A Quantitative and Typological Approach to Correlating Linguistic Complexity

Yoon Mi Oh, François Pellegrino,
Egidio Marsico & Christophe Coupé
Laboratoire Dynamique du Langage,
Université de Lyon & CNRS, France

Co-authors

François Pellegrino

Egidio Marsico

Christophe Coupé

- Laboratoire Dynamique du Langage, Université de Lyon, CNRS

We are grateful to the LABEX ASLAN (ANR-10- LABX-0081) of Université de Lyon for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR).

Overview

- Framework & Objectives
- Methodology
- Results
- Discussion
- Perspectives

Overview

- [x] Framework & Objectives
- [ ] Methodology
- [ ] Results
- [ ] Discussion
- [ ] Perspectives

Framework (I)

- The equal complexity hypothesis:

"All human languages are equally complex."

Discussion about Creole languages: some studies on the morphological complexity of creole languages have shown that the overall complexity of creole languages are not lower than that of non-creole languages such as English.

Framework (I)

- The equal complexity hypothesis:

"All human languages are equally complex."

Discussion about Creole languages: some studies on the morphological complexity of creole languages have shown that the overall complexity of creole languages are not lower than that of non-creole languages such as English.

- Menzerath's law:

illustrates the phenomenon of self-organization (trade-off) in phonology.

➡ "the more sounds in a syllable the smaller their relative length" (Altmann, 1980).

➡ can be applied to morphology as well - "*the longer the word the shorter its morphemes*" (Altmann, 1980).

Framework (II)

- Morphology is a good starting point for complexity computation for its clearness (Bane, 2008; Juola, 1998).

Framework (II)

- Morphology is a good starting point for complexity computation for its clearness (Bane, 2008; Juola, 1998).

- Best known method of calculation:

to take the numbers of linguistic constituents into account (Bane, 2008; Moscoso del Prado, 2011), with different mathematical formula to be applied to these figures.

Two paradigms are commonly employed:
i) Information theory (Blevins, 2013; Fenk et al., 2006; Moscoso del Prado et al., 2004; Pellegrino et al., 2011)
ii) Kolmogorov complexity (Bane, 2008; Juola, 1998)

Framework (III)

- Information-theoretic approach:

(i) Language as a device for transmitting information in the process of human communication

No distinction between "*simple*" and "*elaborated*" languages but what matters is their capacity of information transmission (the rate at which each language conveys information).

Framework (III)

- Information-theoretic approach:

(i) Language as a device for transmitting information in the process of human communication

No distinction between "*simple*" and "*elaborated*" languages but what matters is their capacity of information transmission (the rate at which each language conveys information).

(ii) Language *L* is a source of linguistic sequences composed of syllables (σ) from a finite set (N_L) (Pellegrino, 2012).

Syllabic entropy as a corpus-based measure of phonological complexity:

$$H_L = -\sum_{i=1}^{N_L} p_{\sigma_i} \log_2 \left(p_{\sigma_i} \right)$$

(i) Average quantity of information (amount of surprise) per syllable
(ii) Unconditional entropy: dependencies between adjacent syllables are not taken into account.

Objectives

- What are our objectives?

1) To compare information rates among several languages
2) To explore interactions between phonological and morphological modules in terms of information and complexity.

HERE ➡ Preliminary results obtained from a corpus-based cross-language study.

Overview

- ☐ Framework & Objectives
- ☑ Methodology
- ☐ Results
- ☐ Discussion
- ☐ Perspectives

Methodology (I): Corpus description

- Multilingual oral corpus in 14 languages:

Based on the Multext (Multilingual Text Tools and Corpora) corpus (Campione & Véronis, 1998).

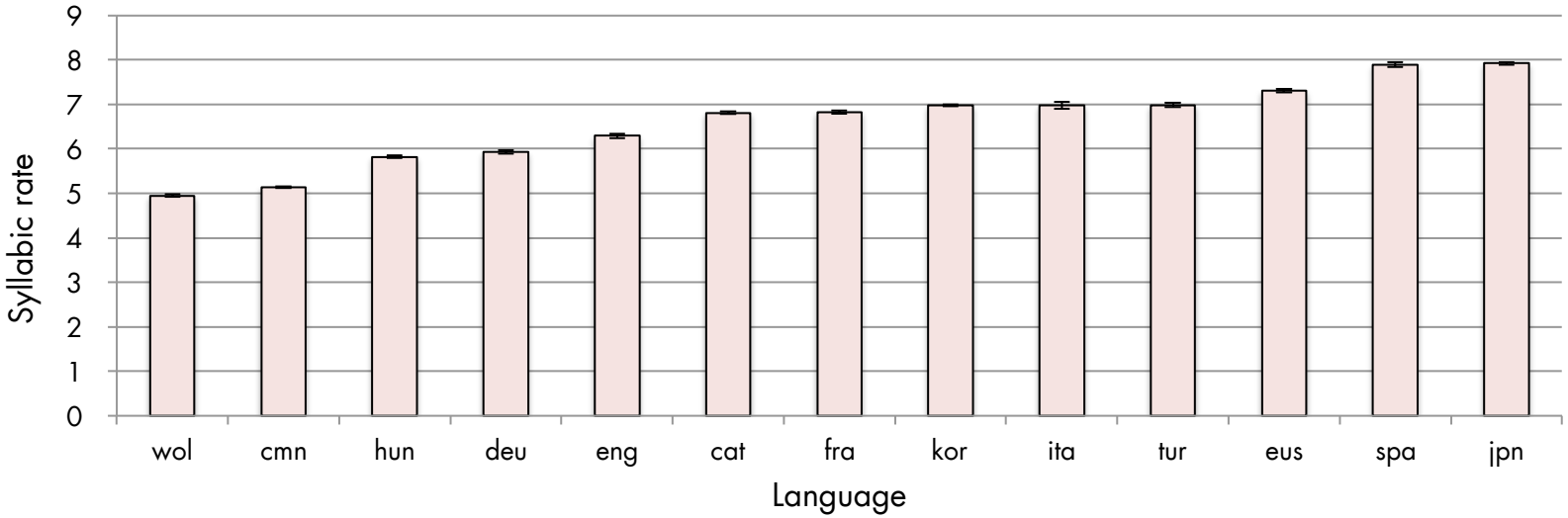
- Oral script for each language:

consists of 15 short texts (of 3-5 semantically connected sentences) translated from British English. 5 female and 5 male native speakers were recorded respectively.

