UNIVERSITÉ
LUMIÈRE
LYON 2

École doctorale Lettres, Langues, Linguistique et Arts

Linguistic Complexity and Information: Quantitative Approaches

THESIS

presented and publicly defended on October 20, 2015

for the degree of

Doctor of the University of Lyon

in Language Sciences

by

Yoon Mi OH

Jury composition

Referees:	Prof. Bart de Boer	Vrije Universiteit Brussel, Belgium
	Prof. Bernd Möbius	Universität des Saarlandes, Germany
Examiners:	Dr. Christophe Coupé	CNRS & Université de Lyon, France
	Dr. Christine Meunier	CNRS & Aix-Marseille Université, France
	Dr. François Pellegrino (Thesis supervisor)	CNRS & Université de Lyon, France

Laboratoire Dynamique du Langage — CNRS - Université Lumière Lyon 2 (UMR 5596)

Mis en page avec la classe thesul.

Acknowledgements

I would like to express my deepest gratitude to my supervisor François Pellegrino for his patient guidance and trust which provided me a lot of support and encouragement for accomplishing my thesis. He has been a role model for me in many aspects, especially with his consideration for others as a director and not to mention his insightful advices. I have been extremely fortunate to work with him who always replied and suggested solutions to so many questions and difficulties I had during my PhD years and I would not have been able to finish this thesis without his help.

I am deeply indebted and grateful to Christophe Coupé for his valuable advices and all the works which enriched my thesis, and to Egidio Marsico for so many things: from his precious advices to his warm-hearted help which made me feel at home throughout my stay at the laboratory. I feel myself very lucky to have worked with three warm-hearted colleagues, François, Christophe, and Egidio and I heartily appreciate this perfect trio lyonnais for their patience and confidence and for all the meetings and discussions which were always harmonious and stimulating.

I am also highly grateful to the rest of my thesis committee, Bart de Boer, Bernd Möbius, and Christine Meunier for their time, attention, and valuable comments on my dissertation. Especially, I would like to thank Bernd Möbius and the other members of the Phonetics and Phonology group at Saarland University (Bistra, Cristina, Erika, Frank, Jeanine, Jürgen, and Zofia) for giving me a wonderful opportunity to collaborate on their project and their warm welcome.

Being a member of the laboratory of Dynamique du Langage was one of the best and unforgettable experiences in my life. Many thanks to Nathalie for her warm and invaluable support and advices which enormously helped and encouraged me throughout my lab life, to Linda and Sophie for their kind help, to Seb for his patient guidance with programming, to Christian for all the fun board games at lunch, to Emilie for her precious help correcting my French, to Esteban, Geny, Rozenn, and Soraya for their good

vibes in our office, to Agathe, Françoise, Hadrien, Jennifer, Ludivine, Maïa, and Marie for social gatherings outside the lab, and all the other members of the laboratory for the great atmosphere at the lab.

This thesis could not have been done without the help of colleagues and native speakers who helped me during data collection and analysis. I am especially grateful to Anna, Borja, Frédéric, Gabriella, Hannu, Jànos, Jean-Léopold, Jordi, Laurent, Maja, Miyuki, Pither, Sakal, and Stéphane for their precious help with translation and data collection. I would also like to thank Ramon Ferrer-i-Cancho for his precious advices during my short stay in Barcelona.

My heartfelt gratitudes go to my family and friends in Korea who have been there for giving me a lot of support. I thank my parents for their unconditional love, my sister and brother for their affection, my aunt Jinhee for her refreshing energy, and Yunhee, Bitna, and Hyesoo, for their friendship regardless of distance.

This work was supported by the LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

To my parents

Abstract

The main goal of using language is to transmit information. One of the fundamental questions in linguistics concerns the way how information is conveyed by means of language in human communication. So far many researchers have supported the uniform information density (UID) hypothesis asserting that due to channel capacity, speakers tend to encode information strategically in order to achieve uniform rate of information conveyed per linguistic unit. In this study, it is assumed that the encoding strategy of information during speech communication results from complex interaction among neurocognitive, linguistic, and sociolinguistic factors in the framework of complex adaptive system. In particular, this thesis aims to find general cross-language tendencies of information encoding and language structure at three different levels of analysis (i.e. macrosystemic, mesosystemic, and microsystemic levels), by using multilingual parallel oral and text corpora from a quantitative and typological perspective.

In this study, language is defined as a complex adaptive system which is regulated by the phenomenon of self-organization, where the first research question comes from: "How do languages exhibiting various speech rates and information density transmit information on average?". It is assumed that the average information density per linguistic unit varies during communication but would be compensated by the average speech rate. Several notions of the Information theory are used as measures for quantifying information content and the result of the first study shows that the average information rate (i.e. the average amount of information conveyed per second) is relatively stable within a limited range of variation among the 18 languages studied.

While the first study corresponds to an analysis of self-organization at the macrosystemic level, the second study deals with linguistic subsystems such as phonology and morphology and thus, covers an analysis at the mesosystemic level. It investigates interactions between phonological and morphological modules by means of the measures of linguistic complexity of these modules. The goal is to examine whether the equal complexity hypothesis holds true at the mesosystemic level. The result exhibits a negative correlation between morphological and phonological complexity in the 14 languages and supports the equal complexity hypothesis from a holistic typological perspective.

The third study investigates the internal organization of phonological subsystems by means of functional load (FL) at the microsystemic level. The relative contributions of phonological subsystems (segments, stress, and tones) are quantitatively computed by estimating their role of lexical strategies and are compared in 2 tonal and 7 non-tonal languages. Furthermore, the internal FL distribution of vocalic and consonantal subsystems is analyzed cross-linguistically in the 9 languages. The result highlights the importance of tone system in lexical distinctions and indicates that only a few salient high-FL contrasts are observed in the uneven FL distributions of subsystems in the 9 languages.

This thesis therefore attempts to provide empirical and quantitative studies at the three different levels of analysis, which exhibit general tendencies among languages and provide insight into the phenomenon of self-organization.

Keywords: complex adaptive system, functional load, information rate, Information theory, language universals, linguistic complexity, quantitative approach, self-organization.

Résumé

La communication humaine vise principalement à transmettre de l'information par le biais de l'utilisation de langues. Plusieurs chercheurs ont soutenu l'hypothèse selon laquelle les limites de la capacité du canal de transmission amènent les locuteurs de chaque langue à encoder l'information de manière à obtenir une répartition uniforme de l'information entre les unités linguistiques utilisées. Dans nos recherches, la stratégie d'encodage de l'information en communication parlée est conçue comme résultant de l'interaction complexe de facteurs neurocognitifs, linguistiques, et sociolinguistiques et nos travaux s'inscrivent donc dans le cadre des systèmes adaptatifs complexes. Plus précisément, cette thèse vise à mettre en évidence les tendances générales, translinguistiques, guidant l'encodage de l'information en tenant compte de la structure des langues à trois niveaux d'analyse (macrosystémique, mésosystémique, et microsystémique). Notre étude s'appuie ainsi sur des corpus oraux et textuels multilingues dans une double perspective quantitative et typologique.

Dans cette recherche, la langue est définie comme un système adaptatif complexe, régulé par le phénomène d'auto-organisation, qui motive une première question de recherche : "Comment les langues présentant des débits de parole et des densités d'information variés transmettent-elles les informations en moyenne ?". L'hypothèse défendue propose que la densité moyenne d'information par unité linguistique varie au cours de la communication, mais est compensée par le débit moyen de la parole. Plusieurs notions issues de la théorie de l'information ont inspiré notre manière de quantifier le contenu de l'information et le résultat de la première étude montre que le débit moyen d'information (i.e. la quantité moyenne d'information transmise par seconde) est relativement stable dans une fourchette limitée de variation parmi les 18 langues étudiées.

Alors que la première étude propose une analyse de l'auto-organisation au niveau macrosystémique, la deuxième étude porte sur des sous-systèmes linguistiques tels que la phonologie et la morphologie : elle relève donc d'une analyse au niveau mésosystémique. Elle porte sur les interactions entre les modules morphologique et phonologique en utilisant les mesures de la complexité linguistique de ces modules. L'objectif est de tester l'hypothèse d'uniformité de la complexité globale au niveau mésosystémique. Les résultats révèlent une corrélation négative entre la complexité morphologique et la complexité phonologique dans les 14 langues et vont dans le sens de l'hypothèse de l'uniformité de la complexité globale d'un point de vue typologique holistique.

La troisième étude analyse l'organisation interne des sous-systèmes phonologiques au moyen de la notion de charge fonctionnelle (FL) au niveau microsystémique. Les contributions relatives des sous-systèmes phonologiques (segments, accents, et tons) sont évaluées quantitativement en estimant leur rôle dans les stratégies lexicales. Elles sont aussi comparées entre 2 langues tonales et 7 langues non-tonales. En outre, la distribution interne de la charge fonctionnelle à travers les sous-systèmes vocaliques et consonantiques est analysée de façon translinguistique dans les 9 langues. Les résultats soulignent l'importance du système tonal dans les distinctions lexicales et indiquent que seuls quelques contrastes dotés d'une charge fonctionnelle élevée sont observés dans les distributions inégales de charge fonctionnelle des sous-systèmes dans les 9 langues.

Cette thèse présente donc des études empiriques et quantitatives réalisées à trois niveaux d'analyse, qui permettent de décrire des tendances générales parmi les langues et apportent des éclaircissements sur le phénomène d'auto-organisation.

Mots-clés : approche quantitative, auto-organisation, complexité linguistique, débit d'information, FL, système adaptatif complexe, théorie de l'information, universaux linguistiques.

Contents

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**

1.1 General framework 2

1.1.1 Multi-level analysis of language universals 2

1.1.2 Language universals and linguistic typology 4

1.2 Measures of linguistic complexity 6

1.2.1 Grammar-based complexity: traditional linguistic approach 6

1.2.2 Usage-based complexity: information-theoretic approach 8

1.3 Information encoding 10

1.3.1 Sociolinguistic factors 10

1.3.2 Neurocognitive factors 11

1.3.3 Trade-off in information encoding 14

1.4 Overview 16

2 Average information rate: macrosystemic analysis **19**

2.1 Introduction 19

2.1.1 Phenomenon of self-organization 19

2.1.2 Information theory: quantifying measures of information 22

2.1.3 Chapter outline 24

2.2 Materials and methods 25

2.2.1 Data and preprocessing 26

2.2.1.1 Oral corpus 26

2.2.1.2 Text corpus and preprocessing 29

2.2.2 Parameters 34

2.2.2.1 SR, ID, and IR . . . 34
2.2.2.2 Syllable complexity . . . 36
2.2.2.3 Information-theoretic measures . . . 37
2.2.3 Language description . . . 43
2.3 Cross-language comparisons of the average information rate . . . 47
2.3.1 Speech rate, information density, and information rate . . . 47
2.3.2 Issues in estimating entropy . . . 54
2.3.3 Entropy . . . 57
2.3.4 Conditional entropy . . . 60
2.3.5 Surprisal . . . 66
2.4 Discussion . . . 72
2.4.1 Effect of contextual information . . . 72
2.4.2 Average information rate and UID hypothesis . . . 74
2.4.3 Conclusion . . . 79

3 Mesosystemic relationship between morphology and phonology 81
3.1 Introduction . . . 81
3.1.1 Holistic typology and equal overall complexity . . . 81
3.1.2 Quantifying linguistic complexity . . . 84
3.1.3 Chapter outline . . . 88
3.2 Method, language, and data description . . . 89
3.2.1 Measures of WID and SID . . . 89
3.2.2 Measures of phonological complexity . . . 90
3.2.3 Measures of morphological complexity . . . 90
3.2.4 Language and data description . . . 94
3.3 Cross-language correlations of linguistic complexity . . . 95
3.3.1 Speech rate, information density, and linguistic complexity . . . 95
3.3.2 Information rate and linguistic complexity . . . 100
3.3.3 Shannon entropy versus conditional entropy . . . 101
3.3.4 Agglutination versus fusion . . . 105
3.3.5 Word order and linguistic complexity . . . 107
3.4 Discussion . . . 110
3.4.1 Equal overall complexity hypothesis: oversimplification or optimization? . . . 110
3.4.2 Sociolinguistic and neurocognitive constraints on complexity . . . 113

3.4.3 Conclusion . . . 115

4 Functional load: microsystemic organization of phonological system 117

4.1 Introduction . . . 119
4.1.1 The concept of functional load . . . 119
4.1.2 Some landmarks on functional load . . . 120
4.1.3 Paper outline . . . 123
4.2 Rationale and methodology . . . 125
4.2.1 Computing functional load . . . 125
4.2.2 Language description . . . 129
4.2.3 Data and preprocessing . . . 132
4.3 Distribution of *FL* for susbsystems of the phonological inventory . . . 135
4.3.1 Contributions of phonological subsystems to *FL* . . . 135
4.3.2 Frequency, morphology, and *FL* . . . 137
4.3.3 Consonontal bias . . . 141
4.4 Distribution of *FL* within phonological subsystems . . . 144
4.4.1 Patterns in *FL* distributions . . . 144
4.4.2 Cross-language trends in preferred phonological features . . . 150
4.5 General discussion . . . 156
4.5.1 *FL* at the level of phonological subsystems . . . 156
4.5.2 *FL* distribution within phonological subsystems . . . 159
4.5.3 Conclusion . . . 160

5 Conclusion 163

Appendix A 171

A.1 Information about oral data . . . 171
A.2 Translations of oral script (text Q1) . . . 176
A.3 Comparison of translations . . . 178
A.4 20 most frequent words in 18 languages . . . 179
A.5 Phonemic inventories of 9 languages . . . 189
A.6 Illustration of different configurations . . . 191
A.7 Contrasting pairs of vowels & consonants . . . 193

Bibliography 195

List of Figures

2.1 Comparison of translations in Mandarin Chinese . . . 28
2.2 Geographic location of the 18 languages studied . . . 44
2.3 *SR*, *ID*, & *IR* . . . 49
2.4 Comparison between the *SR* of female and male speakers . . . 52
2.5 SC_{TOKEN} & ΔSC . . . 53
2.6 Effects of bootstrapping and corpus size . . . 55
2.7 Effects of corpus size in estimating conditional entropy . . . 57
2.8 Correlations among *ID*, *SR*, and Shannon entropy *H(X)* . . . 59
2.9 Correlations among *ID*, *SR*, and conditional entropy $H(X_n/X_{n-1})$ . . . 61
2.10 Conditional entropy & percentage of monosyllabic and bisyllabic words . . . 63
2.11 Residuals of mixed effects models . . . 66
2.12 $IR_{\mathrm{S(X)}}$, $IR_{\mathrm{S(X_n|X_{n-1})}}$, and $IR_{\mathrm{S(X_n|X_{n+1})}}$ . . . 67
2.13 *SR*, *ID*, $IR_{\mathrm{H(X)}}$, & $IR_{\mathrm{S(X)}}$ . . . 70
2.14 Residuals of mixed effects models . . . 71
2.15 Syntagmatic vs. paradigmatic measures of *IR* . . . 76

3.1 SC_{TYPE} & WC_{TYPE} . . . 96
3.2 Morphological complexity & *SR* . . . 98
3.3 SC_{TYPE} & *SID* . . . 99
3.4 *IR* & SC_{TOKEN} . . . 101
3.5 Morphological complexity & conditional entropy . . . 104
3.6 Agglutination vs. fusion . . . 105
3.7 WC_{TOKEN} & *WID* . . . 107
3.8 WC_{TYPE} & $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ in terms of word order . . . 109

4.1 Illustrations of English (RP) vowel system . . . 124
4.2 Functional loads carried by vowels (V), consonants (C), tones (T) and stress (S) and Infra-syllabic *FL* (FL_{VCTS}) . . . 137
4.3 Segmental functional load (FL_{VC}) . . . 139
4.4 *CBias* according to corpus configuration . . . 142
4.5 Distribution of vowel pairs . . . 145
4.6 Distribution of consonant pairs . . . 146
4.7 Simulation of the relative loss of entropy induced by reducing vowel system 148
4.8 Simulation of the relative loss of entropy induced by reducing consonant system . . . 149

4.9 Distribution of FL_E as a function of feature distances of the contrasts . . . 154

A.1 Comparison of translations in French 178
A.2 Comparison of translations in Korean 178
A.13 Vowel inventories of 9 languages 189
A.14 Consonant inventories of 9 languages 190
A.15 Contrasting pairs of vowels 193
A.16 Contrasting pairs of consonants 194

List of Tables

2.1 Description of text corpus . . . 30
2.1 Description of text corpus . . . 31
2.2 List of syllabified word-forms and their frequency . . . 40
2.3 Calculation of Shannon entropy . . . 40
2.4 Conditional entropy: list of bigrams . . . 41
2.5 Calculation of conditional entropy $H(X_n|X_{n-1})$ . . . 41
2.6 Language description . . . 44
2.7 *SR*, *ID*, *IR*, *SC*, & ΔSC . . . 48
2.8 Mixed-effects model of *IR* . . . 51
2.9 *H(X)*, inventory size and $IR_{H(X)}$ . . . 58
2.10 Result of ANOVA taking $H(X)$ as a dependent variable . . . 60
2.11 $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$ . . . 61
2.12 Percentage of monosyllabic, bisyllabic, and trisyllabic words . . . 62
2.13 Result of ANOVA for conditional entropy . . . 64
2.14 *IR* obtained from conditional entropy . . . 65
2.15 Average *IR* obtained from $S(X)$, $S(X_n|X_{n-1})$, and $S(X_n|X_{n+1})$ . . . 67
2.16 Mixed-effects model of $IR_{S(X)}$ . . . 69
2.17 Correlations between syntagmatic and paradigmatic measures of *ID* . . . 73
2.18 Comparison of the AIC scores of mixed effects models . . . 73
2.19 Comparison of the AIC scores of mixed effects models . . . 77

3.1 Measure of morphological complexity . . . 91
3.2 Morphological classification . . . 95
3.3 *WC*, ΔW, *SC*, & ΔS . . . 97
3.4 Correlations among *SR*, *ID*, and linguistic complexity . . . 97
3.5 Correlations between *IR* and linguistic complexity . . . 100
3.6 Morphological and phonological complexity . . . 103
3.7 Correlations between morphological and phonological complexity . . . 103
3.8 Comparison between SOV and SVO . . . 108

4.1 Language and corpus description . . . 130
4.2 Functional loads carried by vowels, consonants, tones and stress and Infra-syllabic $FL_{VCTS.}$ . . . 136
4.3 Functional loads (in %) associated with vowel and consonant inventories . . . 141
4.4 5 Vowel pairs with the highest FL_E . . . 151

4.5 5 Individual vowels with the highest FL_E 152
4.6 5 Consonant pairs with the highest FL_E 153
4.7 5 Individual consonants with the highest FL_E 153

A.1 Speaker description 171
A.2 Fictitious corpus 191
A.3 INF/TOKEN corpus 191
A.4 Intermediate corpus 192
A.5 LEM/TOKEN corpus 192
A.6 INF/TYPE corpus 192

Chapter 1

Introduction

When linguists use the term "language", or "natural human language", they are revealing their belief that at the abstract level, beneath the surface variation, languages are remarkably similar in form and function and conform to certain universal principles [Akmajian et al., 2001].

Human languages have been shaped by dynamic usage in the social interaction between speakers and hearers for tens of thousands of years or more. In the general framework of complex adaptive systems, languages are regarded as non-linear systems with emergent self-organizing behaviors, which result from multi-constrained optimization [Beckner et al., 2009]. Within this framework, it is assumed that universal trends in optimization exist among language structures regardless of language-specific differences, which is often explained by the notion of *self-organization*. The overarching framework of this thesis is thus provided as follows: emergence and self-organization will be explored at the interface between the shape of language systems (i.e. linguistic elements, the internal structure of linguistic subsystems, and linguistic complexity) and language use in speech communication (in particular, in terms of *information rate*). The notions of *complexity* and *information* will hence be extensively used here and potential *universal trends* will be evaluated to assess our hypotheses. The proposed approach is quantitative, cross-linguistic

and multi-level as explained below.

1.1 General framework

1.1.1 Multi-level analysis of language universals

In cognitive and evolutionary linguistics, language is defined as a "a bio-cultural hybrid, a product of intensive gene:culture coevolution over perhaps the last 200 000 to 400 000 years" [Evans & Levinson, 2009]. In their paper *The myth of language universals: Language diversity and its importance for cognitive science*, Evans and Levinson argued that "languages vary radically in sound, meaning, and syntactic organisation" and that language diversity reflects phylogenetic (cultural-historical) and geographical patterns [Evans & Levinson, 2009]. The importance of considering sociocultural factors along with cognitive constraints on the language (co-)evolution was highlighted in their paper.

In the same vein, Beckner and colleagues defined language as a *complex adaptive system* which is characterized by a phenomenon of self-organization [Beckner et al., 2009]. *Self-organization* is defined as a spontaneous emergence of macroscopic system behavior resulting from repeated interactions between simple behaviors of a microscopic scale [de Boer, 2012] [Mitchell, 2009]. In linguistics, the notion of *self-organization* has been applied in particular in phonology and phonetics ([Blache & Meunier, 2004] [Blevins, 2004] [de Boer, 2000] [Lindblom, MacNeilage, & Studdert-Kennedy, 1984] [Liljencrants & Lindblom, 1972] [Oudeyer, 2006] [Wedel, 2012], inter alia). In the framework of *complex adaptive system*, language structures emerge from the interpersonal communication between speakers and hearers and their cognitive processes [Beckner et al., 2009] [Slobin, 1997].

In the present study, language is regarded as a macrosystem which consists of microsystems (i.e. several linguistic subsystems such as phonology, morphology, syntax, and semantics) and the mesosystemic interaction between these microsystems. The main objective of this study is to contribute to the analysis of language universals on a multi-scale

approach. The phenomenon of self-organization (visible through the phenomenon of regulation, trade-off, or the existence of scaling laws) will be assessed at the three different levels of analysis: (i) macrosystemic, (ii) mesosystemic, and (iii) microsystemic levels. Our approach is similar to Greenberg's empirical approach (see below) since it employs the data in 18 languages chosen from 10 language families and it attempts to observe some general tendencies (i.e. statistical and non-implicational universals) among the languages.

The underlying hypothesis of this study is that some general tendencies among the typologically distinct languages are observed at each level of analysis. In the second chapter, the phenomenon of trade-off between speech rate and information density will be examined at the macrosystemic level, which is assumed to result in a relatively stable information rate among the 18 languages. The initial hypothesis was proposed by Pellegrino and colleagues [Pellegrino, Coupé, & Marsico, 2011] and this part of thesis extended their study by adopting information-theoretic approaches.

The third chapter of thesis will be devoted to investigating the correlation between morphological and phonological modules at the mesosystemic level, based on the equal complexity hypothesis ([Fenk & Fenk-Oczlon, 2006] [Hockett, 1958] [Kusters, 2003] [Plank, 1998] [Shosted, 2006], inter alia). The equal complexity hypothesis was popular until the very end of twentieth century along with holistic typology which studies systemic dependencies between linguistic subsystems. However, it has recently been criticized by modern theoretical linguists for lack of evidence and falsifiability [Joseph & Newmeyer, 2012]. In this study, it is assumed that the equal complexity may result from the optimal balance between the sociocultural interaction ([Lupyan & Dale, 2010] [McWhorter, 2001] [Nettle, 2012] [Trudgill, 2011] [Wray & Grace, 2007]) and cognitive constraints ([Beckner et al., 2009] [Bell et al., 2009] [Christiansen & Chater, 2008] [Gregory et al., 1999] [Jurafsky et al., 2001] [Lindblom, 1990]), based on the framework of complex adaptive system and that as a consequence, a negative correlation would exist between morphological and phono-

logical modules.[1]

In the fourth chapter of thesis, the distribution of phonological contrasts will be assessed at the microsystemic level by means of *functional load* which is a tool for measuring the relative importance of phonological contrasts. As argued by [Hockett, 1966], some contrasts play more important role than others in the lexical access and in morphological strategies. The structures of phonological system were previously described by Vitevitch as scale-free networks due to their preferential attachment (i.e. a small number of giant components (hubs) with many other smaller components) [Vitevitch, 2008], based on the growth theory of Barabási-Albert [Barabási & Albert, 1999]. Such property of phonological system (i.e. robustness and resilience to the errors and damages of components) is regarded as the consequences of cognitive optimization for language acquisition, production, and perception. As a consequence, it is estimated that only a few contrasts play an important role in the phonological system in this study.

1.1.2 Language universals and linguistic typology

Linguists have been trying to describe languages based on the assumption that languages share some similarities in common for long time. As Comrie pointed out in *Language universals & linguistic typology*, both the studies of language universals and language typology are related with variation across languages. While the former is focused on the "limits" of variation, the latter is related to the "magnitude" of variation [Comrie, 1989]. In linguistic typology, language universals are generally classified into 4 different types as proposed in [Comrie, 1989]:

i) absolute universals vs. tendencies (i.e. statistical universals): An absolute universal means that there is no exception (e.g. all languages have vowels) whereas a tendency (or a statistical universal) indicates that there are some exceptions (e.g. Greenberg's linguistic universal 4: *With overwhelmingly greater than chance frequency, languages with normal*

[1] See [Fenk & Fenk-Oczlon, 2006] and [Shosted, 2006] for counterargument.

SOV order are postpositional. However, there are some SOV languages with prepositions such as Persian and Latin.).

ii) implicational vs. non-implicational (or unrestricted) univerals: an implicational universal implies the presence of a property on the condition of the presence of some properties, i.e. if p, then q, (e.g. Greenberg's linguistic universal 2: *In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.*). On the contrary, a non-implicational universal refers to a property which does not require any other condition (e.g. *languages in all parts of the world have at least one coronal consonant* [Maddieson, 1991]).

Furthermore, there are two main contrasting approaches to language universals [Comrie, 1989]:

i) Greenberg's empirical approach: a list of 45 language universals in syntax and morphology was proposed by Greenberg, based on a set of 30 languages from different language families [Greenberg, 1966].

ii) Chomsky's generative and formal approach: according to generativism, it was argued that Greenberg's approach only deals with surface syntactic structures while the Universal Grammar of Chomsky is focused on deep syntactic structures, thus, more abstract structures, taking only a single language into account [Chomsky, 1965] [Joseph, 2000].

According to the classification of language universals by Comrie, Greenberg's empirical approach is classified as a tendency (i.e. statistical universal) and an implicational universal, using a wide range of languages while Chomsky's formal approach is considered as an absolute universal, dealing with only one language. In contrast to Greenbergian and Chomskyan approaches to language universals which do not take extra-linguistic factors into account, cognitive and evolutionary approaches emphasize the importance of considering sociolinguistic and neurocognitive factors influencing language evolution, which will be described in Subsection 1.3.

1.2 Measures of linguistic complexity

In linguistic typology, the quantification of linguistic complexity has been used as a tool for describing and comparing languages ([Dahl, 2004] [Fenk-Oczlon & Fenk, 2005] [Fenk & Fenk-Oczlon, 2006] [Maddieson, 2006] [Nichols, 2007] [Shosted, 2006], inter alia). Several measures of linguistic complexity will be employed in the third chapter to assess the equal complexity hypothesis, especially by investigating a negative correlation between morphological complexity and phonological complexity.

There are two different ways of measuring linguistic complexity based on: (i) traditional linguistic approach, (ii) information-theoretic approach. The traditional linguistic method of quantifying linguistic complexity is to count the number of constituents of the linguistic system in question [Bane, 2008] [McWhorter, 2001] [Moscoso del Prado, 2011] [Nichols, 2007] [Shosted, 2006]. For instance, *word complexity* and *syllable complexity* are defined as the average number of syllables per word and the average number of segments per syllable respectively in [Fenk-Oczlon & Fenk, 2005].

