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)

DYNAMIQUE DU LANGAGE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Dynamique Du Langage

UNIVERSITÉ LUMIÈRE LYON 2

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**

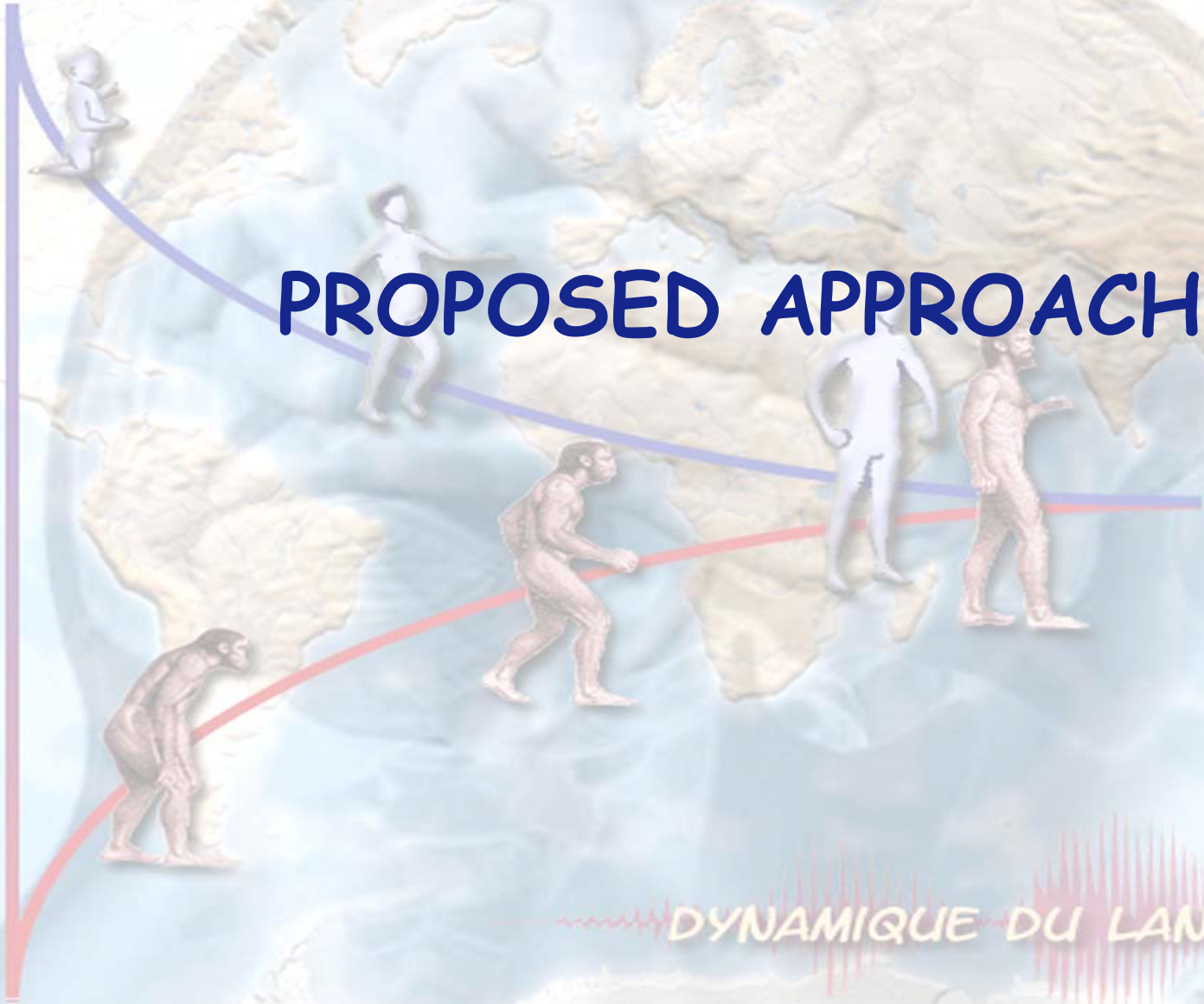
IT and phonology: Recent revival?

- ✓ Goldsmith (1998): *On information theory, entropy, and phonology in the 20th century*, Royaumont
- ✓ 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
    - Among others, Auer (1993), Planck (1998) Maddieson (1986, 2006), Fenk-Oczlon & Fenk (1999, 2005), Shosted (2006)
    - 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



DYNAMIQUE DU LANGAGE

Proposed Approach

- ✓ For several languages, texts conveying an "equivalent" semantic content
- ✓ Estimation of parameters from these texts uttered by several speakers

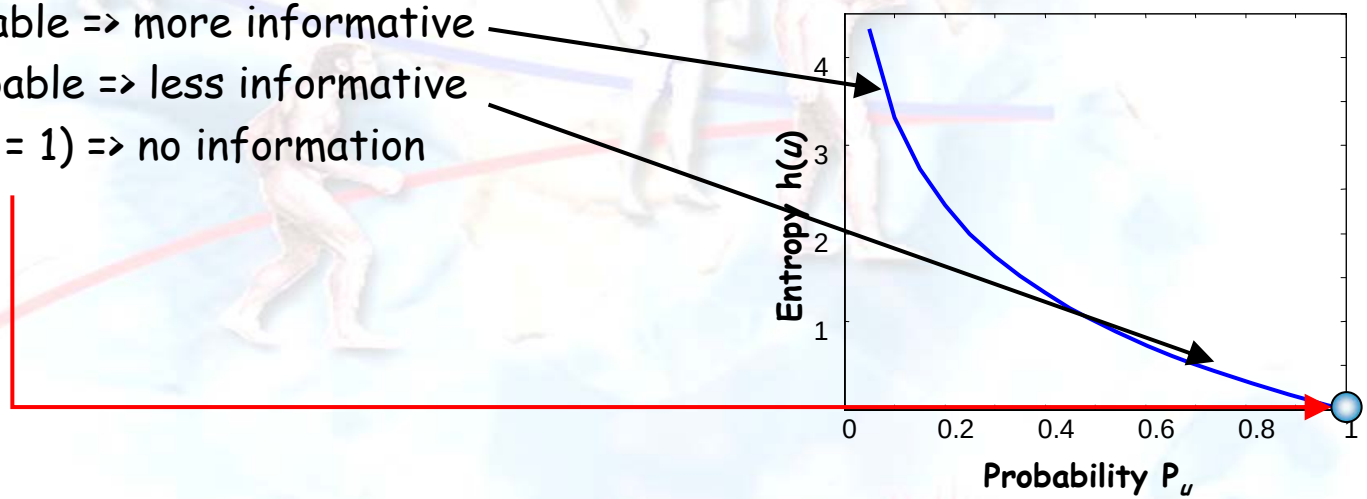
  ⇒ *Increase the number of texts and speakers to "neutralize" within-language variability and get significant cross-linguistic results.*

