Quantifying and Measuring Morphological Complexity

Max Bane
bane@uchicago.edu

Department of Linguistics
University of Chicago

WCCFL 26, April 27–29 2007

The Plan

Linguistic Complexity

An Unpopular Topic

"[A]ll known languages are at a similar level of complexity and detail — there is no such thing as a primitive language." (Akmajian et al 1997)

"In sum, linguists don't even think of trying to rate languages as good or bad, simple or complex." (O'Grady et al 2005)

The Equal Complexity Hypothesis (Truism)

- Received wisdom: Every language is equally (enormously) complex.
- $\Rightarrow$ A language with, say, very simple morphology must compensate with elaborated phonology, syntax, etc.
- A traditionally untested claim
  - "Complexity" is a loaded word.
  - Difficulties of scope.
  - It's not immediately obvious how to approach complexity in a principled, quantitative way.
  - Controversy! McWhorter (2001): "The world's simplest grammars are creole grammars."

An Important Hypothesis

- The equal complexity hypothesis deserves formal articulation and scrutiny.
  - An empirical claim to be tested under particular definitions of complexity.
- Important because of ramifications for:
  - The predictive power of historical linguistics:

    $$\Delta\text{Complexity} = 0$$

  - Theories of grammar and representation
  - Psycholinguistics and Cognitive Science

Goals of This Talk

- Two main points:
  - Let's measure complexity!
  - Information Theory provides powerful formalisms for approaching just this sort of question.

Defining Linguistic Complexity

"How to measure . . . complexity is itself an issue of some complexity." (Nichols 1992)

Existing Metrics

- Have been proposed by Nichols (1992, 2007), McWhorter (2001), Shosted (2005), and others.
- General method: Count the occurrences of a variety of hand-picked, intuitively justified properties of the linguistic system.
- Phonological Complexity
  - Size of phoneme/syllable inventory.
  - Number of “marked” phonemes.
  - Number of rules/alternations.
- Morphological Complexity
  - Number of possible inflection points in a “typical” sentence.
  - Number of inflectional categories, morpheme types.
  - AUTOTYP “synthesis” (Bickel & Nichols 2005).
- Syntactic Complexity
  - Number of parameters deviating from default.

Problems

- Conversion and meaningful comparison of measured variables
  - E.g., how many phonemes is a given rule worth?
- What is the guiding principle in selecting relevant properties?
  - Overspecification? Communicative non-necessity? How do we determine that, exactly?
- Can be impressionistic.

Information Theory

- A guiding principle — McWhorter (2001) hits on it:
  - *[S]ome grammars might be seen to require lengthier descriptions in order to characterize even the basics of their grammar than others.*
- Put another way:
  - We're interested in how much "information" is contained in the most concise description of (the grammar of) a given linguistic system.
- What is information?
  - Bits

The Complexity of Strings

- Assumption: any grammar, linguistic system, module, set of observations, whatever can be encoded systematically as a string over some alphabet.
- Information theory offers a notion of complexity for strings
- Intuitively, which of (1) and (2) is more complex?
  (1) 10101010101010101010
  (2) 11011111000101011010

The Complexity of Strings

- Assumption: any grammar, linguistic system, module, set of observations, whatever can be encoded systematically as a string over some alphabet.
- Information theory offers a notion of complexity for strings
- Intuitively, which of (1) and (2) is more complex?
  (1) 10101010101010101010
  (2) 11011111000101011010
- (1) = "*10* ten times"

The Complexity of Strings

- Assumption: any grammar, linguistic system, module, set of observations, whatever can be encoded systematically as a string over some alphabet.
- Information theory offers a notion of complexity for strings
- Intuitively, which of (1) and (2) is more complex?
  (1) 10101010101010101010
  (2) 11011111000101011010
- (1) = "*10* ten times"
- (2) = ??
  - Need to list the string itself.
- $\Rightarrow$ (2) is more complex.

Kolmogorov Complexity

The length of the shortest description

- Formalized as Kolmogorov Complexity (Solomonoff 1964, Kolmogorov 1965)
- The Kolmogorov complexity $K_L(s)$ of a string s relative to a description language L is the length of the shortest $d \in L$ such that d "describes" s.
- Kolmogorov complexity is measured in *bits*.
- Two issues
  - What's it mean to describe a string?
  - Finding the shortest description.

The Description Language

- English as a description language.
  - Too poorly understood. When does a given statement "describe" something?
- Solution: Programming languages
  - A program P describes a string s iff P outputs s.
  - $\Rightarrow$ Relative to a programming language L, $K_L(s)$ is the length of the shortest program $P \in L$ such that P outputs s.
- But which programming language?
  - Short answer: it (provably) doesn't matter.
  - Long answer: the Gödel numbers of the Universal Turing Machine.

Kolmogorov Complexity is Approximable

- Can never be sure we've found the absolute shortest description.
- This is rarely a problem in practice.
- We can approximate Kolmogorov complexity by computing upper bounds on it.
  - Numerous compression algorithms (Zip, RAR, SIT, etc.)
  - Description Length
- $\Rightarrow$ Kolmogorov complexity serves as an idealized, general purpose definition of complexity as a quantity.
  - Approximable as necessary.

Applying Kolmogorov Complexity to Linguistic Systems

- To craft a complexity metric we must answer three questions:
  - What exactly is the object or system whose complexity we're after?
  - How will we encode that object as a string?
  - What method will we use to approximate the Kolmogorov complexity of that string?