1.2.1 Grammar-based complexity: traditional linguistic approach

This subsection will be devoted to presenting several measures of grammar-based complexity. The information-theoretic measures of linguistic complexity will be covered in the next subsection.

i) Phonological complexity: it can be measured by counting the size of syllable inventory and phonemic inventory [Bane, 2008]. Syllable complexity is calculated as the average number of segments (and tones, if applicable) [Fenk-Oczlon & Fenk, 2005] [Pellegrino, Coupé, & Marsico, 2011]. Furthermore, in [Maddieson, 2006], the degree of syllable complexity is determined based on the maximally complex syllable structure (simple, moderately complex, and complex) [Dryer & Haspelmath, 2013].

ii) Morphological complexity: it can be obtained by counting the number of inflectional

categories which can be marked by verbs [Bane, 2008] [Bickel & Nichols, 2005] [Shosted, 2006] as it was suggested by McWhorter that "inflection always complexifies grammar" [McWhorter, 2001] while derivational morphology was considered more functional and thus, was excluded from the complexity metric. In this study, the measure of morphological complexity proposed in [Lupyan & Dale, 2010] will be employed, where 28 linguistic features accounting for inflectional morphology are chosen from WALS (World Atlas of Language Structures) [Dryer & Haspelmath, 2013].

iii) Syntactic complexity: it is frequently used as a metric for the language proficiency of second language learners and is defined as "the range of forms that surface in language production and the degree of sophistication of such forms" [Ortega, 2003]. It can be measured by the average number of specific syntactic constructions (e.g. passives and nominals) per sentence and the average number of verbs per sentence [Chen & Zechner, 2011].

iv) Semantic complexity: in comparison with other linguistic modules, there are relatively few studies on semantic complexity. In [Fenk-Oczlon, 2013] [Piantadosi, Tily, & Gibson, 2012], semantic complexity was computed by the average number of lemmas per homophones, taking the distribution of word frequencies into account. According to the result presented by Piantadosi and colleagues, high-frequency and short words tend to encompass more meanings, thus, more ambiguity, which is related to the communicative efficiency [Bell et al., 2009] [Zipf, 1949].

In the literature related to linguistic complexity, there are relatively more studies concerning morphological complexity followed by phonological complexity than syntactic and semantic complexity. In a strict sense, syntactic complexity is regarded as a user-based complexity while the other linguistic complexity is considered as a grammar-based complexity, since syntactic complexity is mainly related to the linguistic performance of first and second language learners [Dabrowska, 2010] [Hawkins, 2003]. The grammar-based and user-based complexity has been frequently employed in the studies of linguistic complexity

but its major shortcoming is the absence of justification for selecting complexity indicators [Bane, 2008]. The information-theoretic approach, which will be described in the next subsection, has appeared as an alternative to the quantification of linguistic complexity.

1.2.2 Usage-based complexity: information-theoretic approach

In the framework of Information theory, language is defined as a system consisting of a finite set of linguistic units (e.g. words, syllables, or segments) [Hockett, 1966]. Contrary to grammar-based complexity, the information-theoretic approach takes account of the predictability distribution estimated from a language model based on large corpora for quantifying linguistic complexity.

i) Phonological complexity: it can be obtained by the estimated average amount of information (in bits) contained per linguistic unit [Goldsmith, 2000] [Goldsmith, 2002] [Kello & Beltz, 2009] [Pellegrino, Coupé, & Marsico, 2007] [Villasenor et al., 2012] by means of the information measures such as Shannon entropy $H(X)$ and conditional entropy $H(X|C)$ [Shannon, 1948]. The former quantifies the average amount of information from a unigram language model without context while the latter computes the average amount of information taking contextual information into account. Those two formalized measures of information reduce a message into binary arithmetic coding (i.e. 0s or 1s) and allow us to evaluate how many bits on average are necessary to encode a random linguistic variable [Goldsmith, 2000].

$$H(X) = -\sum_{i=1}^{N_L} p_{\sigma_i} \cdot log_2(p_{\sigma_i}) \qquad H(X|C) = \sum_{c \in C} p(c) \cdot H(X|C = c) \tag{1.1}$$

ii) Morphological complexity: the minimum description length (MDL) [Rissanen, 1984] of inflectional morphology and lexicon can be approximated by means of automatic unsupervised morphological analyzers such as Linguistica [Goldsmith, 2001] and Morfessor [Virpioja et al., 2013]. The lexicon constructed by Linguistica consists of a set of stems, affixes

and signatures. The notion "signature" proposed by Goldsmith refers to a subset of affixes which can be possibly combined with a subset of stems. The metric of morphological complexity (MC) was proposed in [Bane, 2008] as follows, where $DL(x)$ corresponds to the description length of x which is defined as the shortest description (i.e. Kolmogorov complexity) approximated by Linguistica, and morphological complexity is computed as the ratio of the description length of inflectional morphology to the total information encoded by lexicon.

$$MC = \frac{DL(Affixes) + DL(Signatures)}{DL(Affixes) + DL(Signatures) + DL(Stems)} \tag{1.2}$$

The other measure of morphological complexity based on the information-theoretic approach computes the average amount of information per paradigm cell [Ackerman & Malouf, 2013] [Blevins, 2013] [Kostić, 1991] [Moscoso del Prado, Kostić, & Baayen, 2004] [Moscoso del Prado, 2011].

iii) Syntactic complexity: it can be obtained by means of syntactic surprisal and lexicalized surprisal proposed by Demberg and Keller [Demberg & Keller, 2008] [Demberg et al., 2012]. Both measures can be computed using the equation provided below, using an elaborated language model such as probabilistic context-free grammar (PCFG), which computes the probability of grammatical rules obtained from a syntactic tree. Syntactic surprisal quantifies the portion of the structural information between the words W_k and W_{k+1} ignoring the effect of word frequency while lexicalized surprisal employs both the structural information and word frequency distributions.

$$S_{k+1} = \sum_{T} P(T|W_1...W_{k+1}) log \frac{P(T|W_1...W_{k+1})}{P(T|W_1...W_k)} \tag{1.3}$$

iv) Semantic complexity: as Shannon mentioned that "semantic aspects of communication are irrelevant to the engineering problem" [Shannon, 1948], the Information theory does not appear to be directly related to semantic complexity. The alternative method of

counting the average number of lemmas per homophones can be employed, considering the distribution of word frequency [Fenk-Oczlon, 2013] [Piantadosi, Tily, & Gibson, 2012].

1.3 Information encoding

1.3.1 Sociolinguistic factors

In quantitative linguistics and psycholinguistics, it has been argued that human languages are structured for optimal and efficient communication ([Frank & Jaeger, 2008] [Jaeger, 2010] [Levy & Jaeger, 2007] [Mahowald et al., 2013] [Piantadosi, Tily, & Gibson, 2011] [Zipf, 1949], inter alia). In order to assess the way how languages encode and transmit information, language-external factors should be taken into account in addition to linguistic factors. Non-linguistic factors can be distinguished into two types: sociolinguistic and neurocognitive factors.

In sociolinguistics, Lupyan and Dale suggested that language structure is related to social environments such as speaker population size, geographic spread, and the degree of linguistic contact [Lupyan & Dale, 2010]. The result of their study illustrated that languages adapt themselves to the social environments in which they are acquired and spoken. For instance, languages spoken by large population tend to exhibit simple inflectional morphology and use more lexical strategies rather than inflectional morphology. Furthermore, morphological simplification is observed in languages acquired by a large number of adult learners [Trudgill, 2011]. Since languages are shaped by the environments, they are compared to "biological organisms shaped by ecological niche" [Lupyan & Dale, 2010].

Regarding the relationship between phonological complexity and sociolinguistic factors, a positive correlation was found between speaker population size and phoneme inventory size, using a sample of 250 languages by Hay and Bauer [Hay & Bauer, 2007] and their result was further replicated by Atkinson [Atkinson, 2011] and Wichmann and

colleages [Wichmann, Rama, & Holman, 2011], adding more languages to the sample. Contrary to the relationship between morphological complexity and speaker population size, languages spoken by a large population exhibit a large phonemic inventory.

1.3.2 Neurocognitive factors

For the efficient communication and optimal information transmission, words with high frequency tend to be short, simple and contain more meanings [Bell et al., 2009] [Piantadosi, Tily, & Gibson, 2011] [Zipf, 1949]. In other words, high-frequency words require little memory effort and are often used in many different contexts. In this way, speakers reduce their effort in speech production. On the contrary, since high-frequency words are frequently used in different contexts, it requires more disambiguation effort from hearers whereas low-frequency words would require less disambiguation effort from hearers [Ferrer i Cancho & Solé, 2003] [Kello & Beltz, 2009].

While speakers try to economize their articulation effort, hearers also try to reduce their effort of disambiguation and the likelihood of confusion. Thus, "a conflict of interest" is created by the interaction between speakers and hearers and language structures covary by the social interaction between them [Beckner et al., 2009] [Bell et al., 2009] [Christiansen & Chater, 2008] [Gregory et al., 1999] [Jurafsky et al., 2001] [Lindblom, 1990]. In the framework of complex adaptive system, language evolution is not considered as the outcome of the adaptation of brain to the language structures but as the result of the "interpersonal communicative and cognitive process" between speakers and hearers [Christiansen & Chater, 2008] [Slobin, 1997].

After Zipf's law which states that word length is inversely correlated with word frequency [Zipf, 1949], there are several hypotheses which extended Zipf's idea about communicative efficiency. In particular, many studies take an information-theoretic approach into account, which provides the mathematical formalization of the information content transmitted in communication [Shannon, 1948]. In addition to word frequency, the notion

of conditional predictability is taken into account by using contextual information.

Information-theoretic measures such as Shannon entropy and conditional entropy allow us to quantify the cognitive costs of language use for speakers and hearers. Ferrer i Cancho and colleagues suggested that Shannon entropy corresponds to the cognitive effort for both speakers (i.e. memory effort and lexical activation) and hearers (i.e. recognition) and conditional entropy corresponds to the cognitive effort for hearers (i.e. disambiguation) [Ferrer i Cancho & Solé, 2003] [Ferrer i Cancho, 2006] [Ferrer i Cancho & Díaz-Guilera, 2007]. Thus, by comparing Shannon entropy and conditional entropy among typologically distinct languages chosen from 10 different language families in this study, we can observe whether there is a general tendency (i.e. statistical universal) among the languages in terms of their cognitive costs for speakers and hearers. Levinson assumed that hearers' effort of disambiguation is less costly in comparison with speakers' effort of production [Levinson, 2000] [Piantadosi, Tily, & Gibson, 2012].

Based on an information-theoretic approach and statistical mechanics, Ferrer i Cancho and Solé showed that in the distributions of word frequency, there are two kinds of patterns: (i) relatively flat and uniform distribution of probability (i.e. characterized by high entropy which requires less memory effort and more disambiguation effort) vs. relatively unequal and peaked distribution of probability (i.e. characterized by low entropy which requires more memory effort and less disambiguation effort). The authors claimed that the efficient communication results from a balance between these two phases, producing a scaling law in the distribution of word frequency [Ferrer i Cancho & Solé, 2003] [Kello & Beltz, 2009].

"The entropy rate constancy principle" was proposed by Genzel and Charniak, which asserts that speakers tend to maintain the constant rate of conditional entropy given the previous elements during their utterances [Genzel & Charniak, 2002] [Genzel & Charniak, 2003]. It was shown in their results that the Shannon entropy of sentence without considering context increases as the sentence number increases, which supports their hy-

pothesis since conditional entropy can be obtained by subtracting the mutual information between the sentence and the context (which increases as the sentence number increases) from Shannon entropy. Thus, condition entropy remains stable as the sentence number increases. But the replication of this study on Chinese corpora showed that there was no effect of sentence [Qian & Jaeger, 2009] and that Shannon entropy did not increase as a function of sentence number.

In the same vein, "the uniform information density (UID) hypothesis" was proposed by Levy and Jaeger [Jaeger, 2010] [Levy & Jaeger, 2007]. According to the UID hypothesis, speakers modulate the information density of their utterances in order to optimally transmit the information at a uniform rate, near the channel capacity [Frank & Jaeger, 2008] [Jaeger, 2010] [Levy & Jaeger, 2007] [Mahowald et al., 2013] [Piantadosi, Tily, & Gibson, 2011] [Piantadosi, Tily, & Gibson, 2012]. The UID hypothesis is focused on the way how speakers plan and produce their utterances, based on the assumption that they do it efficiently due to several constraints imposed by speakers, hearers, and environments (e.g. channel capacity) [Frank & Jaeger, 2008]. Thus, it is assumed that the efficient and optimal way of transmitting information is to maintain the information density of their utterances uniformly without exceeding the channel capacity.

The UID hypothesis was attested by the results presented in several studies. To begin with, Piantadosi and colleagues suggested that word length is better predicted by information density (obtained by using the previous word as contextual information) than by word frequency, which extended the study of Zipf's law [Mahowald et al., 2013] [Piantadosi, Tily, & Gibson, 2011] [Zipf, 1949]. Furthermore, words with more ambiguity (homophones with more meanings) are short, simple, and highly predictable [Piantadosi, Tily, & Gibson, 2012]. As a consequence, it is expected that speakers would choose their words to balance out the information density of their utterances while minimizing their effort of lexical activation and articulation.

A similar tendency was found for both words [Bell et al., 2009] and syllables [Aylett

& Turk, 2004]. At the syllable level, similarly to the UID hypothesis, Aylett and Turk proposed *The smooth signal redundancy hypothesis* according to which speakers modulate phonetic duration and prosodic prominence as a function of the redundancy which is obtained by word frequency, syllable trigram probability, and givenness (i.e. "how many times a referent has been mentioned") in spontaneous speech [Aylett & Turk, 2004]. In their results, syllable duration is inversely related to language redundancy. At the word level, Bell and colleagues suggested that content words and function words behave differently with respect to the effects of frequency and predictability. The result of their study showed that the duration of content word is affected by their frequency and predictability whereas the duration of function word is not affected and that low-frequency content words have longer duration due to their lower level of lexical activation [Bell et al., 2009].

A correlation between morphosyntactic reduction (e.g. "I am" vs. "I'm") and information density was assessed by Frank and Jaeger [Frank & Jaeger, 2008]. In their result, speakers use a full form to increase the length of their utterances if the elements containing high information (obtained by Shannon entropy) are uttered and they use a reduced form to shorten their utterances if the elements containing low information are uttered. Similarly, the UID hypothesis was also attested by the reduction of syntactic structures [Jaeger, 2010] [Levy & Jaeger, 2007]. In order to maximize the uniformity of the information density of their utterances, speakers omit or add the function word *that* before a relative clause in English sentence, which suggests that information density plays an important role of predicting speakers' preferences during their utterances.

1.3.3 Trade-off in information encoding

A part of the second chapter (cf. 2.3.1) which studies the average information rate at the macrosystemic level is an extended version of the paper *A cross-language perspective on speech information rate* [Pellegrino, Coupé, & Marsico, 2011] where the methodology proposed by the authors was replicated. The following measures, such as *speech*

rate, *information density*, *information rate*, and *syllable complexity*, were adopted from the paper and their initial study was extended by adding more languages and adopting information-theoretic and paradigmatic measures of information. In their paper, by using oral data which contain the equivalent semantic information in the 7 languages (British English, French, German, Italian, Japanese, Mandarin Chinese, and Spanish) translated from British English or French into a target language, Pellegrino and colleagues asserted "the equal overall *communicative* capacity" that languages transmit the information at a relatively similar rate within a limited range of variation, regardless of their specific encoding strategy and linguistic complexity. The underlying hypothesis of their study is that a trade-off exists between speech rate (i.e. the average number of syllables uttered per second) and information density (i.e. the average density of information in speech chunks, obtained by taking vietnamese, the most isolating language, as a reference).

A similar hypothesis was proposed by Fenk-Oczlon and Fenk that a relatively "constant" flow of information transmission results from the complexity trade-offs between linguistic subsystems, without excluding language-specific differences in the trade-offs [Fenk-Oczlon & Fenk, 2014]. However, the conclusion of their paper put emphasis on the difficulty of defining and quantifying the overall complexity of a language, which is the fundamental problem for comparing different languages.

Regarding the UID hypothesis, "the equal overall communicative capacity" proposed by Pellegrino and colleagues differs from the UID hypothesis for the following three reasons: (i) the main concern of the former is focused on a cross-language comparison of a similar rate of information transmission from a typological perspective using the data in several languages from different language families while the UID hypothesis is not related to the typological aspect and the comparison of several languages, (ii) the latter measures information density, using information-theoretic measures while the former quantifies information density, using the vietnamese as a reference, (iii) the main goal of the UID hypothesis is to study the speakers' strategy for optimizing the information transmission

while the main goal of the former is to compare the average rate of information transmitted by speakers in typologically diverse languages, considering language as a complex adaptive system which can be explained by the phenomenon of self-organization.

1.4 Overview

In the present chapter, the main objective of this thesis was presented. Within the framework of *complex adaptive system*, language is defined as an emergent, complex, and non-linear macrosystem which results from the mesosystemic interaction between several microsystems such as phonology, morphology, syntax, and semantics. According to our underlying hypothesis, it is assumed that regardless of language diversity [Evans & Levinson, 2009], a general tendency of self-organization is found among the typologically diverse languages by analyzing linguistic phenomenon at the three different levels of analysis: macrosystemic, mesosystemic, and microsystemic levels. Hence, this thesis can be characterized as a multi-level analysis of language universals (which is summarized as a phenomenon of *self-organization* in this study).

In the second chapter, the average information rate of 18 languages will be compared cross-linguistically at the macrosystemic level by using syntagmatic and paradigmatic (information-theoretic) measures of information rate. First, theoretical framework regarding the notion of self-organization and the Information theory of Claude Shannon will be described. Second, multilingual oral and text corpora in the 18 typologically diverse languages will be illustrated along with the parameters which can be divided into two different types: syntagmatic and paradigmatic measures. Third, the following hypothesis will be assessed: the average information rate is quite stable among the 18 languages, due to a trade-off between speech rate and information density [Pellegrino, Coupé, & Marsico, 2011]. Fourth, the average information rate obtained by means of one syntagmatic measure on a local scale (based on [Pellegrino, Coupé, & Marsico, 2011]) and

several paradigmatic measures on a global scale (based on the Information theory [Shannon, 1948]) of information will be compared, based on the assumption that among the paradigmatic (information-theoretic) measures, conditional entropy which takes contextual information into account is more elaborated and accurate than Shannon entropy.

The mesosystemic interaction between two microsystems, i.e. phonology and morphology, will be examined in the third chapter. Based on the equal overall complexity hypothesis and holistic typology, it is assumed that a negative correlation (i.e. complexity trade-off) exists between linguistic modules [Fenk & Fenk-Oczlon, 2006] [Shosted, 2006]. In this study, a correlation between phonological and morphological modules will be assessed by using the measures of linguistic complexity in 14 languages. However, contrary to the analysis at the macrosystemic level, more cautious approach should be taken to address the phenomenon of complexity trade-off between linguistic modules since there are still ongoing discussions regarding the validity of the equal overall complexity hypothesis (cf. [Fenk-Oczlon & Fenk, 2014] [Joseph & Newmeyer, 2012] [Shosted, 2006]).

In the fourth chapter, the phenomenon of self-organization will be assessed by means of *functional load* within phonology, i.e. at the microsystemic level. Functional load is used as a tool for quantifying the relative importance of a phonological contrast in language. It thus allows us to observe and compare the internal functional organization of phonological systems in 9 languages in this study. As it was suggested by [Oh et al., 2013] [Vitevitch, 2008], it is estimated that only a few phonological contrasts play an important role in each phonological system as a general trend among the 9 languages, regardless of specificities in each phonological system. The uneven distribution of functional load may result from the self-organization of phonological system which adapts itself to be more resilient and robust to the errors and damages of components in speech communication.

Finally, in conclusion, the important results which are related to the main hypothesis of this study (i.e. the phenomenon of self-organization at multilevel) in each chapter will be summed up.

Chapter 2

Average information rate: macrosystemic analysis

2.1 Introduction

2.1.1 Phenomenon of self-organization

Order is created out of disorder, upending the usual turn of events in which order decays and disorder (or entropy) wins out [Mitchell, 2009].

The notion of *self-organization* has been frequently used in the studies of complex systems encompassing various fields: from natural science (biology, chemistry, and physics) to computer science (artificial intelligence, computer modelling, and cybernetics), social and human science (economics, geography, linguistics, psychology, and sociology). The term was first proposed by Ross Ashby and further developed by von Foerster in cybernetics [Ashby, 1947] [von Foerster, 1960]. Ross Ashby asserted that "every isolated determinate dynamic system obeying unchanging laws" goes towards equilibrium by developing "organisms that are adapted to their environments" [Ashby, 1962] and von Foerster proposed the principle of "order from noise" stating that while the internal order of system increases from the interaction with the environment, their external status becomes more

apparent, "changing from unorganized to organized" [von Foerster, 1960]. In thermodynamics, Nicolis and Prigogine explained the self-organization of a non-equilibrium system by the notion of "dissipative structure" according to which the system exports the excess entropy, since entropy can only increase in an isolated system following the second law of thermodynamics [Nicolis & Prigogine, 1977] [Prigogine & Nicolis, 1985].

In linguistics, *self-organization* has been widely applied to account for language evolution and acquisition particularly in phonology ([Blevins, 2004] [de Boer, 2000] [Lindblom, MacNeilage, & Studdert-Kennedy, 1984], [Oudeyer, 2006] [Wedel, 2012], inter alia). Following the definition by de Boer and Mitchell, *self-organization* is a spontaneous emergence of macroscopic system behavior resulting from repeated interactions between simple behaviors of a microscopic scale [de Boer, 2012] [Mitchell, 2009]. Commonly observed characteristics of self-organizing systems can be resumed as follows:

i) Interaction (or *positive feedback*) between two levels of structure: a system is defined as consisting of two levels of structure: microscopic and macroscopic scales (for example, the interaction between individual (Chomsky's *competence*) and population behaviors (Saussure's *parole*) [de Boer, 2012]).

ii) Emergence: complex collective behavior in a large population results from simple microscopic behaviors and the interaction between microscopic and macroscopic scales.

iii) Non-linearity and dissipation: the emergent complex macroscopic behavior of system as a whole cannot be accounted for by assessing the simple microscopic individual patterns (for example, as described in [Wedel, 2011], a cooked egg white cannot be explained by summing the properties of raw egg white proteins).

Along with those properties of self-organizing systems, the following three important features of *complex adaptive system* are described in [Beckner et al., 2009] [Mitchell, 2009]:

i) Collective behavior: a complex adaptive system consists of a large number of individual components (or agents) interacting with the other components of a microscopic level.

ii) Adaptation: by means of the interactions between macroscopic and microscopic scales

(i.e. positive feedback), a complex dynamic system adapts to the environment.
iii) Signaling and information processing: individual microscopic behaviors are produced from both the internal and external environments by a complex adaptive system.

In the present study, language is regarded as a *complex adaptive system* characterized by a phenomenon of self-organization and the three features explained above. Within the framework of language as a complex adaptive system [Beckner et al., 2009], the aim of this chapter is to investigate how 18 typologically distinct languages convey the information on average per second (i.e. information rate) while they exhibit wide-ranging speech rates and information density. Furthermore, language is viewed as a macrosystem composed of microsystems (i.e. linguistic modules) and the ultimate goal of this dissertation is to assess the phenomenon of self-organization at the three different levels of analysis: macrosystemic, mesosystemic, and microsystemic levels. To begin with, this chapter attempts to examine the phenomenon of self-organization at the macrosystemic level.

In *A cross-language perspective on speech information rate*, Pellegrino and his colleagues suggested that there's a phenomenon of self-organization between speech rate and information density with 7 languages [Pellegrino, Coupé, & Marsico, 2011]. This result can be explained by self-organization in a following manner: the recording of each speaker corresponds to individual behavior on a microscopic level which is a consequence resulted from the interactions between individual speaker's behavior and external environments such as information density, sociolinguistic and neurocognitive constraints. Furthermore, a limited range of information rate can be regarded as an emergent macroscopic behavior. As such, the notion of self-organization is used to explain various phenomena of complex non-linear systems with emergent behaviors.

To calculate information rate, information-theoretic measures are used for quantifying information content, based on the notions of Shannon entropy and conditional entropy proposed by Claude E. Shannon [Shannon, 1948]. The results obtained by using these information-theoretic measures are compared with the results of the other parameters of

information proposed in [Pellegrino, Coupé, & Marsico, 2011]: speech rate, information density, and information rate.

2.1.2 Information theory: quantifying measures of information

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. ... The significant aspect is that the actual message is one selected from a set of possible messages [Shannon, 1948].

Before introducing the measure of information, it is crucial to define what is exactly meant by the term "information" in Information theory. As the quote above by Shannon says, *information* does not refer to the semantic content of actual message but denotes what a message *could* contain. Weaver also defined information as "a measure of one's freedom of choice when one selects a message" [Weaver, 1953]. Thus, information is not concerned with its semantic meaning but is regarded as "the amount of surprise", i.e. the number of possible messages one could receive from the information source.

Information content is calculated by means of "entropy" in Information theory. The notion *entropy* was coined by the physicist Clausius in 1865, taken from a greek word τροπέ which means "transformation" [Mitchell, 2009]. The term was used in thermodynamics referring to a heat loss produced by friction, i.e. the amount of energy which cannot be transformed into work but instead is transformed into heat. It was further developed and generalized by Boltzmann, a pioneer of statistical mechanics who defined entropy as a number of possible microscopic properties provided a constant macroscopic behavior. Gibbs later introduced the probability of each microscopic property in the Boltzmann entropy formula, which gave rise to Shannon entropy. Shannon entropy[2] was thus based on the idea of Boltzmann and Gibbs and proposed that entropy is a measure of the collection of all possible messages (i.e. microscopic properties) sent from the information source on

[2]Entropy in Information theory is often termed "Shannon entropy" to be distinguished from the other versions of entropy proposed by Boltzmann and Gibbs.

the macroscopic level. It is often calculated by using the binary logarithm ($log_2 n$) and the unit of Shannon entropy is called *bit* (binary digit). A choice of the base of logarithm depends on the unit used for measuring information content. For example, the logarithm base 2 measures the information in binary digits while the logarithm base 10 measures the information in decimal digits.

In recent years, information-theoretic measures have been frequently used in the study of speech communication. The most commonly employed measures and the principles and hypotheses proposed by a corresponding measure are listed as follows:

i) Shannon and conditional entropy ([Genzel & Charniak, 2002] [Genzel & Charniak, 2003] [Goldsmith, 2000] [Goldsmith, 2002] [Hale, 2003] [Keller, 2004] [Pellegrino, Coupé, & Marsico, 2007] [Piantadosi, Tily, & Gibson, 2009] [Qian & Jaeger, 2012] [Villasenor et al., 2012], inter alia): the *constancy rate principle* was proposed by Genzel and Charniak [2002, 2003] asserting that the Shannon entropy of random variables (i.e. words in a text) is constant on average and the entropy increases as the sentence length increases.

ii) Probability and conditional probability ([Aylett & Turk, 2004] [Bell et al., 2009] [Gahl & Garnsey, 2004] [Gregory et al., 1999] [Jurafsky et al., 2001] [Pluymaekers, Ernestus, & Baayen, 2005] [Tily et al., 2009] [van Son & Pols, 2003], inter alia): the *probabilistic reduction hypothesis* was suggested by Jurafsky and his colleagues that words with a high conditional probability (considering contextual information) are likely to be reduced at the lexical level. At the speech level, the *smooth signal redundancy hypothesis* was presented by Aylett and Turk, which states that prosodic prominence increases syllable duration which is inversely related to language redundancy.

iii) Surprisal (also known as informativity or informativeness)([Cohen Priva, 2008] [Frank & Jaeger, 2008] [Hale, 2001] [Jaeger, 2010] [Levy & Jaeger, 2007] [Mahowald et al., 2013] [Piantadosi, Tily, & Gibson, 2011], [Seyfarth, 2014], inter alia): the *uniform information density (UID) hypothesis* was proposed by Levy and Jaeger. According to the UID hypothesis, speakers modulate the information density of their utterance in order to

optimally transmit the information at an uniform rate, near the channel capacity.

