

# Approaches to language complexity

Symposium Proposal for LSA Annual Meeting, Anaheim 2007

Organizers

K. David Harrison\* and Ryan K. Shosted†

October 4, 2006

## 1 Proposal

It has become almost axiomatic in linguistics that all languages are equally complex (Akmajian et al. (1997: 8), O’Grady et al. (1997: 6), Cipollone et al. (1998: 2), and O’Grady et al. (2005: 7)). This axiom is often coupled with the notion that all languages are “capable of expressing any idea” (Fromkin and Rodman 1988). Beyond linking expressivity to complexity, a number of further assumptions follow, for example: “A language which appears simple in some respects is likely to be more complex in others” (Markowicz 1978). The latter is often popularly expressed as the notion that a language that gains complexity in one part of its grammar necessarily becomes simplified elsewhere, as if regulated by a thermostat. Though this last idea is difficult to find in print, we believe it is nonetheless an unexpressed assumption in the field (cf. Plank (1998) and Shosted (2006)).

Despite the repetition (and potential reification) of these claims, they have seldom been subjected to rigorous tests. This may be due to the fact that linguists have not agreed upon metrics for complexity, though various proposals have been made (Greenberg 1954; Nichols 1992; Kusters 2003). Many still wonder whether complexity can be quantified within a single linguistic domain like phonology, let alone across domains. Moreover, most of the world’s languages remain undescribed or underdescribed, severely limiting the reach of typological approaches.

We feel it is an opportune time to re-examine this constellation of ideas in terms of their intellectual pedigree, their current status (i.e. what motivates the prevailing claims if not empirical evidence?), their impact on linguistics (in research, theory, pedagogy, etc.), and recent attempts to subject them to quantification, modeling, and empirical testing. In general, we hope to approach an answer to the question, “Is equal complexity among languages a reality?”

---

\*Swarthmore College: [dharris2 at swarthmore.edu](mailto:dharris2@swarthmore.edu)

†University of California, Berkeley: [shosted at berkeley.edu](mailto:shosted@berkeley.edu)

In light of the prevailing apprehension towards measurements of complexity, the organizers and panelists openly endorse a quantitative, algorithmic approach. Among the invited speakers, each presents a different quantitative metric for complexity that may be questioned by the linguistic community. Maddieson focuses on the complexity of syllables and syllabic inventories as a window to phonological complexity. Nichols introduces various ‘proxy measures’ that can be used to typify the complexity of many languages in an efficient manner. Pellegrino, Coupé, and Marsico ground their work with speech corpora in Information Theory. Wells-Jensen uses speech errors in a variety of languages as her metric. Finally, Whalen leads the discussion into the realm of brain function. Each speaker employs an experimental approach to the topic and each method is inherently quantitative.

It is hoped that through a lively dialogue on the subject of what to measure and how to measure it, linguists can converge on a set of metrics for complexity and use increasingly sophisticated methods in collecting relevant data.

The proposed symposium ‘Approaches to language complexity’ will be organized for the LSA annual meeting in January 2007. Abstracts have been invited from a broad spectrum of practitioners, with the goal of achieving a diversity of views. A goal is to produce an edited, refereed volume to include papers by panelists and other solicited authors.

## 2 Invited speakers and abstracts

The three-hour symposium will begin with a brief introduction by David Harrison. It will proceed with five talks of twenty-five minutes and five minutes for discussion of each. Opportunity for general discussion or follow-up will be provided at the end. Each of the speakers has agreed to participate in the symposium in Anaheim.

### 2.1 Ian Maddieson

Department of Linguistics  
University of California  
1203 Dwinelle Hall #2650  
Berkeley, California 94720-2650  
`ianm at berkeley.edu`

“COMPLEXITY RELATIONSHIPS IN PHONETIC AND PHONOLOGICAL SYSTEMS”

It is often suggested that languages are likely to ‘compensate’ complexity in one subsystem by simplicity elsewhere. Maddieson (2006a, b) presented some evidence against this idea in examining several basic phonological subsystems in a set of over 600 languages selected to represent genetic and areal diversity. Pairwise relationships between elaboration of the syllable canon, the size of segment inventories and the complexity of tone systems were studied. The languages were classed into three groups on the syllable complexity and tonal complexity variables. Sizes of consonant, total vowel and vowel quality inventories are

