CONSTRUCTING KNOWLEDGE: NOMOTHETIC APPROACHES TO LANGUAGE EVOLUTION

SEÁN G. ROBERTS
Language Evolution and Computation Research Unit
School of Philosophy, Psychology & Language Sciences, The University of Edinburgh
s.g.roberts@sms.ed.ac.uk

JAMES WINTERS
Cardiff School of English, Communication and Philosophy, Cardiff University
wintzis@gmail.com

Nomothetic approaches identify statistical tendencies from large-scale, cross-cultural data sets. Recently, newly available data and statistical techniques have prompted a number of studies which consider the links between language structure and social structure. This can contribute in two ways to constructive approaches that try to explain how cultural forces shape language. First, it can provide new empirical data on which to test models. Secondly, it can suggest phenomena for constructive approaches to aspire to. Agent-based models provide a good way of identifying mechanisms, but are often too abstract to be tested against real data (Vogt & De Boer, 2010). Results from nomothetic research can guide computational models with regards to the relevant factors that affect language evolution. For example, Lupyan and Dale (2010) show that population size predicts morphological complexity. While many models consider demongraphic data (e.g. Nowak et al., 2001), few consider a realistic notion of morphological versus lexical marking. However, caution is required when using nomothetic results to inform constructive approaches. While nomothetic approaches can identify trends, they provide little explanatory power with regards to the mechanisms which brought about these trends. That is, it's not a straightforward task to convert the phenomena that nomothetic approaches discover into abstract features for a model of a mechanism. This paper discusses some limitations of nomothetic studies and how to interpret their results in the light of constructive approaches.

A century ago, Edward Sapir surmised that "even the simplest environmental influence is either supported or transformed by social forces" and that culture and social structure co-evolve (Sapir, 1912, p.26). This co-evolution may have begun before the evolution of modern language (e.g. Dunbar, 1993). Language, then, is a complex adaptive system: it emerges as a product of its underlying speech community, but also adapts to the very dynamics from which it emerged (Beckner et al., 2009). Constructive approaches to language evolution have been emphasising the importance of cultural transmission in shaping the structure of language (e.g. Smith & Kirby, 2008), as opposed to strong nativist claims. Recently, new statistical techniques and real data have been used to address this issue. This is part of a growing field which be broadly categorised as 'nomothetic': Constructing models of general phenomena from large sets of cross-linguistic data. While the use of real data can be an advantage over other constructive approaches, it lacks the careful case-study approach of more traditional 'idiographic' research. This paper discusses how nomothetic approaches fit in with other constructive approaches.

These approaches will be briefly illustrated with examples from studies of colour lexicons. MacKeigan and Muth (2006)'s idiographic research reports that speakers of Tzotzil have hundreds of compound forms for describing colour due to their industry of the hand-weaving colourful textiles. The study tracked the spread of individual colour words within the social network of the weavers to assess the affects of social structure on the lexicon and motivate hypotheses. This careful study of individuals in a particular society is a powerful research method, but it is also complex, costly and time-consuming. The naming game paradigm (Steels & Belpaeme, 2005) is a constructive approach to the transmission of colour words, but relies on abstracted assumptions and is difficult to implement and analyse. Given the availability of new electronic resources and the internet, it is now possible to address the same topics simply, cheaply and quickly.

To demonstrate this point, we conducted a nomothetic analysis of the effect of colour terms on cultural aspects of society. We hypothesised that a society's colour lexicon should effect the range of colours in artifacts whose appearance has been negotiated, such as national flags. Indeed, a multiple regression shows that the number of basic colour categories in a language is a significant predictor of the number of colours in the national flag for speakers of that language ($r=.12$, $F(16,106)=1.86$, $p=.03$, controlling for population size, distance from the equator and per-capita gross domestic product. Data from WALS; Wikipedia, 2011; IMF, 2011). This is an example of a nomothetic study - considering data across many languages and sources (although admittedly using crude statistical techniques). It could be used to support the idea that colour lexicons are grounded in physical objects through negotiation and interaction in a local population (Steels & Belpaeme, 2005). It was done using an ordinary laptop, took less than a day to complete and the data and analysis cost nothing to obtain. The advantages and attraction of such an approach are clear. However, there are many weaknesses, including problems with the data and statistical analyses, the assumptions behind the analyses and the explanatory power of the analyses. These will be discussed below.