Shannon entropy and surprisal both estimate the amount of information by logarithmic equations whereas probability does not serve as a direct measure of information density. Thus, entropy and surprisal measures are mostly employed in recent studies on information density and information rate and are also used as a measure of information in this study. Surprisal is the average predictability of an individual microscopic property in context while Shannon entropy corresponds to the average surprisal obtained from the collection of all possible microscopic properties.

The aim of this chapter is to compare crosslinguistically the information rate in 18 languages computed by different measures of information density, including the quantitative parameters proposed in [Pellegrino, Coupé, & Marsico, 2011]. The research question addressed in this chapter concerns the way how the speakers of different languages convey the information in speech communication. Among the 18 languages which exhibit wide ranges of the average speech rate and information density, the average rate of transmitting information per unit of time (i.e. information rate) is estimated to be in a limited range as suggested in a cross-language study [Pellegrino, Coupé, & Marsico, 2011] with 7 languages. This tendency of relatively similar information rate seems to result from a complex and adaptive behavior of language (i.e. self-organization), along with several external factors such as sociolinguistic and cognitive constraints and the capacity of audio channel.

2.1.3 Chapter outline

Section 2.2 shows the methods and data of the present study. First, the measures of information based on Information theory [Shannon, 1948] and the parameters adopted from [Pellegrino, Coupé, & Marsico, 2011] are displayed in Section 2.2.1. Second, multilingual oral and text corpora in the 18 languages and the preprocessing methods of the data are described in Section 2.2.2. Then, the 18 typologically distinct languages investi-

gated in this study are presented in Section 2.2.3.

In Section 2.3, the results of computed average information rate in the 18 languages are displayed and compared crosslinguistically. First, the results obtained by three quantifying measures based on pairwise comparisons proposed in [Pellegrino, Coupé, & Marsico, 2011] (speech rate, information density, and information rate) are shown in Section 2.3.1. Second, the influence of the size of corpus and the bootstrap simulation is tested in a subset of 4 languages (English, Finnish, French, and Korean) in Section 2.3.2. The next four sections concern the comparison of information rates calculated by information-theoretic measures in the 17 languages: Shannon entropy (Section 2.3.3.), conditional entropy (Section 2.3.4), and surprisal (Section 2.3.5).

The results are further discussed in Section 2.4. First, the importance of considering contextual information for the computation of information density is highlighted by comparing the results of mixed effects models in Section 2.4.1. Second, regarding the UID hypothesis, the results of this study present a different perspective on the optimal encoding strategy (i.e. relatively similar rate of conveying information per second) and the differences between the two perspectives are explained in Section 2.4.2, followed by the conclusion.

2.2 Materials and methods

In this section, multilingual oral and textual corpora are described (Section 2.2.1) along with several measures of information and relevant parameters, such as speech rate and information density (Section 2.2.2) and 18 languages investigated in this study (Section 2.2.3).

2.2.1 Data and preprocessing

2.2.1.1 Oral corpus

A part of oral corpora was initially extracted from the Multext (Multilingual Text Tools and Corpora) project [Campione & Véronis, 1998] and was extended by adding more languages to the initial data. The latest version of oral corpora contains the data in 3 languages (British English, German, and Italian) taken from the Multext project along with the data in 15 languages (Basque, Cantonese, Catalan, Finnish, French, Hungarian, Japanese, Korean, Mandarin Chinese, Serbian, Spanish, Thai, Turkish, Vietnamese, and Wolof) collected by the author, Christophe Coupé, and Eric Castelli.

Among the 15 languages recently added to the initial data, a part of Vietnamese data (4 speakers) was recorded by Eric Castelli at the laboratory of MICA in Vietnam. Furthermore, 4 languages (Cantonese, Mandarin Chinese, Serbian, and Thai) were recorded by Christophe Coupé: Cantonese data was recorded at City University of Hong Kong and Mandarin Chinese (also known as Putonghua) data was collected at Peking University. For Serbian, standard Serbian spoken in Belgrade was recorded in Beijing, Belgrade, and Lyon and Thai data was collected at Chulalongkorn University and Alliance Française in Bangkok.[3]

Following 11 languages (Basque, Catalan, Finnish, French, Hungarian, Japanese, Korean, Spanish, Turkish, Vietnamese, and Wolof) were recorded by the author: Basque data was recorded in Barcelona (with assistance from Euskal Etxea), Donostia, Lyon, and Tolosa. Catalan speakers were recorded at Polytechnic University of Catalonia in Barcelona and Rovira i Virgili University in Tarragona. Finnish, French, Hungarian, Japanese, and Turkish data were collected in Lyon and the recording took place mainly at the laboratory of Dynamique du Langage in Lyon. For Korean, the standard Korean spoken in Seoul was recorded in Seoul. For Spanish, Catalan/Spanish and Basque/Spanish

[3] 12 native speakers of Khmer were also recorded in Phnom Penh by Christophe Coupé but Khmer was discarded in this study due to disfluency in the utterances of native speakers.

bilingual speakers in Barcelona, Donostia, Lyon, Tarragona, and Tolosa were recorded. A part of Vietnamese data (6 native speakers) were recorded in Grenoble and Lyon and Wolof data was collected in Lyon and Paris.[4]

10 native (5 female and 5 male) speakers were recorded in each language reading 15 texts by the RocMe! software [Ferragne, Flavier, & Fressard, 2013] (see Appendix A.1 for the information regarding the 10 native speakers of each language such as the number of texts uttered by speaker, sex, and age of speaker). There was no specific restriction regarding the sociolinguistic profile of native speakers but most of them were students or faculty members at the university.[5] The oral script of corpus consists of 15 short texts containing 3-5 semantically connected sentences translated from British English or French into each target language by a native speaker (see Appendix A.2 for an example of the translations in 18 languages). Most of the native speakers who translated the texts were linguists, except for Catalan, Finnish, Serbian, and Thai. When the texts were not translated by linguists, the translation was checked and verified by other native speakers.

Initially, there were 20 texts in the oral script but 5 of them were discarded due to the semantic content of texts which creates speakers' disfluency or requires a wide range of cultural adaptations in terms of translation. For example, among the 5 discarded texts, there were one passage reporting a traffic summary in England and the other one describing an inventory of a department store. However, since it was not possible to completely discard some proper nouns, the translators were asked to pay more attention to the translation of proper nouns by selecting a corresponding word with a similar number of syllables. For example, in one text, a list of European cities are enumerated in British English script as follows: "Paris, Bruges, Frankfurt, Rome, and Hamburg". In Korean version, it was culturally adapted and translated as "Hong Kong, Shanghai, Beijing, Tokyo, and Kobe".

[4] 5 native speakers of Fang were recorded in Lyon by the author but due to the lack of fluency in speakers' utterance, Fang was discarded.

[5] While there is no imbalance among the data of 15 languages added to the initial Multext corpus, there are less than 15 texts recorded by each speaker in the data for 3 languages (British English, German, and Italian) extracted from the Multext corpus.

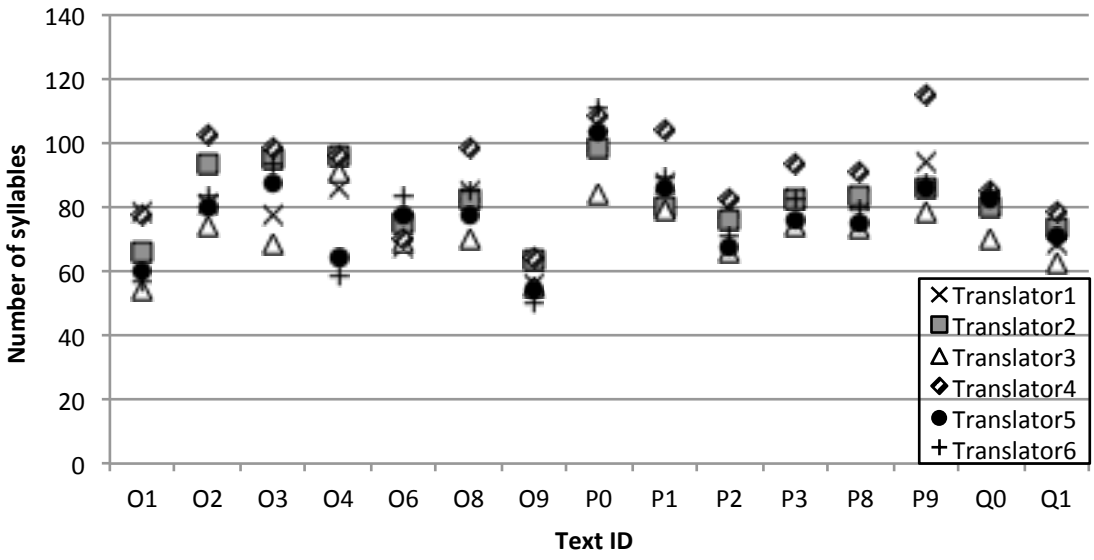

Figure 2.1: Comparison of translations in Mandarin Chinese

Several versions of translation were compared in 3 languages (French, Korean, and Mandarin Chinese).[6] Figure 2.1 displays the variation among 6 different versions of translation in Mandarin Chinese. They are compared in terms of the number of syllables per each text. The version chosen for this study was translated by Translator5 (marked in black circle) which is placed mostly in the middle among the other 5 versions of translation except for the text O4. As expected, it is observed that the range of variation depends on the length of texts: for example, a significant correlation exists between the range of variation (i.e. gap between the maximum and the minimum number of syllables) and the average length of text (i.e. number of syllables)(Pearson's $r = 0.610$*; p-value = 0.016; Spearman's $\rho = 0.731$**; p-value = 0.002; N = 15). A relatively small range of variation is displayed in some texts (O6, O9, P2, P3, P8, Q0, Q1) while a wide range of variation is observed in the others (O1, O2, O3, O4, O8, P0, P1, P9). Thus, some texts seem more prone to individual variation among translators than others.

In a first phase of recording, speakers were asked to read the 15 texts silently which

[6]See Appendix A.3 for the comparison of French and Korean translation.

appear one by one on the screen with a random order.[7] In a second stage, speakers read each text aloud twice before the actual recording. Each text was thus separately recorded one at a time in a random order after being read three times by the speaker, including the silent reading. This process of repeatedly reading the same text before each recording allowed speakers to familiarize themselves with the script and reduce their reading errors. In case of error (such as repetition, omission, or substitution), the recording was conducted again. To measure speech rate (i.e. average number of syllables uttered per text), pauses longer than 150ms were automatically detected by the Praat program and were discarded, after being manually verified by the author. For each language, the recordings whose value of speech rate is below or above 2.5 times standard deviation of each language were considered as outliers. The software R was used to conduct statistical computations [R Core Team, 2013] for detecting those outliers. In total, 24 among 2 438 recordings were detected as outliers after investigating the distribution of speech rate in each language and were filtered out.[8]

2.2.1.2 Text corpus and preprocessing

Text corpora in 18 languages were acquired from various sources as illustrated in Table 2.1. Most of the data were retrieved online except for those in Vietnamese and Wolof collected respectively by Le and his colleagues at the laboratory of IMAG and by Stéphane Robert at the laboratory of LLACAN.

Text corpora were phonetically transcribed into IPA or different phonetic codes from orthographic word-forms except for Wolof.[9] The data were syllabified automatically by a rule-based program written in bash shell script except for the following cases: i) the syllabification was already provided in the data: English, French, and German, ii) the

[7] This procedure was designed to measure silent reading rate to study a cross-language relationship between oral and silent reading rates [Coupé, Oh, Pellegrino, & Marsico, 2014].

[8] Following numbers of recordings were removed in each language: Cantonese (3), Catalan (1), Finnish (6), French (2), Korean (3), Spanish (2), Turkish (3), Vietnamese (3), Wolof (1).

[9] The Wolof data was not transcribed into IPA due to the inconsistency of its writing system and the lack of information for phonetic transcription.

corpus was syllabified by an automatic grapheme to phoneme converter: Catalan, Spanish, and Thai, iii) For Sino-Tibetan languages such as Cantonese and Mandarin Chinese, no syllabification rule was required since one ideogram corresponds to one syllable. For Vietnamese, most of words are monosyllabic due to its exclusively isolating tendency and the syllabification for some non-monosyllabic words is provided in the data. Thus, no syllabification rule was written in particular for Vietnamese.[10]

Table 2.1: Description of text corpus. For each language, a corresponding language code, the reference and the size of corpus (#Types and #Tokens) are provided.

Language	**ISO 639-3 code**	**Corpus**	**#Types**	**#Tokens**
Basque	EUS	E-Hitz [Perea et al., 2006]	100k	4M
British English	ENG	WebCelex (MPI for Psycholinguistics)	160k	17M
Cantonese	YUE	A linguistic corpus of mid-20th century Hong Kong Cantonese	6k	0.13M
Catalan	CAT	Frequency dictionary [Zséder et al., 2012]	63k	442M
Finnish	FIN	Finnish Parole Corpus	125k	15M
French	FRA	Lexique 3.80 [New et al., 2001]	142k	15M
German	DEU	WebCelex (MPI for Psycholinguistics)	84k	5M
Hungarian	HUN	Hungarian National Corpus [Váradi, 2002]	54k	170M
Italian	ITA	The Corpus PAISÀ [Lyding et al., 2014]	16k	181M
Japanese	JPN	Japanese Internet Corpus [Sharoff, 2006]	42k	175M
Korean	KOR	Leipzig Corpora Collection (LCC)	100k	2M
Mandarin Chinese	CMN	Chinese Internet Corpus [Sharoff, 2006]	47k	213M

[10]Automatic syllabification programs (written bash shell script) for syllabifying 8 languages (Basque, Finnish, Hungarian, Italian, Japanese, Korean, Serbian, and Turkish) in this study will be made available online through github for public use.

Table 2.1: Description of text corpus. For each language, a corresponding language code, the reference and the size of corpus (#Types and #Tokens) are provided (continued).

Language	**ISO 639-3 code**	**Corpus**	**#Types**	**#Tokens**
Serbian	SRP	Frequency dictionary [Zséder et al., 2012]	20k	492M
Spanish	SPA	Frequency dictionary [Zséder et al., 2012]	53k	963M
Thai	THA	Thai National Corpus (TNC)	5k	23M
Turkish	TUR	Leipzig Corpora Collection (LCC)	20k	0.96M
Vietnamese	VIE	VNSpeechCorpus [Le et al., 2004]	33k	22M
Wolof	WOL	Corpus collected by Stéphane Robert	3k	0.07M

Regarding the preprocessing, in a first phase, each corpus was cleaned by removing the word-forms with non-alphabetic characters. Most of the text corpora consist of a word frequency list derived from large-scale corpora, except Vietnamese and Wolof for which a raw text data was provided. Some further preprocessing depends on the nature of corpus. Loanwords (e.g. English and Arabic) were discarded as much as possible.

• Basque: The corpus E-Hitz was retrieved online and was provided with transcription and syllabification [Perea et al., 2006]. Since the lexical stresses were not marked in the transcription, for consistency with the other data, the corpus was phonetically transcribed by the speech synthesizer Espeak and was syllabified automatically by a bash shell script.

• British English: The WebCelex corpus was used for English [MPI for Psycholinguistics, 2013], which included syllabification, transcription, and stress assignment.

• Cantonese: Cantonese text corpus was preprocessed by Christophe Coupé at the laboratory of DDL. The Linguistic corpus of mid-20$^{\text{th}}$ century Hong Kong Cantonese [Research Centre on Linguistics and Language Information Sciences, 2013] was downloaded online. To obtain the jyutping transcription, two dictionaries (CantoDict [Sheik, 2013] and JyutDict [Learner, 2013]) were used. If the transcriptions provided by the two dictionaries were divergent, more traditional pronunciation was kept. The word-forms without

corresponding transcription were removed with the help from Prof. Feng Wang at Peking University.

• Catalan: A frequency dictionary derived from large-scale web corpora was retrieved online [Zséder et al., 2012]. For transcription and syllabification, an automatic grapheme-to-allophone converter Segre was used [Pachès et al., 2000].

• Finnish: The Finnish Parole Corpus [Institute for the Languages of Finland, 1996-1998] was retrieved online and was converted into IPA by Espeak. The data was syllabified by a bash shell script. The result of syllabification was verified by Hannu Laaksonen at the laboratory of DDL.

• French: The Lexique 3.80 [New et al., 2001] was used for French. Similarly to the WebCelex corpus, the Lexique 3.80 provides phonetic transcription and syllabification.[11]

• German: The WebCelex corpus was acquired online [MPI for Psycholinguistics, 2013] providing syllabification, transcription, and stress assignment. Several errors in phonetic transcription were corrected with assistance from the colleagues of the Phonetics and Phonology group at Saarland University.

• Hungarian: The Hungarian National Corpus [Váradi, 2002] was obtained online and was transcribed into IPA by Espeak. The transcribed data was automatically syllabified by a bash shell script.

• Italian: The Corpus PAISÀ [Lyding et al., 2014] was downloaded online and was transcribed into IPA based on the dictionary of Italian pronunciation [Canepari, 2009]. For the word-forms without corresponding transcription in the dictionary, an automatic phonemic converter [Carnevali, 2009] was used. The result of automatic transcription was corrected in order to maintain consistency with the transcription rules described in the dictionary of Italian pronunciation. The data was automatically syllabified by a bash shell script.

• Japanese: The Japanese Internet Corpus [Sharoff, 2006] was retrieved online, which

[11] A tendency towards neutralization involving two vowels /e/ and /ɛ/ in some variants of French [Gess, Lyche, & Meisenburg, 2012] was not considered in the transcription.

was already lemmatized. It was then converted into Katakana by an online Kanji converter[12] and was transcribed again into IPA by means of a list of phonemic entities corresponding to morae provided by the National Institute for Japanese Language and Linguistics (NINJAL). The transcribed data was syllabified by a bash shell script.

• Korean: The corpus was downloaded online from the Leipzig Corpus Collection [Biemann et al., 2007] and was converted into romanization using google translate[13]. Based on the Korean pronunciation dictionary [Kim et al., 1993], the romanization was transcribed into IPA and was automatically syllabified by a bash shell script.

• Mandarin Chinese: Mandarin text data was preprocessed by Christophe Coupé. The Chinese Internet Corpus was obtained online [Sharoff, 2006]. To get the pinyin transcription, a dictionary [CC-CEDICT, 2012] was used and when there was no corresponding transcription in the dictionary, the software NJStar Chinese Word Processor [NJStar Software Corp, 2013] was used to obtain the transcription.

• Serbian: A frequency dictionary [Zséder et al., 2012] acquired from large web corpora was converted into IPA by Espeak and was automatically syllabified by a bash shell script.

• Spanish: A frequency dictionary derived from large-scale web corpora [Zséder et al., 2012] was downloaded online. It was transcribed and syllabified by an automatic tool of transcription and syllabification written in perl [López, 2004].

• Thai: A list of the 5 000 most frequent words derived from the Thai National Corpus [Aroonmanakun, Tansiri, & Nittayanuparp, 2009] was downloaded online. The data was automatically transcribed into IPA and syllabified by an online tool[14].

• Turkish: The corpus was retrieved online from the Leipzig Corpora Collection [Biemann et al., 2007]. It was transcribed by Espeak and syllabified automatically by a bash shall script.

• Vietnamese: VNSpeechCorpus was collected by Le and his colleagues at the labora-

[12]http://nihongo.j-talk.com
[13]https://translate.google.com
[14]http://www.thai-language.com

tory of IMAG [Le et al., 2004]. The data was automatically transcribed by a phonetizer vPhon [Kirby, 2008]. Many foreign words in the initial data which do not follow the phonotactics of Vietnamese were automatically detected by vPhon and were discarded.

• Wolof: A small-scale corpus gathered by Stéphane Robert at the laboratory of LLACAN was used. The data was not transcribed into IPA due to the inconsistency of its writing system and the lack of information for phonetic transcription. Graphemic wordforms were automatically syllabified by a bash shell script.

2.2.2 Parameters

2.2.2.1 SR, ID, and IR

Three parameters, i.e. speech rate, information density, and information rate, were proposed in [Pellegrino, Coupé, & Marsico, 2011] with an objective to assess the complexity trade-off between syllabic speech rate and information density. They are measured by using a multilingual oral corpus in 18 languages recorded with the oral script consisting of the 15 short semantically equivalent texts among the 18 languages (cf. Section 2.2.1.1). To begin with, speech rate (*SR*, hereafter) denotes the average number of syllables (σ) pronounced per second where D_L^t is the duration of text t uttered in language L.[15]

$$SR_L = \frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_L^t}{D_L^t} \tag{2.1}$$

We chose syllable as the basic unit of analysis, following many studies ([Aylett & Turk, 2004] [Cholin, Levelt, &Schiller, 2006] [Davis & Zajdo, 2010] [Fenk, Fenk-Oczlon, & Fenk, 2006] [Pellegrino, Coupé, & Marsico, 2011] [Shosted, 2006], inter alia). In comparison with segment, syllable is considered more robust in terms of the reduction of utterance [Greenberg, 1999] [Johnson, 2004] and less ambiguous for counting [Pellegrino, Coupé, &

[15]Pauses longer than 150ms were discarded using Praat. The result of pause detection was manually checked. The syllabic rate considered in this study is hence an articulatory speech rate.

Marsico, 2011]. In *A Course in Phonetics*, syllable is defined as "necessary units in the mental organization and production of utterances" [Ladefoged & Johnson, 2014]. Following this perspective, syllable is used as a basic unit of analysis in the present study.

In order to account for the two parameters, Information density (ID) and Information rate (IR), the average amount of information conveyed per syllable (I_L^t) is defined as the division of the semantic content of text t in language L (S_L^t) by the number of its constituents, i.e. syllables (σ_L^t).

$$I_L^t = \frac{S_L^t}{\sigma_L^t} \tag{2.2}$$

Since the estimates of the amount of semantic content is beyond the scope of this study, ID is measured by a paired comparison using Vietnamese (VIE) as an external reference. Following [Pellegrino, Coupé, & Marsico, 2011], Vietnamese was chosen as a normalizing factor since it is the most isolating language among the 18 languages in the data.

$$ID_L = \frac{1}{T}\sum_{t=1}^{T}\frac{I_L^t}{I_{VIE}^t} = \frac{1}{T}\sum_{t=1}^{T}\frac{S_L^t}{\sigma_L^t} \times \frac{\sigma_{VIE}^t}{S_{VIE}^t} = \frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_{VIE}^t}{\sigma_L^t} \tag{2.3}$$

As a multilingual parallel oral corpus is used in this study, the semantic content of each text is assumed to be equivalent for all languages ($S_L^t = S_{VIE}^t$). Consequently, information density (ID) is computed by a pairwise comparison of the number of syllables of text t in Vietnamese (σ_{VIE}^t) and in a target language (σ_L^t).

Information rate (IR) refers to the average amount of information transmitted per second. IR of an individual speaker of language L (IR_{Spker_L}) is obtained by dividing the semantic information of text t (S_L^t) by the duration of the text t uttered by each native speaker of a target language L ($D^t(Spker_L)$).

$$IR_{Spker_L} = \frac{1}{T}\sum_{t=1}^{T}\frac{S_L^t}{D^t(Spker_L)} \times \frac{D_{VIE}^t}{S_{VIE}^t} = \frac{1}{T}\sum_{t=1}^{T}\frac{D_{VIE}^t}{D^t(Spker_L)} \tag{2.4}$$

Similarly to the equation of *ID*, Vietnamese is used as a reference for the normalization and the semantic information (S_L^t) is considered identical in computing *IR*. Thus, Eq. 2.4 is reduced to a paired comparison between the mean duration of text t uttered by all the speakers in Vietnamese (D_{VIE}^t) and the duration for text t uttered by a native speaker in a target language ($D^t(Spker_L)$). The mean duration for each text was used in Vietnamese since there is no reason to match each speaker of a target language to a specific speaker of Vietnamese.

$$IR_L = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{Spker=1}^{N}\frac{D_{VIE}^t}{D^t(Spker_L)}\right) \tag{2.5}$$

Consequently, the mean *IR* of language L is obtained by averaging the *IR* of each recording in language L as shown in Eq. 2.5 where language L has N native speakers.

2.2.2.2 Syllable complexity

The most common measure of linguistic complexity is to count the number of constituents of the linguistic item under study. Menzerath used this measure of linguistic complexity to investigate a trade-off phenomenon between the size of unit (i.e. word or syllable) and the number of its constituents (i.e. syllables or phonemes) in phonology: "The more sounds in a syllable the smaller their relative length" [Altmann, 1980]. Fenk later used the term *word complexity* and *syllable complexity* in [Fenk-Oczlon & Fenk, 2005] to refer to this method of quantification. Measuring the "richness" of system in terms of the number of its components is related to *system complexity* and phonology [Dahl, 2004].

Syllable complexity is thus computed as the average number of segments per syllable and is used as a traditional measure of linguistic complexity [Fenk-Oczlon & Fenk, 2005] [Maddieson, 2006] [Pellegrino, Coupé, & Marsico, 2011]. In the present study, two measures of syllable complexity, i.e. SC_{TYPE} and SC_{TOKEN}, used in [Pellegrino, Coupé, & Marsico, 2011] are employed with multilingual text corpora in the 18 languages.[16]

[16]Both SC_{TYPE} and SC_{TOKEN} are computed on the 20 000 most frequent words in each language.

SC_{TYPE} refers to a traditional linguistic measure of complexity quantified as the average number of segments (and tones, if applicable) (φ_i) per syllable where language L is considered as a system consisting of a finite set of N syllables from an information-theoretic perspective [Hockett, 1966].

$$SC_{TYPE} = \frac{1}{N_L} \sum_{i=1}^{N_L} \varphi_i \tag{2.6}$$

$$SC_{TOKEN} = \frac{1}{N_L} \sum_{i=1}^{N_L} p_i \cdot \varphi_i \tag{2.7}$$

In contrast to the measure SC_{TYPE}, SC_{TOKEN} is computed from an usage-based approach where each average number of segments and tones (if applicable) per syllable is weighted by the relative frequency of corresponding syllable (p_i) in a large text corpus. As described in [Pellegrino, Coupé, & Marsico, 2011], SC_{TYPE} has been used as a traditional measure of phonological complexity in typological linguistics and psycholinguistics [Maddieson, 2006] [Mueller et al., 2003]. Since it does not take account of the frequency of syllables, the distinction between *TYPE* and *TOKEN* was made in order to assess the impact of the actual usage of syllables. SC_{TOKEN} is considered as a more robust measure since it combines both grammatical and functional approaches [Pellegrino, Coupé, & Marsico, 2011].

2.2.2.3 Information-theoretic measures

The amount of information can be quantified by means of the following information-theoretic measures: Shannon entropy $H(X)$, conditional entropy $H(X|C)$, and surprisal $S(X)$ [Hale, 2001] [Shannon, 1948].