- ✓ 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
  - $s(t) = u_1\, u_2\, u_3 \ldots u_{t-1}\, u_t$
  - $P(u_t)$ *is supposed to be independent of* $s(t-1) = u_1\, u_2\, u_3 \ldots u_{t-1}$
- Quantity of Information of unit u = entropy $h(u) = -\log_2(P(u))$
  - 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 $\log_2(N_L)$

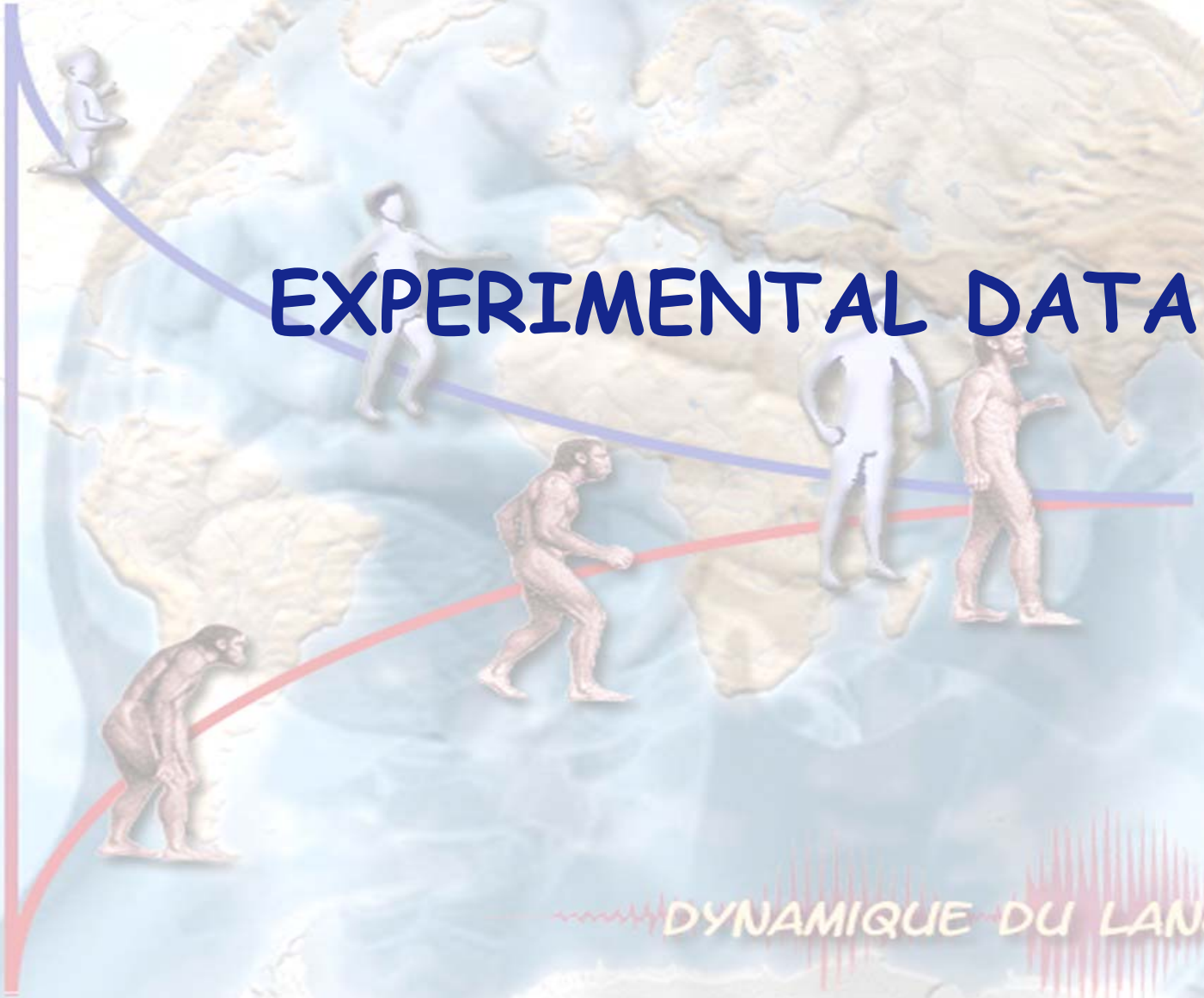
$$H(L) = \sum_{i=1}^{N_L} p_{u_i}.h(u_i) = -\sum_{i=1}^{N_L} p_{u_i} \log_2\left(p_{u_i}\right)$$

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



DYNAMIQUE DU LANGAGE

Raw material
Spoken data and syllable frequencies

- ✓ **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

LANGUAGE	Code	SPEECH DATA	SYLLABLE FREQUENCIES
English	EN	✓	✓
French	FR	✓	✓
German	GE	✓	✓
Italian	IT	✓	✓
Japanese	JA	✓	✓
Mandarin Chinese	MA	✓	✓
Spanish	SP	✓	✓
Dutch	DU	✗	✓
Vietnamese	VI	✓	✗

Speech Data

Language	Source	No of speakers	Total duration
EN	Multext	10	18 min.
FR	Multext	6	14 min.
GE	Multext	10	27 min.
IT	Multext	10	18 min.
JA	Kitazawa, 2002	5	33 min.
MA	Komatsu et al., 2004	9	23 min.
SP	Multext	8	17 min.
VI	Courtesy of E. Castelli MICA, Hanoi	4 (two times each)	38 min.
OVERALL		62	> 3 hours