numeric values. A significant positive correlation exists between the syllable complexity and the size of consonant inventory, but no correlation between syllable complexity and size of total vowel or vowel quality inventories or between consonant and vowel inventory sizes. Complex syllable structure showed some association with absence of tone, but none of the other comparisons indicate that complexity is systematically compensated for by simplicity elsewhere. In this presentation these findings will first be reviewed and updated on the basis of the latest version of the sample of languages. Given that the general absence of compensatory relationships is considered surprising by many linguists, the discussion will be extended to include other factors in the languages' phonetic and phonological domains that might be held to be relevant to complexity, and to an evaluation of whether a different encoding of the variables considered would yield different results. The additional factors to be evaluated include the phonetic complexity of individual segments, the transparency vs opacity of relationships between forms in phonological processes, and phonological patterns at the level of the word, including average word length and phonotactic constraints on word structure. Alternative encodings will particularly be considered for the current syllable type variable, which can be refined into a larger number of classes, divided into separate factors for onset complexity and coda complexity or treated as one or more numeric variables. The data required to include these additional variables or recode previously simpler variables is often unavailable for a major proportion of the total language sample, hence these examinations will be based on smaller subsamples.

Two further issues will be touched on. The first is the question of whether an integrated measure of phonological complexity can be constructed through simultaneous consideration of multiple factors, and, if so, what would this show. Simple integrated measures of phonological complexity, such as the number of distinct syllables allowed, show vast differences between languages (Maddieson 1984; Shosted 2006), but adding in complexity at word-level and in opacity of relationships may conceivably reduce the variance. The second is the question of what explanation the relatively robust correlation between increasing syllable complexity and increasing size of consonant inventory requires. Candidates include chance, in both random and historic-accident formulations, and some form of necessity. It can easily be shown that there is no necessary association of these two variables, and the probability that it arose by random language-to-language variation is unlikely (that is what its statistical reliability means). An explanation that relies on historical accident is thus most probable. The differences between areas of language spread and residual enclaves (Bickel and Nichols 2003) presents a promising means of exploring how the correlation might have emerged.

## **2.2 Johanna Nichols**

Department of Slavic Languages and Literatures  
University of California, Berkeley  
johanna at berkeley.edu

Comprehensive measures of complexity, like that of Greenberg (1960) for morphological complexity, are revealing but time-consuming to assess; developing a comprehensive metric of complexity that would cover at least phonology, morphology, syntax, and lexicon and surveying it for a good-sized standardly-designed sample would be prohibitively time-consuming. This paper proposes a set of proxy and minimal properties that can be surveyed more economically, and surveys it in the Autotyp 230-language genealogically based sample. The worldwide distribution of complexity, measured in this way, is not even; areas and even macroareas have rather clear complexity profiles. The paper also surveys correlations of complexity with some sociological and sociolinguistic variables widely believed to correlate with complexity: literacy of the speech community, size of the speech community (shown by Hyslop (1993) to be a useful proxy for social complexity), known degree of contact with other languages, and known interethnic vs. ethnic-specific status.

### 2.3 François Pellegrino, Christophe Coupé and Egidio Marsico

Laboratoire Dynamique Du Langage  
 UMR 5596  
 Université Lumière Lyon 2  
 Francois.Pellegrino at univ-lyon2.fr  
 Christophe.Coupe at ish-lyon.cnrs.fr  
 Egidio.Marsico at ish-lyon.cnrs.fr

“CROSS-LINGUISTIC COMPARISON OF PHONOLOGICAL INFORMATION RATE”

All human languages are fully functional. Still, linguistic typology provides extensive and obvious evidence that as far as a given component (phonetics, phonology, morphology or syntax) is concerned, languages may be more or less complex. According to Information Theory, this would lead to the conclusion that the functional load associated with each linguistic component is language-dependent. This intuitive statement raises many questions about both the definition and measurement of the linguistic information and the possible correlation or compensation between the complexities of components within a language.

This paper addresses these issues and proposes an approach based on the study of a multilingual speech corpus, mainly focusing on the interaction between the phonetic (speech rate), phonological, and morphological levels (leaving the syntactic level aside).

Comparing languages in terms of complexity is far from straightforward, but Information Theory may provide an elegant answer. In this framework, we propose a corpus-based approach to comparing rates of linguistic information across languages, taking the phonetic level and the phonological inventories into consideration.

The MULTTEXT corpus (Campione and Véronis 1998) consists of short texts (about 100 syllables long) read by several speakers. The texts are available

in seven languages (English, French, German, Italian, Japanese, Mandarin and Spanish), providing comparable semantic material uttered in languages exhibiting different phonological and morphological strategies.