Aside from these problems, the role of the nomothetic approach is often misunderstood. Nomothetic research can be exploratory and used to demonstrate features and interactions of interest that other approaches may test, not necessarily a definitive proof in itself. The candid introduction of Hay and Bauer (2007) demonstrates that researchers are aware of this: "We couldn't resist checking this apparent link ... We do not have well-developed theoretical arguments to offer about why this should be. However the correlation seemed intriguing enough that it was worth simply publishing the result, and leaving it up to readers to draw their own conclusions" (Hay & Bauer, 2007, p.388).

Recently, newly available data and techniques have prompted large-scale statistical investigations of the effect of social structure on language evolution. Two studies will be reviewed here and used as illustrations for the rest of the paper. First, Lupyan & Dale (2010) (LD) show that a community's population size is correlated with the level of morphological complexity in their language . Language typology data from the World Atlas of Language Structures (WALS, Haspelmath et al., 2008) were combined with data on the speaker population and geographic spread from the Ethnologue (Gordon, 2005). A general trend was identified: Languages with a small number of speakers, low geographic spread and few linguistic neighbours tend to have high morphological complexity (esoteric languages). On the other hand, languages with a large number of speakers, large geographic spread and many linguistic neighbours tend to have low morphological complexity (exoteric languages). The Linguistic Niche Hypothesis put forward by LD proposes that there is a pressure for extoteric languages to become more learnable by second-language learners (adults), and so adapt to rely less on complex morphology.This view of language evolution is compatible with the view of language as an organism that adapts to the cognitive niche of human brains (e.g. Christiansen & Chater, 2008; Smith & Kirby, 2008). This is an appealing theory, but its limitations and the most productive methods of testing will be discussed below.

In another study, Atkinson (2011) applies a serial founder effect model to phonemic diversity to infer the most likely origin of modern languages. The analysis shows that languages tend to have a smaller phoneme inventory the further you get from western Africa. Under this model, migrating sub-populations only sample part of the variation in the original population, so that variation is lost through repeated migrations. It is fascinating that such a simple linguistic variable could reflect a key social change after such a long time. However, the analysis has some weaknesses in common with many similar studies, discussed below. The main problem that we see is that, on its own, the explanatory power of nomothetic results are low. There is a temptation, given the relative ease of conducting studies, of coming to rely on such an approach to support the claims of a hypothesis. It is argued that the nomothetic approach is a tool for the construction of hypotheses, not a complete approach in itself. The next sections expand on the problems with the use of statistics, the assumptions made and explanatory power.

1. Use of statistics

Nomothetic approaches typically rely on statistical correlations to support their hypotheses. The amount of data available allows a lot of statistical power, but this should be used cautiously for three reasons. First, the data available for analysis are not necessarily direct measures of the phenomena of interest. Therefore, it is important to control for data quality and covariance. Secondly, the analyses do not necessarily demonstrate *how* the measures are related causally. Furthermore, the explanation is imposed by the researcher, so a careful consideration of competing hypotheses is necessary. While these problems are common to many fields, there are some specific concerns for modern nomothetic approaches, discussed below.

1.1. *Quality of data*

One concern with the nomothetic approach in general is the quality of the data in the analysis. The sampling is poor (LD point out no langauge in WALS has data for all features, and no feature has data for all languages) and there is a lack of temporal depth (Fagyal, Escobar, Swarup, Gasser, & Lakkaraju, 2010), which might obscure a correlation between the rate of change of demographic features and the rate of change in language (Martowicz, 2011). However, longitudinal data is not easily available for analysis. This problem can be addressed using careful statistics or supported with experimental or idiographic approaches.

1.2. *Spurious correlations*

Statistical power can be a weakness. Large datasets will contain spurious correlations which could provide apparent support for any hypothesis against a null-hypothesis. For example, countries in which the acacia tree *Acacia nilotica* grows are significantly more likely to have tonal languages (χ^2=47.1, $p < 0.0001$, data from CPC, 2008; Maddieson, 2008). However, a causal link is, obviously, very unlikely. This problem can be addressed with careful use of statistics or, better yet, testing multiple competing hypotheses against each other.

