Complexity and size of speech community

Kaius Sinnemäki

[This version differs in some minor respects from that to appear in Sampson, Gil, and Trudgill (eds), Language Complexity as an Evolving Variable, OUP].

1 Introduction

A widely accepted presupposition among linguists is that language structure has nothing to do with its geographical or sociocultural setting (e.g. Kaye 1989: 48).[1] However, this claim has not been supported by empirical evidence. On the contrary, there seems to be a growing body of evidence indicating a relationship between e.g. structural complexity of language and its geographical or sociocultural setting. Nichols (1992), for one, has argued that morphological complexity varies geographically. Perkins (1992) has argued that grammatical complexity of deictic categories correlates negatively with cultural complexity. Sociolinguistic aspects of the speech community have also been suggested as causing complexity variation (Nettle 1999; Kusters 2003; Trudgill 1996, 2004a, b). This paper extends the discussion to morphosyntactic parameters by studying whether complexity in core argument marking could vary according to speech community size. I test this relationship statistically with a sample of 50 languages. Complexity is measured as violations of distinctiveness and economy, the two sides of the principle of one-meaning–one-form (discussed in section 2.1). In the following, I formulate the hypothesis (section 2), outline the method (section 3), and present and discuss the results (section 4 and 5, respectively).

2 Formulation of the hypothesis

2.1 Complexity and social typology

I take as a starting point Trudgill's (2004a) suggestion that the small phoneme inventories of Polynesian languages could be explained by their social characteristics. Although this hypothesis is concerned with phonological complexity, I shall argue that it can be fruitfully applied to other domains as well.

The crux of this hypothesis is that languages spoken by small, isolated/low-contact communities with close-knit social networks will likely have either very small or very large phoneme inventories, whereas languages spoken by large communities with a great deal of adult language learning by outsiders and loose social networks will likely have medium size phoneme inventories. The rationale is that the former types of communities can afford and are able to preserve redundancies, whereas those of the latter type tend towards transparency. The underlying complexity factor here is not necessarily size of phoneme inventory but rather its effect on distinctiveness of lexical items. Large inventories may increase the distinctiveness of lexical items redundantly, whereas small

inventories coincide with either greater word lengths or greater confusability of constituents. These consequences increase memory burden, a factor known to inhibit adult language learning (see Trudgill 2004a: 315-16 and references there).

Since such complexity effects can be connected to the general principles of economy and distinctiveness, the thesis can be fruitfully applied outside phonology as well. Economy and distinctiveness are like the two sides of the same coin; they both relate to the principle of one-meaning–one-form (a.k.a. transparency) but from different perspectives. Adherence to the principle of one-meaning–one-form requires adherence to both economy and distinctiveness. Violations of it involve either excessive encoding of distinctions, which increases distinctiveness at the expense of economy, or insufficient encoding of distinctions (when two intended meanings are encoded by non-unique forms), which increases economy at the expense of distinctiveness. The former causes redundancy, while the latter causes homonymy and ambiguity.

Violations of the one-meaning–one-form principle may increase difficulty of processing as well as structural complexity. In this sense, it matters little whether we measure relative difficulty to a user (Kusters 2003) or structural complexity (e.g. Dahl 2004). However, measuring cost/difficulty would require psycholinguistic tests, which is far beyond the scope of this short paper. Following Dahl (2004), complexity is here kept distinct from cost/difficulty and is defined as description length of a phenomenon. In terms of description length, adherence to the one-meaning–one-form principle requires shorter description length than violations of it (Miestamo 2008). Economy violations increase description length by adding rules to the description, whereas distinctiveness violations increase description length by requiring greater contextual specification in the rules in order to disambiguate otherwise identical forms. From the user perspective, adherence to the principle means full transparency, which is easy for most types of language users but especially favoured by adult language learners. Most violations of this principle increase memory load, e.g. redundant agreement or homonymic forms in case marking (Kusters 2003: 52–7). While different types of violation are affordable to different user groups, they are least affordable to adult language learners, who tend to reduce forms and change non-transparent forms to transparent ones (Kusters, ibid.; Trudgill 2004a: 306-7). All in all, because violations of the principle may increase both complexity and cost/difficulty, measures of the former might well approximate measures of the latter.

Trudgill (2004b: 386, personal communication) argues forcefully that the three parameters (size, isolation, network structure) should not be considered independently of one another but rather as jointly effective. Unexpectedly, though, this multifactor scenario seems unnecessary in the light of Hay and Bauer's (2007) cross-linguistic investigation. They tested the relationship between phoneme inventory size and speech community size in 216 languages, and arrived at a statistically very significant positive correlation. This result suggests that it may be fruitful to consider speech community size at least as a tentative parameter of complexity variation: the parameters may be intertwined to such a degree that a univariate

approach (which pays attention to a single societal parameter) could already approximate the phenomenon itself.

