AN INFORMATION THEORY-BASED APPROACH TO THE BALANCE OF COMPLEXITY BETWEEN PHONETICS, PHONOLOGY AND MORPHOSYNTAX

François PELLEGRINO, Christophe COUPÉ & Egidio MARSICO
Dynamics of Language Lab
UMR5596 CNRS - Université de Lyon
FRANCE

LSA Complexity Panel

This research is supported by the French National Research Agency (ANR)
Overview

- General framework
- Proposed IT-based approach
- Data
- Results
Information Theory & Complexity

✓ Information Theory (IT) and Entropy
   - Initiated from C. Shannon’s works in the 1940’s
   - Applied to linguistics, cybernetics and cognitive science
   - Related to notions like functional load, statistics, complexity

✓ Quantitative typology and sciences of Complexity
   - Shed new light on typological issues (and cognition)
   - Quest for correlations and compensations among linguistic components (phonetics, phonology, morphology, syntax, etc.)
   - Problems when dealing with complexity:
     - No straightforward definition
     - Multidimensional problem

✓ IT may provide relevant tools to evaluate complexity
IT and phonology: Recent revival?


- Compression approach (Juola 1998, Kettunen et al., 2006)

- Functional load approach (Surendran and Niyogi, 2004-2006)

- Probabilistic approaches (among others, Goldsmith, 2002; Hume, 2004-2006, ...
Issues and stakes

✓ **Typological approaches involving phonology**
  - Often leave phonetics aside
    - Either because of difficulties or considering it as irrelevant

✓ **Complexity balance between linguistic levels**
  - “Slippery” issue
  - Several studies challenged this statement
    - Different indices lead to different results
    - No universal methodology (yet)

✓ **Goals of this study**
  - Evaluate whether IT is relevant in this context
  - Draw attention on some methodological pitfalls
    - Cross-linguistic comparison based on “tiny” corpora
    - (too) coarse-grained evaluation of indices
    - Interaction between phonetics and phonology (speech rate)
Proposed Approach

- For several languages, texts conveying an “equivalent” semantic content
- Estimation of parameters from these texts uttered by several speakers
  
  Increase the number of texts and speakers to “neutralize” within-language variability and get significant cross-linguistic results.
- Estimation of the information carried by linguistic units in these languages
  - For given units, how to calculate Information?
  - How to choose the units compatible with the chosen approach to estimation of information?
Proposed Approach

How to calculate the Information

Considering that language $L$ is a source of linguistic sequences $s$ composed of units ($u$) from a finite set ($N_L$)

Assuming that the units are independent from each other

- $s(t) = u_1 u_2 u_3 ... u_{t-1} u_t$
- $P(u_t)$ is supposed to be independent of $s(t-1) = u_1 u_2 u_3 ... u_{t-1}$

Quantity of Information of unit $u = \text{entropy} \quad h(u) = -\log_2(P(u))$

- Less probable => more informative
- More probable => less informative
- Certain ($P = 1$) => no information

$H(L)$ Quantity of Information of $L$ (Entropy of $L$)

- Easy to compute from the set of units and their probabilities
- $H(L)$ is always inferior to $\log_2(N_L)$

$$H(L) = \sum_{i=1}^{N_L} P_{u_i} h(u_i) = -\sum_{i=1}^{N_L} P_{u_i} \log_2\left(P_{u_i}\right)$$
Proposed Approach

Information Estimation

✔ What linguistic units?
    ✷ That do not violate too heavily the independence rule?
    ✷ For which the inventory is known and $P_u$ (probability of occurrence) is calculable.

✔ Our choice: the syllable

✎ Pros
    ➢ Higher independence than between phonemes (or features, gestures?)
    ➢ Syllable frequency estimated from big written corpora for several languages
    ➢ Way to somewhat get rid of coarticulation issues

✎ Cons
    ➢ Independence is not absolute
    ➢ Relative frequencies differ from written to speech production
    ➢ Oral syllables resulting from phonological processes (elision, liaison, etc.) are absent from written data
    ➢ Relevance of syllable as a linguistic unit across languages?
EXPERIMENTAL DATA
Need for comparable data across languages

2 types of data

- **Speech data** : subset of MULTEXT corpus
  - 7 languages (5 European languages, 2 East-Asian languages)
  - 20 Passages (texts composed of 5 semantically connected sentences)
  - 4-10 speakers per text
  - Broadly speaking, semantically equivalent texts in each language
    - Last night I opened the front door to let the cat out. It was such a beautiful evening that I wandered down the garden for a breath of fresh air. Then I heard a click as the door closed behind me. I realised I'd locked myself out. To cap it all, I was arrested while I was trying to force the door open!
    - Hier soir, j'ai ouvert la porte d'entrée pour laisser sortir le chat. La nuit était si belle que je suis descendu dans la rue prendre le frais. J'avais à peine fait quelque pas que j'ai entendu la porte claquer derrière moi. J'ai réalisé, tout d'un coup, que j'étais fermé dehors. Le comble c'est que je me suis fait arrêter alors que j'essayais de forcer ma propre porte !