The notion of Shannon entropy has been suggested and used as a quantitative measure of complexity in linguistics ([Ferrer i Cancho & Solé, 2003] [Ferrer i Cancho, 2006] [Ferrer i Cancho & Díaz-Guilera, 2007] [Goldsmith, 2000] [Goldsmith, 2002] [Kello & Beltz, 2009] [Pellegrino, Coupé, & Marsico, 2007], inter alia). The following definitions of Shan-

non entropy were proposed:

i) Measure of the unpredictability of a set of linguistic components: Shannon entropy is considered as "a measure of *complexity* of an analysis" [Goldsmith, 2000].

ii) Measure of the cognitive cost of language use: in particular, it is assumed that "conditional entropy is an effort for the hearer (i.e. disambiguation) and Shannon entropy is an effort for both the speaker (i.e. memory effort) and the hearer (i.e. recognition)" [Ferrer i Cancho & Solé, 2003] [Ferrer i Cancho, 2006] [Ferrer i Cancho & Díaz-Guilera, 2007].

iii) Efficiency of lexicon: this interpretation is in line with the definition by Ferrer i Cancho in (ii) [Kello & Beltz, 2009].

$$H_L = -\sum_{i=1}^{N_L} p_{\sigma_i} \cdot log_2(p_{\sigma_i}) \tag{2.8}$$

Shannon entropy is computed by the equation 2.8 where Language L refers to a finite set composed of N number of syllables (σ) and p_{σ_i} denotes the approximated relative frequency of i^{th} syllable (p_{σ_i}) from a large text corpus which was phonologically transcribed. Shannon entropy would reach its maximum value if each syllable in language L were evenly distributed, i.e. p_{σ_i} were all equal. On the contrary, if p_{σ_i} equaled 1 for one specific syllable, there would be no uncertainty and H_L would become 0. The difficulty of estimating entropy without statistical bias was described in [Paninski, 2003]. For example, since the distribution of syllables is estimated from a large text corpus, the size of corpus plays an essential role in estimating the distribution of syllables. It is assumed that the larger the size of corpus, the more realistic and accurate the approximation of the syllable distribution. However, increasing the size of corpus does not necessarily lead to a better estimation of distribution, as it was demonstrated by the non-convergent behavior of numbers of frequent words estimated from a corpus containing one billion words [Curran & Osborne, 2002]. The effect of corpus size will be tested in Section 2.3.2.

Conditional entropy is defined as *the average amount of uncertainty when contextual information C is known* and is commonly calculated from n-gram language models ob-

tained from a large phonologically transcribed text corpus. Here, language L is considered as a source of words consisting of sequences of syllables x_i drawn from a pool of N_L possible syllables, and not as a mere source of independent syllables.

Conditional entropy is formally defined as:

$$\begin{aligned} H(X|C) &= \sum_{c \in C} p(c) \cdot H(X|C=c) \\ &= -\sum_{c \in C} p(c) \cdot \sum_{i=1}^{N_L} p(X=x_i|C=c) log_2(p(X=x_i|C=c)) \end{aligned} \tag{2.9}$$

where X and C are two random variables respectively corresponding to the syllables and their context. $p(c)$ is the probability of a given context c among the space of possible values taken by C. We propose two ways of measuring conditional entropy. Both are based on a bigram model[17] and the context is defined either as the *preceding* or the $following$ syllable in the sequence. They are respectively noted $H(X_n|X_{n-1})$ and $H(X_n|X_{n+1})$.

The first equation for quantifying $H(X_n|X_{n-1})$ takes the preceding contextual information into account and it is commonly used in psycholinguistics [Piantadosi, Tily, & Gibson, 2011]. However, other studies suggested significant effects of the following context and not the preceding one, at least at the word level [Bell et al., 2009] [Gahl, 2008] [Seyfarth, 2014]. For this reason, taking the following context into account is also proposed in the present study.

In order to take contextual information into account even for monosyllabic word-forms (for which no preceding or following syllable is identified in the lexicon), the random variable C takes its values from the set of N_L possible syllables extended with an asterisk * (resp. a hash #) marking the beginning (resp. the end) of a word-form for computing $H(X_n|X_{n-1})$ (resp. $H(X_n|X_{n+1})$). This process is illustrated with a fictive language example below.

The following example with a toy language is provided in order to illustrate the calcu-

[17]The size of n-gram model is limited to the bigram in this study. However, theoretically it could be extended to larger n-grams if the corpus size is large enough to accurately estimate their probabilities.

lation process of Shannon entropy $H(X)$ and conditional entropy $H(X_n|X_{n-1})$ in detail. This fictive language consists of 5 words presented in Table 2.2 with the corresponding frequencies.

Table 2.2: List of syllabified word-forms and their frequency

Syllabified word-form	**Frequency**	**Probability**
a	20	0.50
a_ba	10	0.25
ba_ka_da	5	0.13
a_ba_fa_ba	3	0.08
ka_a_ba_ga_ha_fa	2	0.05
sum	40	1

We cannot assume any general tendency from this example but comparing Tables 2.2 and 2.3 allows us to observe that the probability distribution of words differ from the probability distribution of syllables. In Table 2.3, the total sum of $-p_{\sigma_i} \cdot \log_2(p_{\sigma_i})$ (2.12) corresponds to the value of Shannon entropy, i.e. the average amount of information of the toy language considered as a source of independent syllables.

Table 2.3: Calculation of Shannon entropy ($N_L = 7$ distinct syllables)

Syllable σ_i	**Frequency**	**Probability**	$-p_{\sigma_i} \cdot \mathbf{\log_2(p_{\sigma_i})}$
a	20+10+3+2 = 35	0.44	0.52
ba	10+5+3+3+2 = 23	0.29	0.52
ka	5+2 = 7	0.09	0.31
da	5	0.06	0.25
ga	2	0.03	0.13
fa	3+2 = 5	0.06	0.25
ha	2	0.03	0.13
sum	79	1.00	H(X) = 2.12

To compute the conditional entropy $H(X_n|X_{n-1})$, bigrams of syllables-in-context are listed. For each syllable $x_i \in \{x_1, ..., x_{N_L}\}$, the list of preceding contexts c in which they are observed are determined, along with their frequencies in the corpus.

Table 2.4: Conditional entropy: list of bigrams

Bigram c_x_i	**Cumulated frequency of** c	**Cumulated frequency of** x_i **given** $C{=}c$	$\mathbf{p}(X{=}x_i \mid C{=}c)$	$-\mathbf{p}(X{=}x_i \mid C{=}c) \cdot \log_2(\mathbf{p}(X{=}x_i \mid C{=}c))$
*_a	20+10+5+3+2 = 40	20+10+3 = 33	33/40 = 0.83	0.23
a_ba	15	15	1	0
*_ba	40	5	0.13	0.38
ba_ka	10	5	0.50	0.50
ka_da	7	5	0.71	0.35
ba_fa	10	3	0.30	0.52
fa_ba	3	3	1	0
*_ka	40	2	0.05	0.22
ka_a	7	2	0.29	0.52
ba_ga	10	2	0.20	0.46
ga_ha	2	2	1	0
ha_fa	2	2	1	0

This table shows, for instance, that in this fictive corpus, the syllable /fa/ is always followed by the syllable /ba/. Hence, the appearance of /ba/ in the preceding context of /fa/ does not carry any information. On the contrary, the syllable /ba/ can be followed by /ka/, /fa/, and /ga/, and each of these syllables carries a significant amount of information in this context.

Table 2.5: Calculation of conditional entropy $H(X_n|X_{n-1})$

Context c	**Cumulated frequency of** c	$\mathbf{p}(c)$	$-\mathbf{p}(c) \cdot \sum_{i=1}^{N_L} \mathbf{p}(X{=}x_i \mid C{=}c) \cdot \log_2(\mathbf{p}(X{=}x_i \mid C{=}c))$
*	40	40/79 = 0.51	0.51×(0.23+0.38+0.22) = 0.42
a	15	0.19	0.19×0 = 0
ba	10	0.13	0.13×(0.50+0.53+0.46) = 0.19
ka	7	0.09	0.09×(0.35+0.52) = 0.08
da	0	0	0
fa	3	0.04	0.04×0 = 0
ga	2	0.03	0.03×0 = 0
ha	2	0.03	0.03×0 = 0
sum	79	1	$H(X_n\|X_{n-1}) = 0.68$

Finally, the weighted entropy of the syllable distribution in each context c is computed in Table 2.5 (last column). This sum over the possible contexts leads to a value of the conditional entropy $H(X_n|X_{n-1})$ of 0.68. While the Shannon entropy computed in Table 2.3

yielded an average value of 2.12 bits of information per syllable, the conditional entropy shows that, when the information carried by the previous context is taken into account, the average information drops to 0.68 bits per syllable.

Piantadosi and his colleagues argued that ambiguity is a "functional property of language that allows for greater communicative efficiency" by proposing that words with more ambiguity (i.e. homophones with more meanings) tend to be short, simple and highly predictable [Piantadosi, Tily, & Gibson, 2012]. A high conditional entropy illustrates a low predictability from the context and hence an important effort for the hearer while Shannon entropy is considered as an effort for both the speaker and the hearer. In the same vein, Levinson suggested that the listener's effort of disambiguating the meanings of word based on contextual information is less costly than the speaker's effort of processing [Levinson, 2000] [Piantadosi, Tily, & Gibson, 2012].

Surprisal $S(X)$ is a measure of the amount of information content per individual linguistic component. There are two ways of measuring surprisal: (i) similar to the calculation of Shannon entropy $H(X)$, it can be measured without considering contextual information as shown in Eq. 2.10, (ii) furthermore, contextual information can be taken into account by means of conditional probability for computing surprisal as shown in Eq. 2.11, using a bigram language model. In Eq. 2.11, X_{n-1} refers to previous context and X_{n+1} corresponds to following context.

$$S(X) = -log_2 P(X) \tag{2.10}$$

$$S(X_n) = -log_2 P(X_n|X_{n-1}) \qquad S(X_n) = -log_2 P(X_n|X_{n+1}) \tag{2.11}$$

In recent studies, the word-level surprisal has been frequently employed in psycholinguistics as a measure for estimating the speaker's "difficulty" or "cognitive effort" of information processing. Since surprisal is inversely related to the probability (or conditional probability) of linguistic components, high-surprisal words are assumed to be longer (i.e.

more difficult to produce) than low-surprisal words, and this tendency was illustrated in [Demberg et al., 2012]. Moreover, two notions of surprisal, i.e. *syntactic surprisal* and *lexicalized surprisal* were further proposed by Demberg and Keller [Demberg & Keller, 2008] [Demberg et al., 2012]. The former was suggested as a measure of syntactic complexity which is quantified as the portion of structural information estimated using an elaborated language model such as probabilistic context-free grammar (PCFG). PCFG computes the probability of grammatical rules obtained from a syntactic tree, ignoring the effect of word frequency. On the contrary, *lexicalized surprisal* combines both the syntactic structural information and the lexical effect of word frequency. As a result, it was observed that syntactic surprisal better predicts word duration than surprisal obtained from simple trigram probabilities. However, a distinction between lexicalized and syntactic surprisal does not seem to be applicable in the present study where the syllable-level surprisal is employed.

2.2.3 Language description

In order to crosslinguistically compare the 18 languages, some information are provided in Table 2.6. First, the typological diversity of the 18 languages analyzed in this study is illustrated by displaying the language family and genus of each language.[18] In total, there are 10 different language families, including 3 languages (Basque, Japanese, and Korean) often considered language isolates[19]: Basque, Indo-European (Germanic, Romance, and Slavic), Sino-Tibetan, Uralic (Finnic and Ugric), Japanese, Korean, Tai-Kadai, Altaic, Austro-Asiatic, and Niger-Congo. In addition, at the genus level, there are 13 distinct types of languages.

[18]The revelant information was acquired from WALS [Dryer & Haspelmath, 2013].

[19]According to Glottolog [Hammarström et al., 2015], Japanese and Korean are also considered as a part of Japonic and Koreanic language families respectively.

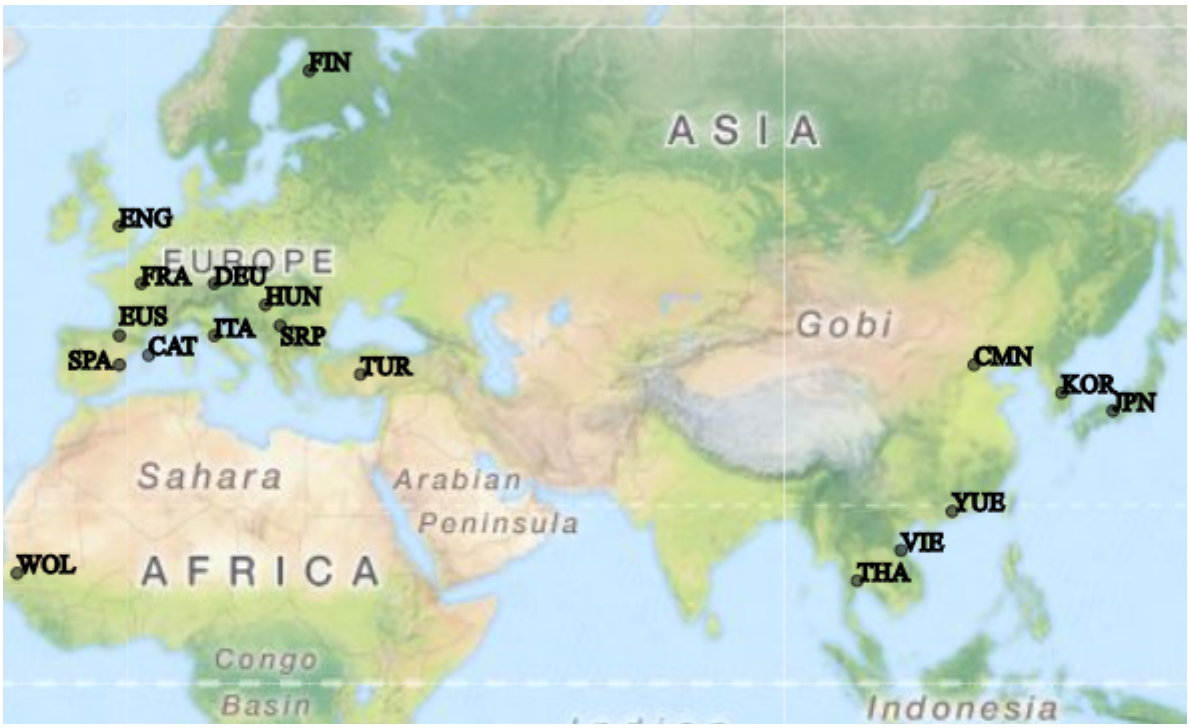

Figure 2.2: Geographic location of the 18 languages studied

Table 2.6: Language description. The phonological system of each language is illustrated. In case of English and German, diphthongs are included in the vowel inventory as they were coded separately from vowels in the WebCelex corpus. The size of syllable inventory is calculated from the 20 000 most frequent words.

Language	**Family/ Genus**	**Phonological system**		**Inventory size** (# Syllables)	**Syllable structure/** LAPSyD index	**Vowel harmony**
Basque	Basque/ Basque	C V S	29 7 2	2 082	complex/ 4	×
British English	Indo-European/ Germanic	C V S	30 24 2	6 949	complex/ 8	×
Cantonese	Sino-Tibetan/ Chinese	C V T	19 13 6	1 298	moderately complex/ 3	×
Catalan	Indo-European/ Romance	C V S	25 8 2	3 600	moderately complex/ 4	×
Finnish	Uralic/ Finnic	C V S	19 19 2	3 844	moderately complex/ 3	◯

Table 2.6: Language description. The phonological system of each language is illustrated. In case of English and German, diphthongs are included in the vowel inventory as they were coded separately from vowels in the WebCelex corpus. The size of syllable inventory is calculated from the 20 000 most frequent words (continued).

Language	**Family/ Genus**	**Phonological system**		**Inventory size** (# Syllables)	**Syllable structure/** LAPSyD index	**Vowel harmony**
French	Indo-European/ Romance	C V	22 15	2 949	complex/ 7	×
German	Indo-European/ Germanic	C V S	27 32 1	5 100	complex/ 8	×
Hungarian	Uralic/ Ugric	C V S	24 15 2	4 325	complex/ 8	◯
Italian	Indo-European/ Romance	C V S	27 7 1	2 729	complex/ 6	×
Japanese	Japanese/ Japanese	C V	17 10	643	moderately complex/ 4	×
Korean	Korean/ Korean	C V	22 8	1 104	moderately complex/ 3	◯
Mandarin Chinese	Sino-Tibetan/ Chinese	C V T	25 7 5	1 274	moderately complex/ 4	×
Serbian	Indo-European/ Slavic	C V S	26 11 2	3 831	complex/ 8	×
Spanish	Indo-European/ Romance	C V S	27 5 1	2 778	moderately complex/ 5	×
Thai	Tai-Kadai/ Kam-Tai	C V T	21 18 5	2 438	moderately complex/ 4	×
Turkish	Altaic/ Turkic	C V S	27 19 2	3 260	moderately complex/ 3	◯
Vietnamese	Austro-Asiatic/ Viet-Muong	C V T	23 12 6	5 156	moderately complex/ 4	×
Wolof	Niger-Congo/ Northern Atlantic	C V	24 15	2 776	complex/ 4	◯

Second, the phonological system, i.e. the number of consonants, vowels, stress, and tones (if applicable), is described as well as the size of syllable inventory (i.e. number of distant syllables).[20] Similarly to the study of Ian Maddieson [Maddieson, 2006] presenting an overall positive correlation between syllable complexity and the size of consonant inventory among a large number of languages, it is observed that the size of phonological system (number of vowels and consonants) is positively correlated with the size of syllable inventory among the 18 languages (Pearson's $r = 0.748$**; p-value < 0.001; Spearman's $\rho = 0.688$**; p-value $= 0.002$; N $= 18$).

Third, the degree of complexity of syllable structure in WALS [Dryer & Haspelmath, 2013] and syllabic index in LAPSyD [Maddieson et al., 2013] based on the classification method of Ian Maddieson [Maddieson, 2013] are displayed.[21] In WALS, 486 languages were classified into 3 types in terms of the language's maximal syllable structure:

(i) simple syllable structure: (C)V

(ii) moderately complex syllable structure: (C)(C)V(C)

(iii) complex syllable structure: (C)(C)(C)V(C)(C)(C)(C)

For example, the maximal syllable structure of English is represented as (C)(C)(C)V(C)(C)(C)(C). The sum of the maximum number of consonants in onset and coda, and the number of vowel per syllable, i.e. $3 + 4 + 1 = 8$, corresponds to the value of syllabic index provided in LAPSyD. Thus, languages with moderately complex syllable structure are assumed to exhibit a syllabic index ranging from 3 to 4. However, some inconsistencies are found between the two measures of syllable complexity in the cases of Basque, Spanish, and Wolof: Basque and Wolof should be considered as moderately complex and Spanish

[20]The information regarding the phonological system was obtained from transcribed corpora which inevitably reflect the use of loanwords (such as Arabic, English, French, and Spanish), except for Wolof. In case of Wolof, the size of phonological system was based on the information provided in LAPSyD [Maddieson et al., 2013]. Thus, the size of phonological system may diverge from the traditional description for 17 languages.

[21]No corresponding information was found for Italian and Serbian in WALS, and for Serbian in LAPSyD, which was completed by the author following the method of Ian Maddieson. For example, the syllable index of 6 was given for Italian in LAPSyD. Since the languages displaying a syllable index ranging from 5 to 8 are considered as those with complex syllable structure in WALS, Italian accordingly belongs to the category of complex languages.

as complex according to the criteria provided in [Maddieson, 2013].

Finally, the presence of vowel harmony is examined in each language since it is assumed to be significantly related to the reduction of information caused by the knowledge of context. Among the 18 languages, there are 5 languages including 2 Uralic languages (Finnish, Hungarian, Korean, Turkish, and Wolof), which exhibit clear evidence of vowel harmony.[22]

2.3 Cross-language comparisons of the average information rate

This section presents the results of the assessment of the three following aspects: (i) Subsection 2.3.1 analyzes the results of the extension of the previous study on speech information rate [Pellegrino, Coupé, & Marsico, 2011], (ii) Subsection 2.3.2 covers some methodological aspects concerning the computation of Shannon entropy and conditional entropy, and (iii) Subsections 2.3.3, 2.3.4, and 2.3.5 are devoted to the assessment of the entropy-based estimation of information rate, by observing whether there is a faithful relationship between *IR* in the sense of the Information theory and *IR* computed by using a pairwise comparison with Vietnamese as a reference.

2.3.1 Speech rate, information density, and information rate

This subsection provides the results of investigating the *SR*, *ID*, and *IR* of the 18 languages obtained by adding 11 more languages to the previous study with the 7 languages [Pellegrino, Coupé, & Marsico, 2011].[23] The results obtained with the 18 languages

[22] 3 Romance languages (Ascrea Italian, Valencian Catalan, and Eastern Andalusian Spanish) display some examples of vowel harmony in their dialects [Lloret, 2007].

[23] It should be noted that the initial oral scripts in French, Japanese, and Mandarin Chinese were modified and the recordings were done again based on the new version. Furthermore, 5 texts were discarded among the 20 texts initially chosen due to the unnaturalness of the oral scripts which degraded the fluency of speakers.

are shown in Table 2.7.

Table 2.7: Speech rate (SR), information density (ID), information rate (IR), syllable complexity (SC), & difference between SC_{TYPE} and SC_{TOKEN} (ΔSC). The maximum and minimum values are marked in green and blue, respectively.

Language	SR	ID	IR	SC_{TYPE}	SC_{TOKEN}	ΔSC
CAT	7.07	0.63	0.85	3.20	2.25	0.96
CMN	5.86	1.03	1.15	3.97	3.69	0.28
DEU	6.09	0.77	0.90	3.38	2.59	0.79
ENG	6.34	0.90	1.08	3.46	2.50	0.96
EUS	7.54	0.65	0.95	2.92	2.06	0.85
FIN	7.22	0.71	0.97	3.33	2.46	0.88
FRA	6.85	0.80	1.05	3.29	2.14	1.15
HUN	5.87	0.71	0.80	3.07	2.33	0.75
ITA	7.16	0.72	0.99	3.09	2.23	0.87
JPN	8.03	0.53	0.82	2.83	2.04	0.79
KOR	6.93	0.63	0.84	2.95	2.39	0.56
SPA	7.71	0.63	0.95	3.11	2.29	0.82
SRP	7.15	0.68	0.94	3.36	2.26	1.10
THA	4.70	0.90	0.81	4.02	3.85	0.18
TUR	7.00	0.65	0.87	3.00	2.35	0.65
VIE	5.25	1.00	1.00	3.99	3.89	0.10
WOL	5.02	0.85	0.83	2.99	2.39	0.60
YUE	5.57	0.91	0.98	4.00	3.70	0.30

Among the 18 languages, the language with the highest SR, Japanese (8.03) is 1.7 times faster than Thai (4.70), the language with the lowest SR. In terms of ID, Mandarin Chinese displays the highest value (1.03) and is 1.9 times denser than Japanese (0.53), the language with the lowest ID. ID does not differ among several languages while their SR and IR vary: (i) Catalan, Korean, and Spanish, (ii) Basque and Turkish, (iii) Finnish and Hungarian, (iv) English and Thai. Hence, it can be suggested that there may exist linguistic factors, such as syllable complexity (SC) and the size of syllable inventory, as well as external factors (i.e. sociocultural and cognitive constraints) which come into play in shaping the encoding strategy of information.

In comparison with SR and ID, a relatively low range of variation exists between the maximum IR (1.15, CMN) and the minimum IR (0.80, HUN) where the IR of Mandarin

Chinese is 1.4 times faster than the *IR* of Hungarian. Furthermore, tonal languages (CMM, THA, VIE, and YUE) exhibit higher SC_{TYPE} and SC_{TOKEN} than non-tonal languages. By observing the values of those parameters, particularly in Japanese, some correlation patterns are expected, such as a negative relationship between *SR* and *ID*, *SR* and SC_{TYPE}, and *SR* and SC_{TOKEN}.

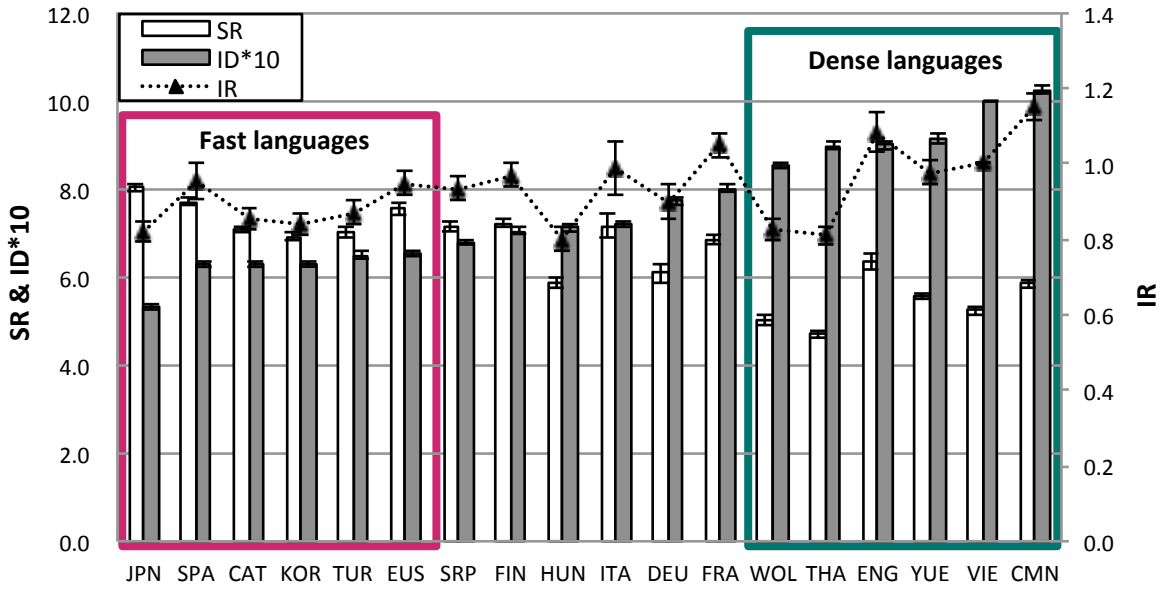

Figure 2.3: *SR* (average number of syllables uttered per second) and *ID* (average amount of information contained per syllable, unitless) multiplied by 10 on y-axis on the left and *IR* (average amount of information conveyed per second on y-axis on the right), 95% confidence intervals displayed. Languages are ordered by increasing *ID* from left to right.

The main finding in [Pellegrino, Coupé, & Marsico, 2011] consisted of a trade-off between *SR* and *ID*, which is confirmed by the result in this study since a negative correlation exists between *SR* and *ID* among the 18 languages (Pearson's $r = -0.812$**; p-value < 0.001; Spearman's $\rho = -0.798$**; p-value < 0.001; N $= 18$).[24][25] Furthermore, the 18 languages can be divided into three different types according to their encoding strategy in Figure 2.3: 6 "fast" languages on the left side of the figure exhibit a relatively

[24]In the previous result, a negative correlation between *SR* and *ID* was found among 7 languages (British English, French, German, Italian, Japanese, Mandarin Chinese, and Spanish) (Pearson's $r = -0.81$*; p-value $= 0.02$; Spearman's $\rho = -0.86$*; p-value $= 0.02$; N $= 7$).