Complexity of Grammars

- Object: the *grammar* that generates/accepts the language.
  - Or some component thereof ...
  - Phonological grammar
  - Morphological grammar
  - etc.
- $D \rightarrow G \rightarrow \lambda$
  - G generates the language λ, a set of licit strings.
  - G is a grammar devised by linguists for λ.
  - D is the shortest description of (i.e., computer program that outputs) G.
  - Our ideal complexity metric is $|D|$.

Why (Inflectional, Affixal) Morphology?

- A common domain for existing complexity metrics (Nichols 1992, Juola 1998, McWhorter 2001, Shosted 2005)
- McWhorter (p.c. in Shosted 2005): "usually richer and more widespread interaction with syntax, this interaction being of note in covering the general issue of complexity more widely."
- A simple model (string encoding) of affixal morphology:
  - A lexicon
  - $\Rightarrow$ All morphemes (stems & affixes) and descriptions of their distribution ("signatures")
- . . . As produced by an automatic morphological analyzer ("lemmatizer").
  - Linguistica (Goldsmith 2001, 2006; Yu 2007).

Example

- French:

Stem	*Suffixal Signature*
accompli	∅.e.t.r.s.ssent.ssez
académi	cien.e.es.que
académicien	∅.s

finale+*ment l'*+**ange** me **montr**+*a* le fleuve de la vie limpide comme du cristal qui **jaillissai**+*t* du trône de dieu et de *l'*+**agneau**
au milieu de *l'*+**avenue** de la ville **entr**+*e* deux bras du fleuve se **trouv**+*e l'*+**arbre** de vie il **produi**+*t* douze **récolte**+*s* chaque mois il **port**+*e* son fruit ses **feuill**+*es* **serve**+*nt* à **guéri**+*r* les **nati**+*ons* (Apocalypse 22)

A Complexity Metric for Affixal Morphology

- Linguistica tries to minimize the K. complexity of the grammars it induces.
  - Kolmogorov approximation: “Description Length” (DL)
  - Gives us enough information to construct a simple metric
- Complexity (DL) of the morphological lexicon distributed between:
  - Stems
  - Affixes
  - Signatures
- A morphologically simple language will have. . .
  - Most words analyzed as monomorphemic (stems)
  - Few affixes, few signatures
- A morphologically complex language will have. . .
  - Fewer monomorphemic words
  - More affixes, more signatures

A Complexity Metric for Affixal Morphology

- Metric:

$$\frac{DL(Affixes)+DL(Signatures)}{DL(Affixes)+DL(Signatures)+DL(Stems)}$$

  where $DL(x)$ = complexity (description length) of x (bits).
- The proportion of total lexicon complexity due to affixes and description of their distribution
  - A ratio of bits
- *Not* just counting *number* of affixes, etc.
- Caveats
  - Very poor for non-affixal morphology (templatic, infixal, reduplicative, etc.)
  - Linguistica isn't perfect

Exploratory Measurements: Method

- Two sets of written corpora
  - Bible translations (“fixed” semantics)
  - Texts collected from the web (by Kevin Scannell’s language-targeting web-crawler)
- Corpora analyzed by Linguistica
  - Produces lexicon and description length figures
- Metric computed

Exploratory Measurements: Bible Corpora

Language	*Metric*	*Types*	*Tokens*
Latin	35.51%	46,722	604,305
Hungarian	33.98%	63,046	597,084
Italian	28.34%	35,232	774,946
Spanish	27.50%	29,021	664,108
Icelandic	26.54%	34,911	667,363
French	23.05%	31,684	728,191
Danish	22.86%	24,280	653,036
Swedish	21.85%	23,964	675,315
German	20.40%	24,692	729,853
Dutch	19.58%	21,242	727,489
English	16.88%	15,570	737,241
Maori	13.62%	8,271	977,565
Haitian Creole	2.58%	7,307	920,332
Vietnamese	0.05%	7,144	837,733

- Metric-Types Correlation: 0.89 ($p = 0.002\%$)
- Metirc-Tokens Correlation: -0.79 ($p = 0.07\%$)

Exploratory Measurements: All Corpora

Language	*Metric*	*Language*	*Metric*
Latin	35.51%	English	16.88%
Hungarian	33.98%	Maori	13.62%
Italian	28.34%	Papiementu*	10.16%
Spanish	27.50%	Nigerian Pidgin*	9.80%
Icelandic	26.54%	Tok Pisin*	8.93 %
French	23.05%	Bislama*	5.38%
Danish	22.86%	Kituba*	3.40%
Swedish	21.85%	Solomon Pijin*	2.91%
German	20.40%	Haitian Creole*	2.58%
Dutch	19.58%	Vietnamese	0.05%

* = Creole/Pidgin

Summary

- The Equal Complexity Hypothesis could have important ramificatioins.
- Information Theory offers a general notion of complexity, and a methodology.
- Applied to morphological grammars
  - Via an automatic lemmatizer (Linguistica)
- Preliminary measurements
  - Reasonable relative rankings
  - Creoles are least complex

The Future

- Applying to more principled models of morphological grammar
- Other information theoretic approaches to morphological complexity
  - Juola 1998, to appear
- Correlation with other quantities?
  - Production errors
  - Speech rate
- Metrics for other grammatical domains
- System complexity vs processing complexity
- Information theoretic typology of complexity
  - Test the Equal Complexity Hypothesis

Thank You

Special thanks to Alan Yu, Jason Riggle, Salikoko Mufwene, Jason Merchant, Kevin Scannell, and Patrick Juola for valuable discussion and advice.