- **Syllable Frequency data**
  - Computed from large text resources (newspapers, books, etc.)
  - Different resources depending on the languages
## Available Data for this study

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>Code</th>
<th>SPEECH DATA</th>
<th>SYLLABLE FREQUENCIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>EN</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>French</td>
<td>FR</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>German</td>
<td>GE</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Italian</td>
<td>IT</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Japanese</td>
<td>JA</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mandarin Chinese</td>
<td>MA</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Spanish</td>
<td>SP</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dutch</td>
<td>DU</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>VI</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>
## Speech Data

<table>
<thead>
<tr>
<th>Language</th>
<th>Source</th>
<th>No of speakers</th>
<th>Total duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN</td>
<td>Multext</td>
<td>10</td>
<td>18 min.</td>
</tr>
<tr>
<td>FR</td>
<td>Multext</td>
<td>6</td>
<td>14 min.</td>
</tr>
<tr>
<td>GE</td>
<td>Multext</td>
<td>10</td>
<td>27 min.</td>
</tr>
<tr>
<td>IT</td>
<td>Multext</td>
<td>10</td>
<td>18 min.</td>
</tr>
<tr>
<td>JA</td>
<td>Kitazawa, 2002</td>
<td>5</td>
<td>33 min.</td>
</tr>
<tr>
<td>MA</td>
<td>Komatsu et al., 2004</td>
<td>9</td>
<td>23 min.</td>
</tr>
<tr>
<td>SP</td>
<td>Multext</td>
<td>8</td>
<td>17 min.</td>
</tr>
<tr>
<td>VI</td>
<td>Courtesy of E. Castelli MICA, Hanoi</td>
<td>4 (two times each)</td>
<td>38 min.</td>
</tr>
</tbody>
</table>

**OVERALL** 62 > 3 hours

Small but not tiny corpus
# Syllable Frequencies

<table>
<thead>
<tr>
<th>Language</th>
<th>Source</th>
<th>No of different syllables</th>
<th>Total No of syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>DU</td>
<td>WebCelex</td>
<td>6 486</td>
<td>1.4 M</td>
</tr>
<tr>
<td>EN</td>
<td>WebCelex</td>
<td>7 931</td>
<td>1.0 M</td>
</tr>
<tr>
<td>FR</td>
<td>Lexique3</td>
<td>5 685</td>
<td>1.3 M</td>
</tr>
<tr>
<td>GE</td>
<td>WebCelex</td>
<td>4 207</td>
<td>0.8 M</td>
</tr>
<tr>
<td>IT</td>
<td>PhD Massimiliano Pone, 2005</td>
<td>2 719</td>
<td>27.0 M</td>
</tr>
<tr>
<td>JA</td>
<td>Tamaoka and Makioka, 2004</td>
<td>416</td>
<td>575.7M</td>
</tr>
<tr>
<td>MA</td>
<td>PhD Peng Gang, 2005, (incl. tones)</td>
<td>1 191</td>
<td>138.0 M</td>
</tr>
<tr>
<td>SP</td>
<td>PhD Massimiliano Pone, 2005</td>
<td>1 593</td>
<td>0.9 M</td>
</tr>
</tbody>
</table>
RESULTS

✓ Speech corpus
✓ Syllabic Entropy
✓ Back to initial issues
CAVEAT

✓ Very Few languages => no typological dimension

✓ Differences between oral and written syllables may be significant (but invisible here)

✓ Descriptive and not explanatory results so far
Speech Corpora Comparison

Parameters

✓ **Raw parameters for each Passage**
  - Passage Duration (in seconds, Silences >= 150 ms are discarded)
  - Number of Syllables (from canonical pronunciation)
  - Number of Words (according to language-specific standards)

✓ **Normalized parameters**
  - Significant differences in length among Passages for a given language (e.g. from 62 to 104 syllables in the English corpus)
  - BUT Passages are matched among languages
  - Normalization procedure
    - Normalized Length (Syllables)
      - Ratio between the number of syllables in each Passage in language $L$ and the matched Passage in English
      - Median Value calculated among Passages for each language
    - Normalized Length (Words)
    - Normalized Duration (Time)

☐ **Additional parameters**
  - Syllabic Rate: Number of syllables per second (average value among Passages)
  - ASW = Average Number of Syllables per Word (average value among Passages)
High within-language variation

Cross-linguistic comparison CANNOT be done with just one utterance...