Methodology (II): Parameters – Syllabic rate

- Four types of parameters are taken into account in this study.

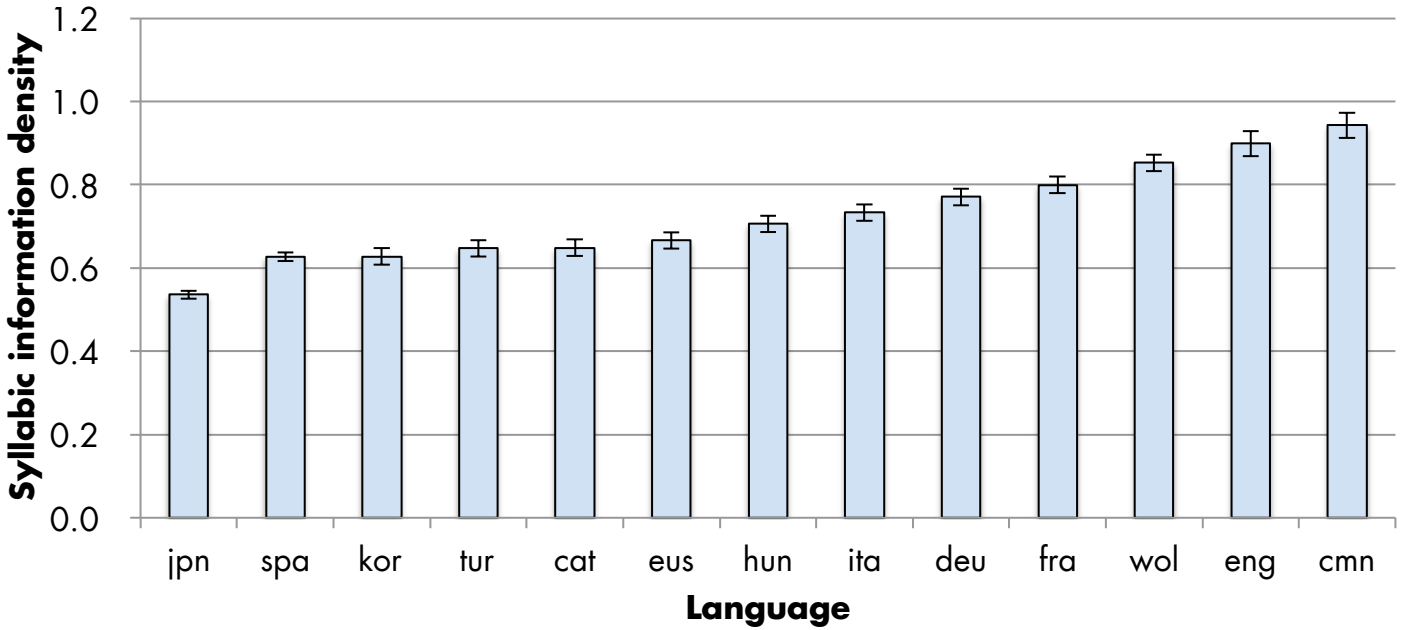
(1) Syllabic rate:
Average number of syllables pronounced per second (#σ/s)



Syllabic rates of 13 languages

Methodology (II): Parameters – Syllabic information density

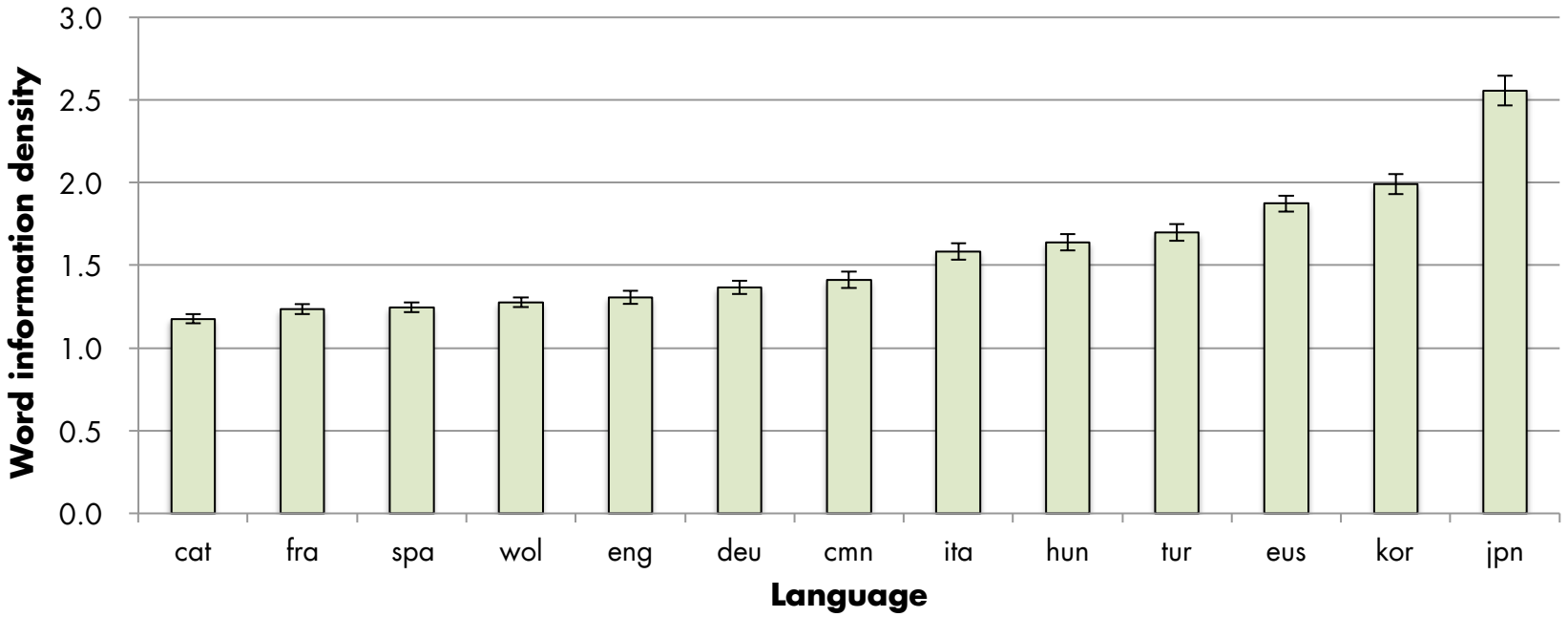
(2) Syllabic information density:
The average ratio between the total number of syllables in a text in Vietnamese (used as an external reference) and the number of syllables of this text translated in the target language. (unitless)