1.3. *Alternative hypotheses*

When testing a hypothesis with potentially low-quality, covarying data, it is important to consider alternative explanations. For instance, with regards to Atkinson's discovery of a correlation between phonemic diversity and distance from Africa, only the 'out of Africa' explanation is considered. However, considering alternative hypotheses can become complicated. A toy example will be used to illustrate the difficulty of interpreting nomothetic results. Suppose we wanted to test the hypothesis that languages are more morphologically complex the further their communities are from Europe.We can test this using data on number of grammatical categories a verb can take in a given language and the geographic locations of the societies of those languages. Indeed, morphological complexity of a language is significantly correlated with the physical distance of its community from center of the European Union (partial correlation controlling for speaker population, $r = 0.12$, $df = 141$, $p= 0.04$, data from WALS). This evidence rejects a null hypothesis, and so appears to support the hypothesis that linguistic typologies are Euro-centric by showing that a linguistic feature has an 'origin' in Europe.

However, the crucial question is how to interpret the statistic. Using distance from a single point does not allow other hypotheses to compete. A more appropriate baseline may be to consider the probabilities for regularly spaced points across the earth. Doing this, we find that there *is* a region around Europe where the correlation fits. However, there are also similar regions at most extreme latitudes. One possible explanation is that the correlation is significant for locations which maximises the variation in distances, (i.e. locations far away form the co-ordinates of the languages: Extreme latitudes, oceans etc.). This is tested by comparing the current map of correlations to the same map where the values of the data points are randomised. The map of correlations was re-sampled with random feature values using a Monte Carlo method. These maps were compared to the map of true values using a mantel test. 47% of the random maps produced significantly similar ($p < 0.05$) maps to the map with true values. That is, even when the linguistic data is randomised, there is a likelihood that an 'origin' map will look the same. This suggests that there is some amount of structure in the geographic locations of languages in addition to their linguistic variation that could be inadvertently harnessed by an analysis. This argument requires more analysis, however, it does give a flavour of the complexity of handling large statistical analyses. Rejecting a null hypothesis is not enough: considering alternate hypotheses is always necessary.

2. Assumptions

Studies differ with respect to the assumptions they make. For instance, although Atkinson (2011) and LD both incorporate broad demographic variables, their theoretical rationale for the application of these variables differs: whereas the former is interested in how these population dynamics impact upon the distribution of variation, the latter sees social structure as a substantial feature of the linguistic environment in which a language adapts. There are three basic challenges for the nomothetic approach to overcome in terms of the assumptions it makes: Choosing the right typological framework, the linearity of relationships between variables and consideration of the relatedness of languages.

2.1. *Typological frameworks*

One of the challenges for nomothetic approach is that, in comparing feature data across languages, a single analytical framework is required. Measuring complexity across many features of many languages is susceptible to two problems, according to (Miestamo, 2008): First, no single measure can capture the relative complexity of all languages. Secondly, comparing different aspects within a language and how they contribute to overall complexity is very difficult. Dale, Dietrich, and Chemero (2009) would agree with this, however, arguing that many frameworks and levels of explanation are necessary to account for complexity. Despite this, other studies have conflicting conclusions about the effects of population size on grammatical complexity. Nichols (2009), Sinnemäki (2009) and Hyslop (1993) find negative correlations between population size and other measures of grammatical complexity (see also Martowicz, 2011). A consensus on a theory of language complexity is required to make progress, but we suggest that, rather than further nomothetic studies, other approaches could be used to test the different predictions that each study makes against each other.

2.2. *Linearity*

Linearity is an unrealistic assumption for complex adaptive systems. Even if a linear relationship is currently observable, the current snapshot may be unable to detect a more complex relationship over time. For instance, technological and cultural complexity are shown to be drivers of population growth (Richerson, Boyd, & Bettinger, 2009), so the relationship we see between language features and demography could be an artefact of technological sophistication and cultural complexity. Also, although the nomothetic studies above suggest that large-scale changes in the social structure are intricately linked with phonemic diversity and morphological complexity, these features are part of a wider linguistic environment. Changes in social structure, then, may not only pattern variation at multiple-levels of linguistic organisation, but also lead to synergistic adaptation in a language: Here, we would expect to see signatures of adaptive patterning between phonemic diversity, morphological complexity and syntactic structuring. LD already hint at such a relationship when they linked decreasing morphological complexity with an increasing degree of form-meaning compositionality. One way to decrease a language's reliance on inflectional morphology is to shorten word lengths. An information theoretic perspective will show that more phonemes allow for shorter word lengths, as evident in laboratory experiments (Selten & Warglien, 2007) and statistical studies (Nettle, 1999).