No significant correlation between Syllabic and Word Lengths
Cross-linguistic variation of Syllabic Rate

- Not only a speaker-specific parameter!
- Values are pretty high (compared to spontaneous speech)
- Inter-speaker variation is pretty low in this task (reading)

Somewhat linked to syllable structure (shell complexity)

But not only

- MA and VI exhibit low Syllabic Rates though their syllable structures is moderately complex
- Tone dimension may not be "orthogonal" to syllabic structure complexity
Speech Corpora Comparison (cont’d)
Are Syllabic Rate and Text Length linked?

- Normalized Duration does NOT correlate with Syllabic Rate
  - Speak faster does not mean speak shorter!
- Normalized Length (Syllable) correlates with Syllabic Rate
  - Is this just an artifact (Duration linked to Number of syllables?)
  - Is there any causality?
Cross-linguistic variations

How much of the potential offered by a given syllabic inventory is used?

Redundancy = difference between the maximum possible entropy for each inventory and the observed entropy in a written corpora

Pretty similar redundancy across languages

Comment: Sizes of Syllabic Inventories (calculated from corpora) are much lower than those computed just from phonotactic rules
Very high correlation between Syllabic Entropy and Normalized Length (Syllable)

- Syllabic entropy (or Information load of syllables) is efficient to quantify the linguistic amount of information

- Size of syllable inventory is not efficient to do so

=> Syllabic inventory (without frequency) is probably not informative enough for cross-linguistic comparison
Back to initial questions (cont’d)

Is there any trade-off between Syllabic Rate and Syllabic Entropy?

- Tendency to negatively correlate
  - Left-skewed distribution (Syllabic Rates)
  - Normalization through transformation
  - Significant correlation with exp(Syllabic Rate)
    - \( R^2 = 0.61 \) (p<.05)

Consequence on Syllabic Information Rate
- SIR = Syllabic Entropy x Syllabic Rate
- Amount of syllabic information per second
  - Syllabic Information Rate is not predictable from syllable inventory and probabilities

- Possible balance between syntagmatic and paradigmatic information
  - Cognitive (memory and process) load?

- Speech Rate matters when looking for correlations!
- Information RATE may be more important than Information LOAD
Back to initial questions (still cont’d)
What about morphosyntax?

✓ Limitations

- No explicit knowledge on morphology and syntax in the corpora
- Hypotheses: indirect indices related to morphosyntax
  - Normalized Length (Word) \( \Rightarrow \) not correlated to any index
  - ASW (Average Number of Syllables per Word)
    - Trade-off to limit word informational load (or complexity)

\[
\text{Normalized Length (Word) vs. SIR}
\]

\[
\text{ASW vs. Syllabic Entropy}
\]

\[ R^2 = 0.90 \; p < 0.01 \]
Conclusions & Perspectives

- **Results**
  - Methodology assessed on a small set of languages => no definitive conclusion
  - Syllabic Entropy seems to be relevant in terms of linguistic information
  - High Syllabic Entropy does not automatically result in high Syllabic Information Rate!

- **Take Home message**
  - If Information is what matters,
  - Just looking at descriptions or inventories of linguistic systems is not enough:
    - Pay attention to the SPEECH dimension!

- **Perspectives**
  - Take more phonetic and phonological factors of speech into account
  - Add more languages
    - Related languages to track historical trajectories (e.g. Romance languages)
    - Typologically distant languages
  - Rank these languages on morphological and syntactic scales of complexity
Thanks

Ching-pong AU, Solène BLANDIN, Eric CASTELLI,
Gang PENG, Miyuki ISHIBASHI,
Shogo MAKIOKA and Katsuo TAMAOKA
for their help with data processing

Ian MADDIESON

THANK YOU FOR YOUR ATTENTION
References


Karlsgen, H., 1961, “Speech Rate and Information Theory”, in proc. of 4th ICPhS, pp. 671-677


Maddieson, I., 2006, “Correlating phonological complexity: Data and validation”, Linguistic Typology, 10-1