Small but not tiny corpus

Syllable Frequencies

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

RESULTS

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

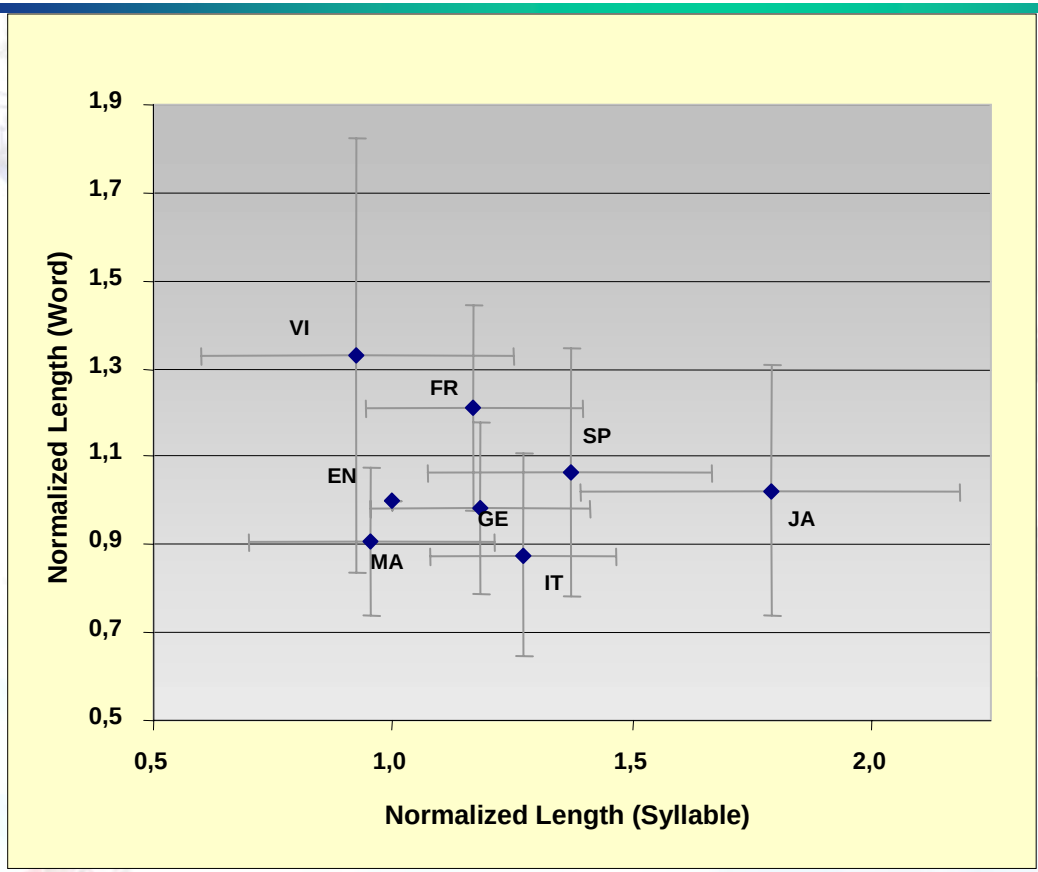
CAVEAT

- ✓ 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)

Speech Corpora Comparison
Normalized Lengths

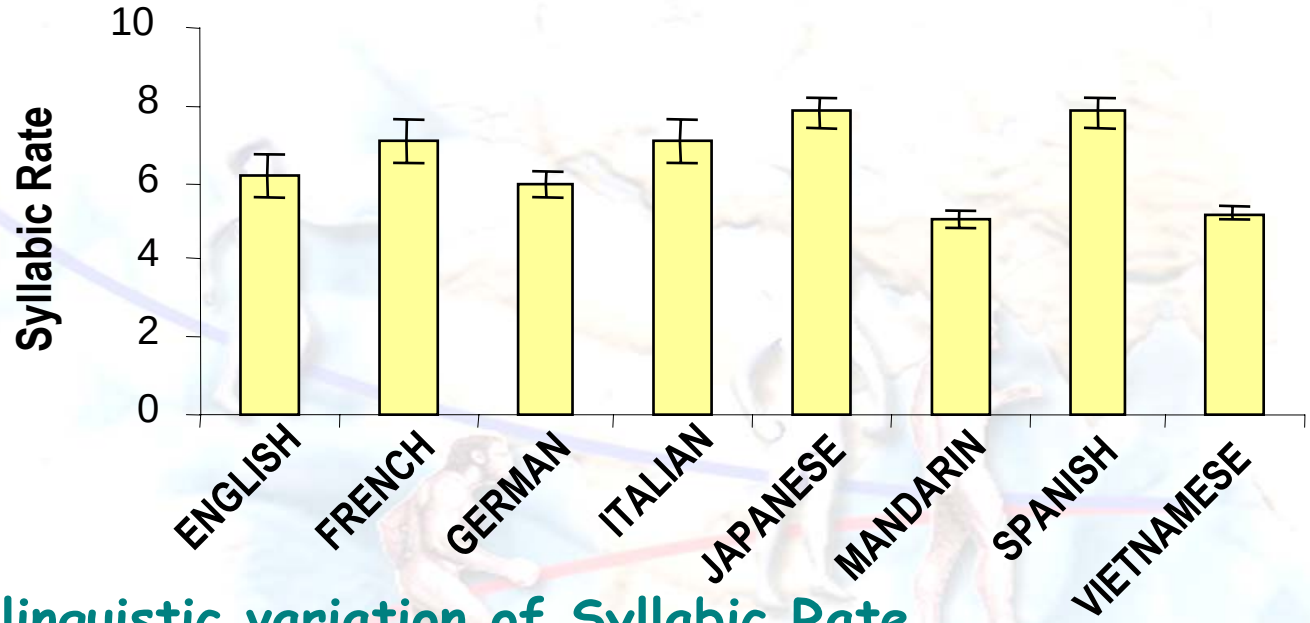


Average values and interquartile ranges

- ✓ **High within-language variation**
  - ↳ Cross-linguistic comparison CANNOT be done with just one utterance…
- ✓ **No significant correlation between Syllabic and Word Lengths**