Phonemic diversity, then, can in some cases provide the raw material with which a language will systematically restructure itself in accordance to a particular niche. Of course, this is hypothesised to be reasonably path-dependent, in so much as an increase in the phoneme inventory size opens up the potential to explore new adaptive avenues. One general finding places high phonological complexity (large phoneme inventory and high syllable complexity) as feeding through other levels of languages: morphological complexity lowers whilst the introduction of greater degrees of homonymy and polysemy increases semantic complexity (Fenk-Oczlon & Fenk, 2008, see Winters, 2011 for further details).

2.3. *Lineages*

The geographic spread of linguistic features is not an accident - languages and societies expand in lineages. This issue is not addressed directly in LD, in contrast with some previous research. Nichols (1992) finds no significant correlation between population size and their measure of complexity within each macro-continent and suggests that the correlation is an accident of geography. Hay and Bauer (2007) also find that language family explains some of the variance in phoneme inventory size, in addition to population size. Martowicz (2011), looking at the complexity of clause linkers, finds more evidence for the influence of geographic distribution and transmission lineage than population size. Phylogenetic techniques provide important new tools for addressing these issues. For instance, Dunn et al. (2011) use a phylogenetic analysis of real typological data to test nativist theories of word order universals against cultural transmission theories. This is a very productive use of the nomothetic approach as it combines work from constructive approaches, formalist approaches and empirical data.

3. Explanatory power

The relationship between population size and morphological complexity reported above is a trend - there are languages that go against this tendency. However, LD claim that their approach is nomothetic (see LD, supplementary Material S3). That is, the results show a strong tendency which suggests a hypothesis that can go on to be tested directly by other approaches. Although nomothetic approaches are a powerful tool for generating hypotheses, the amount of explanatory power they produce is weak. This can make it difficult to incorporate the findings into an abstract model. As an example, LD suggest the difference between exoteric and esoteric languages is related to the difference between child and adult learning abilities. Although it is generally true that ultimate attainment in language acquisition is correlated with age of first exposure, the cause of this is difference is contested. For instance, the linguistic output of a second language learner (and so the input to the next generation) may be affected by social aspects such as embarrassment (Schachter, 1974), lack of motivation, low self-esteem or anxiety (Krashen, 1982). Furthermore, adults and children interact differently in terms of number of interlocutors, frequency, context and function (Kuhl et al., 1997). At the society level, ultimate attainment is effected by the linguistic distance between the first and second language (e.g. Lado, 1957). This means that it's not just the number of other languages a community has contact with that should affect its evolution, but the properties of those languages.

Thus, identifying the causal factor driving the relationship between second language learners and morphological complexity is not straightforward, meaning that extracting the relevant features of the mechanism for a model is difficult. However, nomothetic approaches can work with idiographic approaches to obtain a better idea of the processes involved. For instance, proponents of the idiographic approach also argue that the amount of language contact has an effect on language complexity (e.g. Trudgill, 2004), although opinions differ on the importance of this factor. Trudgill (2004) argues that population size, network structure and language contact need to be considered together while McWhorter (2008) suggests that the level of non-native language acquisition is the only factor that affects the simplification of grammar over generations. It's worth noting that nomothetic approaches do not always contradict conclusions from other approaches, and are not always the most extreme.Recent experimental evidence also supports the hypothesis that second language learners find morphology more difficult to master than native speakers (Clashen et al, 2010). Also, Little (2011) uses a human language game experiment to test the hypothesis that forigner-directed speech is a driving factor. In conclusion, although the proportion of second language speakers in a community is related to the morphological complexity of its language, the proximate causes of this relationship are complex. This problem holds for many nomothetic studies. Several approaches from many disciplines are required for a full explanation.

4. Conclusion

Through the research reviewed in this paper, a picture is emerging which emphasises a complex interaction between the language of a society and its social structure. This work, together with theoretical and experimental work on cultural transmission, weakens the necessary effects of innate properties of language and demonstrates the importance of the way people interact (e.g. Smith & Kirby, 2008; Dunn et al. 2011). This paper has discussed the nomothetic approach and argued that, while it may inform other approaches, it has a limited explanatory power. A pluralistic approach involving constructive approaches to language evolution is needed. Given the high rate at which cultural processes can change cultural structures (Christiansen & Chater, 2008), there could be a cut-off point in history where the current data can't inform us of aspects of culture in the past in a straightforward way. Constructive and experimental approaches can help support this weakness in the nomothetic approach.

References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, *332*(6027), 346-349.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489-508; discussion 509-58.

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, *60*(1), 21–43.