Syllabic information density of 13 languages

Methodology (II): Parameters – Word information density

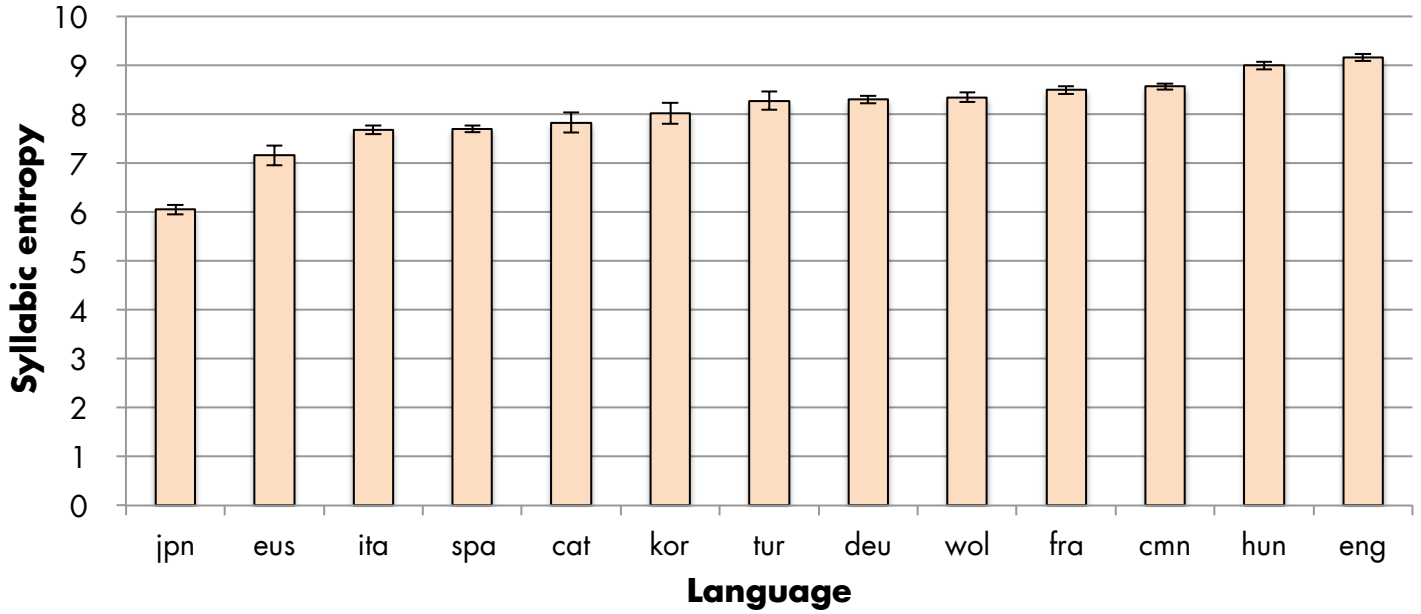
(3) Word information density:
The average ratio between the total number of words in a text in Vietnamese and the number of words of this text translated in the target language. (unitless)



Word information density of 13 languages

Methodology (II): Parameters – Syllabic entropy

(4) Syllabic entropy:
The unconditional entropy calculated from the distribution of syllable frequencies (a measure of phonological complexity) (bits/syllable)



Syllabic entropy of 13 languages

Methodology (III): Quantitative measure of morphological complexity (Lupyan & Dale, 2010)

Language structure is partly determined by social structure. *PLoS ONE, 5(1).*

Morphological complexity:

For 28 linguistic features obtained from WALS (World Atlas of Language Structures), a complexity score is calculated by distinguishing the semantic and syntactic encoding strategies (lexical strategies (-1) vs inflectional morphology (0)).

Feature (WALS code)	Description
1. Fusion of selected inflectional formatives (20A)	the degree to which grammatical markers (called formatives) are phonologically connected to a host word or stem.
2. Prefixing vs. suffixing in inflectional morphology (26A)	the overall extent to which languages use prefixes versus suffixes in their inflectional morphology.
3. Number of cases (49A)	numerical variation in the productive case paradigms of substantives.
4. Case syncretism (28A)	case syncretism is identified when a single inflected form corresponds to two or more case functions.
5. Alignment of case marking of full noun phrases (98A)	the ways in which core argument noun phrases are marked - by means of morphological case or adpositions - to indicate which particular core argument position they occupy.
6. Inflectional synthesis of the verb (22A)	grammatical categories like tense, voice, or agreement can be expressed either by individual words or by affixes attached to some other word (or the stem of a word).

Methodology (III): Quantitative measure of morphological complexity (Lupyan & Dale, 2010)

Feature (WALS code)	Description
7. Alignment of verbal person marking (100A)	The term *alignment* may be intuitively understood as reflecting how the two arguments of the transitive verb, the agentive argument and the more patie nt-like argument, align with the sole argument of the intransitive verb.
8. Person marking on verbs (102A)	the number and identity of the arguments of a transitive clause which dis-play person marking on the verb.
9. Person Marking on Adpositions (48A)	the major function of an adposition is to relate its object, i.e. the noun phra-se with which it forms a constituent, to another nominal or a verbal consti-tuent on the basis of a more or less specific semantic relationship.
10. Syncretism in verbal person/number marking (29A)	instances of syncretism in the inflectional marking of subject person in verbs
11. Situational possibility (74A)	the ways in which core argument noun phrases are marked to indicate which particular core argument position they occupy.
12. Epistemic possibility (75A)	the strategies used to express epistemic possibility in positive main clauses.
13. Overlap between situational and epistemic modal marking (76A)	to what extent languages have identical markers for situational and episte-mic modality.
14. Semantic distinction of evidentiality (77A)	the presence of grammatical markers of evidentiality which express the evi-dence a speaker has for his/her statement.
15. Negative morphemes (112A)	the nature of morphemes signalling clausal negation in declarative senten-ces.
16. Occurrence of nominal plurality (34A)	the extent to which plural markers on full nouns are used in a language.
17. Associative plural (36A)	associative plural constructions consist of a noun and some other material (an affix, a clitic, or a word).