[25]Both Spearman and Pearson correlations are given since *SR* follows a normal distribution while *ID* does not.

high *SR* and a low *ID* whereas 6 "dense" languages on the right side of the figure are characterized by a low *SR* and a high *ID*. In addition, there are 6 languages placed in the middle between the fast and dense languages.[26] Based on the language description in Table 2.6, the 6 fast languages refer to the languages with moderately complex syllable structure except for Basque whereas 5 languages in the middle and 2 dense languages display complex syllable structure except for Finnish and 4 tonal languages. Although this should be confirmed with languages with simple syllable structure lacking in this study, an overall tendency toward a negative relationship between the complexity of syllable structure and *SR* is found among the 18 languages. This tendency is further confirmed by significant correlations between *SR* and SC_{TOKEN} (Pearson's $r = -0.751$**; p-value < 0.001; Spearman's $\rho = -0.763$**; p-value < 0.001;N = 18) and *SR* and SC_{TYPE} (Pearson's $r = -0.674$**; p-value = 0.002; Spearman's $\rho = -0.579$*; p-value = 0.012; N = 18), where *SC* refers to the average number of segments (and tones if applicable) per syllable.

Regarding *ID*, *ID* and *IR* (Pearson's $r = 0.553$*; p-value = 0.017; Spearman's $\rho = 0.512$*; p-value = 0.030; N = 18), *ID* and SC_{TOKEN} (Pearson's $r = 0.820$**; p-value < 0.001; Spearman's $\rho = 0.725$**; p-value = 0.001; N = 18), *ID* and SC_{TYPE} (Pearson's $r = 0.848$**; p-value < 0.001; Spearman's $\rho = 0.759$**; p-value < 0.001; N = 18) are positively correlated. On the contrary, in comparison to *SR* and *ID*, *IR* exhibits a relatively few number of significant correlations and the magnitude of correlation coefficient is lower: no significant correlation is detected between *IR* and *SR* (Pearson's $r = 0.023$; p-value = 0.928; Spearman's $\rho = 0.014$; p-value = 0.955; N = 18), *IR* and SC_{TOKEN} (Pearson's $r = 0.267$; p-value = 0.284; Spearman's $\rho = 0.177$; p-value = 0.483; N = 18), and *IR* and SC_{TYPE} (Pearson's $r = 0.452$; p-value = 0.059; Spearman's $\rho = 0.447$; p-value = 0.063; N = 18) while *IR* and *ID* are significantly correlated.

[26]This arbitrary distinction is made in order to describe a general tendency appeared among the 18 languages.

Table 2.8: Mixed-effects model of *IR*. The effects of fixed factors and random factors are displayed on the left and right sides of the table respectively. (Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.')

Fixed factor				**Random factor**			
Predictor	**Coefficient**	**t-value**	**Sig**	**Predictor**	X^2**(df)**	p**-value**	**Sig**
Model: *IR* (dependent variable) ∼ ID * SR + Sex + Language + (1|Speaker) + (1|Text)							
Intercept	0.1110	4.043	**	Speaker	0 (1)	1	
SR	0.8100	748.650	***	Text	6309.4 (1)	< 0.001	***
ID	1.1690	836.913	***				
Sex_{Male}	0.0003	0.232					
$\text{Language}_{\text{CAT}}$	−0.0116	−2.692	**				
$\text{Language}_{\text{CMN}}$	0.0028	0.851					
$\text{Language}_{\text{DEU}}$	0.0013	0.298					
$\text{Language}_{\text{ENG}}$	0.0038	0.853					
$\text{Language}_{\text{EUS}}$	−0.0105	−2.382	*				
$\text{Language}_{\text{FIN}}$	−0.0029	−0.702					
$\text{Language}_{\text{FRA}}$	−0.0030	−0.780					
$\text{Language}_{\text{HUN}}$	0.0008	0.211					
$\text{Language}_{\text{ITA}}$	0.0046	0.908					
$\text{Language}_{\text{JPN}}$	−0.0062	−1.212					
$\text{Language}_{\text{KOR}}$	0.0027	0.623					
$\text{Language}_{\text{SPA}}$	−0.0183	−4.036	***				
$\text{Language}_{\text{SRP}}$	−0.0136	−3.260	**				
$\text{Language}_{\text{THA}}$	−0.0143	−4.274	***				
$\text{Language}_{\text{TUR}}$	0.0005	0.126					
$\text{Language}_{\text{WOL}}$	0.0097	2.856	**				
$\text{Language}_{\text{YUE}}$	0.0007	0.211					
SR:*ID*	0.2073	261.296	***				

A statistical model containing both fixed and random effects, i.e. a mixed-effects model, is used to further assess the relationships among *SR*, *ID*, *IR*, and other potential factors. The statistical package lme4 [Bates et al., 2015] was used to compute mixed-effect models with R and numeric variables were transformed into z-scores for the comparison of variables with different magnitudes. A model presented in Table 2.8 takes *IR* as a dependent variable while language and sex are treated as fixed variables and speaker and text are considered as random variables. Since *SR* and *ID*, *ID* and *IR* are significantly correlated, a test of collinearity was conducted by means of variance inflation factor (VIF) and the result exhibiting the values smaller than 5 revealed that there was no problem of collinear-

ity.

In the table, it is observed that text has a significant effect on *IR*. Regarding the effect of language treated as a fixed predictor, Vietnamese is taken as a baseline from which 6 languages significantly differ in terms of *IR*. In comparison with the average *IR* presented in Figure 2.3 where the average *IR* of Wolof (0.83 obtained by averaging 149 data points) is lower than Vietnamese (147 data points), considering individual data points reveals that the *IR* of Wolof is positively correlated with the *IR* of Vietnamese. As for the effect of *SR* and *ID*, they are significant predictors of *IR*, as well as their interaction (*SR*:*ID*) which presents a better model fit if it is included into the model. Since the effects of *ID*, *SR* and their interaction (*SR*:*ID*) turn out significant toward *IR*, the initial hypothesis that *IR* is assumed to be explained by the interaction between *SR* and *ID* is confirmed by this result.

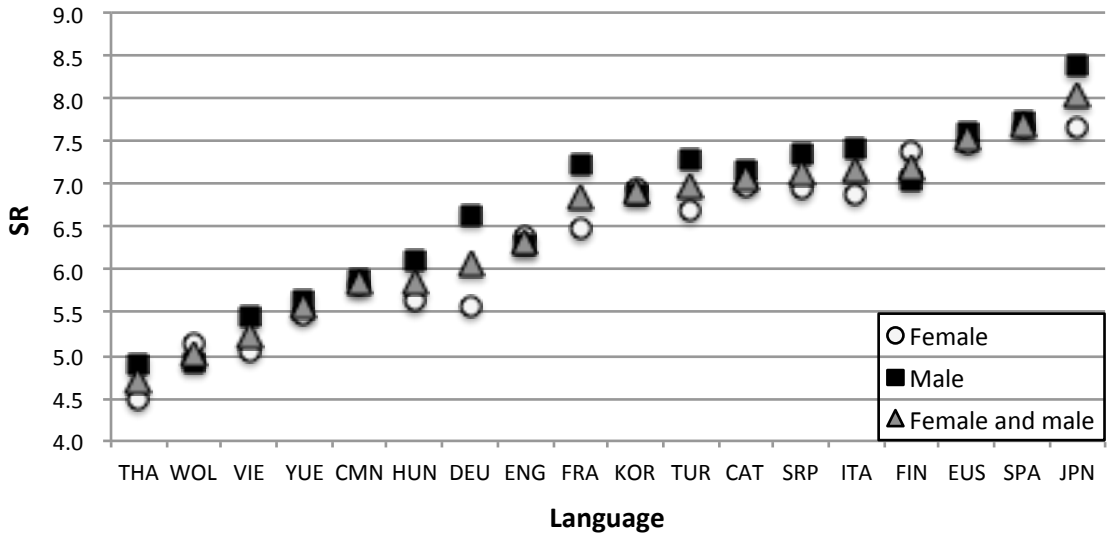

Figure 2.4: Comparison between the *SR* of female and male speakers. The average *SR* of 5 female, 5 male, and a total of 10 speakers are displayed. Languages are ordered by increasing average *SR* of 10 speakers from left to right.

The effect of sex is only found in the model in which *SR* is taken as a dependent variable.[27] As depicted in Figure 2.4, it is observed that the average *SR* of 5 male speakers

[27]The model (SR $\sim$ ID * IR + Sex + Language + (1|Speaker) + (1|Text)) is not presented in this subsection. The coefficient estimate of a fixed variable sex: 0.0671**, p-value = 0.002007.

is faster than female speakers in 14 among the 18 languages, except in English, Finnish, Korean, and Wolof.[28] Such a difference between the *SR* of female and male speakers has already been investigated in phonetics and sociolinguistics (see [Byrd, 1994] among many others).

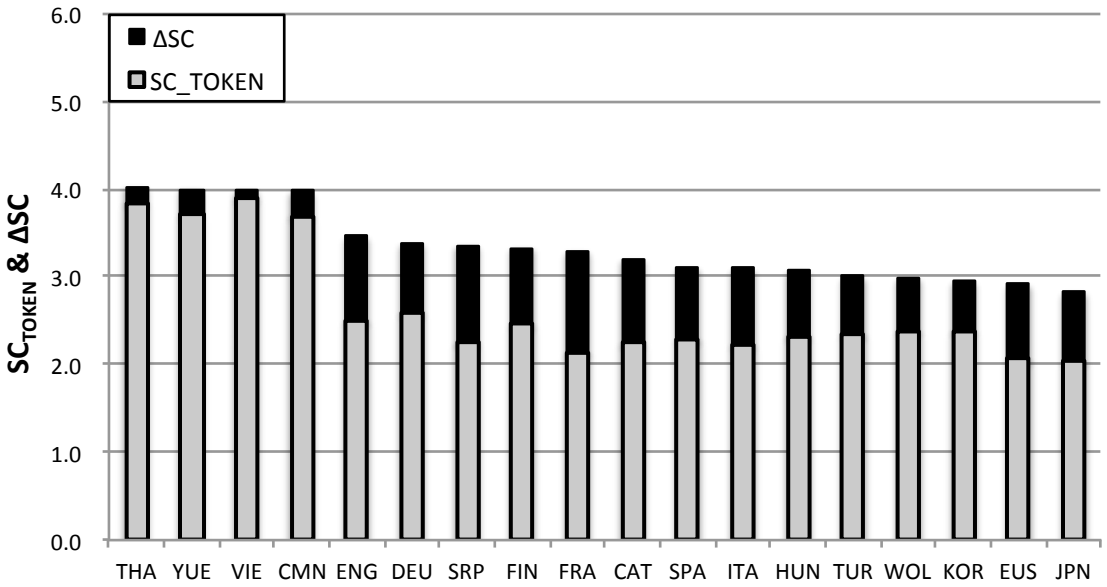

Figure 2.5: SC_{TOKEN} (average number of segments (and tones, if applicable) per syllable) and ΔSC (difference between SC_{TYPE} and SC_{TOKEN}) on y-axis on the left. Languages are ranked by increasing SC_{TYPE} from left to right.

On the one hand, the results obtained by the oral corpus are computed by syntagmatic measures such as *SR*, *ID*, and *IR* on the local scale. On the other hand, the results obtained by the multilingual text corpora are calculated by paradigmatic measures such as SC_{TOKEN}, SC_{TYPE}, and the information-therotic measures (i.e. Shannon entropy, conditional entropy, and surprisal) on the global scale. Figure 2.5 displays the SC_{TOKEN}, ΔSC, and SC_{TYPE} of 18 languages.[29] It is observed that tonal (and slow[30]) languages show a

[28]Regarding the 4 languages with lower male *SR* than female *SR*, it can be explained by the small number of speakers (recorded) in each language, leading to a high sensitivity to individual variation. For example, in Wolof, there's one female speaker with exceptionally fast speech rate (6.24) among 5 female speakers. If the recordings of this speaker is discarded, the average *SR* of male speakers (4.92) becomes higher than female speakers (4.87).

[29]SC_{TYPE} corresponds to the sum of SC_{TOKEN} and ΔSC and thus, is displayed in the cumulative graph in Figure 2.5.

[30]As previously mentioned, SC_{TYPE} and SC_{TOKEN} are significantly correlated with *SR*. Most of slow

relatively little difference between SC_{TYPE} and SC_{TOKEN} (i.e. ΔSC) in comparison with non-tonal languages. It can be explained by the following argument that Chinese languages (including Mandarin and Cantonese) have *a uniform syllable structure of three segments* (one slot for the onset and two slots for the rime) [Duanmu, 1990]. This argument corresponds with our result since the value of SC consists of the number of segments and tones in case of tonal languages. Nevertheless, it suggests that a phenomenon of self-organization exists in speech communication, as illustrated in Figure 2.3 that most of the dense languages with a high ID and a low SR are tonal.

2.3.2 Issues in estimating entropy

In this section, two problems frequently arising in estimating Shannon and conditional entropy are investigated: (i) the effect of corpus size, (ii) the influence of bootstrap simulation which has been proposed as one method of statistical inference, especially for dealing with small-sized data.

In order to test the effect of corpus size, corpora of various sizes (in terms of total number of words) in 4 languages (English, Finnish, French, and Korean) are used to compute Shannon entropy and conditional entropy.[31] In addition, Shannon entropy $H(X)$ is calculated by using both estimated distribution obtained from bootstrap simulation and real distribution of data in 4 languages as presented in Figure 2.6. Bootstrap sampling was performed by a function *bootstrp* in [MATLAB (R2011a)] software, taking 1 000 bootstrap samples into account. The number of samples was chosen by following one rule of thumb suggested for a 95% confidence interval [Zoubir, A. M. & Iskander, 2007]. Bootstrap method is considered as a useful and robust measure for statistical inference of small-scale data, allowing to obtain confidence intervals. The values of Shannon entropy

languages are tonal except for Wolof since many native speakers of Wolof are not familiar with reading written texts in Wolof. Thus, Wolof displays a relatively low SR, which may be due to sociocultural factors.

[31] This subset of languages was chosen to offer some linguistic diversity, without the intention to provide a comprehensive study for the whole corpus.

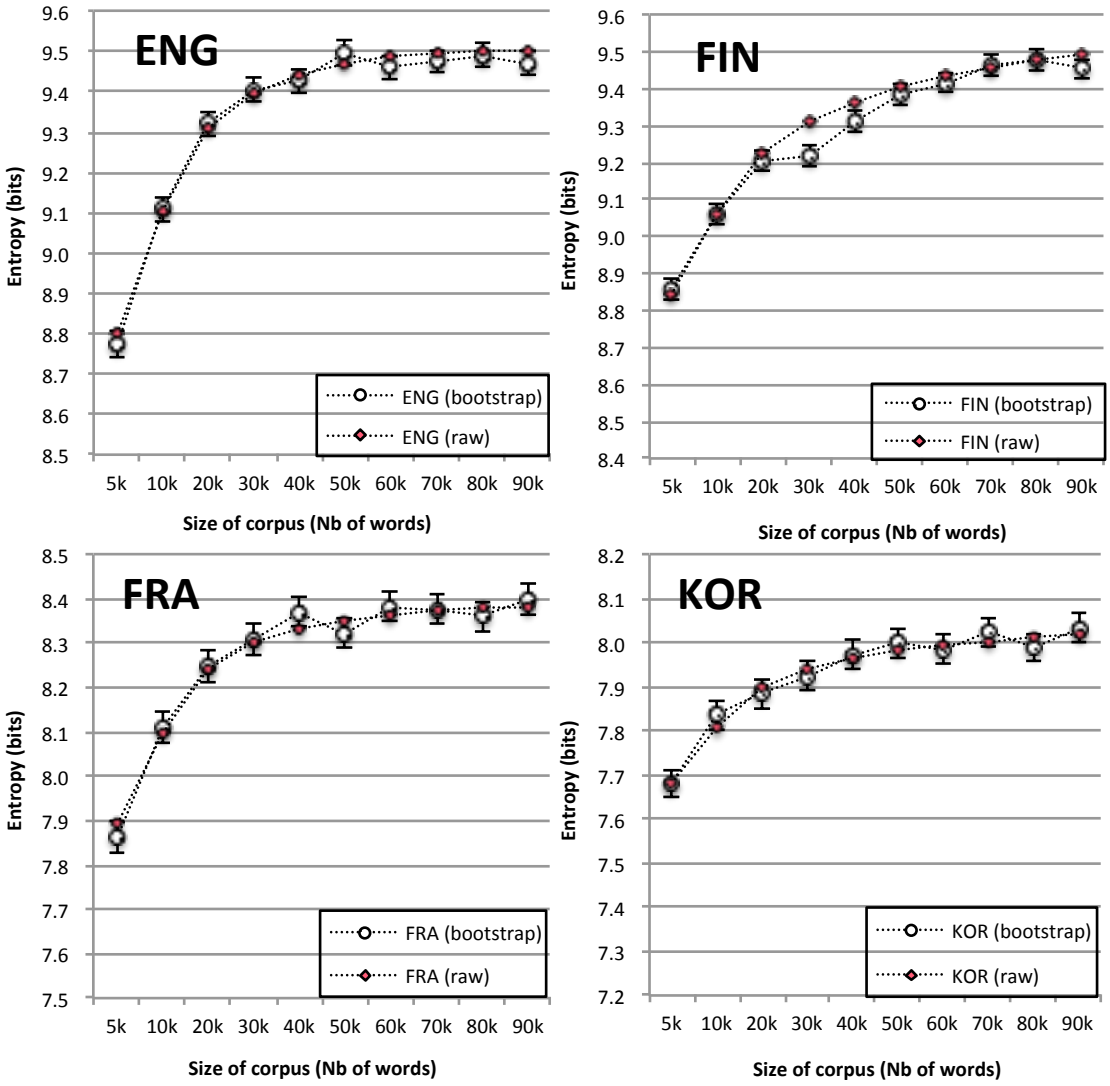

Figure 2.6: Effects of bootstrapping and corpus size in estimating Shannon entropy. Shannon entropy (in bits) on y-axis and the size of corpus (number of words) on x-axis.

computed with and without bootstrap sampling method are compared.

In Figure 2.6, the Shannon entropy values obtained from the estimated distribution by means of bootstrap method are marked in white circles and those obtained from the real distribution are displayed in red diamonds. Contrary to a general assumption based on "maximum-entropy principle" [Jaynes, 19] which states that the probability distribution yielding the highest Shannon entropy value is the one which best reflects the observed or realistic distribution, it is found that the Shannon entropy values obtained from the esti-

mated distribution by bootstrap sampling are sometimes lower and sometimes higher than those calculated from the real distribution. This can be due to the specific distribution of syllable frequency which is a long-tail (somewhat similar to a power-law) distribution, far from a normal distribution. As a consequence, the bootstrapping strategy may be highly sensitive to the individual (high-frequent) syllables resulting in an unstable estimation. Due to this inconsistency, the comparison between the values of Shannon entropy computed by two different methods leads to the following assumption that bootstrap method may not be a robust measure for estimating Shannon entropy in the present study.

In terms of the effect of corpus size, it is shown that the value of Shannon entropy tends to be more robust if the size of corpus becomes larger. In English and French, the values of Shannon entropy start to converge with the corpus containing 60 000 words while in Korean and Finnish, the convergence of Shannon entropy values starts with the data having 50 000 and 70 000 words respectively. Thus, a convergence threshold seems to vary among the 4 languages between 50 000 and 70 000 words.

Conditional entropy $H(X_n|X_{n-1})$ is also computed with different-sized corpora. On the one hand, a convergence of conditional entropy values is displayed in English and French with the data containing 60 000 words. However, on the other hand, in Finnish and Korean, the conditional entropy values do not seem to converge. Compared to Shannon entropy, conditional entropy requires a larger data to estimate more robust values of conditional entropy. As a consequence, rather than taking the same size of corpus for each language, the largest possible data are used in the computation of information-theoretic measures such as Shannon entropy and conditional entropy.

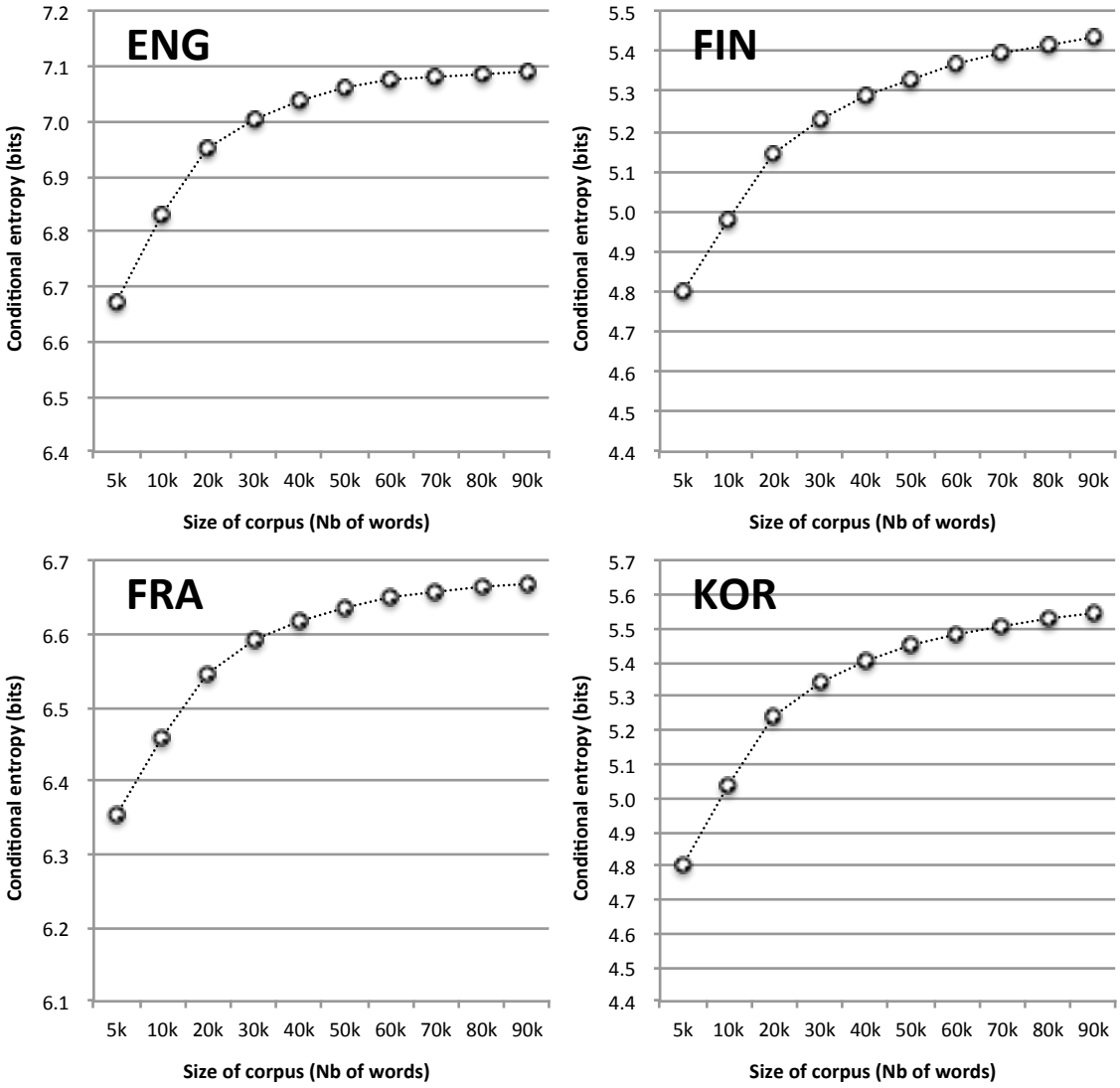

Figure 2.7: Effects of corpus size in estimating conditional entropy. Conditional entropy (in bits) on y-axis and the size of corpus (number of words) on x-axis.

2.3.3 Entropy

The values of Shannon entropy *H(X)*, the size of syllable inventory, and $IR_{H(X)}$ for the 17 languages are displayed in Table 2.9.[32]

[32]The result for Wolof was discarded during the analysis phase because the calculation based on graphemes was assumed to overestimate Shannon entropy and conditional entropy.

Table 2.9: Shannon entropy *H(X)*, inventory size and information rate $IR_{H(X)}$. The maximum and minimum values are marked in green and blue.

Language	H(X)	Inventory	$\mathbf{IR_{H(X)}}$
CAT	8.10	3600	57.30
CMN	8.69	1274	50.90
DEU	9.30	5100	56.66
ENG	9.51	6949	60.28
EUS	8.32	2082	62.74
FIN	9.54	3844	68.83
FRA	8.39	2949	57.48
HUN	9.83	4325	57.69
ITA	8.32	2729	59.59
JPN	6.07	643	48.77
KOR	8.05	1104	55.79
SPA	8.32	2778	64.15
SRP	8.79	3831	62.88
THA	9.13	2438	42.92
TUR	9.19	3260	64.31
VIE	9.72	5156	51.00
YUE	7.97	1298	44.36

$IR_{H(X)}$ is calculated by dividing Shannon entropy *H(X)* multiplied by the number of syllables contained in a text σ_t, by the duration of the utterance of the corresponding text D_t: $IR_H = \frac{H(X) \cdot \sigma_t}{D_t}$. Since Shannon entropy refers to the average amount of information (unpredictability) of a finite set of syllables, $IR_{H(X)}$ represents the average amount of information (in bits) conveyed per second. While it is observed that the range of Shannon entropy varies from 6.07 (JPN) to 9.83 (HUN), Shannon entropy is positively correlated with the size of syllable inventory among the 17 languages (Pearson's $r = 0.736$**; p-value $= 0.001$; Spearman's $\rho = 0.808$**; p-value < 0.001; N $= 17$). For example, the language with the smallest syllable inventory, Japanese, displays the lowest value of $H(X)$ and the language with the largest syllable inventory, English, exhibits the highest value of $H(X)$ disregarding Hungarian, Vietnamese, and Finnish.

Shannon entropy is also regarded as a measure of the cognitive cost in speech communication for both speaker (in terms of memory effort) and hearer (in terms of recognition) [Ferrer i Cancho & Solé, 2003] [Ferrer i Cancho, 2006] [Ferrer i Cancho & Díaz-

Guilera, 2007]. As a consequence, $IR_{H(X)}$ can be interpreted as the average amount of cognitive cost or information transmitted during communication, computed by means of Shannon entropy. Among the 17 languages, the values of $IR_{H(X)}$ range from 42.92 (Thai) to 68.83 (Finnish). In a strict sense, this measure of $IR_{H(X)}$ does not take two following factors into account: (i) contextual information and (ii) probability of individual syllable. Therefore, Shannon entropy may be considered as a less accurate measure of information in comparison with the other measures such as surprisal and conditional entropy.

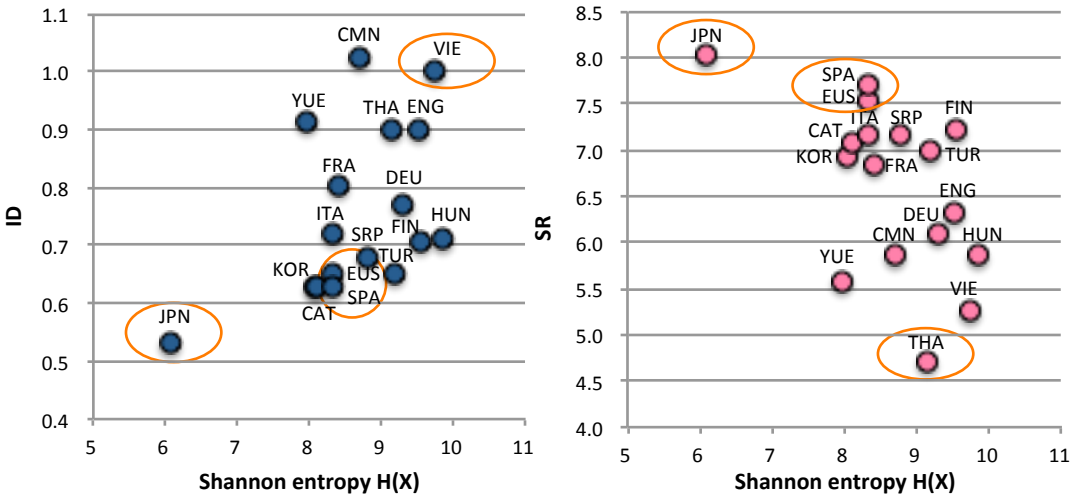

Figure 2.8: Correlations among *ID*, *SR*, and Shannon entropy *H(X)*.