Additional slides

- Normalization Procedure
- Suitable linguistic units?
- How to estimate Syllabic Entropy
- Methodology for Japanese and Chinese Mandarin
- Zipf-like curves
## NORMALIZATION

<table>
<thead>
<tr>
<th>Language</th>
<th>Data</th>
<th>Passages</th>
<th>MEDIAN</th>
<th>IQTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Passage EN</td>
<td>O1</td>
<td>O2</td>
<td>O3</td>
</tr>
<tr>
<td>EN</td>
<td>Nb Words</td>
<td>51</td>
<td>57</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Nb Syllables</td>
<td>72</td>
<td>86</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Mean duration</td>
<td>10,7</td>
<td>13,7</td>
<td>13,0</td>
</tr>
<tr>
<td></td>
<td>Normalized Length(Syl)</td>
<td>1,00</td>
<td>1,00</td>
<td>1,00</td>
</tr>
<tr>
<td></td>
<td>Normalized Length (Wd)</td>
<td>1,00</td>
<td>1,00</td>
<td>1,00</td>
</tr>
<tr>
<td></td>
<td>Normalized duration</td>
<td>1,00</td>
<td>1,00</td>
<td>1,00</td>
</tr>
<tr>
<td>FR</td>
<td>Nb Words</td>
<td>62</td>
<td>77</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Nb Syllables</td>
<td>81</td>
<td>106</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Mean duration</td>
<td>11,3</td>
<td>14,5</td>
<td>9,4</td>
</tr>
<tr>
<td></td>
<td>Normalized Length(Syl)</td>
<td>1,1</td>
<td>1,2</td>
<td>0,8</td>
</tr>
<tr>
<td></td>
<td>Normalized Length (Wd)</td>
<td>1,22</td>
<td>1,35</td>
<td>0,83</td>
</tr>
<tr>
<td></td>
<td>Normalized duration</td>
<td>1,06</td>
<td>1,06</td>
<td>0,72</td>
</tr>
</tbody>
</table>
Candidate linguistic units

What linguistic units?

- That do not violate too heavily the independence rule?
- For which inventory is known and $P_u$ (probability of appearance) is calculable.

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Independence</th>
<th>Calculable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme</td>
<td>✗ phonotactics</td>
<td>✓ from corpora</td>
</tr>
<tr>
<td>Syllable</td>
<td>✓ no too bad</td>
<td>✓ from corpora</td>
</tr>
<tr>
<td>Morpheme</td>
<td>✗ not really!</td>
<td>✗ morpheme count?</td>
</tr>
<tr>
<td>Phrase</td>
<td>✓ Not too bad</td>
<td>✗ from HUGE corpora?</td>
</tr>
<tr>
<td>Feature</td>
<td>✗</td>
<td>✓ from corpora?</td>
</tr>
<tr>
<td>Gesture</td>
<td>✓ To be explored...</td>
<td>✓ from corpora?</td>
</tr>
</tbody>
</table>
Estimation of Syllabic Entropy

How to estimate the information carried by syllables?
- Using syllable inventories AND syllable frequencies

How to estimate syllable inventories and frequencies?
1. Phonotactic constraints from language description (.CV, .CVC, etc.)
   => “skeleton” inventory, rough frequency
2. Lexicon or dictionary ([a], [pa], etc. from lexicon)
   => written/oral syllable inventory, “type” frequency
3. Written Corpora ([a], [pa], etc. from written production)
   => written syllable inventory, “token” frequency
4. Oral Corpora ([a], [pa], etc. from oral spontaneous data)
   => written syllable inventory, “token” frequency

Comparison of 2 methods in French
- From (written) syllable frequencies
  - Statistical evaluation (only the number of syllables present in a text is considered)
  - From syllabification (the observed syllables are individually taken into account)

* Except from very few languages
Estimation of Syllabic Entropy (cont'd)

How to evaluate the information carried by syllables?
1. From distribution of syllable types and phonological inventories
2. From distributions of syllables in corpora (.a., .e., .dø., .la. ...)