In addition to Hay and Bauer's (2007) results, there are three other reasons why, in an exploratory typological study, community size may serve as a feasible starting point and a springboard for further research. First, if we assume that the three criteria (size, isolation/amount of contact, network structure) must operate jointly, then the hypothesis will make predictions only about two of eight logically-possible classes of language – languages which are small, isolated, and socially close-knit, and languages which are large, non-isolated, and socially loose-knit. (Here I assume for the sake of argument that each criterion is bivalent – in reality the measure of each criterion is more likely a matter of degree, Trudgill 2004b: 384–5.) Strictly speaking, the hypothesis makes no prediction for the six other possible combinations of the criteria; consequently, a sample of 50 languages might contain perhaps no more than a dozen for which the hypothesis can be tested. However, at least some of the criteria seem more or less interconnected. For instance, tight social networks are more characteristic of isolated communities (Milroy and Milroy 1985) and of small communities (Allcott et al. 2007). Large languages, on the other hand, are more likely to attract adult language learning by outsiders for e.g. socioeconomic reasons, but they also generally have loose networks, which potentially bring about rapid change compared to communities with fewer weak ties in the network (Trudgill 2004b: 385; Milroy and Milroy 1985: 375, 380). This actually makes large, non-isolated languages with loose-knit networks prototypical large languages, because the combination of large community size and tight social network is practically impossible.

Secondly, it may be difficult to characterize speech communities with respect to all three criteria. Languages spoken by large communities pose particular problems, since a large community will be composed of smaller communities which may vary greatly in terms of size, isolation, and network structure. Speakers can also belong to many diverse communities simultaneously. Should large communities, therefore, be categorized according to the overall (macro) level or according to the smaller communities they consist of (micro-level)? For instance in the Russian speech community, many micro-level communities are small, isolated, and relatively tight-knit, whereas the macro-level community is large, has loose networks, and has a good deal of adult language learning by outsiders (Kusters 2003: 44). Small speech communities have their particular classification problems as well. As one reviewer pointed out, culturally specific habits such as exchanging women between otherwise isolated tribes may introduce adult language learning into the speech community and thus simplify the language. Simplification is certainly possible in this situation, but it is not inevitable; the outcome may equally well be that gender-specific registers begin to diverge, or that there is an increase in distinctiveness but no decrease in redundancy (see the discussion in section 5). These scenarios exemplify the complexity of talking about the issue at large when classifying languages according to the three criteria.

Thirdly, many languages lack reliably documented social histories. This forces one to sample languages for which documented histories are

available, which may bring unwanted bias to the sample. Speech community size is much more readily available and does not bias the sample from the outset. I suggest here that in an exploratory typological study it may be worthwhile to focus on speech community size alone. The result will necessarily remain suggestive and require further research, but will do as an initial approximation.

2.2 A hypothesis about core argument marking

In this section I formulate my hypothesis about core argument marking. I focus on the morphosyntactic strategies found in simple main clauses with affirmative polarity and indicative mood (note that this definition covers some pragmatically marked clauses, which may include e.g. focused elements as in (3)).

In core argument marking, three morphosyntactic strategies – head marking, dependent marking, and word order – interact in distinguishing "who does what to whom". They differentiate the arguments of a prototypical two-place transitive predicate, one more agent-like (A) and the other more patient-like (P) (Comrie 2005: 398). As defined by Nichols (1992), head and dependent marking are morphological strategies that indicate syntactic relations either on the head (as in (1) below) or the dependent of the constituent (as in (2)). (There is some head marking in (2) as well, but this will be discounted in the analysis (see below).) In the clause as a constituent, the predicate is the head and the arguments are its dependents.

(1) Yimas (Lower Sepik; Foley 1991: 193)[2]
Payum narmaŋ *<u>na-mpu</u>-tay*.
man.PL woman.SG 3SG.P-3PL.A-see
'The men saw the woman.'

(2) Kannada (Southern Dravidian; Sridhar 1990: 86)
Cirate mariy-<u>annu</u> nekkutti-de.
leopard cub-ACC lick.PRS-3SG.N
'The leopard is licking the cub.'

The role of word order is considered in clauses in which the arguments are part of the clause proper, that is, where they are not separated from the rest of the clause e.g. by a pause, or when there is no pronoun in situ replacing a transposed argument. Word order has a role in distinguishing the arguments if the position of the argument relative to the verb and to the other argument expresses its role – in other words, if reversible word order pairs (APV/PAV, AVP/PVA, and VAP/VPA) are disallowed. In the English sentence *John hit Mike*, *John* can only be interpreted as A and *Mike* as P: the opposite interpretation is disallowed.

Word order may occasionally have a role even when a reversible word order pair is allowed. In these cases, a change in word order is parallelled by a change in the morphological properties of the clause and (as we analyse the situation) the former would not be allowed without the latter. Slave (Rice 1989) obligatorily marks the head with the co-indexing pronoun ye- when the object occurs clause-initially (3):

Slave (Athapaskan, Rice 1989: 1197)

(3) a. *łį ʔehkee kayįhshu.*
dog (A) boy (P) 3.bit
'The dog bit the boy.'

b. *ʔehkee łį kayeyįhshu.*
boy (P) dog (A) 3.bit.4
'The boy, a dog bit him.'

It is understood that word order in Slave helps to distinguish the arguments at least in the canonical APV word order.

