorical. Crucially, it was the frequent presence (vs. absence) of following object nouns which led to this transitive (vs. intransitive) reinterpretation of forms without *-y*.

This development was a form of reanalysis (Harris and Campbell 1995: 30) or *exaptation* (Lass 1990; 1997): formerly meaningless differences ended up being employed to make meaningful distinctions. Exaptation can lead "to the development of new grammatical categories", and the "conceptual invention" that it represents means that "what was once a predictable alternation […] can be […] reanalysed as a new primary categorical marker" (Lass 1997: 319) – which is precisely what happened with the southwest of England infinitives.

In a second example of the same kind, certain North Frisian dialects have developed a distinction between two different definite articles (Ebert 1971; Ebert and Keenan 1973; Walker 1990: 14–15), an innovation which has no parallel at all in any of the standard or urban forms of any of the Germanic languages. In the Mooring dialect of the Bökingharde area the two sets of forms are:

masc.	*di*	*e*
fem.	*jü*	*e*
neut.	*dåt*	*et*
pl.	*da*	*e*

The usage of the two forms is grammatically and semantically complex (see Markey 1981: 228), but typically the *-e/-et* forms are proximal and/or refer to a unique referent, as in *e moune* 'the moon', *e wjard* 'the truth', whereas the *di/ jü/dåt/da* forms are distal and/or are context-bound and apply to definite but nonunique referents. For example, 'I have spoken to the village-mayor' can be rendered in two ways:

(3) a. *ik hääw ma e bürgermäister snååked*
I have with the mayor spoken
b. *ik hääw ma di bürgermäister snååked*
I have with the mayor spoken
'I have spoken with the mayor' (Walker 1990: 14–15)

In (3a) the reference is to the mayor of one's own village, whereas in (3b) it is to the mayor of some other village. Similarly, in the Fering dialect (the dialect of the island of Feer/Föhr), as Ebert (1984) points out, a question posed by some non nativespeaking outsider in a particular village of the type:

(4) *Huar wenet di bürgermäister?*
Where live the mayor
'Where does the mayor live?' (Ebert 1984: 235)

would produce the answer:

(5) *Hün bürgermäister mänst dü?*
Which mayor mean you
'Which mayor do you mean?' (Ebert 1984: 235)

Any question asking for information about the local mayor ought instead to use the article *e*.

How did the development of this distinction come about? The forms in the lefthand column come directly from the nominative definite article forms of Old

Frisian: *thi, thiu, thet, tha* (Bremmer 2009: 54). As Hoekstra (2001: 777) indicates, the forms in the right-hand column are in origin "weak variants" of the forms in the left-hand column i. e. they were originally phonologically reduced forms which occurred in unstressed environments. So in origin they were simply phonologically conditioned and therefore predictable variants, as they still are in many dialects. However, in the Mooring and other innovating dialects, the two types of variant have been reanalysed as having distinct semantic-grammatical functions, and refunctionalised as applying to non-unique (etc.) vs. unique (etc.) referents (Löfstedt 1968).

Given that this type of diachronic exaptational refunctionalization of variants is common enough in linguistic change, the question then is: why might reanalysis, leading to morphological-category development, be more likely to occur in some sociolinguistic environments than others? The answer would appear to be that this is the type of change which naturally takes place when a language does not have much contact of the relevant type (see section 2 above) with other languages – changes which Bailey (1982) referred to as *connatural*. I hypothesize that this form of complexification is more common in smaller, stable communities with no significant involvement of adult L2 outsiders because it requires a number of generations of uninterrupted native-speaker development to go to completion. Croft (2000) writes that "linguistic isolation allows for processes of change to evolve to an elaborate degree that would otherwise be curtailed by leveling or simplification in a larger, more loose-knit society" (Croft 2000: 193).

If it is true, then, that linguistic changes leading to complexification require long periods of time to go to completion, this is equivalent to saying that the endpoints of these changes are linguistic forms which can be described, as Dahl (2004) says, as *mature phenomena*. Mature linguistic phenomena are linguistic features which imply a lengthy period of historical development – they "presuppose a non-trivial prehistory" (Dahl 2004: 2). For example, Dahl (2004) singles out syntactic agreement as "belonging to the later stages of maturation processes" (Dahl 2004: 197). According to Dahl (2004), "linguistic phenomena have life cycles in the sense that they pass through a number of successive stages, during which they 'mature', that is, *acquire properties that would not otherwise be possible*" [my italics] (Dahl 2004: 2).

My contention is that in large, high-contact, unstable communities with loose social networks, such lengthy periods are less likely to be available. It is also relevant that mature phenomena are very vulnerable to being lost if exposed to high contact situations. This is because, according to Dahl (2004), "there is a significant overlap" (Dahl 2004: 286) between mature phenomena and "those linguistic features which are most recalcitrant in second language learning" (Dahl

2004: 286). They are therefore highly prone to being filtered out in suboptimal language acquisition (Dahl 2004: 286).