Crop Protection Compendium. (2008). *Acacia confusa. online, accessed 18-04-2011. http://www.cabi.org.*

Dale, R., Dietrich, E., & Chemero, A. (2009). Explanatory pluralism in cognitive science. *Cognitive Science*, *33*(5), 739–742.

Dunbar, R. (1993). Coevolution of neocortical size, group-size and language in humans. *Behavioral and Brain Sciences*, *16*(4), 681–694.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*(7345), 79-82.

Fagyal, Z., Escobar, A., Swarup, S., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: network roles in language change. *Lingua*, *120*(8), 2061-2079.

Fenk-Oczlon, G., & Fenk, A. (2008). Complexity trade-offs between the subsystems of language. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), (Vols. Language Complexity: Typology, Contact, Change, p. 43-65). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gordon, R. (2005). *Ethnologue: Languages of the world, 15th edition.* SIL International.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2008). *World atlas of language structures* (Vol. Available online at http://wals.info/feature/22 Accessed on 2011-04-18.). Munich: Max Planck Digital Library.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *2*, 388-400.

Hyslop, C. (1993). *Towards a typology of spatial deixis.* Unpublished doctoral dissertation, Honours thesis, Australian National University.

International Monetary Fund. (2011). *World economic outlook, april 2011.*

Krashen, S. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., , & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684-686.

Lado, R. (1957). *Linguistics across cultures.* Ann Arbor: University of Michigan Press.

Little, H. (2011). The role of foreigner-directed speech in the cultural transmission of language and the resulting effects on language typology. In *30th Language at Edinburgh Lunch. University of Edinburgh, UK.*

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559.

MacKeigan, T., & Muth, S. Q. (2006). A grammatical network of tzotzil mayan colour terms. In C. P. Biggam & C. J. Kay (Eds.), (Vol. Progress in Colour Studies, p. 23-36). Amsterdam: John Benjamins Publishing Company.

Maddieson, I. (2008). Tone. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), (Vol. The World Atlas of Language Structures Online). Munich: Max Planck Digital Library.

Martowicz, A. (2011). *The origin and functioning of circumstantial clause linkers: a cross-linguistic study.* Unpublished doctoral dissertation, The University of Edinburgh.

McWhorter, J. (2008). Why does a language undress? strange cases in indonesia. In F. Karlsson, M. Miestamo, & K. Sinnemäki (Eds.), (Vols. Language Complexity: Typology, Contact, Change, p. 167-190). John Benjamins, Amsterdam.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. Amsterdam: John Benjamins Publishing Company.

Nettle, D. (1999). *Linguistic diversity.* Oxford: Oxford University Press.

Nichols, J. (1992). *Linguistic diversity in space and time.* University of Chicago Press.

Nichols, J. (2009). Linguistic complexity: a comprehensive definition and survey. In P. T. Sampson, D. Gil (Ed.), (Vol. Language complexity as an evolving variable, p. 110-125). Oxford University Press, Oxford.

Nowak, M., Komarova, N., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*(5501), 114.

Richerson, P. J., Boyd, R., & Bettinger, R. L. (2009). Cultural innovations and demographic change. *Hum Biol*, *81*(2-3), 211-35.

Sapir, E. (1912). Language and environment. *American Anthropologist*, *14*(2), 226–242.

Schachter, J. (1974). An error in error analysis. *Language Learning*, *24*(2), 205–214.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, *104*(18), 7361-7366.

Sinnemäki, K. (2009). Complexity and size of speech community. In P. T. Sampson, D. Gil (Ed.), (Vol. Language complexity as an evolving variable, p. 126-141). Oxford University Press, Oxford.

Smith, K., & Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3591-3603.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469-489.

The "Five Graces Group", Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*, 1-26.

Trudgill, P. (2004). Linguistic and social typology: the austronesian migrations and phoneme inventories. *Linguistic Typology*, *8*(305-320).

Vogt, P., & De Boer, B. (2010). Editorial: Language evolution: computer models for empirical data. *Adaptive Behavior-Animals, Animats, Software Agents, Robots, Adaptive Systems*, *18*(1), 5–11.

Wikipedia. (2011). *List of flags by number of colors. online, accessed 12-04-2011. http://en.wikipedia.org/wiki/list_of_flags_by_number_of_colors.*

Winters, J. (2011). *Crossing those curious parallels: Investigating the relationships between segment inventory size, demography and subsystem tradeoffs.* Unpublished master's thesis, Cardiff University.