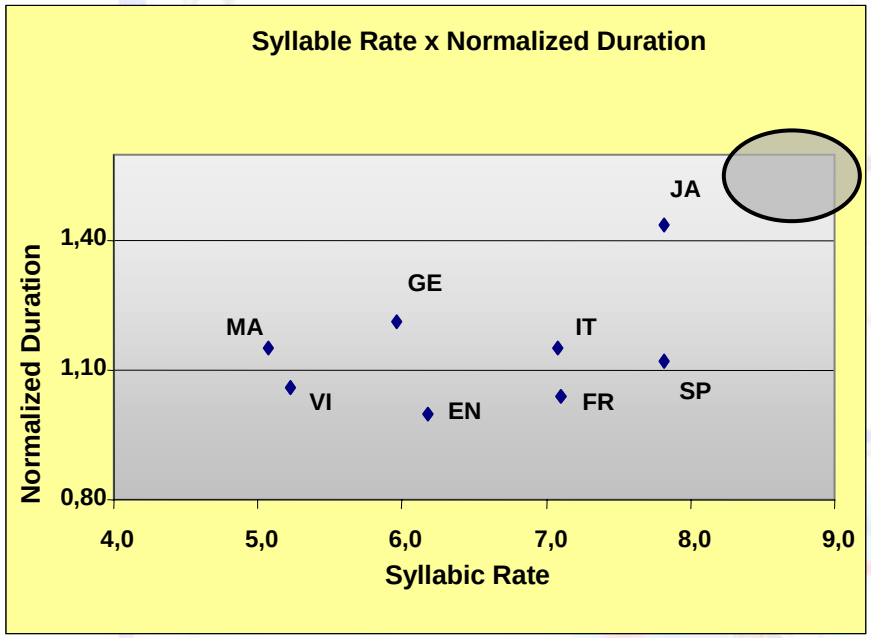
Dynamique Du Langage

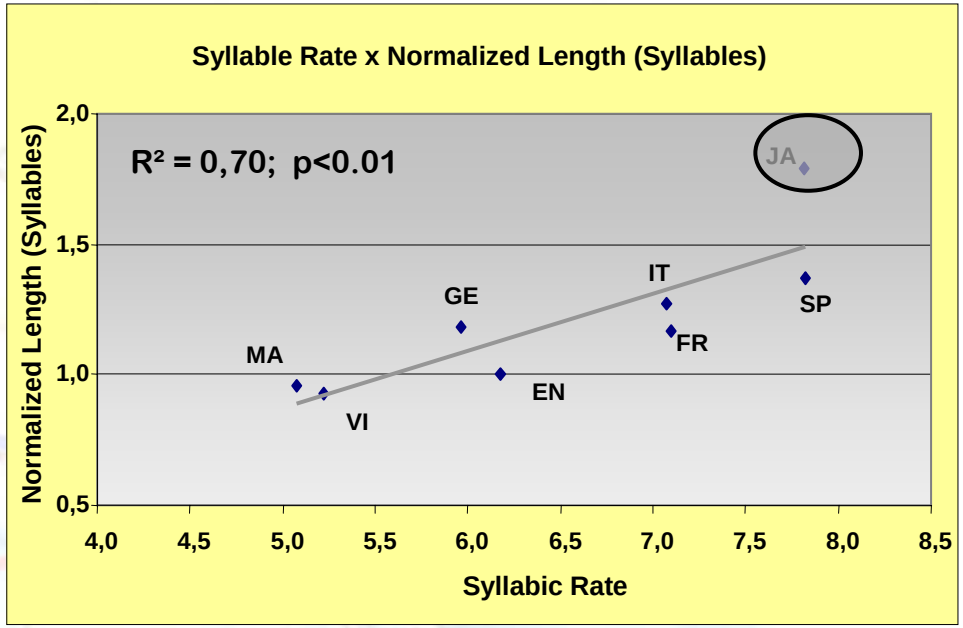
Speech Corpora Comparison
Speech Rate



- ✓ **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?



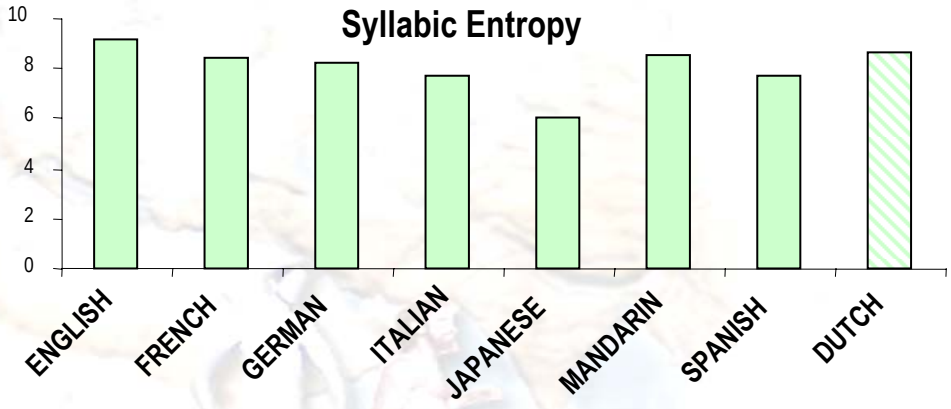


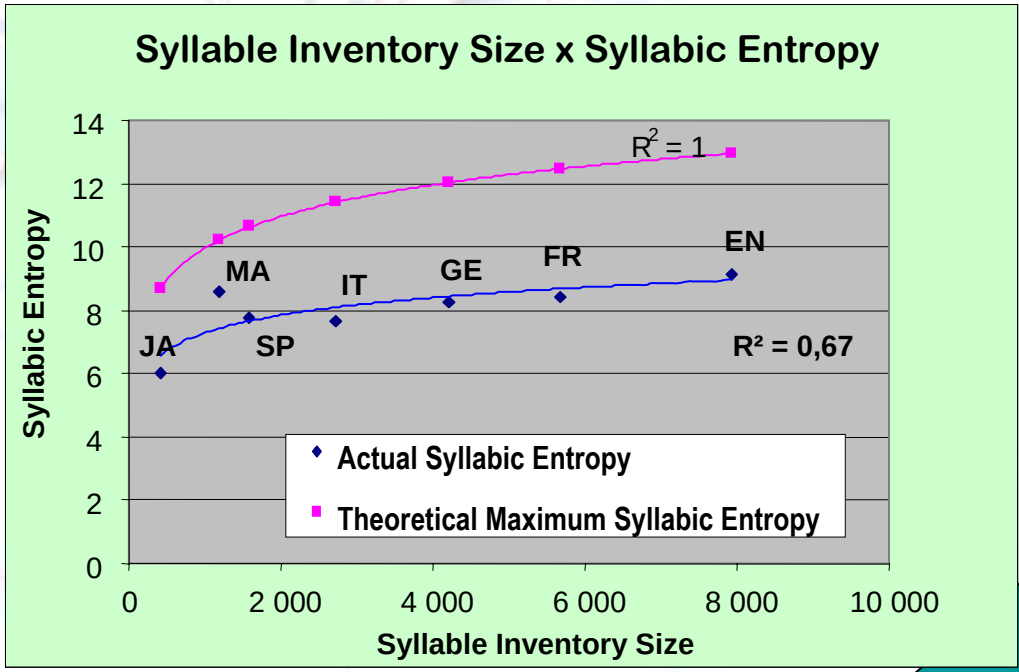
- ✓ **Normalized Duration does NOT correlate with Syllabic Rate**
  - Speak faster does not mean speak shorter!
- ✓ **Normalized Length (Syllable) correlates with Syllabic Rate**
  - Is this just an artifact (Duration linked to Number of syllables?)
  - Is there any causality?