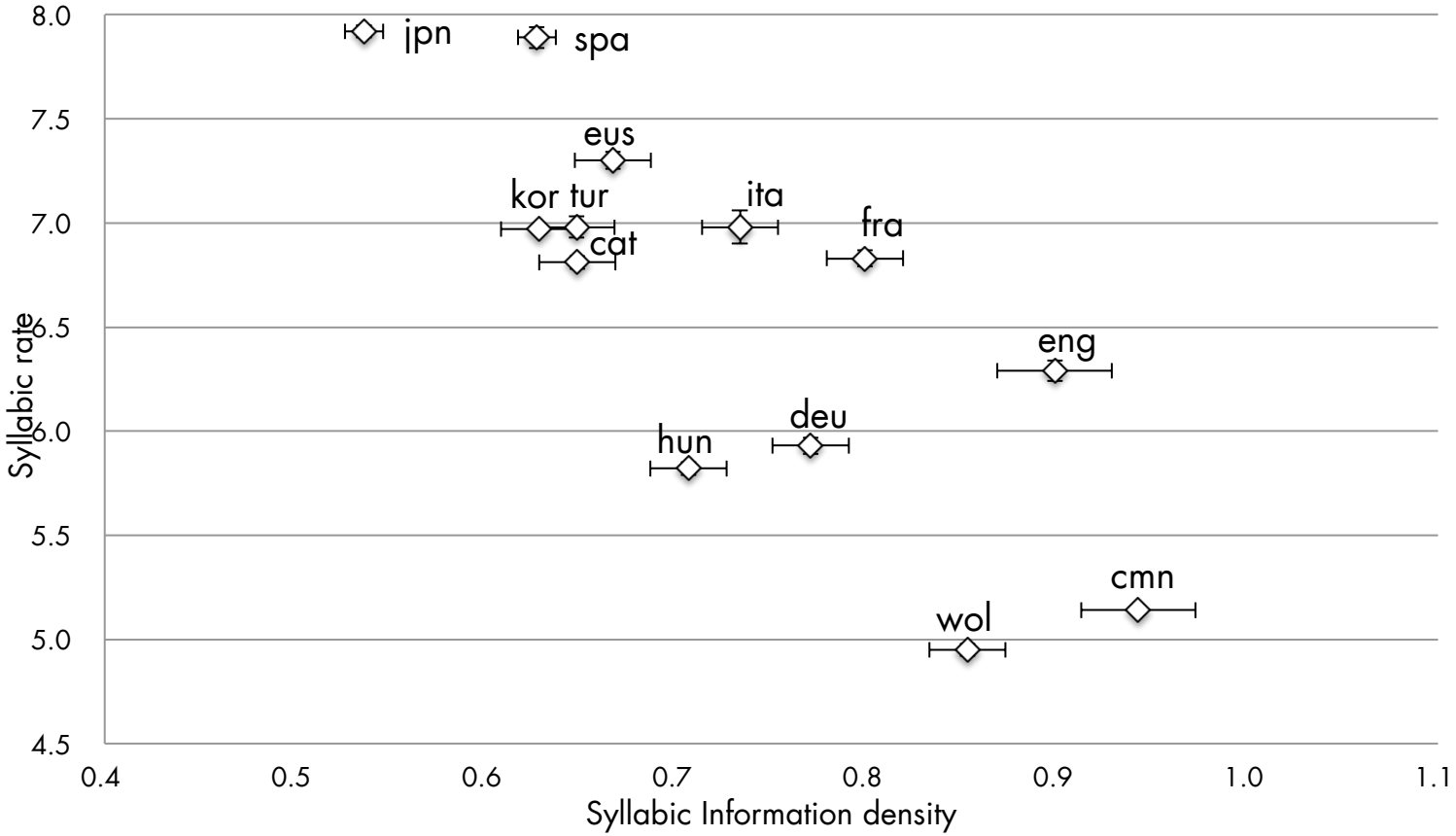
Methodology (III): Quantitative measure of morphological complexity (Lupyan & Dale, 2010)

Feature (WALS code)	Description
18. Position of polar question particles (92A)	the position of question particles in polar questions (questions that elicit the equivalent o f a 'yes' or 'no' response).
19. Future tense (67A)	the distinction between languages in which there is inflectional marking of future time reference and those where there is not.
20. Past tense (66A)	the distinction between languages that mark the past/non-past distinction grammatical -ly (including marking by periphrastic constructions) and those which do not.
21. Perfective/Imperfective aspect (65A)	the distinction between imperfective and perfective plays an important role in many verb systems and is commonly signaled by morphological means.
22. Morphological imperative (70A)	to what extent languages have second person singular and plural imperatives as dedi- cated morphological categories.
23. Position of pronominal possessive affixes (57A)	the position of possessive affixes on nouns.
24. Possessive classification (59A)	the contrast of two formal types of possession, determined by the possessed noun.
25. Optative (73A)	the term *optative* refers to an inflected verb form dedicated to the expression of the wish of the speaker.
26. Definite/indefinite article (37A/38A)	a definite article is a morpheme which accompanies nouns and which codes definite- ness or specificity. A morpheme is considered here to be an indefinite article if it accom -panies a noun and signals that the noun phrase is pragmatically indefinite in the sense t-hat it denotes something not known to the hearer.
27. Distance contrasts in demonstratives (41A)	the relative distance of a referent in the speech situation vis-à-vis the deictic center (which is roughly equivalent to the speaker's location at the time of the utterance).
28. Expression of pronominal subjects (101A)	A pronominal subject is involved in a simple sentence where there is no nominal sub- ject and where the subject is expressed at most by a morpheme or morphemes coding se-mantic or grammatical features of the subject, such as person, number, or gender.

Overview

- ☐ Framework & Objectives
- ☐ Methodology
- ☑ Results
- ☐ Discussion
- ☐ Perspectives

Results (I): Syllabic rate and Syllabic information density



Syllabic rate and syllabic information density
(Error bars indicate standard error.)