Next we may formulate the hypothesis to be tested. I aim to test whether there is a relationship between complexity in core argument marking of a language and social typology of the community speaking that language. This hypothesis is broken down into a pair of interrelated hypotheses about core argument marking:

(i) languages spoken by small speech communities are likely to violate the principle of one-meaning–one-form by either redundant or insufficient morphosyntactic marking of core arguments;

(ii) languages spoken by large speech communities are likely to adhere to the principle of one-meaning–one-form.

3 Method

3.1 Sample

The two hypotheses were tested against a random sample of 50 languages. I follow Dryer (1992, 2005) in sampling genera rather than languages. A genus is a grouping of languages with a time-depth of circa 3000–4000 years, corresponding roughly to e.g. the Germanic or the Romance languages (Dryer 1992: 83-5).

The sample is genealogically stratified so that no two languages come from the same genus and no two genera come from the same language family (with minor deviations at the highest strata in Africa and Australia–New Guinea). The sample is areally stratified so that the number of genera chosen in each macro-area is represented in the same proportion to the total number of genera in each macro-area (Miestamo 2005: 31-9). Macro-areas correspond roughly to continent-sized geographical areas: Africa, Eurasia, Southeast Asia–Oceania, Australia–New Guinea, North America, and South America (Dryer 1992). See the Appendix for the classification and sources of the sample languages.

3.2 Speech community size

The number of speakers for each language was obtained from grammar descriptions whenever possible. When the grammar description gave no estimate or an unreliable estimate (e.g. including number of speakers of closely related variants not described in the grammar), I took the number

of speakers from the Ethnologue (R. Gordon 2005) (more specifically, number of speakers for all countries). Table 1 provides the figures for each language.

Basing language-size estimates on the grammar descriptions rather than the Ethnologue has some advantages. For one, neither speech community size nor complexity remains constant over time, although the former seems to change more readily than the latter (Hay and Bauer 2007: 398). Consequently, the most recent estimate of community size may differ from the community size during the writing of the grammar description. It is therefore desirable to use the count whose date most closely matches the grammar description.

Community size during the formation time of the grammatical phenomena recorded in the grammar descriptions may also differ from community size when the grammar was written (and may differ even more compared to the most recent estimates). One sign of this discrepancy could be a disproportionately large ethnic population no longer speaking the language. According to the Ethnologue, Lakhota has 6000 speakers but the ethnic population is 20,000 people. This raises the question how accurately the present-day figure would represent Lakhota as a "small" or a "large" language. It would be tempting to make the leap and include the ethnic group in the speech community size, but this would introduce new problems that are impossible to address within the limits of this paper.

3.3 Complexity metric

Complexity is here measured as adherence to v. deviation from the principle of one-meaning–one-form. Two types of deviation are recognized: violations of economy and violations of distinctiveness. No account is taken of different degrees of deviation.

As a starting point, the number of morphosyntactic strategies interacting in core argument marking was counted for each language. Note that, for head marking, only languages which mark both A and P on the verbal head were counted. Head marking of only one of the arguments would require other strategies in order fully to distinguish the arguments from one another – at least when both participants are third person and identical in number and gender. Also, head marking of just A, in particular, is very widespread cross-linguistically, and does not seem to correlate with anything else in languages (Nichols with Barnes and Peterson 2006: 97).

Languages were analysed as adhering to the principle of one-meaning–one-form if a single strategy distinguishes the arguments in all or nearly all contexts. As an example, word order in Ngiti (Kutsch Lojenga 1994) distinguishes the arguments from one another in all contexts. Various minor deviations were discounted in categorizing certain languages as adhering to the principle. In Imbabura Quechua (Cole 1985), the arguments are distinguished from one another in all contexts by dependent marking. In addition, head marking of A is obligatory, but that of P occurs optionally and only for the first person singular. This slight redundancy of optionally using both dependent marking and head marking of both A and P when P is first person singular was treated as a negligibly small deviation from the principle. Some languages use more

than one morphosyntactic strategy in roughly complementary manner. In Kannada (Sridhar 1990), dependent marking distinguishes the arguments obligatorily when P has a human referent or when it is emphasized, and optionally elsewhere, but word order is used when both arguments are inanimate. This kind of complementary distribution is interpreted as adhering to the principle.

The two types of deviations are treated next. Languages violate economy if they use more than one strategy and not in a complementary manner. Maricopa uses both head and dependent marking in all relevant contexts:

(4) Maricopa (Yuman; Gordon 1986: 74)
Va vany-a nyip '-n'ay-sh chew-k.
house DEM-VAUG me 1-father-SBJ 3>3.make-REAL
'My father built *that house*.'

Languages violate distinctiveness if they use only one strategy but in limited contexts, or allow a lot of syncretism in the head or dependent marking paradigms. As an example, Iau (Bateman 1986) uses neither head marking nor word order but uses some type of dependent marking in limited contexts. Noun phrases "which are part of the new information being predicated about the topic" and are preceded by a non-A are marked with a postpositional particle *be* whose tone indicates the role of the NP in question (Janet Bateman, personal communication):

(5) Iau (Lakes Plain; Janet Bateman, personal communication)
Da⁷ Das⁷⁻⁸ be-⁸ di³.
dog Das FOC-A kill.TOT.DUR
'Das killed the dog.'