With respect to *SR* and *ID*, Shannon entropy is not significantly correlated with *ID* (Pearson's $r = -0.481$; p-value $= 0.051$; Spearman's $\rho = 0.443$; p-value $= 0.075$; N $= 17$) whereas a significant Pearson correlation is found between *SR* and *H(X)* (Pearson's $r = -0.547$*; p-value $= 0.023$; Spearman's $\rho = -0.405$; p-value $= 0.106$; N $= 17$). However, this correlation exists due to Japanese as displayed in Figure 2.8. If Japanese is discarded from the language samples, the correlation no longer exists between *SR* and *H(X)* (Pearson's $r = -0.409$; p-value $= 0.115$; N $= 16$). Nevertheless, few languages exhibit a following tendency that slow (with respect to *SR*) and dense (with respect to *ID*) languages (for example, Thai and Vietnamese) display higher Shannon entropy than fast and sparse languages (for example, Basque, Japanese, and Spanish).

Table 2.10: Result of ANOVA taking $H(X)$ as a dependent variable (Df = degrees of freedom, Sum Sq = sum of squares, F value = ANOVA statistic, Pr = probability, % of variance = $\frac{Sum\ Sq\ explained}{Sum\ Sq\ total}$, % (Mono, Bi, Tri) W = percentage of monosyllabic, bisyllabic, or trisyllabic words in terms of token)

Source	Df	Sum Sq	F value	Pr	% of variance
SC_{TOKEN}	1	1.3737	4.8524	0.0498405*	8.59
Inventory	1	9.0260	31.8833	0.0001496**	56.41
% Mono W	1	2.4806	8.7624	0.0129720*	15.50
% Bi W	1	0.0053	0.0189	0.8932176	0.03
% Tri W	1	0.0004	0.0013	0.9713766	0.003
Residuals	11	3.1140			19.46

In order to analyze the effect of different variables on $H(X)$, a one-way analysis of variance (ANOVA) is conducted using the software R taking $H(X)$ as a dependent variable. According to the result presented in Table 2.10, the size of syllable inventory is the factor which most influences $H(X)$ (56.41%) and is followed by the percentage of monosyllabic words (15.50%) and the syllable complexity SC_{TOKEN} (8.59%). This result will be compared with the results of ANOVA taking conditional entropy and mutual information as a dependent variable in the following subsections.

2.3.4 Conditional entropy

As mentioned in Subsection 2.2.2.3., there are two different methods used for calculating conditional entropy. The first method consists of marking the initial position of word by an asterisk (*) in order to handle monosyllabic words which cannot be taken into account in a bigram language model otherwise. Hence, it allows us to compute $H(X_n|X_{n-1})$, i.e. the average amount of uncertainty given by the previous context. The second method of computing conditional entropy is to add a hash (#) to the final position of each word and it enables us to obtain $H(X_n|X_{n+1})$, i.e. the amount of information given by the following context.

Table 2.11: $H(X_n|X_{n-1})$ and $H(X_n|X_{n+1})$. The maximum and minimum values are marked in green and blue.

Language	$H(X_n\|X_{n-1})$	$H(X_n\|X_{n+1})$
CAT	5.49	5.53
CMN	6.96	6.99
DEU	6.08	6.13
ENG	7.09	7.10
EUS	4.83	5.05
FIN	5.49	5.86
FRA	6.68	6.76
HUN	5.90	5.95
ITA	5.29	5.26
JPN	5.03	5.07
KOR	5.56	5.53
SPA	5.43	5.41
SRP	5.47	5.99
THA	7.19	7.13
TUR	5.34	5.18
VIE	8.02	8.04
YUE	6.53	6.59

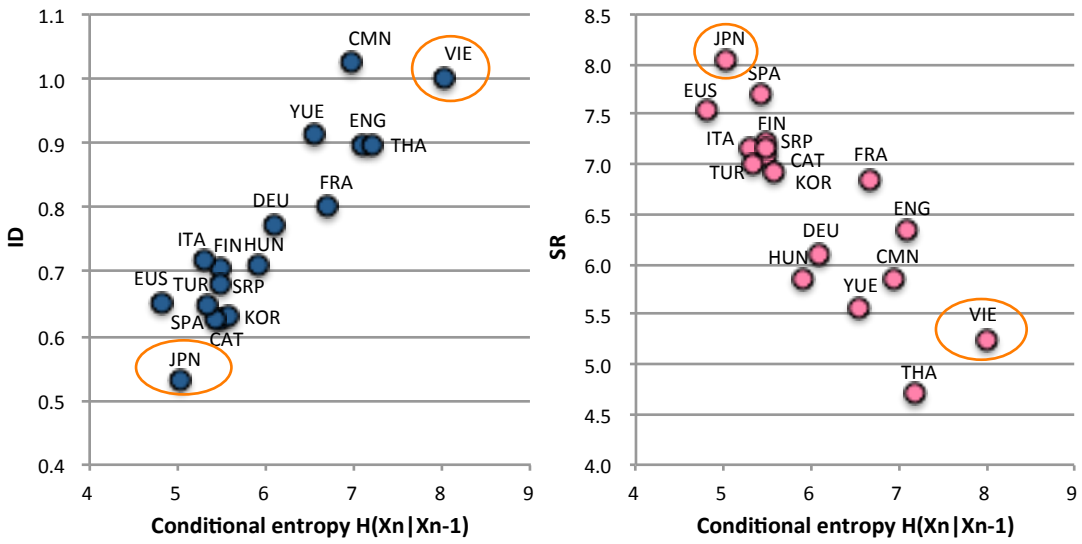

Figure 2.9: Correlations among *ID*, *SR*, and conditional entropy $H(X_n/X_{n-1})$.

Regarding *ID* and *SR*, conditional entropy $H(X_n|X_{n-1})$ is significantly correlated with both *ID* (Pearson's $r = 0.912$**; p-value < 0.001; Spearman's $\rho = 0.796$**; p-value < 0.001; N $= 17$) and *SR* (Pearson's $r = -0.837$**; p-value < 0.001; Spearman's $\rho = -0.887$**;

p-value < 0.001; N $= 17$) as displayed in Figure 2.9, contrary to Shannon entropy which is neither significantly correlated with *ID* nor with *SR* (if Japanese is discarded from the language sample). It is revealed that fast (in terms of *SR*) and sparse (in terms of *ID*) language (e.g. Japanese) exhibit higher conditional entropy than slow and dense language (e.g. Vietnamese).[33]

Table 2.12: Percentage of monosyllabic, bisyllabic, and trisyllabic words in terms of type and token. The maximum and minimum values (% token) are marked in green and blue.

Language	**Monosyllabic**		**Bisyllabic**		**Trisyllabic**	
	% type	**% token**	**% type**	**% token**	**% type**	**% token**
CAT	3	55	20	19	30	13
CMN	9	53	69	44	14	3
DEU	4	51	24	29	33	12
ENG	15	71	38	19	28	7
EUS	1	9	6	34	21	28
FIN	1	15	14	35	30	26
FRA	7	68	30	21	40	9
HUN	5	39	24	27	32	19
ITA	3	41	19	28	34	18
JPN	2	41	22	30	38	19
KOR	1	11	11	35	34	34
SPA	1	45	18	28	38	18
SRP	5	38	28	30	38	21
THA	44	75	40	19	11	4
TUR	5	20	24	32	33	27
VIE	27	75	70	24	3	0.2
YUE	27	79	63	20	5	1

To understand conditional entropy, it is crucial to take account of the percentage of monosyllabic and non-monosyllabic words. There are 8 out of 17 languages in which monosyllabic words take up more than 50% of token (Catalan, Mandarin Chinese, German, English, French, Thai, Vietnamese, and Cantonese). Among the 17 languages in Table 2.12, Basque with the lowest coverage of monosyllabic words (9%) exhibits the lowest value of $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$ while Vietnamese with the second largest coverage of monosyllabic words (75%) after Cantonese shows the highest value of $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$

[33] A very similar result is found with conditional entropy $H(X_n|X_{n+1})$.

and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$. As we can assume by comparing the values presented in Tables 2.11 and 2.12, conditional entropy is positively correlated with the percentage of monosyllabic words ($H(X_n|X_{n-1})$: Pearson's $r = 0.790^{**}$; p-value < 0.001; Spearman's $\rho = 0.740^{**}$; p-value $= 0.001$, $H(X_n|X_{n+1})$: Pearson's $r = 0.764^{**}$; p-value < 0.001; Spearman's $\rho = 0.745^{**}$; p-value $= 0.001$; N $= 17$).

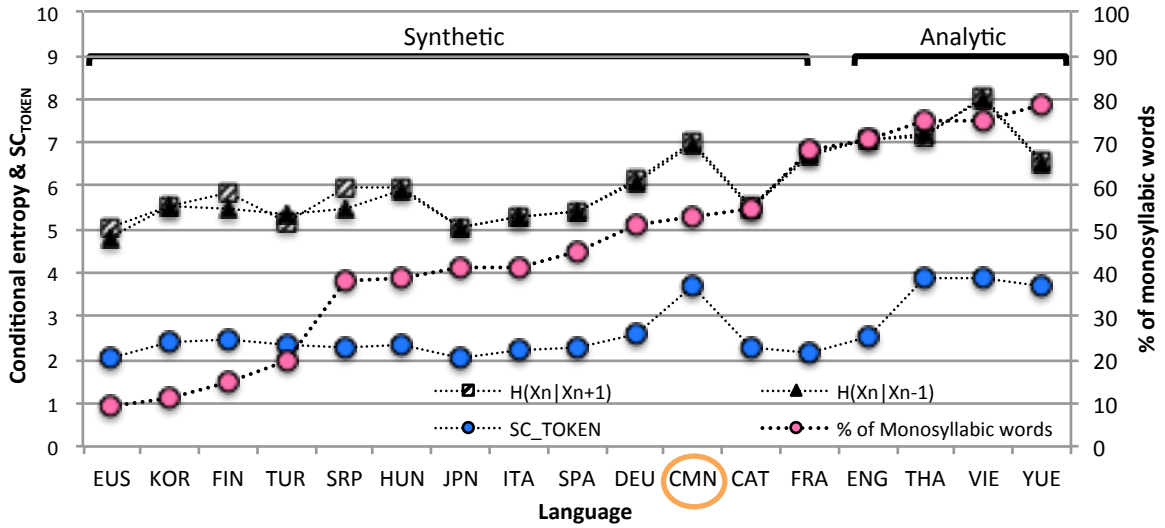

Figure 2.10: Conditional entropy and SC_{TOKEN} on the left y-axis & percentage of monosyllabic words on the right y-axis. Language are ordered by increasing % of monosyllabic words from left to right.

In Figure 2.10, languages can be divided into two types in terms of their morphological classification: (i) synthetic and (ii) analytic languages, except for Mandarin Chinese.[34] English, Thai, Vietnamese, and Cantonese are classified as analytic languages which are characterized by having one morpheme per word in word formation. On the contrary, synthetic languages contain more than one morphemes per word. Thus, the contextual information of synthetic languages is more informative than analytic languages and the values of $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$ are higher for analytic languages than synthetic languages. The result leads to a following assumption that conditional entropy is strongly connected with the patterns of affixation and word formation of languages.

[34]Mandarin Chinese is regarded as an analytic language although it has many words containing more than one morpheme per words, despite its lack of affixation.

Table 2.13: Result of ANOVA taking $H(X_n|X_{n-1})$ as a dependent variable (Df = degrees of freedom, Sum Sq = sum of squares, F value = ANOVA statistic, Pr = probability, % of variance = $\frac{Sum\ Sq\ explained}{Sum\ Sq\ total}$, % (Mono, Bi, Tri) W = percentage of monosyllabic, bisyllabic, or trisyllabic words in token)

Source	Df	Sum Sq	F value	Pr	% of variance
SC_{TOKEN}	1	10.1231	58.2094	< 0.001***	63.27
Inventory	1	2.4991	14.3700	0.002991**	15.62
% Mono W	1	1.0771	6.1937	0.030104*	6.73
% Bi W	1	0.1796	1.0329	0.331305	1.12
% Tri W	1	0.2081	1.1964	0.297409	1.30
Residuals	11	1.9130			11.96

A one-way analysis of variance (ANOVA) is conducted taking $H(X_n|X_{n-1})$ as a dependent variable.[35] Contrary to the result presented in Table 2.10 concerning $H(X)$, SC_{TOKEN} is the factor with the highest impact on $H(X_n|X_{n-1})$ (63.27%) and is followed by the size of inventory (15.62%) and the percentage of monosyllabic words (6.73%). In particular, syllable complexity in terms of token, i.e. SC_{TOKEN}, appears to be strongly related to $H(X_n|X_{n-1})$ while it does not exhibit such an effect on $H(X)$ (8.59%). The size of syllable inventory (56.41%) and the percentage of monosyllabic words (15.50%) seem to be more concerned with $H(X)$.

Regarding the average amount of information conveyed per second computed by means of conditional entropy,[36] no particular tendency among the languages is found. However, it is observed that they are significantly correlated with the size of corpus (in terms of the number of types, cf. Table 2.1) ($IR_{\mathrm{H(X_n|X_{n-1})}}$: Pearson's $r = 0.550$*; p-value = 0.022; Spearman's $\rho = 0.444$; p-value = 0.074; N = 17, $IR_{\mathrm{H(X_n|X_{n+1})}}$: Pearson's $r = 0.573$*; p-value = 0.016; Spearman's $\rho = 0.511$*; p-value = 0.036; N = 17). French with the second largest number of words (142k) in the text corpus after English (160k) exhibits the highest values of $IR_{H(X_n|X_{n-1})}$ and $IR_{H(X_n|X_{n+1})}$ and Thai with the smallest number of words (5k) in the data has the lowest values of $IR_{H(X_n|X_{n-1})}$ and $IR_{H(X_n|X_{n+1})}$.

[35] After comparing the F-statistic scores of 2 conditional entropy ($H(X_n|X_{n-1})$: 16.2, $H(X_n|X_{n+1})$: 13.82), $H(X_n|X_{n-1})$ with a higher F-statistic score was chosen for the comparison with $H(X)$.

[36] $IR_{H(X_n|X_{n-1})} = \frac{H(X_n|X_{n-1})\cdot\sigma_t}{D_t}$ and the same equation applies to $IR_{H(X_n|X_{n+1})}$.

Table 2.14: *IR* obtained from conditional entropy ($IR_{\mathrm{H(X_n|X_{n-1})}}$ & $IR_{\mathrm{H(X_n|X_{n+1})}}$). The maximum and minimum values are marked in green and blue.

Language	$\mathbf{IR_{H(X_n|X_{n-1})}}$	$\mathbf{IR_{H(X_n|X_{n+1})}}$
CAT	38.83	39.12
CMN	40.75	40.95
DEU	37.06	37.34
ENG	44.94	45.00
EUS	36.40	38.08
FIN	39.62	42.28
FRA	45.76	46.32
HUN	34.62	34.92
ITA	37.89	37.67
JPN	40.40	40.74
KOR	38.56	38.32
SPA	41.89	41.71
SRP	39.14	42.85
THA	33.81	33.52
TUR	37.35	36.25
VIE	42.10	42.19
YUE	36.37	36.68

The residuals of 4 mixed effects models are compared in Figure 2.11. The first model on the top left corresponds to the one presented in Table 2.8: *IR* (dependent variable) ∼ ID * SR + Sex + Language + (1|Speaker) + (1|Text). The second mixed effects models on the top right takes $IR_{H(X)}$ as a dependent variable instead of *IR* as follows: $IR_{H(X)}$ ∼ ID * SR + Sex + Language + (1|Speaker) + (1|Text). The two mixed effects models on the bottom left and right are those taking $IR_{\mathrm{H(X_n|X_{n-1})}}$ and $IR_{\mathrm{H(X_n|X_{n+1})}}$ as a dependent variable respectively. It is observed that the model with *IR* calculated by means of the syntagmatic measure on the local scale displays the best estimation of *IR* in comparison with the other 3 models with *IR* computed by the paradigmatic measures on the global scale (i.e. $H(X)$, $H(X_n|X_{n-1})$, and $H(X_n|X_{n+1})$). Among those 3 models, a better estimation is obtained by the model with *IR* obtained by Shannon entropy $IR_{H(X)}$ rather than *IR* obtained by two conditional entropy $IR_{\mathrm{H(X_n|X_{n-1})}}$ and $IR_{\mathrm{H(X_n|X_{n+1})}}$.

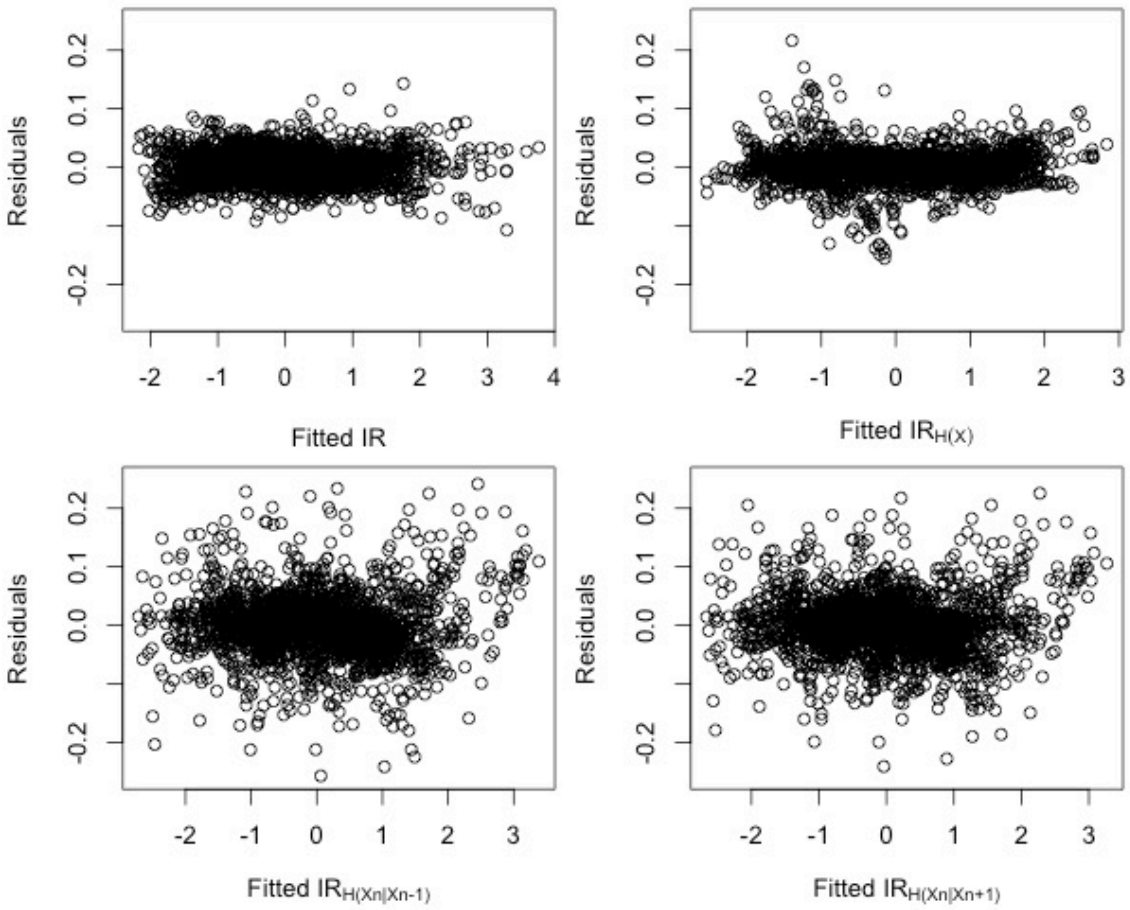

Figure 2.11: Residuals of mixed effects models. Residuals on the y-axis and fitted values of *IR* on the x-axis

2.3.5 Surprisal

Surprisal $S(X)$ takes account of the individual probability of syllable for computing the average *IR*. It is thus expected to be more accurate in comparison with Shannon entropy $H(X)$. The amount of information conveyed per second $IR_{\mathrm{S(X)}}$ is obtained by dividing the sum of surprisal $\sum S(X)$ of text t by the duration of utterance of the corresponding text D_t: $IR_{S(X)} = \frac{\sum S(X)_t}{D_t}$. In order to take contextual information into account, two bigram language models are used: (i) $S(X_n|X_{n-1})$ is obtained from a bigram language model where the initial position of each word is marked with an asterisk (*) and (ii) $S(X_n|X_{n+1})$ is calculated from a bigram language model where the final position of each word is marked with a hash (#). The average information rate is obtained in the same way

as $IR_{S(X)}$: $IR_{S(X_n|Xn-1)} = \frac{\sum S(X_n|Xn-1)_t}{D_t}$ and $IR_{S(X_n|Xn+1)} = \frac{\sum S(X_n|Xn+1)_t}{D_t}$. The results are presented in Table 2.15.

Table 2.15: Average IR obtained from $S(X)$, $S(X_n|X_{n-1})$, and $S(X_n|X_{n+1})$: $IR_{S(X)}$, $IR_{S(X_n|X_{n-1})}$, and $IR_{S(X_n|X_{n+1})}$. The maximum and minimum values are marked in green and blue.

Language	$IR_{S(X)}$	$IR_{S(X_n\|X_{n-1})}$	$IR_{S(X_n\|X_{n+1})}$
CAT	63.01	81.47	83.97
CMN	50.33	67.23	67.80
DEU	59.32	69.60	67.92
ENG	63.28	70.12	70.54
EUS	64.67	87.96	91.83
FIN	69.83	97.01	94.85
FRA	59.89	72.46	74.49
HUN	59.88	75.81	84.61
ITA	65.44	95.83	88.17
JPN	56.19	111.58	108.41
KOR	62.40	102.45	103.44
SPA	68.57	92.95	93.14
SRP	66.00	89.20	92.67
THA	47.85	58.29	55.92
TUR	67.46	89.64	89.44
VIE	50.82	55.43	55.81
YUE	51.38	53.05	54.26

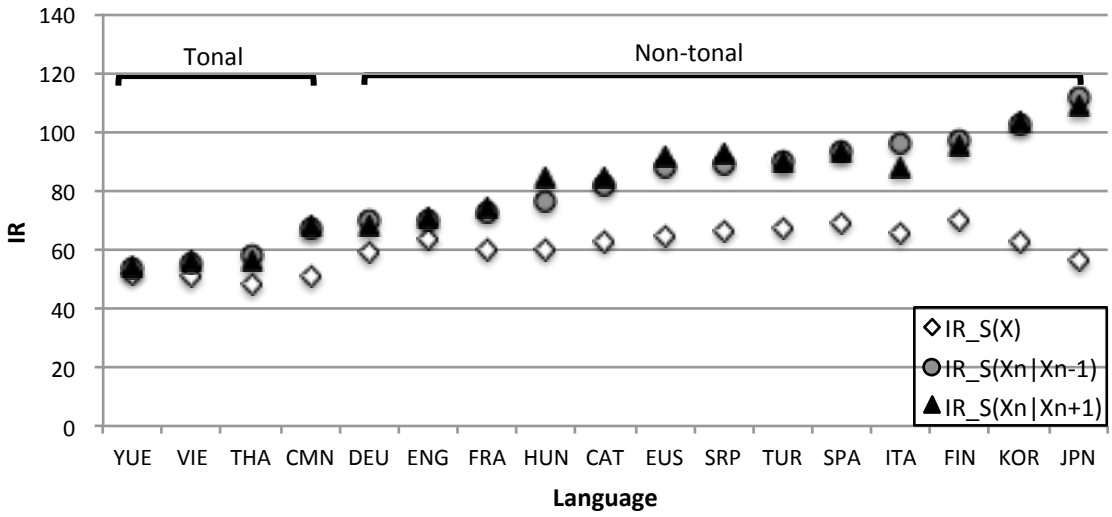

Figure 2.12: $IR_{S(X)}$, $IR_{S(X_n|X_{n-1})}$, and $IR_{S(X_n|X_{n+1})}$. Languages are ordered by increasing $IR_{S(X_n|X_{n-1})}$ from left to right.

In comparison with $IR_{S(X_n|X_{n-1})}$ and $IR_{S(X_n|X_{n+1})}$, the values of $IR_{S(X)}$ vary within a relatively limited range, i.e. from 47.85 (Thai) to 69.83 (Finnish). $IR_{S(X_n|X_{n-1})}$ and $IR_{S(X_n|X_{n+1})}$ take contextual information into account and their maximum value of IR (Japanese) is 2 times faster than the minimum value of IR (Cantonese). The ranges of variation in terms of three different IR are displayed and compared in Figure 2.12 where the languages can be divided into tonal (isolating) and non-tonal (fusional/agglutinative) languages. There are 4 tonal languages (Cantonese, Vietnamese, Thai, and Mandarin Chinese) in the language samples and regarding the two IRs obtained from bigram language models, they transmit a lower amount of information per second on average than non-tonal languages. In particular, agglutinative languages such as Korean and Japanese exhibit the largest gap between $IR_{S(X)}$ and $IR_{S(X_n|X_{n-1})}$ and $IR_{S(X)}$ and $IR_{S(X_n|X_{n+1})}$.

Since surprisal is a kind of hybrid measure which combines the syllable distribution obtained at the global scale (from a bigram language model) and the information of the individual syllable in the oral scripts at the local level (from the 15 short texts), it is assumed to be more sensitive to (i) the size of a large text corpus, (ii) the syllable distribution estimated from the corpus, and (iii) the individual syllables and their context in the oral scripts (Multext), i.e. bigrams. As the length of each oral script is limited to $3-5$ short sentences, if the script contains some bigrams (or syllables) which are rarely used or unobserved in a language model, it may lead to the overestimation of IR.[37] Depending on the probability of an unseen bigram (or syllable) estimated by the SGT algorithm, the sum of surprisal may become larger if there are more unobserved or low-frequency bigrams (or syllables) in the oral scripts. As a consequence, it is supposed that larger corpora (a text corpus for creating a language model and an oral corpus for computing IR) may provide a better estimation of syllable distribution and more stable IR across the 17 languages.

[37]In order to deal with some unobserved bigrams (or syllables) in the 15 oral scripts by the language models created with a large text corpus, the Simple Good Turing (SGT) algorithm [Gale & Sampson, 1995] was used. This algorithm provides an estimation for the probability of an unobserved bigram (or syllable) by analyzing the distribution of probabilities (i.e. the frequency of frequency) in language model.

Table 2.16: Mixed-effects model of $IR_{S(X)}$. The effects of fixed factors and random factors are displayed on the left and right sides of the table respectively. (Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.')