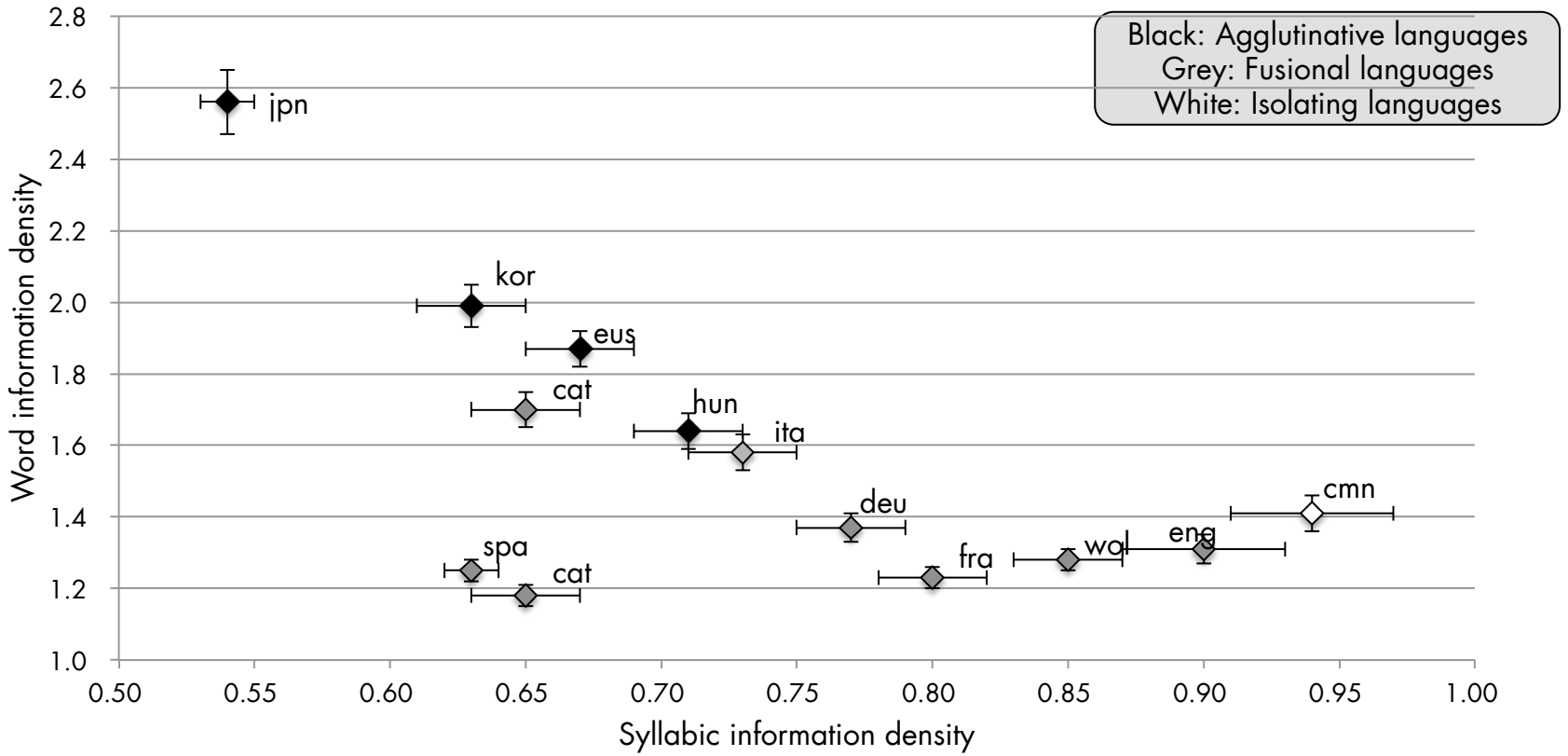
Negative correlation (cor= -0.8, p-value= 0.001) between two phonological factors, i.e. a trade-off between syllabic rate and information density

Results (II): Morphological classification (Greenberg, 1960)

At the morphological level, the languages of our corpus can be classified into three categories.

Category	Languages
Agglutinative languages	Basque, Hungarian, Japanese, Korean, Turkish
Fusional languages	Catalan, English, French, German, Italian, Spanish, Wolof
Isolating languages	Mandarin Chinese, Vietnamese