In one of his examples, Dahl (2004) writes that "reviewing the candidates for inclusion in the class of mature linguistic phenomena, we find that the most obvious one is inflectional morphology" (Dahl 2004: 111). He focusses on the role of fusion in producing mature phenomena in the long term, and points out that phonological change is crucial in the development of fusional opacity: "structural condensation would depend on phonological condensation – the fusion of two words into one is conditioned by their having been phonologically integrated" (Dahl 2004: 179). He also cites irregularity as a mature phenomenon: "lexical idiosyncrasy" (Dahl 2004: 112) occurs when a rule applies to lexical items in an unpredictable way – and includes also the presence of different inflectional classes. Highly fusional languages such as Latin, then, are the supreme example of the outcome of linguistic complexification, as well as the supreme demonstration of the nature of mature phenomena that require long periods of time to come into being.

How long is "a long period of time"? The theory of the *morphological cycle* (Hodge 1970, Dixon 1997) suggests that languages may change in such a way that they, as it were, move around a typological circle: from isolating to agglutinating, from agglutinating to fusional, and eventually back to isolating again. If we presented this circle graphically with purely isolating languages like Classical Chinese at 4 o'clock, agglutinating at 8, and fusional at 12, we could say that modern Indo-European languages represent a movement from Proto-Indo-European at an approximately 12 o'clock position towards a more isolating type at 2 o'clock or, in the more advanced cases, 3 o'clock (Dixon 1997). "Present-day agglutinative languages may have had an ancestor of more isolating profile, with what were distinct words having developed into grammatical affixes (e. g. postpositions into cases). The Dravidian family is roughly of this type, and here one can successfully recover a good deal of the proto-language" (Dixon 1997: 42).

According to Dixon (1997), Proto-Finno-Ugric was at about 9 o'clock in the cycle – highly agglutinating – while the modern languages have moved to 10 or 11 o'clock, with Estonian clearly having moved further than Finnish. Now, if proto-Finno-Ugric was at "9 o'clock", and modern Finnish is at "11 o'clock", then we can obtain a chronological estimate of how long this kind of change takes. Proto-Finno-Ugric dates back to about 4000 BC (Campbell 1997) i. e. 6,000 years ago. Even if modern Finnish has been at "11 o'clock" for as long as 1,000 years, this would mean it took 5,000 years to "travel" from 9 to 11, and arithmetic therefore indicates that for a language to transform from fully isolating to fully fusional (i. e. from 4:00 till 12:00), it would take, if the same trajectory speed was maintained, 20,000 years. I claim absolutely NO reality at all for this figure! I merely observe that it suggests that the development of at least certain mature phenomena requires very long periods of time indeed. The fact that millennia-long periods are involved makes clear the extent to which mature phenomena depend for their development on very long passages of time with relatively little interruption or "punctuation" (Dixon 1997) of the type that results from significant periods of social instability and/or simplifying adult language contact.

What implications does this have from a sociolinguistic-typological perspective? Faarlund (2005) writes about "the well-known drift from a synthetic to an analytic type" (Faarlund 2005: 1149) of language structure. This tendency is indeed well known, and was pointed out for English and the Romance languages by Schlegel (1846). And, as Lass (1997) says of the linguistics community, "there is apparently a traditional intuition of evolutionary direction [...] we prefer morphological complexity to decrease" (Lass 1997: 253), which might suggest that we could establish a uniformitarian directional principal, namely that complexity gives way to simplicity. There is a tendency amongst linguists, that is, to regard simplification as "normal". This is surely because we are so familiar with changes of this type in the histories of many of the betterknown Indo-European languages, as well as the Semitic languages, that it is very easy for us to think of simplification as simply "what happens". When we compare English with Old English, German with Old High German (Faarlund 2001), French with Latin, Bulgarian with Old Slavonic, we find it "normal" to note features such as reduction in overt case-marking; reduction in conjugations, declensions, and inflections; loss of the dual number; increase in periphrastic verb forms; and so on. As I argued long ago (Trudgill 1983), we may even have been tempted to regard such developments as a kind diachronic universal. As Lass (1997) says, however, this "traditional intuition" is in fact not at all well supported by the evidence (Lass 1997: 253).

My claim, indeed, is that simplification is actually not normal. If it were normal, all languages in the world would by now have been highly regular and maximally transparent. I suggest that it is actually complexification that is, in an important sense, more normal. If languages are "left alone", the natural tendency is for them to accrue more and more complexity, as a result of connatural changes (Bailey 1982), not to simplify. Linguistic complexification arises as a result of linguistic processes such as fusion, reanalysis, refunctionalization, exaptation and grammaticalization. These are all processes which require considerable periods of time in order to develop undisturbed and go to completion – something which it seems is most likely to be the case in communities characterized by "isolation" – stability, low contact, small size, and tight social networks.

And, in the modern world, communities with these characteristics are of course becoming fewer and fewer.

References

Bailey, Charles-James 1982: *On the Yin and Yang Nature of Language*. Ann Arbor: Karoma.

Bremmer, Rolf 2009: *An Introduction to Old Frisian*. Amsterdam: Benjamins.

Campbell, Lyle 1997: On the linguistic prehistory of Finno-Ugric. In: Raymond Hickey and Stanisław Puppel (eds.), *Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak on his 60th Birthday*, 829–862. Berlin/New York: Mouton de Gruyter.