<table>
<thead>
<tr>
<th>Syllable type</th>
<th>% of occurrence (tokens)</th>
<th>Number of V</th>
<th>Number of C</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>60,4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td>12,5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CCV</td>
<td>9,2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CVC</td>
<td>11,6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>VC</td>
<td>1,6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CCVC</td>
<td>1,4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CVCC</td>
<td>1,4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>CCCV</td>
<td>0,4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Other</td>
<td>1,5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then, considering that vowels (resp. consonants) are equiprobable, syllable's information is the sum of consonantal and vocalic information:

\[ h(.CCVC.) = - (1 \times \log_2(1/N_v) + 3 \times \log_2(1/N_c)) \]

\[ H(L) = \text{sum of } h(.XXX.) \text{ weighted by } \% \text{ of occurrence} \]

⇒ Several significant approximations

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Number of Occurrences per M. of σ</th>
<th>Probability</th>
<th>Syllable's information</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>44 255</td>
<td>0,04</td>
<td>4,50</td>
</tr>
<tr>
<td>e</td>
<td>31 889</td>
<td>0,03</td>
<td>4,97</td>
</tr>
<tr>
<td>dø</td>
<td>30 472</td>
<td>0,03</td>
<td>5,04</td>
</tr>
<tr>
<td>la</td>
<td>21 752</td>
<td>0,02</td>
<td>5,52</td>
</tr>
<tr>
<td>el</td>
<td>18 200</td>
<td>0,02</td>
<td>5,78</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zlø</td>
<td>0,05</td>
<td>~ 0,00</td>
<td>24,19</td>
</tr>
<tr>
<td>zla</td>
<td>0,05</td>
<td>~ 0,00</td>
<td>24,19</td>
</tr>
<tr>
<td>zuk</td>
<td>0,05</td>
<td>~ 0,00</td>
<td>24,19</td>
</tr>
<tr>
<td>zwa</td>
<td>0,05</td>
<td>~ 0,00</td>
<td>24,19</td>
</tr>
<tr>
<td>zyrc</td>
<td>0,05</td>
<td>~ 0,00</td>
<td>24,19</td>
</tr>
</tbody>
</table>
Estimation of Syllabic Entropy (cont’d)

Evaluation from distribution of syllables in corpora

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Number of Occurrences per M. of σ</th>
<th>Probability</th>
<th>Syllable’s information</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>44 255</td>
<td>0.04</td>
<td>4.50</td>
</tr>
<tr>
<td>e</td>
<td>31 889</td>
<td>0.03</td>
<td>4.97</td>
</tr>
<tr>
<td>o</td>
<td>30 472</td>
<td>0.03</td>
<td>5.04</td>
</tr>
<tr>
<td>u</td>
<td>21 752</td>
<td>0.02</td>
<td>5.52</td>
</tr>
<tr>
<td>i</td>
<td>18 200</td>
<td>0.02</td>
<td>5.78</td>
</tr>
<tr>
<td>z</td>
<td>0.05</td>
<td>~ 0.00</td>
<td>24.19</td>
</tr>
<tr>
<td>zlo</td>
<td>0.05</td>
<td>~ 0.00</td>
<td>24.19</td>
</tr>
<tr>
<td>zla</td>
<td>0.05</td>
<td>~ 0.00</td>
<td>24.19</td>
</tr>
<tr>
<td>zuk</td>
<td>0.05</td>
<td>~ 0.00</td>
<td>24.19</td>
</tr>
<tr>
<td>zwa</td>
<td>0.05</td>
<td>~ 0.00</td>
<td>24.19</td>
</tr>
<tr>
<td>zygp</td>
<td>0.05</td>
<td>~ 0.00</td>
<td>24.19</td>
</tr>
</tbody>
</table>

Syllabification available?

YES

• Exact Calculation
• Text of n syllables
• $H(\text{Text}) = \sum \text{h(σ)}$, i from 1 to n

NO

• Statistical Estimation
• Text of n syllables
• $H(\text{Text}) = n \times H(L) = n \times \text{mean of h(σ)}$
French: Methodology comparison

<table>
<thead>
<tr>
<th>Syllabic Entropy</th>
<th>from syllable types</th>
<th>from syllable frequencies (statistical)</th>
<th>from syllable frequencies (exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRENCH</td>
<td>9.05</td>
<td>8.43</td>
<td>9.21</td>
</tr>
</tbody>
</table>

\[ y = 0.8638x + 27.93 \]

\[ R^2 = 0.9261 \]
Japanese and Mandarin Data

- **Japanese Word segmentation**

- **Mandarin syllabic frequencies**
  - Corpus of character frequencies (6526 different logographs)
  - Transposition to pinyin using the (1st rank) pronunciation for each character (software NJStar Chinese Word Processor)
  - 0.006% of the characters were not recognized by NJStar and left apart.
  - Syllabic frequencies estimation (1191 different syllables)
Relation between Frequency and Rank of Syllables

Log Frequency (per million of syl.) vs. Log Rank

Languages represented:
- EN
- FR
- GE
- IT
- JA
- MA
- SP

Dynamique Du Langage