Fixed factor				Random factor			
Predictor	**Coefficient**	**t-value**	**Sig**	**Predictor**	**X^2(df)**	**p-value**	**Sig**
$\boldsymbol{IR_{S(X)}}$ **(dependent variable)**							
Intercept	0.0360	0.453		Speaker	129.74 (1)	< 0.001	***
SR	0.9913	47.242	***	Text	212.45 (1)	< 0.001	***
ID	0.1736	8.432	***				
Sex_{Male}	0.0112	0.385					
$\text{Language}_{\text{CAT}}$	−0.0347	−0.375					
$\text{Language}_{\text{CMN}}$	−0.6102	−7.452	***				
$\text{Language}_{\text{DEU}}$	0.3272	3.615	***				
$\text{Language}_{\text{ENG}}$	0.3966	4.219	***				
$\text{Language}_{\text{EUS}}$	−0.2925	−3.089	**				
$\text{Language}_{\text{FIN}}$	0.4737	5.170	***				
$\text{Language}_{\text{FRA}}$	−0.3140	−3.569	***				
$\text{Language}_{\text{HUN}}$	0.6255	7.232	***				
$\text{Language}_{\text{ITA}}$	0.0669	0.668					
$\text{Language}_{\text{JPN}}$	−1.4795	−14.435	***				
$\text{Language}_{\text{KOR}}$	0.0302	0.328					
$\text{Language}_{\text{SPA}}$	−0.0194	−0.201					
$\text{Language}_{\text{SRP}}$	0.1564	1.706	.				
$\text{Language}_{\text{THA}}$	0.2583	3.141	**				
$\text{Language}_{\text{TUR}}$	0.4725	5.147	***				
$\text{Language}_{\text{YUE}}$	−0.1510	−1.850	.				
SR:*ID*	0.0078	0.668					

Contrary to the previous result of a mixed-effects model of *IR* presented in Section 2.3.1, (i) $IR_{\text{S(X)}}$ is not significantly related to the interaction between *SR* and *ID* and (ii) speaker is considered as a significant random factor. As a consequence, this result does not fit the initial hypothesis and the previous result obtained by the *IR* measured by a pairwise comparison, which supports the hypothesis. It will be further discussed in the next section.

In Figure 2.13, the average *IR* calculated from Shannon entropy $H(X)$ is compared with the average *IR* calculated from surprisal $S(X)$. Except for Mandarin Chinese, the values of $IR_{\text{S(X)}}$ are always greater than the values of $IR_{\text{H(X)}}$, which can be explained by following reasons: (i) $H(X)$ may underestimate the average *IR* as it corresponds to

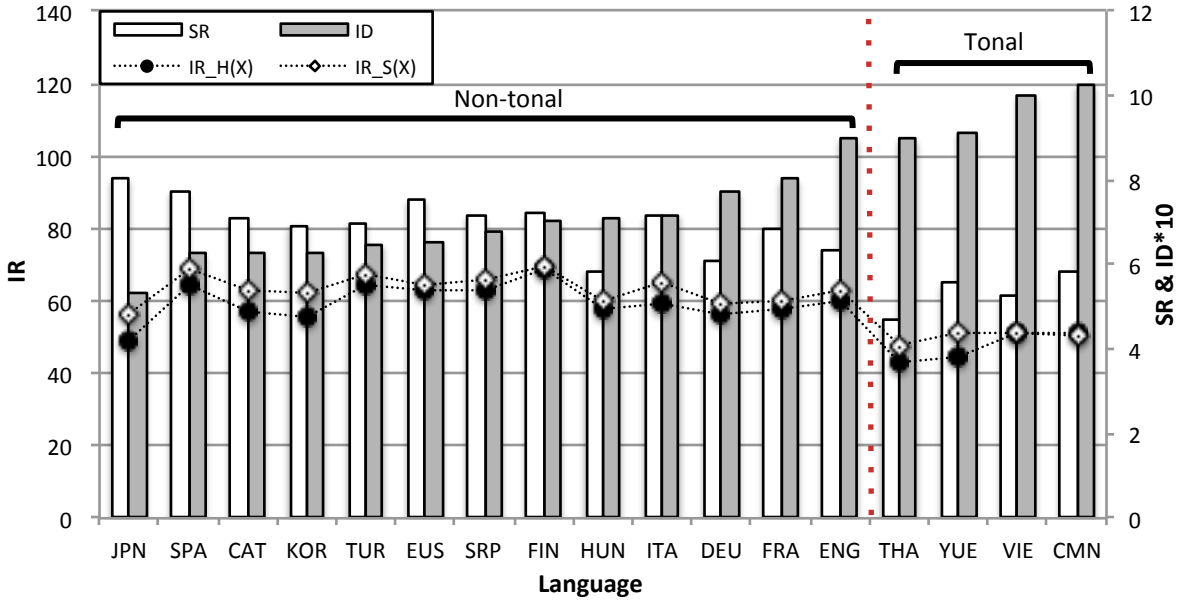

Figure 2.13: $IR_{\mathrm{H(X)}}$ and & $IR_{\mathrm{S(X)}}$ on the left y-axis and SR and ID*10 on the right y-axis. Language are ordered by increasing ID*10 from left to right.

the average uncertainty of a finite set of syllables, (ii) the Simple Good Turing (SGT) algorithm [Gale & Sampson, 1995] was used to deal with some unobserved syllables in the 15 oral scripts by the language models . Thus, the sum of surprisal may become larger if there are more unobserved syllables in the oral scripts, which leads to the greater value of $IR_{\mathrm{S(X)}}$ than $IR_{\mathrm{H(X)}}$. Furthermore, different patterns are found between tonal and non-tonal languages. Non-tonal languages seem to transmit more information per second on average than tonal languages. However, within each group of tonal or non-tonal languages, the average $IR_{\mathrm{H(X)}}$ and $IR_{\mathrm{S(X)}}$ remain stable across languages.

4 mixed effects models are compared by means of their residuals in Figure 2.14. The model on the top left takes IR as a dependent variable: $IR \sim$ ID * SR + Sex + Language + (1|Speaker) + (1|Text). The mixed effects models on the top right takes $IR_{S(X)}$ as a dependent variable as follows: $IR_{S(X)} \sim$ ID * SR + Sex + Language + (1|Speaker) + (1|Text). The two mixed effects models on the bottom left and right take $IR_{\mathrm{S(X_n|X_{n-1})}}$ and $IR_{\mathrm{S(X_n|X_{n+1})}}$ as a dependent variable respectively. The model with the best estimation is the one with the syntagmatic measure of IR on the local scale. Among the 3 models with

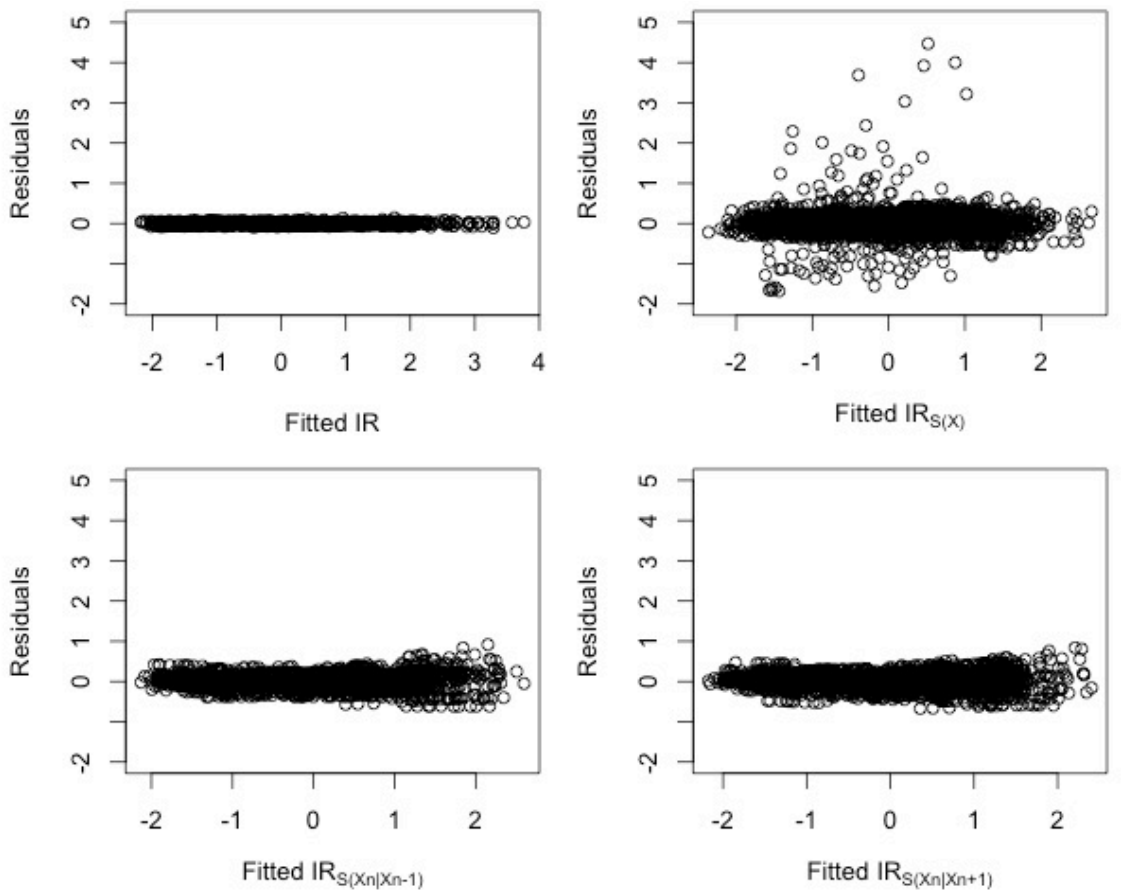

Figure 2.14: Residuals of mixed effects models. Residuals on the y-axis and fitted values of *IR* on the x-axis

the paradigmatic (information-theoretic) measures of *IR* on the global scale, the 2 models with *IR* obtained by surprisal with the contextual information provide a better estimation than the model on the top right which did not take the context into account.

Since there are only 17 (or 18 including Wolof presented in Section 2.3.1) languages analyzed in this study, the present result can be further improved by adding more typologically distinct languages from different language families and furthermore, by adding languages with simple syllable structure, which are lacking in this study (for example, Hawaiian and Maori). A major consideration for selecting a language concerns sociolinguistic aspects. During the data analysis, 2 languages with oral tradition (Khmer and Fang) were discarded from the language samples due to the lack of fluency in oral data.[38]

[38]If native speakers of language are not accustomed to read written texts naturally in their native

In addition, the data availability for creating a statistical language model is a crucial issue. Since information-theoretic approaches are wholly corpus-dependent, the size and characteristic of corpus almost determines the nature and quality of study. If the corpus is too small or domain-specific, the result is expected to be biased, except for the studies in specific domains.

2.4 Discussion

2.4.1 Effect of contextual information

The effects of contextual information can be assessed by comparing the AIC score of the mixed effects models in which one of their fixed factors, *ID* (syntagmatic measure of the average amount of information on the local scale, using the oral corpus) is replaced by the paradigmatic measures of information density on the global scale (using the large text corpus) such as syllable complexity, Shannon entropy, conditional entropy, and surprisal.[39][40] Surprisal can be considered as a hybrid measure combining both global and local aspects since the probability distribution is obtained on the global scale while the mean value of surprisal is acquired by averaging the total sum of the surprisal of individual syllables in each text on the local scale. Such distinction between syntagmatic and paradigmatic measures of information density allows us to observe some different patterns between the local and global scales.[41]

language, it requires them a lot of efforts. In case of Wolof which is also a language with oral tradition, the data was collected by the author during the conference SENELANGUES where some linguists from Senegal gathered at the laboratory of DDL in Lyon. The recording took longer than other languages, even with the linguists whose native language is Wolof.

[39]The average value of surprisal was obtained by dividing the sum of surprisal by the number of the syllables in each text and averaging each of those values.

[40]It should be noted that *ID* used in this subsection is the average *ID* obtained from the 15 texts, which differs from *ID* in the mixed effects model presented in Table 2.8 since the latter corresponds to a text-dependent *ID* for the corresponding data point.

[41]Surprisal obtained from conditional probability based on a bigram language model is not considered in this section due to its strong dependency on the size of text corpus which differs among the 17 languages.

Table 2.17: Correlations between syntagmatic and paradigmatic measures of *ID*

Parameter	ID (N=17)
$\mathbf{H(X)}$	r= 0.481; p= 0.051, ρ= 0.443; p= 0.075
$\mathbf{SC_{TYPE}}$	r= 0.912**; $p<$ 0.001, ρ= 0.833**; $p<$ 0.001
$\mathbf{SC_{TOKEN}}$	r= 0.849**; $p<$ 0.001, ρ= 0.704**; p= 0.002
$\mathbf{H(X_n|X_{n-1})}$	r= 0.912**; $p <$ 0.001, ρ= 0.796**; $p<$ 0.001
$\mathbf{H(X_n|X_{n+1})}$	r= 0.912**; $p <$ 0.001, ρ= 0.840**; $p<$ 0.001
$\mathbf{S(X)}$	r= 0.418; p= 0.095, ρ= 0.406; p= 0.106

In Table 2.17, a strong positive correlation is found between *ID* and SC_{TYPE}, SC_{TOKEN}, and conditional entropy whereas no significant correlation is observed between *ID* and Shannon entropy and surprisal obtained from a unigram language model. Thus, it is assumed that the paradigmatic measures taking the contextual information into account are more strongly correlated with the syntagmatic measure of information density (i.e. *ID*) than those based on a unigram language model without considering the context, i.e. $H(X)$ and $S(X)$.

Table 2.18: Comparison of the AIC scores of mixed effects models

Mixed effects model	AIC (ML)
$IR \sim \mathbf{ID}$ * SR + Sex + (1|Speaker) + (1|Text)	2815.376
$IR \sim \mathbf{SC_{TYPE}}$ * SR + Sex + (1|Speaker) + (1|Text)	3175.321
$IR \sim \mathbf{SC_{TOKEN}}$ * SR + Sex + (1|Speaker) + (1|Text)	3241.944
$IR \sim \mathbf{H(X)}$ * SR + Sex + (1|Speaker) + (1|Text)	3320.334
$IR \sim \mathbf{H(X_n|X_{n-1})}$ * SR + Sex + (1|Speaker) + (1|Text)	3210.66
$IR \sim \mathbf{H(X_n|X_{n+1})}$ * SR + Sex + (1|Speaker) + (1|Text)	3192.307
$IR \sim \mathbf{S(X)}$ * SR + Sex + (1|Speaker) + (1|Text)	3336.989

The result presented in Table 2.18 displays the AIC score of mixed effects models. The model taking the syntagmatic measure of information density (*ID*) as one of its fixed factors exhibits the lowest score of AIC. Hence, it can be considered as the best-fit model among the models in the table above. The other models take the paradigmatic measures of information density as one of their fixed factors instead of *ID* in the model. It is thus suggested that the syntagmatic measure of information density (*ID*) provides a better

model fit to the data than the paradigmatic measures of information density, i.e. SC_{TYPE}, SC_{TOKEN}, $H(X)$, $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$, $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$, and $S(X)$. Among the models taking the paradigmatic measures, the one with SC_{TYPE} displays the lowest AIC score and is followed by the model with $H(X_{\mathrm{n}}|_{\mathrm{n+1}})$, $H(X_{\mathrm{n}}|_{\mathrm{n-1}})$, SC_{TOKEN}, $H(X)$, and $S(X)$. The last two models with $H(X)$ and $S(X)$ exhibiting the highest AIC scores are those without the consideration for context. Consequently, by comparing the AIC scores, it appears that the mixed effects model exhibits a better fit if the contextual information is taken into account by the measures of information density.

Furthermore, the effects of contextual information are reflected by conditional entropy (cf. Figure 2.10 in Subsection 2.3.4). According to our result, the contextual information of synthetic languages is more predictable than analytic languages and thus, conditional entropy (uncertainty) is higher for analytic languages than for synthetic languages. It leads us to the following assumption that conditional entropy is strongly related to the morphological strategies of languages (e.g. the patterns of affixation and the word formation).

2.4.2 Average information rate and UID hypothesis

The Uniform Information Density (UID) hypothesis states that speakers modulate the information density of their utterance in order to optimally transmit the information at an uniform rate, near the channel capacity [Levy & Jaeger, 2007] [Frank & Jaeger, 2008] [Jaeger, 2010], based on the assumption that speech communication occurs through a noisy channel with a limited bandwidth. Thus, it is compatible with the hypothesis that human languages are organized for optimal and efficient communication in the framework of Information theory [Shannon, 1948]. This study is not directly connected with the UID hypothesis since it aims to investigate a cross-language tendency for the information rate, i.e. the average amount of information conveyed per second, among typologically diverse languages while the UID hypothesis is focused on "speakers' choices about structuring

their utterances" in order to maximize the uniformity of information density in the production of utterances. As such, "the uniform rate" mentioned in the UID hypothesis does not exactly correspond with "the stable information rate" in our result. In a strict sense, the first could be considered as "the uniform rate of information density per linguistic unit (word)[42]" while the latter refers to "the average rate of information transmission per second" using syllable as the unit of analysis.

The underlying hypothesis of this study is that human languages are self-organizing complex systems [Beckner et al., 2009] and that they exhibit a relatively stable *IR* resulted from a trade-off between *SR* and *ID* [Pellegrino, Coupé, & Marsico, 2011]. In this subsection, *IR* calculated by the paradigmatic measures of information density is compared with *IR* obtained from the syntagmatic measure. *IR* obtained by the syntagmatic measure of *ID* corresponds to the average amount of information transmitted per second on the local scale, using the oral corpus while *IR* obtained by the paradigmatic measures refers to the average amount of information transmitted per second on the global scale by estimating the syllable distribution based on a unigram or bigram language model, using the large text corpus. In Figure 2.15, *IR* obtained from the syntagmatic measure is marked in pink and the other values correspond to those obtained from the paradigmatic measures such as Shannon entropy, conditional entropy, and surprisal.

The figure above illustrates that *IR* obtained from both syntagmatic and paradigmatic measures reveal quite stable across languages, except for the following cases: Mandarin Chinese (for a relatively high syntagmatic *IR*) and Thai and Cantonese (for a relatively low paradigmatic *IR*, in particular, $IR_{H(X)}$ and $IR_{S(X)}$). In case of Mandarin Chinese, it could be related to the sociolinguistic factor since the value of *IR* increased from 0.94 to 1.15 after modifying the initial oral scripts and re-recording 10 speakers who were students at the department of linguistics at Peking University. Their utterance is faster (*SR*: 5.86) than the initial Multext corpus (*SR*: 5.18). Furthermore, in comparison with the

[42]Most of the studies were conducted at the word level, except for [Aylett & Turk, 2004] at the syllable level.

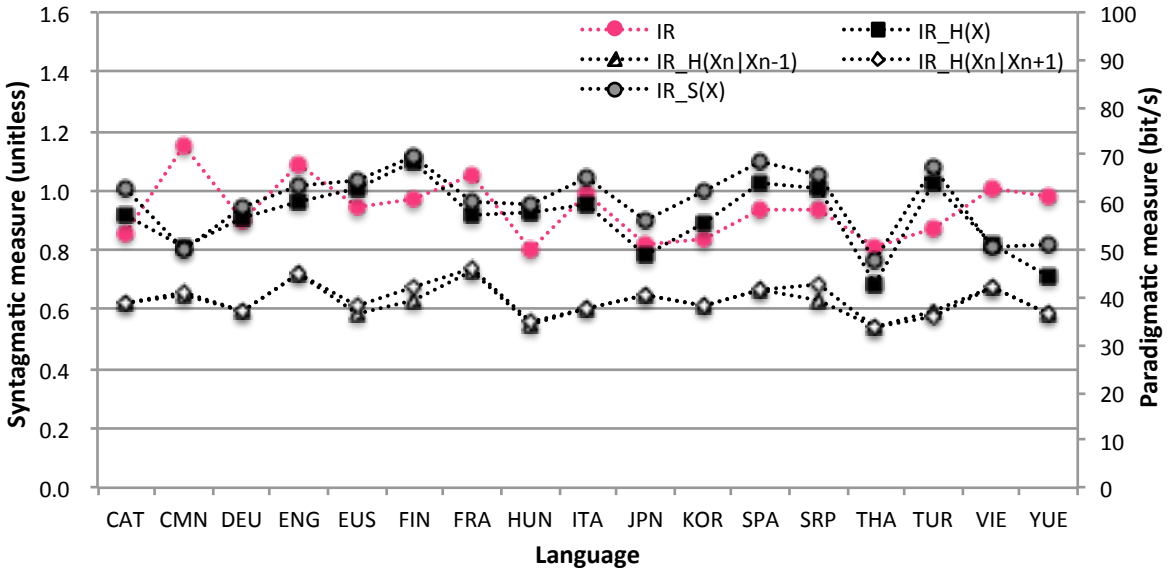

Figure 2.15: Syntagmatic vs. paradigmatic measures of *IR*. Syntagmatic measure of *IR* on the left y-axis and paradigmatic measures of *IR* on the right y-axis

data in the other languages, the size of text corpus is relatively small in Cantonese and Thai (6 000 types and 130 000 tokens in Cantonese and 5 000 types and 960 000 tokens in Thai). It is assumed that a more stable distribution of *IR* among the 17 languages can be obtained by means of a larger corpus, as the result presented in Subsection 2.3.2 displays that the values of Shannon entropy and conditional entropy start to converge after a certain number of words.[43]

In contrast with *IR* obtained by Shannon entropy and surprisal ($IR_{\mathrm{H(X)}}$ and $IR_{\mathrm{S(X)}}$), those obtained by conditional entropy ($IR_{\mathrm{H(X_n|X_{n-1})}}$ and $IR_{\mathrm{H(X_n|X_{n+1})}}$) reveal the most stable distribution of *IR* across the 17 languages and are much lower than the other two measures, $IR_{\mathrm{H(X)}}$ and $IR_{\mathrm{S(X)}}$. Hence, the two following arguments are suggested: (i) once context is taken into account, languages differ in *IR* are leveled out, (ii) the lower $IR_{\mathrm{H(X_n|X_{n-1})}}$ and $IR_{\mathrm{H(X_n|X_{n+1})}}$ can be explained in relation with some previous studies in psycholinguistics. In the previous studies by Ferrer i Cancho and his colleagues [Ferrer i Cancho & Solé, 2003] [Ferrer i Cancho, 2006] [Ferrer i Cancho & Díaz-Guilera, 2007],

[43]See [Curran & Osborne, 2002] for some counterarguments, which suggests that increasing the size of corpus does not result in a better estimation of distribution.

conditional entropy is suggested as the disambiguation effort for the hearer. On the contrary, Shannon entropy is regarded as the effort for both the speaker and the hearer involving the cognitive effort of memory and recognition. This linkage between them can be accounted by the argument of Levinson that the hearer's cognitive effort of the inference involved in the disambiguation process costs less in comparison with the speaker's effort of articulation [Levinson, 2000] and also by the argument of Piantandosi and his colleagues regarding the importance of ambiguity "as a functional property of language allowing for greater communicative efficiency" [Piantadosi, Tily, & Gibson, 2012]. As a conclusion, the lower $IR_{\mathrm{H(X_n|X_{n-1})}}$ and $IR_{\mathrm{H(X_n|X_{n+1})}}$ in comparison with $IR_{\mathrm{H(X)}}$ and $IR_{\mathrm{S(X)}}$ can be explained by the argument that the former corresponds to the hearer's effort of disambiguation while the latter refers to the efforts of both the speaker and the hearer in terms of memory and recognition process.

Table 2.19: Comparison of the AIC scores of mixed effects models

Model	Predictor	Coefficient	Sig
$\boldsymbol{IR}$ ~ SR * ID + Sex + Language + (1|Speaker) + (1|Text)	SR	0.776	***
	ID	1.191	***
	SR*ID	0.201	***
$\boldsymbol{IR_{H(X)}}$ ~ SR * ID + Sex + Language + (1|Speaker) + (1|Text)	SR	1.061	***
	ID	−0.006	***
	SR*ID	0.005	***
$\boldsymbol{IR_{H(X_n/X_{n-1})}}$ ~ SR * ID + Sex + Language + (1|Speaker) + (1|Text)	SR	1.343	***
	ID	−0.002	
	SR*ID	0.021	***
$\boldsymbol{IR_{H(X_n/X_{n+1})}}$ ~ SR * ID + Sex + Language + (1|Speaker) + (1|Text)	SR	1.305	***
	ID	0.001	
	SR*ID	0.020	***
$\boldsymbol{IR_{S(X)}}$ ~ SR * ID + Sex + Language + (1|Speaker) + (1|Text)	SR	0.991	***
	ID	0.174	***
	SR*ID	0.008	

In order to further confirm the hypothesis that the stable *IR* is resulted from a trade-off between *SR* and *ID*, the result of mixed effects models taking several values of *IR* as a dependent variable are presented in Table 2.19. According to the result, *IR* obtained

by the syntagmatic measure and $IR_{\mathrm{H(X)}}$, $IR_{\mathrm{H(X_n|X_{n-1})}}$, and $IR_{\mathrm{H(X_n|X_{n+1})}}$ acquired by the paradigmatic measures are significantly predicted by the interaction between *SR* and *ID*. In those cases, if the interaction between *SR* and *ID* is added to the model, it improves the model's fit. On the contrary, no significant correlation is found between $IR_{\mathrm{S(X)}}$ and the interaction between *SR* and *ID*. Since surprisal is kind of a hybrid measure combining both syntagmatic and paradigmatic approaches, in comparison with the other measures, it is more strongly dependent on the size of corpus and the syllable distribution estimated from the corpus. One of the limits of our study is that the size of corpus in some languages are very small and not large enough to provide a quite accurate estimation of syllable distribution. Such data-dependance is the weakness of the information-theoretic approach. It appears that surprisal can reflect both local and global scales of the analysis and fits in well with language-specific studies in psycholinguistics but it may not be suitable for a cross-language study with the corpora of various sizes. Nevertheless, since a similar pattern is found between $IR_{\mathrm{H(X)}}$ and $IR_{\mathrm{S(X)}}$ among the 17 languages in Figure 2.15[44], if the size of data is very small, the average value of surprisal (i.e. Shannon entropy) can be used instead in typological comparative studies.

The three information-theoretic measures are compared with one syntagmatic measure in this subsection. In conclusion, our result suggests that (i) among the 17 languages, *IR* is significantly predicted by the interaction between *ID* and *SR*, and that (ii) both syntagmatic (local scale) and paradigmatic (global scale) measures of *IR* are corpus-dependent and yield a similar result that *IR* remains quite stable across the languages. Within the framework of language as a complex adaptive system [Beckner et al., 2009], the result of this chapter supports the argument that language is structured by the phenomenon of self-organization at the macrosysteic level, which results from the cognitive efficiency and the optimization during language learning and speech communication.

[44]It is confirmed by a strong positive correlation between them (Pearson's $r = 0.942^{**}$; p-value < 0.001, Spearman's $\rho = 0.941^{**}$; p-value < 0.001; N = 17).

2.4.3 Conclusion

In this chapter, the effect of context is observed by comparing the information-theoretic measures, i.e. Shannon entropy, conditional entropy, and surprisal, as conditional entropy better predicts *IR* than Shannon entropy and surprisal which do not consider the context. The hybrid measure, i.e. surprisal, is more data-sensitive than other measures and it may not be suitable for the data which is not large enough to estimate an accurate syllable distribution. A relatively stable *IR* is obtained by means of the syntagmatic and paradigmatic measures of information density and it allows us to assume that the phenomenon of self-organization exists at the macroscopic level of linguistic analysis. In the next chapter, this phenomenon will be assessed at the mesosystemic level by correlating the linguistic complexity at the morphological and phonological modules.

Chapter 3

Mesosystemic relationship between morphology and phonology

3.1 Introduction

3.1.1 Holistic typology and equal overall complexity

Chaque langue forme un système où tout se tient... [Meillet, 1915].

In the present study, a language is defined as a macrosystem consisting of microsystems (i.e. linguistic modules such as syntax or phonology). The notion *mesosystem* refers to the interactions between those microsystems [Bronfenbrenner, 1979] and the aim of this chapter is to assess the mesosystemic relationship between linguistic modules which differ in level of representation. The previous study in Chapter 2 presented negative correlations between information density and speech rate and interpreted a limited range of information rate as a result of trade-off between information density and speech rate. In this chapter, the equal overall complexity hypothesis (or equi-complexity, in Kuster's terminology [Kusters, 2003]) is evaluated at the mesosystemic level, by means of multilingual text and oral corpora in 14 typologically diverse languages (Basque, Cantonese, Catalan, British English, French, German, Hungarian, Italian, Japanese, Korean, Mandarin Chi-

nese, Spanish, Turkish, and Vietnamese).