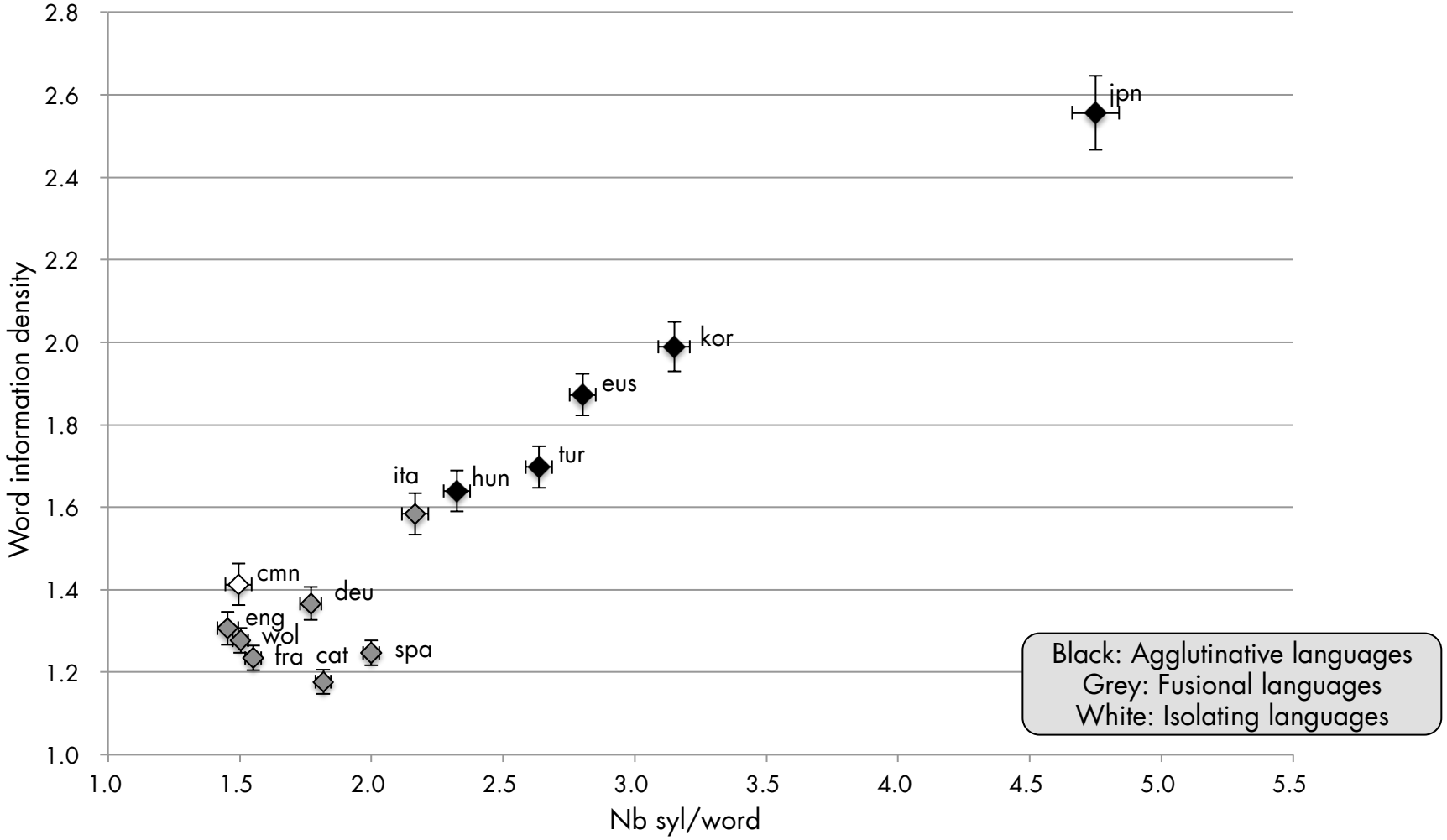
Results (III): Word information density and syllabic information density



Word information density and syllabic information density
(Error bars indicate standard error)

- Negative correlation (cor= -0.61, p-value= 0.03)
- The higher word information density, the lower syllabic information density: Agglutinative > fusional > isolating languages

Results (IV): Word information density and mean number of syllables per word

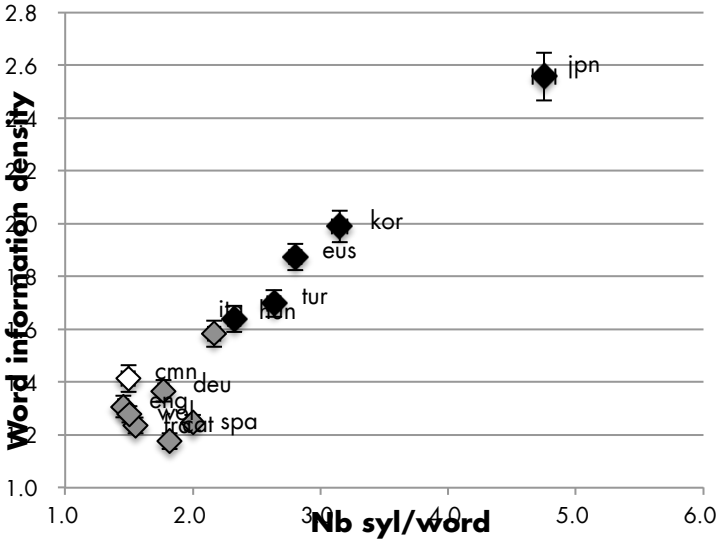


Word information density and mean number of syllables per word
(Error bars indicate standard error)

Results (IV): Word information density and mean number of syllables per word

- A strong positive correlation (cor=0.96, p-value=0.0000001) exists between the average number of syllables per word and the information density at the word level.

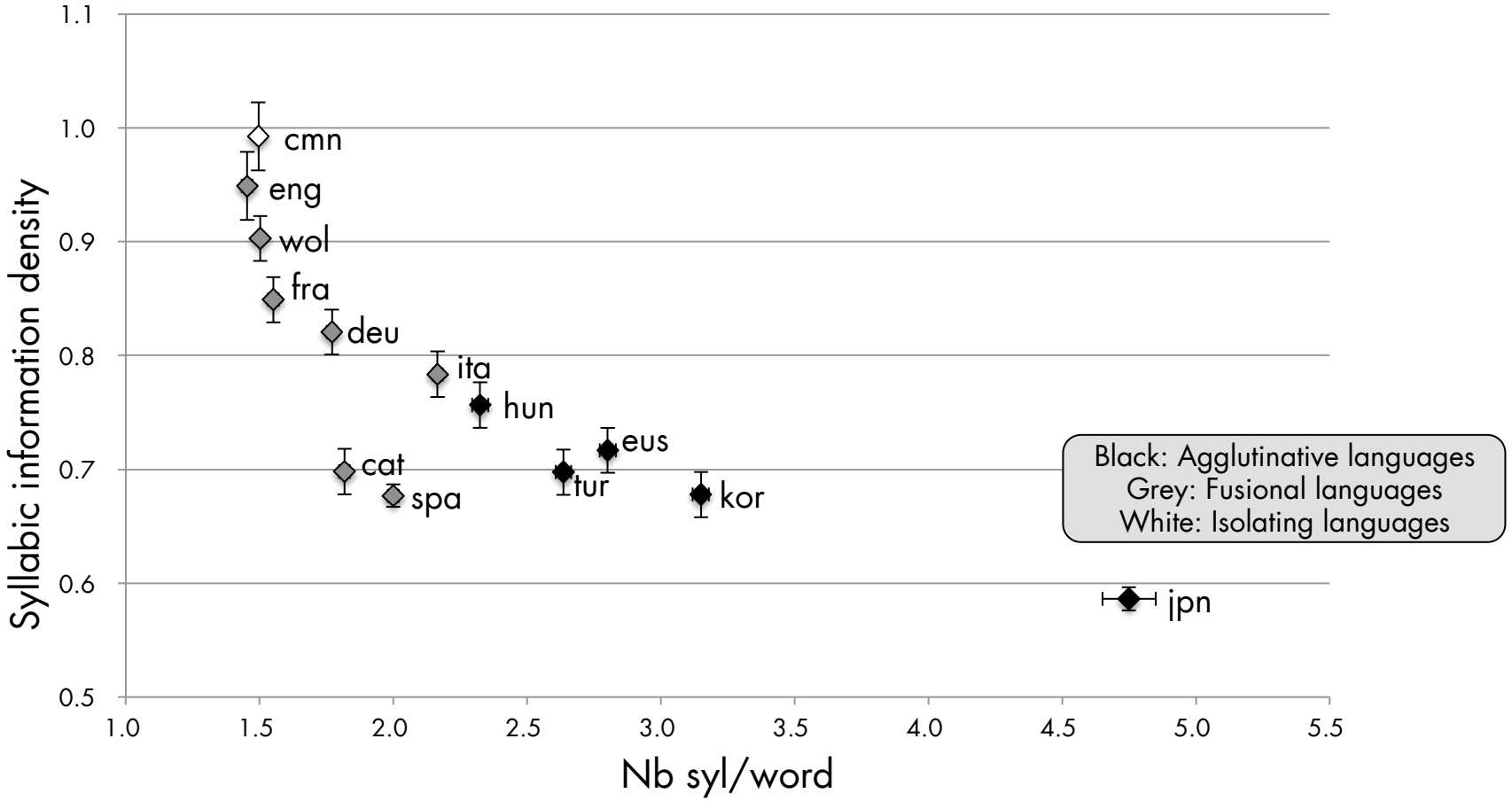
→ the longer the word, the more information it contains.