In most contexts, therefore, Iau uses no morphosyntactic strategy for distinguishing the arguments from one another.

One might ask whether languages which violate distinctiveness to a great extent have actually grammaticalized the distinction at all. Although not all sample languages seemed to identify A and P uniquely, at the least they all seemed to distinguish A from non-A, which suffices for the present purposes. Table 1 presents the complexity analysis for each language.

Table 1. Complexity analysis and speech community sizes.

Language	Strategies used	Type	Population
Adang	DM (limited)	VD	7000
Alawa	HM; DM	VE	30
Arapesh	WO; HM	VE	5000
Babungo	WO (limited) ; DM (some pronouns only)	VD	14,000
Berber (Middle Atlas)*	WO; DM	VE	3,000,000
Cora*	WO; HM; DM	VE	8000
Cree (Plains)*	HM	A	34,100
Daga*	HM	VD	6000
Georgian	DM; HM	VE	2,734,393
Gooniyandi*	DM; HM	VE	100
Greenlandic (West)	WO and DM (complementary); HM	VE	43,000
Hixkaryana	WO; HM	VE	350

Hungarian	DM; HM	VE	13,150,000
Iau	DM (limited)	VD	400
Ika	DM; HM	VE	7000
Indonesian*	WO; DM (opt. for 3rd singular pronoun)	A	23,143,354
Jaqaru	DM; HM	VE	1500
Kannada	WO and DM (complementary)	A	25,000,000
Khoekhoe*	DM	A	233,701
Kisi	WO; DM	VE	500,000
Koasati	DM; HM	VE	400
Kombai	WO	A	4000
Kuot	WO; HM	VE	1500
Lakhota*	HM	A	6000
Lavukaleve	WO; HM	VE	1700
Maricopa*	DM; HM	VE	181
Maybrat	WO	A	22,000
Mien*	WO	A	8,186,685
Miwok (Southern Sierra)	DM; HM	VE	20
Namia	DM (pronouns only)	VD	3500
Ngiti	WO	A	100,000
North Slave*	WO; HM	VE	790
Nubian (Dongolese)*	DM	A	280,000
Nuuchahnulth*	(HM only for A)	VD	200
Pirahã	WO	A	110
Pitjantjatjara	DM	A	3500
Qiang	DM; HM	VE	70,000
Quechua (Imbabura)	DM	A	40,000
Semelai	DM (limited)	VD	4103
Shipibo-Konibo	DM	A	23,000
Sko	WO; DM	VE	700
Somali*	DM; HM	VE	12,653,480
Thai*	WO	A	20,229,987
Trumai	WO; DM	VE	51
Tzutujil	HM	A	50,000
Urubú-Kaapor	DM (optional)	VD	500
Warao	DM (pronouns only)	VD	15,000
Welsh*	WO; DM	VE	536,258
Yagua	WO; HM	VE	3000
Yimas*	HM	VD	300

* Number of speakers taken from the Ethnologue (Gordon 2005).
WO = Word order.
DM = Dependent marking.
HM = Head marking.
A = Adherence to the principle of one-meaning–one-form.
VD = Violation of distinctiveness.
VE = Violation of economy.

4 Results

To begin with, table 2 shows the distribution of adherences to v. deviations from the principle of one-meaning–one-form across different speech community sizes. Two general tendencies can be observed from the figures. First, the number of languages adhering to the principle increases as the number of speakers increases, and most of these languages (N = 12, 75%) are larger than 10,000. Secondly, languages spoken by 10,000

speakers or less tend to violate either economy or distinctiveness (N = 24, 71%) and the number of these languages decreases as the number of speakers increases. However, the number of languages violating economy jumps up again for languages spoken by more than 100,000 speakers; but all of these are spoken in the Old World.[3] These observations suggest greater support for hypothesis (i) than for hypothesis (ii).

Table 2. Distribution of complexity values across different community sizes.

	<=1000	1001–10,000	10,001–100,000	>100,001	Total
Viol. dist.	4	4	2	0	10
Adherence	1	3	6	6	16
Viol. econ.	9	7	2	6	24
Total	14	14	10	12	50

Since it is very difficult to determine what constitutes a "small" or "large" speech community, several different community size thresholds were used (following Pericliev 2004), beginning from 250 and doubling the threshold size for each consecutive test. The hypotheses were rewritten for each threshold size accordingly:

(i') languages spoken by speech communities smaller than or equal to the threshold size are likely to violate distinctiveness or economy (cell C in table 3);

(ii') languages spoken by speech communities larger than the threshold size are likely to adhere to the principle of one-meaning–one-form (cell B).

Table 3. Contingency table for the statistical tests.

	Threshold size	
	<=	>
Adherence to 1M:1F	A	B (ii')
Violation of 1M:1F	C (i')	D

The hypotheses were tested in the open-source statistical computing environment R (R Development Core Team 2007) with chi-square and with Fisher's Exact Test where chi-square was not a valid test due to low expected counts. Table 3 shows the contingency table and table 4 the results.

Table 4. Results for different threshold sizes (absence of chi-value in column 2 indicates use of Fisher's Exact test).