Considering the inventories of phonological segments and the distribution of syllabic structures, it is possible to estimate the average quantity of information conveyed by each phoneme, syllable, and word using relevant levels of labeling of the corpus.

Given that the texts express the same semantic content in the seven languages, it is thus possible to estimate to what extent a language relies on the phonetic, phonological, or morphological level to convey linguistic information.

For instance, it is possible to assess whether a language with a small phonological inventory will compensate by producing speech at a faster speech rate or not. Obviously, the very limited language inventory does not permit drawing general conclusions, but the proposed methodology may be extended to other languages with available lexical data or corpora.

This methodology has been evaluated with the English and Japanese subsets of MULTTEXT and will be extended to the five other languages for the presentation. Speech rates were estimated using an automatic detection algorithm (Pellegrino et al. 2004). Phonological information rate was defined as the entropy of the syllables (Shannon 1948), taking both the phonological inventories and the frequencies of occurrence of the syllable structures in large speech corpora (Arai and Greenberg 1997) into account, and weighted by the estimated speech rates (Karlsgren 1961). Preliminary results have shown that:

1. The average information load of the syllables is higher for English than Japanese;
2. The average syllabic speech rate is lower for English than Japanese;
3. The overall average information rates are similar in the two languages, showing a compensation effect.

In the full paper, this study will be extended (a) to the five other languages and (b) to the morphological level, through morphological labeling of the data. It will thus enable us to estimate the rate of morphological information and to consider phonological-morphological interaction effects as well.

## 2.4 Sheri Wells-Jensen

English Department  
Bowling Green State University  
423 East Hall  
Bowling Green, OH 43403  
swells at bgnet.bgsu.edu

“A COMPARATIVE, PSYCHOLINGUISTIC INVESTIGATION OF LANGUAGE COMPLEXITY”

This paper approaches the topic of language complexity from a strictly psycholinguistic perspective, relying on measurable production phenomena (speech errors) as its data. It is a systematic, cross-linguistic examination of speech errors in English, Hindi, Japanese, Spanish and Turkish. (See Fromkin (1973, 1980); Baars (1992) for discussions of the relevance of speech error data to linguistics.)

Participants narrated a fast-paced silent film. Their narrations were recorded and the 1,300 resulting errors were categorized. The speech error corpora that resulted are fully parallel, i.e., conditions were held constant while data from each language were gathered and analyzed. Thus, these corpora, unlike corpora of naturally-occurring speech errors, can be directly compared with one another.

The data were used to examine two interrelated hypotheses about the relationship between language structure and the speech production system. Both of the following hypotheses were supported:

**HYPOTHESIS A:** “As measured in this way, languages are equally complex.” No overall differences were found in the numbers of errors made by speakers of the five languages in the study. This supports the common assumption that no language is functionally more difficult than any other. If there are differences in complexity, either these differences are strictly formal, rather than functional, or they are functionally insignificant, such that the human language production system deals with them in any quantity without noticeable difficulty. (See MacKay (1982); Navon and Gopher (1979); Schweizer (1996) for discussions of how more complicated tasks cause more human error).

**HYPOTHESIS B:** “The patterns of distribution of different types of errors will be distinct from one language to another.” In the five languages studied, errors tended to cluster around loci of apparent formal complexity within each language. Languages such as Turkish and Spanish, which have more inflectional morphology, exhibited more errors involving inflected forms, while languages such as Japanese, with rich systems of closed-class forms, tended to have more errors involving closed-class items. This supports two widely-expressed beliefs about language. First, morphosyntactic phenomena do represent real tasks with which the speech production system must cope. Second, it appears that a language which is simple in one aspect tends to be more complex in another. For example, languages in the study with relatively few errors involving closed-class items tended to exhibit more errors in other areas such as inflectional morphology. On the other hand, amounts of errors in phonology were roughly the same across the languages studied.

## **2.5 Douglas H. Whalen**

Haskins Laboratories  
300 Georget Street  
New Haven, CT 06511  
whalen at haskins.yale.edu

“BRAIN ACTIVATIONS RELATED TO CHANGES IN SPEECH COMPLEXITY”