- Obviously, there are more syllables per word in agglutinative languages (in black) than in fusional languages (in grey).

- Values of languages in the same morphological category are quite dispersed: Large dispersion in agglutinative languages and weak dispersion in fusional languages

Results (V): Syllabic information density and mean number of syllables per word

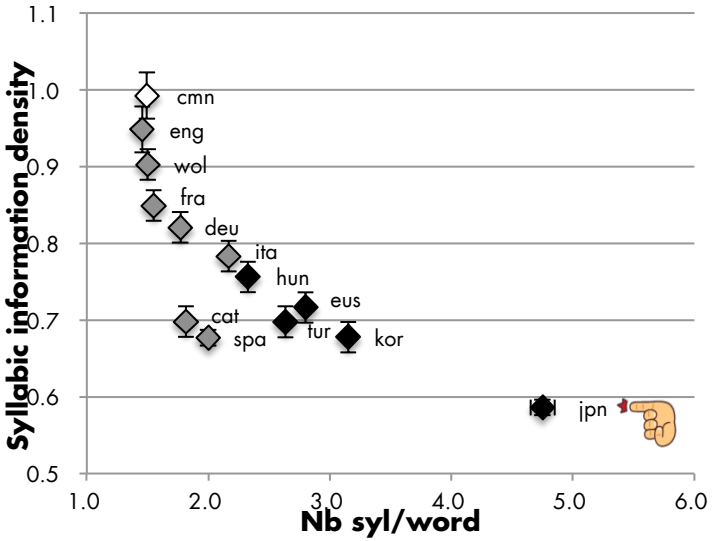


Syllabic information density and mean number of syllables per word
(Error bars indicate standard error)

Results (V): Syllabic information density and mean n umber of syllables per word

- At the syllable level, fusional languages have a tendency towards higher information density compared to agglutinative languages

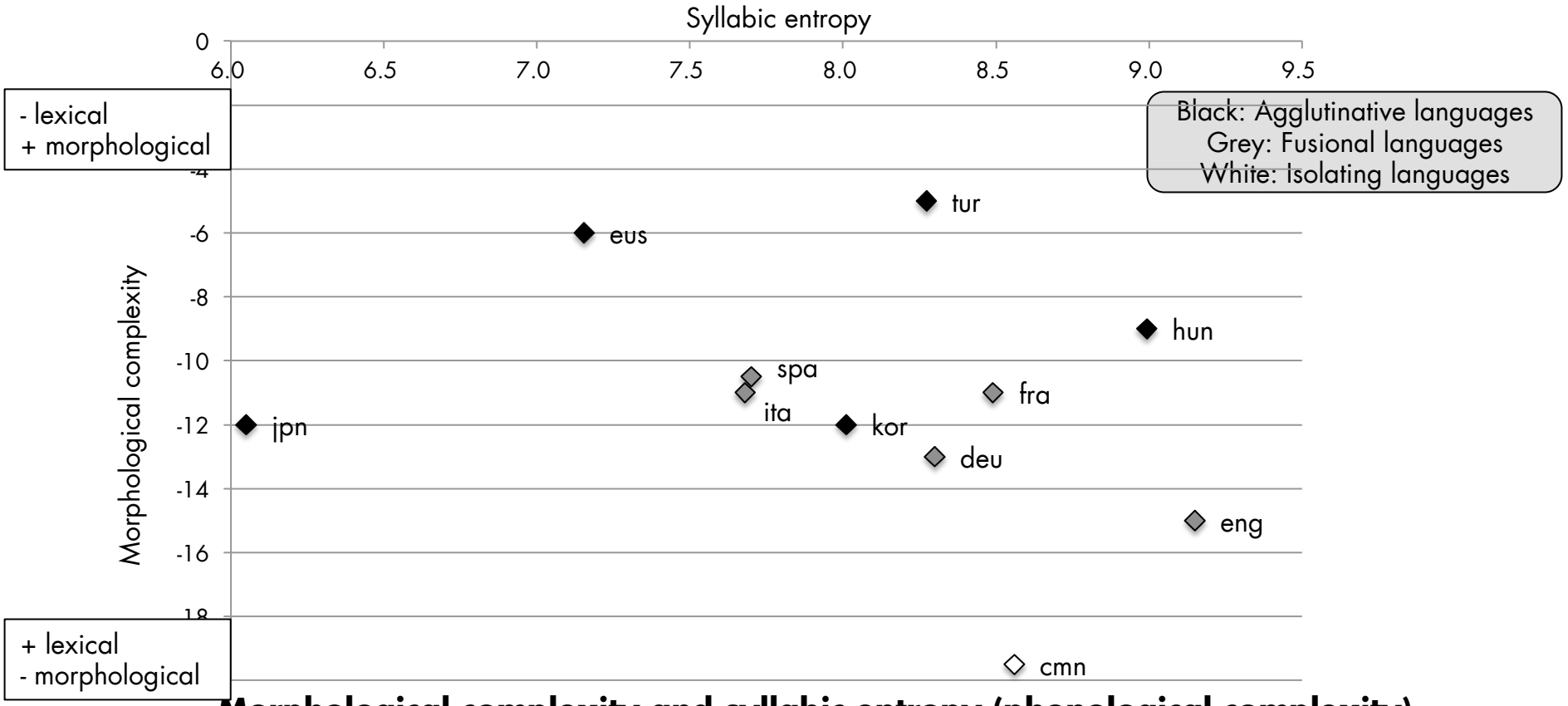
- Japanese, which has a relatively simple phonological system, has the largest number of syllables per word and transmits the least amount of information per syllable.