Croft, William 2000: *Explaining Language Change*. London: Longman.

Dahl, Östen 2004: *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins.

Dixon, Robert M. W. 1997: *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Ebert, Karen Heide 1984: Zur grammatischen Differenziertheit des Dialektes (am Beispiel Fering). *Nordfriesisches Jahrbuch* 20: 227–238.

Ebert, Karen Heide and Keenan, Ed 1973: A note on marking transparency and opacity. *Linguistic Inquiry* 4: 421–424.

Faarlund, Jan Terje 2001: From Ancient Germanic to modern Germanic languages. In: Martin Haspelmath, Ekkehard König, Wulf Oesterreicher and Wolfgang Raible (eds.), *Language Typology and Language Universals: An International Handbook*, 1706–1719. Berlin/New York: de Gruyter.

Faarlund, Jan Terje 2005: Syntactic developments from Old Nordic to Early Modern Nordic. In: Oscar Bandle, Kurt Braunmüller and Ernst H. Jahr (eds.), *The Nordic Languages: An International Handbook of the History of the North Germanic Languages*, 1149–1160. Berlin/New York: de Gruyter.

Faarlund, Jan Terje 2010: Review of Geoffrey Sampson, David Gil and Peter Trudgill (eds.), Language complexity as an evolving variable (review). *Language* 86: 748–752.

Gachelin, Jean-Marc 1991: Transitivity and intransitivity in the dialects of Southwest England. In: Peter Trudgill and Jack K. Chambers (eds.), *Dialects of English: Studies in Grammatical Variation*, 218–228. London: Longman.

Harris, Alice and Campbell, Lyle 1995: *Historical Syntax in a Cross-Linguistic Perspective*. Cambridge: Cambridge University Press.

Hockett, Charles F. 1958: *A Course in Modern Linguistics*. New York: Macmillan.

Hodge, Carleton T. 1970: The linguistic cycle. *Language Sciences* 13: 1–7.

Hoekstra, Jarich 2001: Comparative aspects of Frisian morphology and syntax. In: Horst Haider Munske et al. (eds.), *Handbuch des Friesischen/Handbook of Frisian Studies*, 775–786. Tübingen: Niemeyer.

Ihalainen, Ossi 1991: On grammatical diffusion in Somerset folk speech. In: Peter Trudgill and Jack K. Chambers (eds.), *Dialects of English: Studies in Grammatical Variation*, 148–60. London: Longman.

Kusters, Wouter 2003: *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Leiden: Leiden University Press.

Lass, Roger 1997: *Historical Linguistics and Language Change*. Cambridge: Cambridge University Press.

Lenneberg, Eric 1967: *Biological Foundations of Language*. New York: Wiley.

Löfstedt, Ernst 1968: *Beiträge zu einer nordfriesischen Grammatik. I. Das Substantiv und das Adjektiv, das Zahlwort und der bestimmte Artikel*. Uppsala: Almquist & Wiksell.

Markey, Thomas 1981: *Frisian*. The Hague: Mouton.

Mithun, Marianne 1999: *The Languages of Native North America*. Cambridge: Cambridge University Press.

Mithun, Marianne 2007: Grammar, contact, and time. *Journal of Language Contact* 1: 333–155. www.jlc-journal.org.

Sampson, Geoffrey, Gil, David and Trudgill, Peter (eds.) 2009: *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Schlegel, August Wilhelm von 1846: *OEuvres de M. Auguste-Guillaume de Schlegel.* Leipzig: Weidmann.

Shosted, Ryan 2006: Correlating complexity: A typological approach. *Linguistic Typology* 10: 1–40.

Sinnemäki, Kaius 2009: Complexity in core argument marking and population size. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 125–140. Oxford: Oxford University Press.

Torp, Arne 2003: Frekvens, trykkletthet, reduksjon. In: Jan Terje Faarlund (ed.), *Språk i endring: Indre norsk språkhistorie*, 219–254. Oslo: Novus.

Trudgill, Peter 1983: *On Dialect: Social and Geographical Perspectives*. Oxford: Blackwell.

Trudgill, Peter 2001: Greek dialects: Linguistic and social typology. In: Angela Ralli, Brian Joseph and Mark Janse (eds.), *Proceedings of the First International Conference of Modern Greek Dialects and Linguistic Theory*, 263–272. Patras: Patras University Press.

Trudgill, Peter 2002: Sociolinguistic typology. In: Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change*, 707–728. Oxford: Blackwell.

Trudgill, Peter 2009: Sociolinguistic typology and complexification. In: Geoffrey Sampson, David Gil and Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*, 98–109. Oxford: Oxford University Press.

Trudgill, Peter 2010: Contact and sociolinguistic typology. In: Raymond Hickey (ed.), *Handbook of Language Contact*, 299–319. Oxford: Blackwell.

Trudgill, Peter 2011: *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford: Oxford University Press.

Walker, Alastair 1990: Frisian. In: Charles Russ (ed.), *The Dialects of Modern German: A Linguistic Survey*, 1–30. London: Routledge.