Speech is perhaps the most complex sound that humans pay attention to. Its complexity allows for the transmission of an impressive amount of information over a relatively weak sensory channel. Our previous functional Magnetic Resonance Imaging study found evidence for regions in speech-specific areas that increase activation with increases in what we called “complexity”. The remaining question is which aspects of the stimulus are contributing to this effect. In a follow-up study, syllables were presented auditorily in an event-related paradigm, and the hemodynamic response to each was measured. Ten syllables were used: /a, ta, la, ka, da, sa, sta, tag, skla, sklag/. Small areas within the most posterior portion of the speech area in superior temporal gyrus increased activation with increasing complexity defined in several different ways. Of particular interest, contrasting /sta/ with /tag/ showed increased activation for /sta/ (here number of segments was constant but number of syllable slots varied). This may be due to the greater complexity of the timing (phasing) relationships of segments within the onset versus the simpler timing relationships when both onset and coda are filled. To explore this further, a new study is being performed that contrasts more syllables with clusters and changes in number of syllable slots used. Clusters will either contain /s/, as before, or /l/. Codas as well as onsets will contain clusters. The correlation of brain activation with the complexity that can be associated with these changes to number of segments and composition of the syllable will then be assessed. The use of a simple passive listening task holds the promise of indicating which linguistic structures exemplify greater complexity.

## References

- Akmajian, A., R. A. Demers, A. K. Farmer, and R. M. Harnish (1997). *Linguistics: An Introduction to Language and Communication* (4th ed.). Cambridge: MIT Press.
- Arai, T. and S. Greenberg (1997). The temporal properties of spoken Japanese are similar to those of English. In *Proceedings of Eurospeech*, Rhodes, Greece, pp. 1011–1014.
- Baars, B. J. (Ed.) (1992). *Experimental Slips and Human error: Exploring the Architecture of Volition*, New York. Plenum.
- Campione, E. and J. Véronis (1998). A multilingual prosodic database. In *Proceedings of ICSLP 98*, Sydney, Australia, pp. 3163–3166.
- Cipollone, N., S. H. Keiser, and S. Vasisht (1998). *Language Files: Materials for an Introduction to Language and Linguistics* (7th ed.). Columbus, OH: Ohio State University Press.
- Fromkin, V. A. (1973). *Speech Errors as Linguistic Evidence*. The Hague: Mouton and Company.

- Fromkin, V. A. (Ed.) (1980). *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*, New York. Academic Press.
- Fromkin, V. A. and R. Rodman (1988). *An Introduction to Language*. New York: Holt, Rinehart, and Winston.
- Greenberg, J. (1954). A quantitative approach to the morphological typology of language. In R. F. Spencer (Ed.), *Method and Perspective in Anthropology: Papers in Honor of Wilson D. Wallis*, pp. 192–220. Minneapolis, MN: University of Minnesota Press.
- Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *IJAL* 26, 178–194.
- Hyslop, C. (1993). Towards a typology of spatial deixis. Honours thesis, Australian National University.
- Karlgren, H. (1961). Speech rate and information theory. In *Proceedings of 4th ICPHS*, pp. 671–677.
- Kusters, W. (2003). *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht, Netherlands: Landelijke Onderzoekschool Taalwetenschap (LOT).
- MacKay, D. (1982). The problems of flexibility, fluency and speed-accuracy trade-off in skilled behavior. *Psychological Review* 89, 483–506.
- Maddieson, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.
- Markowicz, H. (1978). *American Sign Language: Fact and Fantasy*. Gallaudet College. Online version (2001) accessed, <http://facstaff.gallaudet.edu/harry.markowicz/asl/myth4.html>, August 2005.
- Navon, D. and D. Gopher (1979). On the economy of the human processing system. *Psychology Review* 86, 214–255.
- Nichols, J. (1992). *Linguistic Diversity in Space and Time*. Chicago, IL: University of Chicago Press.
- O’Grady, W., J. Archibald, M. Aronoff, and J. Rees-Miller (2005). *Contemporary Linguistics: An Introduction* (5th ed.). New York: St. Martin’s Press.
- O’Grady, W., M. Dobrovsky, and M. Aronoff (1997). *Contemporary Linguistics: An Introduction* (3rd ed.). New York: St. Martin’s Press.
- Pellegrino, F., J. Farinas, and J.-L. Rouas (2004). Automatic estimation of speaking rate in multilingual spontaneous speech. In *Proceedings of Speech Prosody 2004*, Nara, Japan, pp. 517–520.

- Plank, F. (1998). The co-variation of phonology with morphology and syntax: A hopeful history. *Journal of Linguistic Typology* 2, 195–230.
- Schweizer, K. (1996). The speed/accuracy transition due to task complexity. *Intelligence* 22(2), 115.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal* 27, 379–423.
- Shosted, R. K. (2006). Correlating complexity: A typological approach. *Journal of Linguistic Typology* 10. To appear.