Dynamique Du Langage

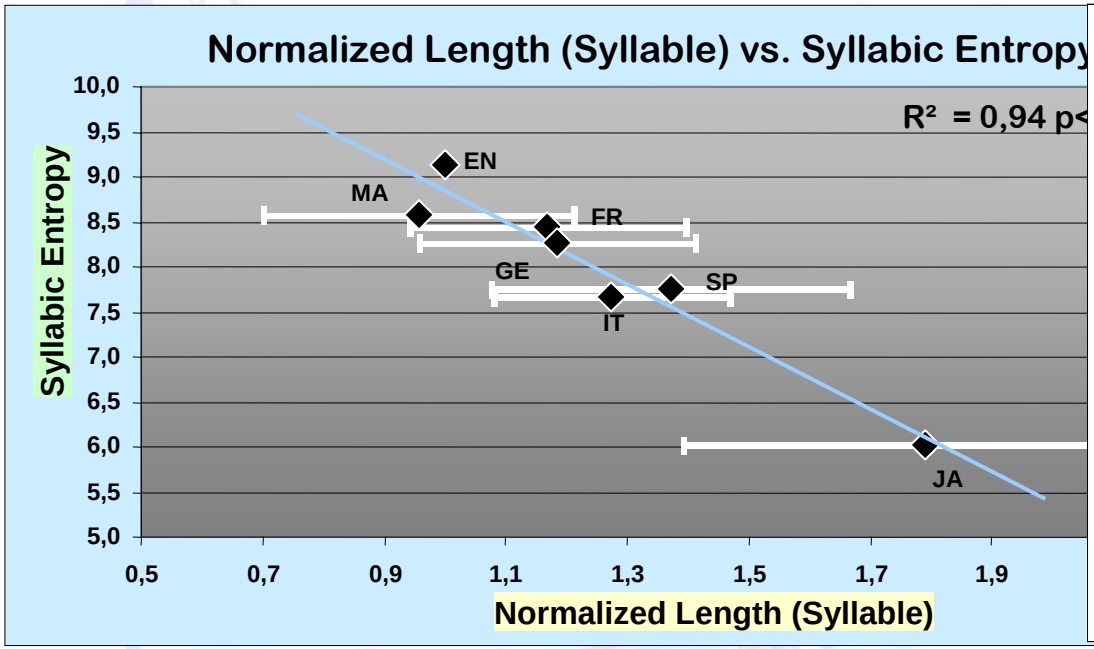
Syllabic Entropy

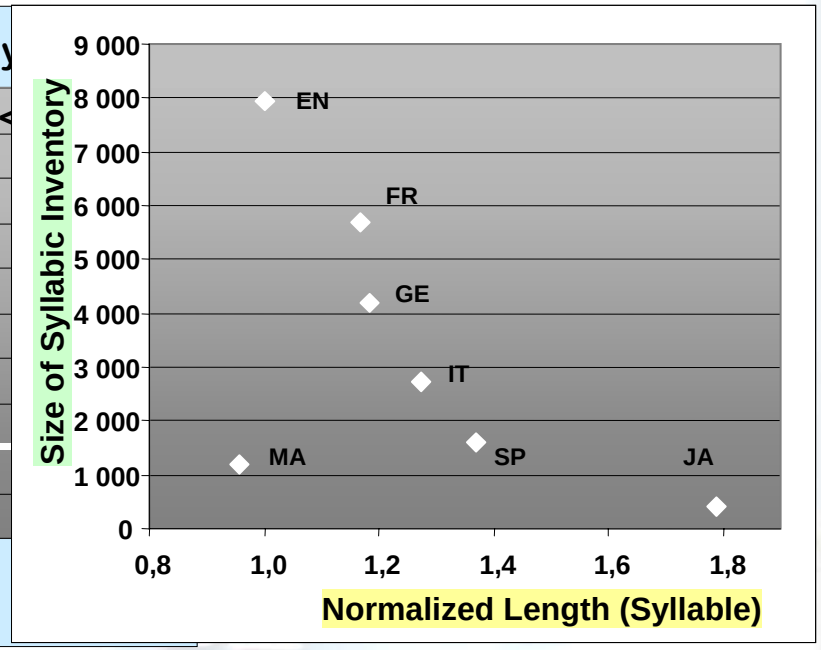
- ✓ **Cross-linguistic variations**
- ✓ **How much of the potential offered by a given syllabic inventory is used?**
  - Redundancy = difference between the maximum possible entropy for each inventory and the observed entropy in a written corpora
  - Pretty similar redundancy across languages
- ✓ ***Comment: Sizes of Syllabic Inventories (calculated from corpora) are much lower than those computed just from phonotactic rules***





Back to initial questions
Is Syllabic entropy relevant?