Demarcator	X^2	p
250		.41
500		.07
1000		.02
2000	8.1	.0045
4000	5.2	.022
8000	9.2	.0025
16,000	12.0	.00053
32,000	5.2	.023
64,000		.105
128,000		.163
256,000		.297

The test was statistically significant for a number of threshold sizes, from 1000 up to 32,000 speakers (table 4). Since the median of the sample was 6500 (7000 in the Ethnologue), these threshold sizes grouped around it rather evenly. The strongest association between the variables was for threshold size 16,000 ($X^2 = 12.0$, $p = .00053$, d.f. = 1). Two follow-up tests were further performed to assess the reliability of the result. First, the reliability of the association was tested by observing the contribution of individual cells to the chi-value (Arppe, forthcoming). If the contribution of an individual cell by itself exceeds the critical value which makes a table with the same degrees of freedom statistically significant ($X^2 = 3.84$; $p = .05$; d.f. = 1), this confirms the reliability of the result but also shows the locus of greatest deviation. For threshold size 16,000, the contribution of the expected value of languages larger than 16,000 speakers that adhere to the principle of one-meaning–one-form was 4.9, which by itself exceeds the minimum of the critical value. This confirms the reliability of the result and provides evidence especially for hypothesis (ii).

Secondly, the vulnerability of the association to potential misclassifications was tested by following the procedure in Janssen et al. (2006: 435–7). Keeping the sample size fixed, the margins (the sums of columns and rows) are altered until the p-value is no longer statistically significant. If statistical significance is lost by altering the values by one case, one should be careful in interpreting the results. The association for threshold size 16,000 loses significance when four languages at a minimum were misclassified. Consequently, the result seems relatively resistant to potential misclassifications.

As a preliminary conclusion, a statistically significant association occurred between speech community size and structural complexity of core argument marking for many different threshold sizes, corroborated by the follow-up tests. The high number of large Old World languages which violate economy (table 2) indicates that hypothesis (i) is better supported than hypothesis (ii), while the size of deviations from expected values suggests greater support for hypothesis (ii). As indicated by the data in table 2, the distributions of adherences v. deviations from the principle of one-meaning–one-form across different speech community sizes largely conform to our hypotheses, except for the high number of large languages violating economy. Since it is plausible that there are differences between languages violating distinctiveness v. economy, one further test was performed in which the effect of large languages violating economy was bypassed: languages which adhere to the principle of one-meaning–one-form were compared to those that violate distinctiveness. The association was statistically very significant for threshold size 16,000 ($p = .0002$, Fisher's Exact Test). Since the association was resistant to three misclassifications at a minimum, its reliability is also confirmed. Consequently, when languages violating economy are discarded, there is rather strong support for both hypotheses (i) and (ii). These results suggest that large languages tend to avoid violations of distinctiveness, but do not mind violations of economy.

5 Discussion

There is a statistically relatively strong association between community size and complexity in core argument marking, measured as adherence to v. deviation from the principle of one-meaning–one-form. That there should be such an association is unexpected and calls for an explanation.

One might claim on methodological grounds that this association is due to chance, insisting that an absolute definition of "small" and “large" ought to be used. According to this objection, rather than defining "small" and "large" relative to various threshold sizes, they should be defined e.g. relative to the world median or perhaps to some upper limit of community size which a community with tight network structure could still have. However, this would not necessarily be helpful, since the definitions of "small" and "large" depend to some extent on geographical areas: a language spoken by e.g. 10,000 people is small in Europe but already relatively large in the Pacific (the respective medians are 220,000 and 800 according to the Ethnologue). Moreover, the association was statistically significant ($X^2 = 8.0$; $p = .0048$; d.f. = 1) even when using e.g. the world median (7000) as the threshold size.

On the other hand, one could argue that the association is due to chance because of the uneven geographical distribution of small/large languages across the globe. One way of answering this is to consider geographical areas independently of one another, e.g. Old and New World separately. For the sixteen Old World languages in the sample, the association was not statistically significant with any threshold size (Fisher’s Exact Test, $p > .05$). For the new world, however, the association was statistically significant for the median (2300 speakers) as threshold size (Fisher’s Exact Test, $p = .017$) and even more significant for 16,000 speakers as threshold size ($p = .0024$). The association for threshold size 16,000 loses significance when two languages at a minimum were misclassified. Since the result was resistant to at least one misclassification, its reliability is tentatively confirmed. According to these results, the association does not seem to have arisen by chance, but it does seem areally confined, which somewhat undermines the generality of the association.

However, the lack of association in the Old World might be a consequence of large languages tolerating different kinds of violations of the one-meaning–one-form principle compared to small languages. The data suggests that adherence to distinctiveness may be more important than adherence to economy in languages with more than 16,000 speakers, most of which are spoken in the Old World. Consequently, the data for the 16,000 threshold were scrutinized areally. According to the results, there was a statistically significant association in both Old and New World ($p = .028$ and $p = .029$, respectively, Fisher's Exact Test). This suggests that large languages in both Old and New world tend to avoid violations of distinctiveness but do not mind violations of economy. But why should large languages tolerate violations of economy more than violations of distinctiveness? One reason might be that because speakers of large languages generally have little shared background information, transparency and distinctiveness are especially needed for mutual understanding. Redundancy can be beneficial in these situations as well,

because it can increase distinctiveness. But because the transparency of morphosyntactic strategies varies across languages, languages probably differ in how affordable redundancy is for them in these situations.