- Compared to Japanese, Mandarin Chinese, an isolating language with a relatively complex phonological system, shows completely opposite values.

- Non linear relationship: large dispersion in syllabic information density among fusional languages.

Results (VI): Morphological complexity and syllabic entropy



Morphological complexity and syllabic entropy (phonological complexity)

- No correlation (cor= -0.25, p-value= 0.46)
- Difficult to find a clear tendency between the level of morphological complexity and morphological categories but regarding morphological complexity:

Agglutinative > Fusional > Isolating languages

Overview

- ☐ Framework & Objective
- ☐ Methodology
- ☐ Results
- ☑ Discussion
- ☐ Perspectives

Discussion (I)

Fenk, A., Fenk-Oczlon, G. and Fenk, L. 2006.Syllable complexity as a function of word complexity. *In The VIII-th International Conference "Cognitive Modeling in Linguistics" Vol. 1*, 324-333.

- Previous work by Fenk et al. (2006):

defined "word complexity" as the mean number of syllables per word and "syllable complexity" as the mean number of phonemes per syllable.

➡ and found a negative linear correlation between these two figures.

- Our result:

shows a negative correlation between word complexity and information density at the syllable level.

➡ i.e. the less complex a word, the more information per syllable.

➡ Similar to Fenk et al. (2006)

Discussion (II)

Can a general tendency be found?

Some differences are observed between morphological categories:

(i) Large dispersion among agglutinative languages and weak dispersion among fusional languages and vice versa…

(ii) Agglutinative languages tend to have more syllables per word than fusional languages
-> Lower syllabic information density & higher word information density than fusional languages.

-> self-organization and the law of Menzerath

Discussion (III)

- According to Lupyan & Dayle (2010), morphologically rich languages are over-specified whereas morphologically simple language are less specified.

-> But no visible relation was found between morphological complexity and morphological categories. Why?

- This method of binary selection between -1 (lexical strategy) and 0 (inflectional system) seems to have limits for the morphological complexity calculation since it does not allow us to compute the degree of specificity of the morphological system of a given language.

- The absence of inflectional marker in certain linguistic features does not necessarily mean that the morphological system of a given language is less complex. It could well be compensated by other specified linguistic features.

-> Overall morphological complexity

Overview

- ☐ Framework & Objective
- ☐ Methodology
- ☐ Results
- ☐ Discussion
- ☑ Perspectives

Perspectives

In further studies, the relation between morphological and phonological complexity will be investigated in more details by enlarging our multilingual parallel data and by adding more isolating languages to observe their patterns.

We are working on unsupervised morpheme segmentation in order to compute morphemic entropy (morphological complexity) and to compare our multilingual data at the morphological level.

Finally, we aim to measure the conditional entropy of syllables and morphemes in order to take the contextual information of these linguistic units into account.

References

- Altmann, G. 1980. Prolegomena to Menzerath's law. *Glottometrika, 2*, 1–10.
- Bane, M. 2008. Quantifying and measuring morphological complexity. *In Proc. of the 26th West Coast Conference on Formal Linguistics*, 69-76.
- Blevins, J. 2013. Morphological Structure and Complexity, *Collegium de Lyon*.
- Campione, E. and Véronis, J. 1998. A multilingual prosodic database. *In Proc. of ICSLP98*, Sydney, Australia, 3163-3166.
- Fenk, A., Fenk-Oczlon, G. and Fenk, L. 2006. Syllable complexity as a function of word complexity. *In The VIII-th International Conference "Cognitive Modeling in Linguistics" Vol. 1*, 324-333.
- Fenk-Oczlon, G., and Fenk, A. 1985. The mean length of propositions is 7 plus minus 2 syllables–but the position of languages within this range is not accidental. *In Proc. of the XXIII International Congress of Psychology: Selected/Revised Papers*, Vol. 2, 355–359.
- Forns, N. and Ferrer-i-Cancho, R. 2009. The self-organization of genomes. *Complexity, 15(5)*, 34-36.
- Frank, A., and Jaeger, T. F. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *In Proc. of the Cognitive Science Society.*
- Greenberg, J. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics, 26(3)*, 178-194.
- Juola, P. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics, 5(3)*, 206–213.
- Lupyan, G. and Dale, R. 2012. Language structure is partly determined by social structure . *PLoS ONE*. 20;5(1):e8559.
- Moscoso del Prado, F., Kostić, A., and Baayen, R.H. 2004. Putting the bits together: An information theoretical perpective on morphological processing. *Cognition, 94(1)*, 1-18.
- Moscoso del Prado, F. 2011. The Mirage of morphological complexity, In Proc. of the 33rd Annual Conference of the Cognitive Science Society, 3524-3529.
- Moscoso del Prado, F. 2013. The grammatical complexity of Tok Pisin: A quantitative assessment, 19th *International Congress of Linguists.*
- Pellegrino, F., Coupé, C., and Marsico, E. 2011. A cross-language perspective on speech information rate. *Language*, 87(3), 539–558.
- Pellegrino, F. 2012. Syllabic information rate: a cross-language approach. Dartmouth College, September, 27 2012.
- Teupenhayn, R., and Altmann, G. 1984. Clause length and Menzerath's law. *Glottometrika 6,* 127-138.

감사합니다!
Hartelijk bedankt!
Merci beaucoup!