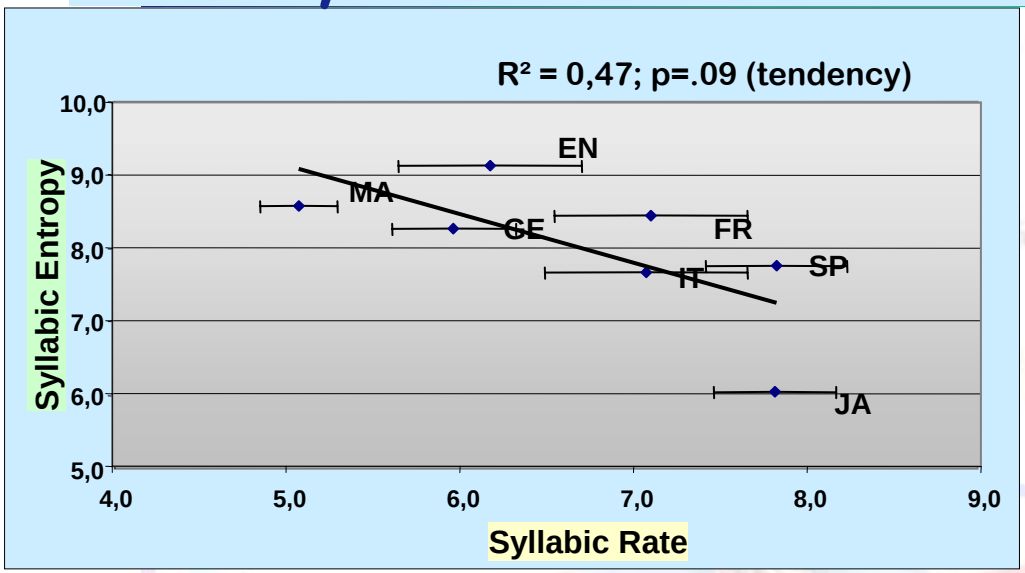


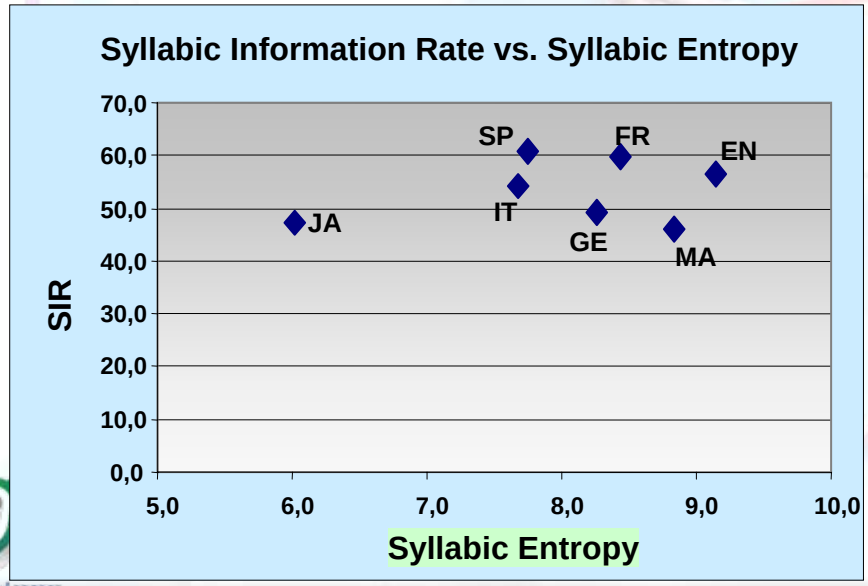


- ✓ **Very high correlation between Syllabic Entropy and Normalized Length (Syllable)**
  - Syllabic entropy (or Information load of syllables) is efficient to quantify the linguistic amount of information
- ✓ **Size of syllable inventory is not efficient to do so**
  - => Syllabic inventory (without frequency) is probably not informative enough for cross-linguistic comparison

Dynamique Du Langage

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

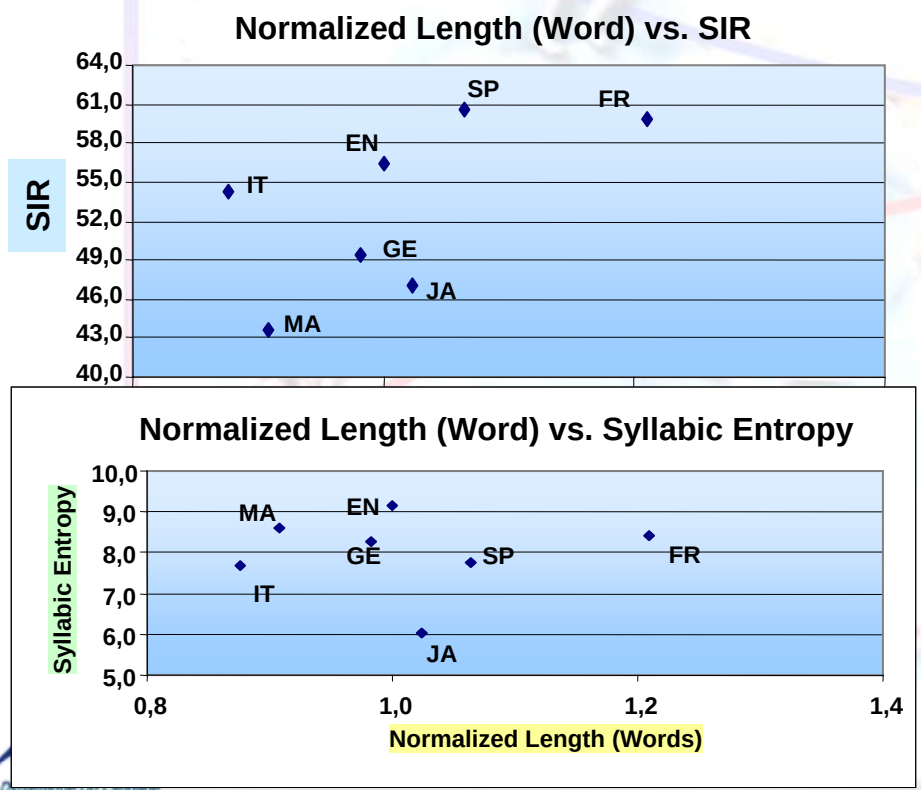
Syllabic Information Rate is not predictable from syllable inventory and probabilities