In section 2.1 I proposed that the criteria used by Trudgill (2004a) – size of speech community, amount of adult language learning by outsiders, and tightness of network structure – may be more or less intertwined. Although the matter has not been studied in great detail, it is not totally implausible to assume that small community size would tend to combine with tight network structure and little or no adult language learning by outsiders rather than with loose network structure and/or large amount of adult language learning by outsiders. In case of large community size, it seems even more plausible that one particular combination of the criteria, namely that which combines loose network structure and large amount of adult language learning by outsiders, would represent a prototypical combination of the three criteria (cf. section 2.1).

If these generalizations are correct, the present results could be seen as an attempt to single out one variable from a multivariate phenomenon and to study its effect in isolation – but maybe, also, as an attempt to approximate the multivariate phenomenon. Interpreted in this way, Trudgill's (2004a) adapted model could provide an explanation here as well, that is, small and isolated languages with tight networks can afford and preserve complexities thanks to great amounts of shared background information, whereas large languages with loose network structure and much adult language learning by outsiders tend towards greater transparency.

By and large, the present paper indicates that language complexity is not necessarily independent of sociolinguistic properties such as speech community size. Future research should study the phenomenon with a multivariate cross-linguistic approach, paying more attention to geographical areas as well as to neighbouring languages with different sociolinguistic and typological profiles.

Appendix: the language sample

Africa

Babungo (Bantoid; Schaub 1985), Dongolese Nubian (Nubian; Armbruster 1960), Khoekhoe (Central Khoisan; Hagman 1977), Kisi (Southern Atlantic; Childs 1995), Middle Atlas Berber (Berber; Penchoen 1973), Ngiti (Lendu; Kutsch Lojenga 1994), Somali (Eastern Cushitic; Saeed 1999).

Eurasia

Georgian (Kartvelian; Harris 1981, Hewitt 1996). Hungarian (Ugric; Rounds 2001, Kiss 2002), Kannada (Southern Dravidian; Sridhar 1990), Welsh (Celtic; King 1993).

Southeast Asia-Oceania

Indonesian (Sundic; Sneddon 1996), Mien (Hmong-Mien; Court 1985), Qiang (Qiangic; LaPolla 2003), Semelai (Aslian; Kruspe 1999), Thai (Kam-Tai; Iwasaki and Ingkaphirom 2005).

Australia-New Guinea

Adang (Timor-Alor-Pantar; Haan 2001), Alawa (Maran; Sharpe 1972), Arapesh (Kombio-Arapesh; Conrad and Wogiga 1991), Daga (Dagan; Murane 1974), Gooniyandi (Bunuban; McGregor 1990), Iau (Lakes Plain; Bateman 1986), Kombai (Awju-Dumut; de Vries 1993), Kuot (Kuot; Lindström 2002), Lavukaleve (Solomons East Papuan; Terrill 2003),

Maybrat (North-Central Bird's Head; Dol 1999), Namia (Yellow River; Feldpausch and Feldpausch 1992), Pitjantjatjara (Pama-Nyungan; Bowe 1990), Sko (Western Sko; Donohue 2004), Yimas (Lower Sepik; Foley 1991).

North America
Cora (Corachol; Casad 1984), Koasati (Muskogean; Kimball 1991), Lakhota (Siouan; van Valin 1977), Maricopa (Yuman; L. Gordon 1986), Nuuchahnulth (Southern Wakashan; Nakayama 2001), Plains Cree (Algonquian; Dahlstrom 1991), Slave (Athapaskan; Rice 1989), Southern Sierra Miwok (Miwok; Broadbent 1964), Tzutujil (Mayan; Dayley 1985), West Greenlandic (Eskimo-Aleut; Fortescue 1984).

South America
Hixkaryana (Cariban; Derbyshire 1979), Ika (Aruak; Frank 1990), Imbabura Quechua (Quechuan; Cole 1985), Jaqaru (Aymaran; Hardman 2000), Pirahã (Mura; Everett 1986), Shipibo-Konibo (Panoan; Valenzuela 1997), Trumai (Trumai; Guirardello 1999), Urubú-Kaapor (Tupi-Guaraní; Kakumasu 1986), Warao (Warao; Romero-Figueroa 1997), Yagua (Peba-Yaguan; Payne and Payne 1990).

Notes

1 I am grateful to Fred Karlsson, Matti Miestamo, and the editors (especially Geoffrey Sampson) for their many helpful comments on earlier versions of this paper. I alone am responsible for any remaining errors. Research for this article has been funded by Langnet, the Finnish Graduate School in Languages Studies, whose support is gratefully acknowledged.

2 The following abbreviations are used in morphemic glossing: 1 first person, 3 third person, A agent, ABS absolutive, ACC accusative, DEM demonstrative, DUR durative, ERG ergative, FOC focus, N neuter, NDPST non-distant past, P patient, PL plural, PRS present, REAL realis, SBJ subject, SG singular, TOT totality of action, VAUG augmentative vowel.

3 Old World covers Africa, Eurasia, and Southeast Asia, whereas New World covers the Pacific (Australia, New Guinea, Melanesia, Micronesia, and Polynesia) and the Americas (cf. Nichols 1992: 12-3, 25-8).

Data sources

Armbruster, Carl 1960. *Dongolese Nubian: A Grammar*. Cambridge: The University Press.

Bateman, Janet 1986. *Iau Verb Morphology*. NUSA: Linguistic Studies of Indonesian and Other Languages in Indonesia, 26. Jakarta: Universitas Katolik Indonesia Atma Jaya.

Bowe, Heather 1990. *Categories, Constituents and Constituent Order in Pitjantjatjara. An Aboriginal Language of Australia*. London: Routledge.

Broadbent, Sylvia 1964. *The Southern Sierra Miwok Language*. Berkeley: University of California Press.

Casad, Eugene 1984. Cora. In *Studies in Uto-Aztecan Grammar, vol. 4: Southern Uto-Aztecan Grammatical Sketches*, Ronald Langacker (ed.), 153-459. Dallas: Summer Institute of Linguistics.

Childs, G. Tucker. 1995. *A Grammar of Kisi: A Southern Atlantic Language*. Berlin: Mouton de Gruyter.

Cole, Peter 1985. *Imbabura Quechua*. London: Croom Helm.

Conrad, Robert and Wogiga, Kepas 1991. *An Outline of Bukiyip Grammar*. Canberra: Australian National University.

Court, Christopher 1985. *Fundamentals of Iu Mien (Yao) Grammar*. PhD Dissertation, University of California at Berkeley.

Dahlstrom, Amy 1991. *Plains Cree Morphosyntax*. New York: Garland Publishing.
Dayley, Jon 1985. *Tzutujil Grammar*. Berkeley: University of California Press.
Derbyshire, Desmond 1979. *Hixkaryana*. Amsterdam: North-Holland.
Dol, Philomena 1999. *A Grammar of Maybrat: A Language of the Bird's Head, Irian Jaya*. PhD Dissertation, University of Leiden.
Donohue, Mark 2004. *A Grammar of the Skou Language of New Guinea*. Unpublished manuscript, National University of Singapore. <http://www.papuaweb.org/dlib/tema/bahasa/skou/>.
Everett, Daniel 1986. Pirahã. In Derbyshire and Pullum (eds), 200-325.
Derbyshire, Desmond and Pullum, Geoffrey 1986. *Handbook of Amazonian Languages*, vol. 1. Berlin: Mouton de Gruyter.
Feldpausch, Tom and Feldpausch, Becky 1992. Namia Grammar Essentials. In *Namia and Amanab Grammar Essentials, Data Papers on Papua New Guinea Languages*, vol. 39, John Roberts (ed.), 2-97. Ukarumpa: Summer Institute of Linguistics.
Foley, William 1991. *The Yimas Language of New Guinea*. Stanford: Stanford University Press.
Fortescue, Michael 1984. *West Greenlandic*. London: Croom Helm.
Frank, Paul 1990. *Ika Syntax*. Arlington: Summer Institute of Linguistics and the University of Texas at Arlington.
Gordon, Lynn 1986. *Maricopa Morphology and Syntax*. Berkeley: University of California Press.
Guirardello, Raquel 1999. *A Reference Grammar of Trumai*. PhD Dissertation, Rice University, Houston, Texas.
Haan, Johnson 2001. *The Grammar of Adang: A Papuan Language Spoken on the Island of Alor, East Nusa Tenggara – Indonesia*. PhD Dissertation, University of Sidney. <http://www-personal.arts.usyd.edu.au/jansimps/haan/adang-index.htm>.
Hagman, Roy 1977. *Nama Hottentot Grammar*. Bloomington: Indiana University Press.
Hardman, Martha 2000. *Jaqaru*. München: Lincom Europa.
Harris, Alice 1981. *Georgian Syntax: A Study in Relational Grammar*. Cambridge: CUP.
Hewitt, George 1996. *Georgian: A Learner's Grammar*. London: Routledge.
Iwasaki, Shoichi and Ingkaphirom, Preeya 2005. *A Reference Grammar of Thai*. Cambridge: CUP.
Kakumasu, James 1986. Urubu-Kaapor. In Derbyshire and Pullum (eds), 326-403.
Kimball, Geoffrey 1991. *Koasati Grammar*. Lincoln: University of Nebraska Press.
King, Gareth 1993. *Modern Welsh: A Comprehensive Grammar*. London: Routledge.
Kiss, Katalin 2002. *The Syntax of Hungarian*. Cambridge: CUP.
Kruspe, Nicole 1999. *A Grammar of Semelai*. PhD Dissertation, University of Melbourne.
Kutsch Lojenga, Constance 1994. *Ngiti: A Central-Sudanic Language of Zaire*. Köln: Rüdiger Köppe Verlag.