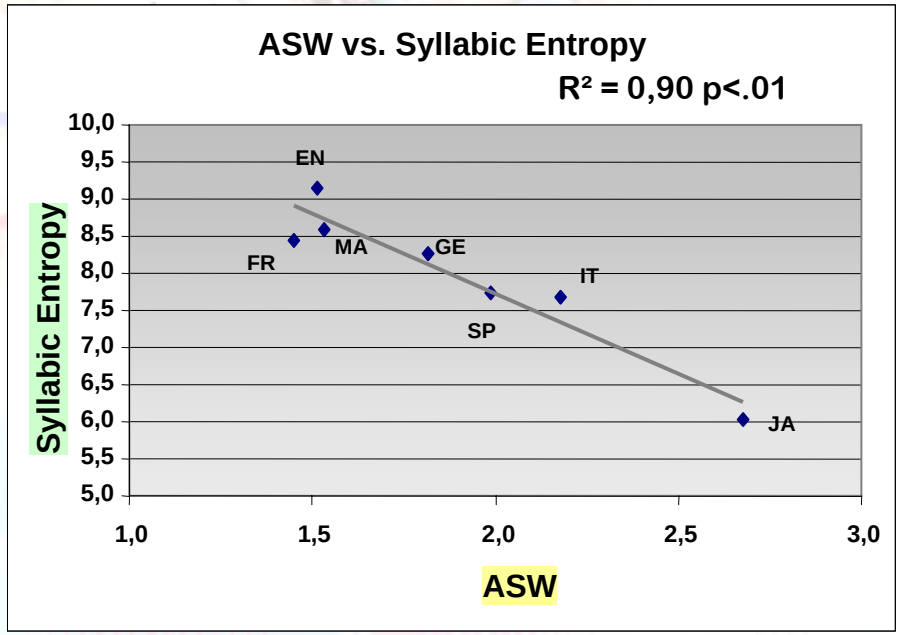
- ✓ Possible balance between syntagmatic and paradigmatic information
  - Cognitive (memory and process) load?
- ✓ Speech Rate matters when looking for correlations!
- ✓ Information *RATE* may be more important than Information *LOAD*

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) => not correlated to any index
  - ASW (Average Number of Syllables per Word)
    - Trade-off to limit word informational load (or complexity)





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

Auer, P. 1993, Is a rhythm-based typology possible? “A study of the role of prosody in phonological typology”. *KontRI Working Paper* (Universitat Konstanz) 21. 96p.

Campione, E. and Véronis, J., 1998, “A multilingual prosodic database”, *in proc. of ICSLP98*, Sydney, Australia, pp. 3163-3166

Dahl, O. 2004. *The growth and maintenance of linguistic complexity*, Studies in Language Companion Series, John Benjamins.

Fenk-Oczlon G. and Fenk A. 1999, “Cognition, quantitative linguistics, and systemic typology”, *Linguistic Typology*, vol. 3-2, pp. 151-177

Fenk-Oczlon G. and Fenk A. 2005, “Crosslinguistic correlations between size of syllables, number of cases, and adposition order”, in *Sprache und Natürlichkeit, Gedenkband für Willi Mayerthaler*, G. Fenk-Oczlon & Ch. Winkler (eds), Tübingen.

Goldsmith, J., 1998, “On information theory, entropy, and phonology in the 20th century”, *Royaumont CTIP II Round Table on Phonology in the 20th Century*, Royaumont (June 26, 1998)

Goldsmith, J., 2002, “Probabilistic models of grammar: phonology as information minimization”, in *Phonological Studies,* Vol. 5, pp. 21-46

Greenberg, J., 1960, “A quantitative approach to the morphological typology of language”, in *International Journal of American Linguistics,* Vol. 26:3, pp. 178-194

Hume, E., 2006, “Language specific and universal markedness: an information-theoretic approach”, *Linguistic Society of America Annual Meeting* (January 7, 2006)

Juola, P., 1998, "Measuring Linguistic Complexity : The Morphological Tier" (1998). *Journal of Quantitative Linguistics* 5(3):206-213

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

Kettunen, K., Sadeniemi, M., Lindh-Knuutila, T. & Honkela, T. 2006. Analysis of EU Languages through Text Compression. In T. Salakoski et al. (Eds.): *Advances in Natural Language Processing*, LNAI 4139, pp. 99 - 109, 2006. Springer-Verlag Berlin Heidelberg.

Maddieson, I., 1986, “The size and structure of phonological inventories: analysis of UPSID”, in *Experimental phonology*, J. J. Ohala and J. Jaeger (eds), Academic Press, New York, pp. 105-123

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

Plank, F., 1998, “The co-variation of phonology with morphology and syntax: a hopeful history”, *Linguistic Typology* 2:2 pp. 195-230.

Shannon, C.E., 1948, “A Mathematical Theory of Communication”, *Bell Syst. Tech. J. 27*, pp. 379-423

Shosted, R. K., 2006, “Correlating complexity: a typological approach”, Linguistic Typology, 10-1

Surendran D. and Niyogi P. to appear, “Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals”, in *Current trends in the theory of linguistic change. In commemoration of Eugenio Coseriu (1921-2002)*, O. Nedergaard Thomsen (ed), Amsterdam & Philadelphia: Benjamins.

Additional slides

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

NORMALIZATION

Language	Data	Passages						MEDIAN	IQTL
Reference	*Passage EN*	*O1*	*O2*	*O3*	*O4*	*O6*	*...*		
EN	Nb Words	51	57	48	53	47	58	52,3	4,5
	Nb Syllables	72	86	84	86	66	86	80,0	8,8
	Mean duration	10,7	13,7	13,0	14,7	11,0	13,6	13,3	1,6
	Normalized Length(Syl)	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00
	Normalized Length (Wd)	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00
	Normalized duration	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00
FR	Nb Words	62	77	40	84	63	69	65,8	15,2
	Nb Syllables	81	106	67	120	96	99	94,8	18,7
	Mean duration	11,3	14,5	9,4	17,6	13,6	13,7	13,7	2,8
	Normalized Length(Syl)	1,1	1,2	0,8	1,4	1,5	1,2	1,19	0,22
	Normalized Length (Wd)	1,22	1,35	0,83	1,58	1,34	1,19	1,28	0,15
	Normalized duration	1,06	1,06	0,72	1,20	1,24	1,01	1,06	0,14

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.

Candidate	Independence	Calculable
Phoneme	× phonotactics	✓ from corpora
Syllable	✓ no too bad	✓ from corpora
Morpheme	× not really!	× morpheme count?
Phrase	✓ Not too bad	× from HUGE corpora?
Feature	×	✓ from corpora?
Gesture	✓ To be explored…	✓ from corpora?

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 — **COARSE-GRAINED**
  2. Lexicon or dictionary ([a], [pa], etc. from lexicon)
     => written/oral syllable inventory, "type" frequency — **BIASES (e.g. morphology)**
  3. Written Corpora ([a], [pa], etc. from written production)
     => written syllable inventory, "token" frequency — **BIAS (≠ from oral production)**
  4. Oral Corpora ([a], [pa], etc. from oral spontaneous data)
     => written syllable inventory, "token" frequency — **UNAVAILABLE***
- **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. ...)

Syllable type	% of occurrence (tokens)	Number of V	Number of C
CV	60,4	1	1
V	12,5	1	0
CCV	9,2	1	2
CVC	11,6	1	2
VC	1,6	1	1
CCVC	1,4	1	3
CVCC	1,4	1	3
CCCV	0,4	1	3
Other	1,5		

Then, considering that vowels (resp. consonants) are equiprobable, syllable's information is the sum of consonantal and vocalic information:
h(.CCVC.) = - (1 x log2(1/Nv) + 3 x log2(1/Nc))
H(L) = sum of h(.XXX.) weighted by % of occurrence

⇒Several significant approximations

Syllable	Number of Occurrences per M. of σ	Probability	Syllable's information
a	44 255	0,04	4,50
e	31 889	0,03	4,97
dø	30 472	0,03	5,04
la	21 752	0,02	5,52
ɛl	18 200	0,02	5,78
...			
zlø	0.05	~ 0,00	24,19
zla	0.05	~ 0,00	24,19
zuk	0.05	~ 0,00	24,19
zwa	0.05	~ 0,00	24,19
zyʀp	0.05	~ 0,00	24,19

Estimation of Syllabic Entropy (cont'd)

- ✓ **Evaluation from distribution of syllables in corpora**

Syllable	Number of Occurrences per M. of σ	Probability	Syllable's information
a	44 255	0,04	4,50
e	31 889	0,03	4,97
dø	30 472	0,03	5,04
la	21 752	0,02	5,52
ɛl	18 200	0,02	5,78
...			
zlø	0.05	~ 0,00	24,19
zla	0.05	~ 0,00	24,19
zuk	0.05	~ 0,00	24,19
zwa	0.05	~ 0,00	24,19
zyʀp	0.05	~ 0,00	24,19

Syllabification available?

YES

- Exact Calculation
- Text of *n* syllables
- H(Text) = sum of $h(\sigma_i)$, i from 1 to *n*

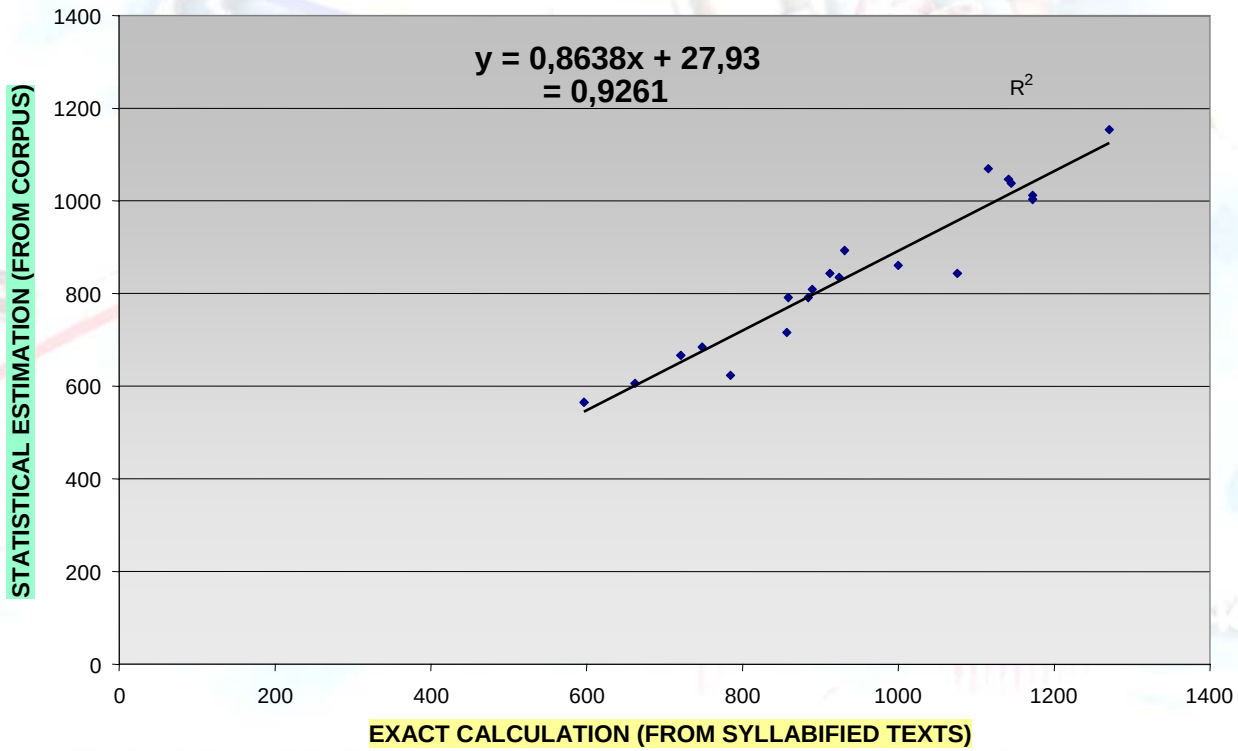
NO

- Statistical Estimation
- Text of *n* syllables
- H(Text) = *n* x H(L) = *n* x mean of h(σ)

Dynamique Du Langage

French: Methodology comparison

	Syllabic Entropy		
	from syllable types	from syllable frequencies (statistical)	from syllable frequencies (exact)
FRENCH	9,05	8,43	9,21



Japanese and Mandarin Data

- **Japanese Word segmentation**
  - Masuoka, T. & Takubo, Y. (1992) Kiso nihongo bunpo [Basic Japanese Grammar]. Tokyo, Kurosio (2003)

- **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