LaPolla, Randy with Chenglong Huang 2003. *A Grammar of Qiang*. Berlin: Mouton de Gruyter.
Lindström, Eva 2002. *Topics in the Grammar of Kuot, a Non-Austronesian Language of New Ireland, Papua New Guinea*. PhD Dissertation, Stockholm University.
McGregor, William 1990. *A Functional Grammar of Gooniyandi*. Amsterdam: Benjamins.
Murane, Elizabeth 1974. *Daga Grammar, from Morpheme to Discourse*. Norman: Summer Institute of Linguistics.
Nakayama, Toshihide 2001. *Nuuchahnulth (Nootka) Morphosyntax*. Berkeley: University of California Press.
Payne, Doris and Payne, Thomas 1990. Yagua. In *Handbook of Amazonian Languages*, vol. 2, Desmond Derbyshire and Geoffrey Pullum (eds), 249-474. Berlin: Mouton de Gruyter.
Penchoen, Thomas 1973. *Tamazight of the Ayt Ndhir*. Los Angeles: Undena Publications.
Rice, Keren 1989. *A Grammar of Slave*. Berlin: Mouton de Gruyter.
Romero-Figueroa, Andres 1997. *A Reference Grammar of Warao*. München: Lincom Europa.
Rounds, Carol 2001. *Hungarian: An Essential Grammar*. London: Routledge.
Saeed, John 1999. *Somali*. Amsterdam: Benjamins.
Schaub, William 1985. *Babungo*. London: Croom Helm.
Sharpe, Margaret 1972. *Alawa Phonology and Grammar*. Canberra: Australian Institute of Aboriginal Studies.
Sneddon, James 1996. *Indonesian: A Comprehensive Grammar*. London: Routledge.
Sridhar, Shikaripur 1990. *Kannada*. London: Routledge.
Terrill, Angela 2003. *A Grammar of Lavukaleve*. Berlin: Mouton de Gruyter.
Valenzuela, Pilar 1997. *Basic Verb Types and Argument Structures in Shipibo-Konibo*. MA Thesis, University of Oregon.
van Valin, Robert Jr. 1977. *Aspects of Lakhota Syntax: A Study of Lakhota (Teton Dakota) Syntax and its Implications for Universal Grammar*. PhD Dissertation, University of California, Berkeley.
de Vries, Lourens 1993. *Forms and Functions in Kombai, an Awyu Language of Irian Jaya*. Canberra: Australian National University.

References

Allcott, Hunt, Karlan, Dean, Möbius, Markus, Rosenblat, Tanya, and Szeidl, Adam 2007. Community Size and Network Closure. *American Economic Review Papers and Proceedings* 97 (2): 80-85.
Arppe, Antti (forthcoming). Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy. PhD Dissertation, University of Helsinki.
Comrie, Bernard 2005. Alignment of case marking. In Haspelmath et al., 398-405.
Dahl, Östen 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: Benjamins.

Dryer, Matthew 1992. The Greenbergian word order correlations. *Language* 68: 81-138.
Dryer, Matthew 2005. Genealogical language list. In Haspelmath et al., 584-644.
Gordon, Raymond Jr. (ed.) 2005. *Ethnologue: Languages of the World*. 15th edn. Dallas: SIL International. <http://www.ethnologue.com>.
Haspelmath, Martin, Dryer, Matthew Gil, David, and Comrie, Bernard (eds) 2005. *The World Atlas of Language Structures*. Oxford: OUP.
Hay, Jennifer and Bauer, Laurie 2007. Phoneme inventory size and population size. *Language* 83 (2): 388-400.
Janssen, Dirk, Bickel, Balthasar, and Zúñiga, Fernando 2006. Randomization tests in language typology. *Linguistic Typology* 10: 419-440.
Kaye, Jonathan 1989. *Phonology: A Cognitive View*. Hillsdale (NJ): Lawrence Erlbaum.
Kusters, Wouter 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. PhD Dissertation, University of Leiden.
Miestamo, Matti 2005. *Standard Negation: The Negation of Declarative Verbal Main Clauses in a Typological Perspective*. Berlin: Mouton de Gruyter.
Miestamo, Matti 2008. Grammatical complexity from a cross-linguistic point of view. In *Language Complexity: Typology, Contact, Change*, Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson (eds), 23-41. Amsterdam: Benjamins.
Milroy, James and Milroy, Lesley 1985. Linguistic change, social networks and speaker innovation. *Journal of Linguistics* 21: 339-384.
Nettle, Daniel 1999. *Linguistic Diversity*. Oxford: OUP.
Nichols, Johanna 1992. *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press.
Nichols, Johanna with Jonathan Barnes and David A. Peterson 2006. The robust bell curve of morphological complexity. *Linguistic Typology* 10: 96-106.
Pericliev, Vladimir 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology* 8: 376-383.
Perkins, Revere 1992. *Deixis, Grammar, and Culture*. Amsterdam: Benjamins.
R Development Core Team 2007. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>.
Trudgill, Peter 1996. Dialect typology: isolation, social network and phonological structure. In *Towards a Social Science of Language*, Gregory R. Guy, Crawford Feagin, Deborah Schiffrin, and John Baugh (eds), 3-21. Amsterdam: Benjamins.
Trudgill, Peter 2004a. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8: 305-320.
Trudgill, Peter 2004b. On the complexity of simplification. *Linguistic Typology* 8: 384-388.