According to the equal complexity hypothesis, all languages are considered equal in terms of overall complexity as depicted by Hockett: "the total grammatical complexity of any language, counting both morphology and syntax, is about the same as any other" [Hockett, 1958]. Joseph and Newmeyer described how the equal complexity hypothesis became an indisputable consensus in the mid-twentieth century [Joseph & Newmeyer, 2012]. The supporting arguments are summed up in the three following points:

i) Humanism: the equal complexity hypothesis was employed as a counterargument in respect to the race and culture superiority. The underlying implication is that language was identified as culture and that complexity was regarded as a kind of hierarchy.

ii) Language processing: the constraints on the use of language balance out the overall complexity (law of compensation).

iii) Universal grammar: Chomsky's idea of innate and universal grammar implies that languages are comparable [Chomsky, 1959].

However, since the end of the 20^{th} century until recently, many linguists - especially in sociolinguistics and typology - have demonstrated the weakness of the equal overall complexity hypothesis and expressed their skepticism. Among them, in sociolingusitics, the simplicity of creole grammars in comparison with non-creole grammars has been argued in depth in a special issue of the *Linguistic typology, vol. 5* (cf. [DeGraff, 2001] [McWhorter, 2001]). In typology, Shosted suggested that no significant correlation was found between the number of potential syllables and the number of verbal inflectional markers in 32 languages and asserted that there was no previous work which attempted to demonstrate the negative correlations between linguistic modules using a large number of world's languages [Shosted, 2006]. In the same vein, Fenk-Ozclon and Fenk presented several studies on complexity trade-off using a quantitative approach [Fenk-Oczlon & Fenk, 1999, 2004, 2005, 2006, 2014]. In [Fenk & Fenk-Oczlon, 2006], they combined morphological, phonological and syntactic measures to assess the cross-language variation patterns among the

two groups of languages divided by their rhythmical structure: syllable-timed vs. stress-timed rhythm, using both metric and non-metric variables. Contrary to [Shosted, 2006], their results showed significant negative correlations between linguistic modules such as phonology, morphology and syntax. Nevertheless, they claimed that the negative correlations and the self-organizing trade-off do not provide any convincing evidence towards the equal overall complexity hypothesis.

The equal overall complexity hypothesis has been challenged by many sociolinguists and typologists for the following reasons:

i) It is almost impossible to define and quantify the overall complexity of language (i.e. holistic typology).

ii) This hypothesis can easily be falsified by finding a counterexample due to the diversity of world's languages.

iii) A common problem with a large-scaled data concerns the likelihood of getting a spurious correlation [Roberts & Winters, 2013].

In the paper *The co-variation of phonology with morphology and syntax: A hopeful history* [Plank, 1998], the author enumerated a list of works conducted from a holistic perspective.

> In recent times, typologists have often confined themselves to seeking dependencies among variable language-parts WITHIN syntax, WITHIN morphology, or WITHIN phonology. As to dependencies BETWEEN levels or modules, syntax and morphology were considered essentially the only candidates showing some real typological promise [Plank, 1998].

While modern typological studies were mainly focused on comparing specific *INTER-level* grammatical properties, for example, word order or inflectional morphology, there were few studies whose goal was to analyze systemic dependencies between linguistic levels as listed in [Plank, 1998]. Comparing variations within a part of grammar is the mainstream of current typological studies and corresponds to *partial typology*. The other

type of typology is defined as *holistic typology* (or *systemic* using the terminology of [Fenk & Fenk-Oczlon, 2006]). Holistic typology gained its popularity in the 19th century with the influence of natural science but gave its place to partial typology in the 20th century [Song, 2014] for the very similar reasons which brought on the fall of the equal overall complexity hypothesis.

The equal overall complexity hypothesis and holistic typology are complementary to each other and share the following fundamental research question: a problem of defining and quantifying linguistic parameters which describe a language as a whole system from a functional and systemic perspective. While it seems feasible to provide a corroborating evidence of the *overall equal "communicative" complexity* [Pellegrino, Coupé, & Marsico, 2011] by means of the limited range of information rate on the macrosystemic level as discussed in the previous chapter, addressing the same hypothesis on the mesosystemic level is more complex due to the difficulty of defining a null hypothesis (i.e. the overall complexity) and there are still ongoing discussions regarding this issue [Fenk-Oczlon & Fenk, 2014] [Shosted, 2006]. Hence, the present work tries to approach the equal overall complexity hypothesis with great caution and attempts to present a preliminary result of the empirical observation in the 14 languages.

3.1.2 Quantifying linguistic complexity

The main reason of quantifying linguistic complexity is to compare languages. From Chomsky's Universal Grammar [Chomsky, 1959] to the evolutionary linguistics, there are linguistic features considered *comparable* across languages and there is few doubt about the importance of quantifying linguistic complexity as demonstrated by the discussion among linguists in the previous section. A list of questions that researchers from various fields frequently ask regarding how to quantify complexity is provided by a physicist Lloyd [Lloyd, 2001] as follows:

i) Difficulty of description (measured in bits)

ii) Difficulty of creation (measured in time, energy, etc.)

iii) Degree of organization: a) difficulty of describing organizational structure

b) information shared between the parts of a system

In historical and comparative linguistics, several studies proposed *a stability metric* (measuring a change rate) of typological linguistic features to compare languages from a diachronic perspective [Croft, 1996] [Greenberg, 1978] [Nichols, 1995] [Sapir, 1970] [Wichmann & Holman, 2009]. In those studies, some linguistic modules are assumed to be more prone to change than others (for example, phonology changes faster than morphology [Sherard, 1985, p.199], syntax is more stable than morphology [Mithun, 1984]), due to different levels of representation.

In addition to the *comparability* of languages and the *stability* of linguistic features, there arises also the question of *opacity* or *clearness* of features. In their work on the morphological complexity, [Bane, 2008] and [Juola, 1998] asserted that morphology is a good starting point for complexity computation for its clearness, compared to other more abstract and higher levels such as syntax and semantics. This explains the fact that there are more works on morphological complexity than phonological, semantic, and syntactic complexity. In his paper *Quantitative approach to morphological typology of language*, Greenberg employed the term *complexity* while referring to one of the criteria for morphological distinction defined by [Sapir, 1970].

> One such axis distinguished by Sapir may be said to relate to the gross *complexity* of the word, i.e., the degree of *complexity* exhibited on the basis of the number of subordinate meaningful elements it contains. The terms employed here by Sapir are *analytic*, *synthetic*, and *polysynthetic*, in ascending order of *complexity*. [Greenberg, 1960, p.182]

The same definition of *complexity* is still applied to the recent studies on linguistic complexity in which the most frequently used method of quantifying *complexity* is to calculate the number of constituents of the linguistic system at issue [Bane, 2008] [McWhorter,

2001] [Moscoso del Prado, 2011] [Nichols, 2007] [Shosted, 2006]. In addition to the previous traditional linguistic measure, Information-theory based measure is also used to compute linguistic complexity ([Ackerman & Malouf, 2013] [Blevins, 2013] [Goldsmith, 2000] [Goldsmith, 2001] [Goldsmith, 2002] [Juola, 1998] [Kello & Beltz, 2009] [Kostić, 1991] [Moscoso del Prado, Kostić, & Baayen, 2004] [Moscoso del Prado, 2011] [Pellegrino, Coupé, & Marsico, 2007] [Pellegrino, Coupé, & Marsico, 2011] [Villasenor et al., 2012] and many others). In this study, both traditional and information-theoretical measures are used to account for interactions between morphological and phonological modules. On the one hand, small and large corpora are used to compute phonological complexity by means of metric variables from both information-theoretic and traditional approaches. On the other hand, morphological complexity is calculated using both metric and non-metric variables from the traditional grammar-based method.

Dahl distinguished two notions of linguistic complexity in *The Growth and Maintenance of Linguistic Complexity* [Dahl, 2004].

> Given that a language as a system can be seen as involving both resources and regulations, it follows that a language could be characterized as more or less complex with respect to both these notions [Dahl, 2004, p.42].

The first notion of linguistic complexity regards language as a system (*system complexity*) and measures the "richness" of a system in terms of its *resources*. The second notion applies to the structure of expressions (*structural complexity*). This distinction of linguistic complexity accounts for the differences between the methodologies used to measure morphological and phonological complexity. According to Dahl, system complexity could be measured at the phonological level and structural complexity could be calculated at the morphological level of analysis, but not exclusively. Most of the previous large-scaled cross-language studies on the correlation between different linguistic modules ([Dahl, 2004] [Fenk-Oczlon & Fenk, 2004] [Fenk & Fenk-Oczlon, 2006] [Shosted, 2006]) were only focused on measuring the system complexity even at the morphological level. However,

since morphology investigates the structure and form of words, it should be crucial to take morphological coding strategies, i.e. structural complexity, into account for measuring morphological complexity.

Ackerman and Malouf distinguished two levels of analysis in investigating morphological complexity: Enumerative complexity (E-complexity) and Integrative complexity (I-complexity) [Ackerman & Malouf, 2013]. E-complexity denotes morphological coding strategies for both the internal structure of word and the global organization of inflectional system. On the contrary, I-complexity is based on an information-theoretic approach where the cost of learning inflectional grammar is taken into account by means of the average conditional entropy of individual paradigm cell. The authors suggested that languages which differ in E-complexity can exhibit similar patterns of I-complexity, i.e. "low conditional entropy among (patterns of) words" [Ackerman & Malouf, 2013, p.454]. The measure of morphological complexity used in our study is more in line with E-complexity which takes inflectional morphological strategies into account, without considering speaker's learning effort.

The methodology used for calculating morphological complexity in this study was first proposed by [Lupyan & Dale, 2010] and later also presented in [Nettle, 2012] in sociolinguistics. Lupyan and Dale calculated the morphological complexity score by distinguishing two different coding strategies, i.e. lexical versus morphological strategies (see Section 3.2.3 for details).

To sum up several notions and distinctions presented in this section, the method for quantifying linguistic complexity differs as a function of linguistic module in question and perspective: bottom-up or usage-based approach to reflect phonological complexity and on the contrary, top-down or grammatical approach to assess morphological complexity.

3.1.3 Chapter outline

In Section 3.2, the measures for quantifying phonological and morphological complexity are introduced and then, the 14 languages are classified based on the morphological typology illustrated in [Sapir, 1970] along with the description of the data.

Section 3.3 displays a number of cross-language correlations between morphological and phonological complexity, extending the previous result presented in [Oh et al., 2013]. Furthermore, the groups of languages classified according to morphological typology [Sapir, 1970] are compared. First of all, the correlations among speech rate, information density, information rate, and linguistic complexity are investigated in the 14 languages in Sections 3.3.1 and 3.3.2. Second, two measures of phonological complexity are compared, i.e. Shannon entropy and conditional entropy, in terms of their trade-off relationship with morphological complexity in Section 3.3.3. Third, some general tendencies among the languages of the same morphological group (in particular, agglutinative and fusional languages) are analyzed in Section 3.3.4. Finally, the effect of word order (i.e. Subject-Verb-Object versus Subject-Object-Verb) on morphological and phonological modules is assessed in Section 3.3.5.

In the discussion section (3.4), the results are interpreted as a supporting evidence for the equal overall complexity hypothesis from functional and cognitive perspectives and lead to a conclusion that the equal overall complexity hypothesis should not be considered as a mere oversimplification but rather as a cognitive optimization. To support this view, the importance of sociolinguistic and neurocognitive factors which come into play and interact with linguistic factors during the process of language evolution is emphasized.

3.2 Method, language, and data description

3.2.1 Measures of WID and SID

Three parameters, *SR*, *ID*, and *IR* were described previously in Section 2.2.2.1. In Chapter 2, syllable is used as the basic unit of analysis but in the present chapter, the two following parameters take both word and syllable as the unit of analysis: i) the average length of unit, i.e. *WC* and *SC*, and ii) the average amount of information per unit, i.e. *WID* and *SID*.

In order to account for the two parameters, *WID* and *SID*, the average information conveyed per word (*WI*) or per syllable (*SI*) is defined as the division of the semantic content of text t in language L (S_L^t) by the number of its constituents, either words (w_L^t) or syllables (σ_L^t), the latter being identical to the information density (*ID*) considered in the previous chapter.

$$WI_L^t = \frac{S_L^t}{w_L^t} \qquad SI_L^t = \frac{S_L^t}{\sigma_L^t} \tag{3.1}$$

Information density (*ID*) is computed at both word and syllable levels respectively to assess general trade-off tendencies among the morphologically classified languages (see Section 3.2.3 for the description of classification).

$$WID_L = \frac{1}{T}\sum_{i=1}^{T}\frac{WI_L^t}{WI_{VIE}^t} = \frac{1}{T}\sum_{i=1}^{T}\frac{S_L^t}{w_L^t} \times \frac{w_{VIE}^t}{S_{VIE}^t} = \frac{1}{T}\sum_{i=1}^{T}\frac{w_{VIE}^t}{w_L^t} \tag{3.2}$$

$$SID_L = \frac{1}{T}\sum_{i=1}^{T}\frac{SI_L^t}{SI_{VIE}^t} = \frac{1}{T}\sum_{i=1}^{T}\frac{S_L^t}{\sigma_L^t} \times \frac{\sigma_{VIE}^t}{S_{VIE}^t} = \frac{1}{T}\sum_{i=1}^{T}\frac{\sigma_{VIE}^t}{\sigma_L^t} \tag{3.3}$$

Since $S_L^t = S_{VIE}^t$ (see Section 2.2.2.1 for a detailed description of equation), word information density (*WID*) and syllable information density (*SID*) are computed by a pairwise comparison of the number of words or syllables of text t in Vietnamese (w_{VIE}^t, σ_{VIE}^t) and in a target language (w_L^t, σ_L^t) respectively.

3.2.2 Measures of phonological complexity

Two measures of syllable complexity were presented previously in Section 2.2.2.2, which were considered as the most common measures of linguistic complexity. In addition to SC_{TYPE} and SC_{TOKEN}, two measures of word complexity, i.e. WC_{TYPE} and WC_{TOKEN}, are used in this chapter.

$$WC_{TYPE} = \frac{1}{N_L}\sum_{i=1}^{N_L}\frac{\sigma_i}{w_i} \qquad WC_{TOKEN} = \frac{1}{N_L}\sum_{i=1}^{N_L} p_{w_i}\frac{\sigma_i}{w_i} \tag{3.4}$$

WC_{TYPE} corresponds to the mean number of syllables (σ_i) per word (w_i) where language L is considered as a system consisting of finite set of N words while WC_{TOKEN} is computed from an usage-based approach where each average number of syllables per word is weighted by the relative frequency of corresponding linguistic units (p_w) in a large text corpus.

In addition to the traditional measures of SC and WC, the following two measures of phonological complexity are employed in this section: Shannon entropy $H(X)$ and conditional entropy $H(X_n|X_{n-1})$ and $H(X_n|X_{n+1})$ (see Section 2.2.2.3 for a detailed description of the equations). These two information-theoretic measures are considered as a "measure of *complexity* of an analysis" [Goldsmith, 2000] which allows us to compare the complexity of phonological system of languages.

3.2.3 Measures of morphological complexity

Inflectional morphology can vary considerably across languages and is defined as "an effective tool for complexity reduction" which optimizes the grammar of language by "reducing uncertainty and simplifying the description of whole grammar" [Ackerman & Malouf, 2013] [Moscoso del Prado, 2011]. In this study, the measure of morphological complexity is adopted from the methodology proposed in [Lupyan & Dale, 2010]. In their paper *Language structure is partly determined by social structure*, Lupyan and Dale chose

28 linguistic features[45] accounting for the inflectional morphology from WALS (World Atlas of Language Structures) [Dryer & Haspelmath, 2013]. The score of morphological complexity was calculated by dichotomically distinguishing between lexical and inflectional morphological coding strategies and summing assigned values (-1 for lexical and 0 for morphological strategies) to the 29 linguistic features displayed in Table 3.1.

Table 3.1: Measure of morphological complexity. Features are chosen and classified following [Lupyan & Dale, 2010] with descriptions taken from WALS [Dryer & Haspelmath, 2013]

Feature (WALS code)	Description
Morphological type	
Fusion of selected inflectional formatives (20A)	The degree to which grammatical markers (*formatives*) are phonologically connected to a host word or stem
Prefixing vs. suffixing (26A)	The degree to which languages use prefixes or suffixes in their inflectional morphology
Cases	
Number of cases (49A)	The number of case categories represented in a language's inflectional system
Case syncretism (28A)	The ways in which a single inflected form represents two or more case functions
Alignment of case marking of full noun phrases (98A)	The ways in which core argument noun phrases are marked to indicate which particular core argument position they occupy
Verb morphology	
Inflectional synthesis of the verb (22A)	The strategies of expressing grammatical categories either by individual words or by affixes attached to some other words
Alignment of verbal person marking (100A)	The ways in which the two arguments of the transitive verb aline with the sole argument of the intransitive verb
Agreement	
Person marking on verbs (102A)	The number and identity of the arguments of a transitive clause which display person marking on the verb
Person marking on adpositions (48A)	The strategies of person marking used to relate an object to another nominal or verbal constituent on the basis of a more or less specific semantic relationship

[45]It should be noted that two features, Definite articles (37A) and Indefinite articles (38A), are considered together as one linguistic feature in [Lupyan & Dale, 2010] but are separately taken into account in the present study.

Table 3.1: Measure of morphological complexity. Features are chosen and classified following [Lupyan & Dale, 2010] with descriptions taken from WALS [Dryer & Haspelmath, 2013] (continued)

Feature (WALS code)	Description
Syncretism in verbal person/number marking (29A)	The ways in which multiple person values underlie a single form in the inflectional marking of subject person in verbs
Possibility and evidentials	
Situational possibility (74A)	The strategies used to express situational possibility in positive main clauses
Epistemic possibility (75A)	The strategies used to express epistemic possibility in positive main clauses
Overlap between situational and epistemic modal marking (76A)	The extent to which languages have identical markers for situational and epistemic modality
Semantic distinctions of evidentiality (77A)	The presence of grammatical markers of evidentiality which express the evidence a speaker has for his/her statement
Negation, plurality, interrogatives	
Negative morphomes (112A)	The nature of morphemes signaling clausal negation in declarative sentences
Occurrence of nominal plurality (34A)	The extent to which plural markers on full nouns are used in a language
Associative plural (36A)	It consists of a noun X and some other materials referring to 'X and other people associated with X'.
Position of polar question particles (92A)	The position of question particles in polar questions (questions that elicit the equivalent of a 'yes' or 'no' response)
Tense, possession, aspect, mood	
Future tense (67A)	The distinction between languages with and without inflectional marking of future time reference
Past tense (66A)	The ways in which past/non-past distinction is marked grammatically
Perfective/Imperfective aspect (65A)	The distinction between languages with and without the perfective/imperfective grammatical marking
Morphological imperative (70A)	The extent to which languages have second person singular and plural imperatives as dedicated morphological categories
Position of pronominal possessive affixes (57A)	The distinction between languages with and without possessive suffixes and prefixes on noun
Possessive classification (59A)	The forms of possessive marking whose choice is conditioned lexically by the possessed noun

Table 3.1: Measure of morphological complexity. Features are chosen and classified following [Lupyan & Dale, 2010] with descriptions taken from WALS [Dryer & Haspelmath, 2013] (continued)

Feature (WALS code)	Description
Optative (73A)	An inflected verb form dedicated to the expression of the wish of the speaker
Articles, demonstratives, pronouns	
Definite articles (37A)	A morpheme which accompanies nouns and codes definiteness or specificity
Indefinite articles (38A)	A morpheme which accompanies a noun and signals that the noun phrase denotes something not known to the hearer
Distance contrasts in demonstratives (41A)	The ways in which deictic expressions indicating the relative distance of a referent in the speech situation vis-à-vis the deictic center are marked
Expression of pronominal subjects (101A)	The ways in which a pronominal subject is expressed by a morpheme or morphemes coding semantic or grammatical features of the subject.

The relevant information for each linguistic feature is, for the most part, taken from WALS. However, WALS does not provide all the information regarding the features presented in Table 3.1 and in such cases, the missing information was completed by the author. This task was feasible due to much smaller number of languages in comparison with 2 236 languages analyzed in [Lupyan & Dale, 2010] where the complexity score was solely obtained from the information provided in WALS if a language had relevant description for at least 3 linguistic features. The complexity score was then calculated by dividing the overall score by the proportion of available linguistic features.

The features are distinguished into two types of variables: metric (quantitative) and non-metric (categorical or qualitative) variables. The measure applied in this paper differs from the method used in [Lupyan & Dale, 2010] in a way that the latter converted non-metric, categorical variables with multiple values into dichotomous variables by assigning two possible values for each feature, -1 for lexical and 0 for inflectional morphological strategy. On the contrary, some features are considered as continuous variables in this study. To reflect the quantitative variables, such as the number of case categories (49A)

and the number of grammatical categories expressed by the inflectional synthesis of the verb (22A), all the values are normalized between 0 and -1, including those attributed to continuous variables. Taking normalized values of continuous variables into account is assumed to better represent the degree of morphological complexity since they specify the evaluation criteria.

3.2.4 Language and data description

Oral and textual corpora in 14 typologically diverse languages are used to measure phonological complexity while no additional data is required for measuring morphological complexity. Regarding the oral corpora, the data of 3 languages (British English, German, and Italian) are taken from the Multext (Multilingual Text Tools and Corpora) project [Campione & Véronis, 1998] and the data of 11 languages (Basque, Cantonese, Catalan, French, Hungarian, Japanese, Korean, Mandarin Chinese, Spanish, Turkish, and Vietnamese) are collected by the author and her colleague Christophe Coupé (see section 2.2.1.1 for more description of the oral corpora). Text corpora used for computing Shannon entropy and conditional entropy were acquired mostly online from various sources and the relevant information regarding the data was previously described in Table 2.1 as well as the preprocessing steps for each data in Section 2.2.1.2.

Table 3.2 provides the morphological types of each language which was classified by the traditional morphological typology. In his book *Language*, Sapir proposed a morphological classification of languages based on five parameters [Croft, 2002] [Greenberg, 1960] [Sapir, 1970]. Among them, the following two parameters (*synthesis* and *technique*, according to Sapir's terminology) are employed to classify the 14 languages in this study.

(i) Degree of synthesis (i.e. number of morphemes per word)

  (a) Analytic - one morpheme per word

  (b) Synthetic - a small number of morphemes per words

  (c) Polysynthetic - a large number of morphemes per words

(ii) Degree of morphophonemic alternation (i.e. how elements are related)

(a) Isolating - no affixes and no modification of elements

(b) Agglutinative - simple and transparent affixation

(c) Fusional - morphophonemic alternation and complex affixation

(d) Symbolic - internal changes of the radical element

Table 3.2: Morphological classification

Morphological type	**Abbreviation**	**Languages**
Analytic/Isolating	AI	CMN, VIE, YUE
Analytic/Fusional	AF	ENG
Synthetic/Agglutinative	SA	EUS, HUN, JPN, KOR, TUR
Synthetic/Fusional	SF	CAT, DEU, FRA, ITA, SPA

Among 12 possible combinations obtained from the parameters, the 14 languages are classified into 4 groups: *Analytic/Isolating*, *Analytic/Fusional*, *Synthetic/Agglutinative*, and *Synthetic/Fusional*, as displayed in Table 3.2. In the next section, cross-language patterns of variation among the 4 types of languages divided according to their morphological coding strategies are compared and general tendencies within each group are investigated.

3.3 Cross-language correlations of linguistic complexity

3.3.1 Speech rate, information density, and linguistic complexity

This section aims to investigate correlations among *SR*, *ID*, and morphological and phonological complexity. The most common and traditional measure of phonological complexity is the average number of components per unit. 4 types of system complexity measures were previously described: WC_{TYPE}, WC_{TOKEN}, SC_{TYPE}, and SC_{TOKEN}. As depicted in Figure 3.1, a significant negative correlation (Pearson's $r = -0.894$**; p-value < 0.001; Spearman's $\rho = -0.789$**; p-value $= 0.001$; N $= 14$) is found between WC_{TYPE} and SC_{TYPE}. This negative correlation can be interpreted as a phenomenon of compensation between word length and number of phonemes and tones per syllable, which refers to

Menzerath's law [Altmann, 1980] [Fenk, Fenk-Oczlon, & Fenk, 2006].

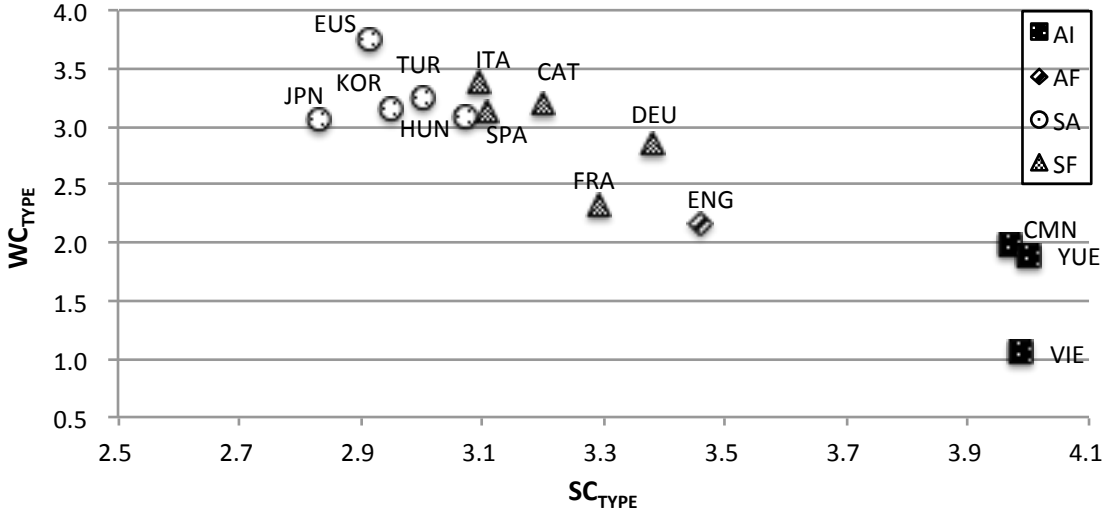

Figure 3.1: SC_{TYPE} (average number of segments per syllable) on x-axis and WC_{TYPE} (average number of syllables per word) on y-axis

If WC and SC values are weighted by relative frequency, a weaker negative correlation is found between WC_{TOKEN} and SC_{TOKEN} (Pearson's $r = -0.681^{**}$; p-value = 0.007; Spearman's $\rho = -0.572^{*}$; p-value = 0.033; N = 14). In Figure 3.1, the languages classified as the same morphological group are clustered together, exhibiting a similar pattern between each other. Since WC and SC are computed on the 20 000 most frequent words in each language, when frequency effect is taken into account, the complexity values decrease as demonstrated by ΔW and ΔS in Table 3.3. This can be explained by the fact that high-frequency words tend to be shorter [Bell et al., 2009] [Zipf, 1949].

On average, SF languages exhibit the largest gap between WC_{TYPE} and WC_{TOKEN} followed by SA, AF, and AI languages in decreasing order. The distinction between synthetic and analytic languages is illustrated by such a pattern. At the syllable level, the largest gap between SC_{TYPE} and SC_{TOKEN} exists in SF languages followed by AF, SA, and AI languages on average. Contrary to the word level, the result can be associated with the distinction between isolating, agglutinative, and fusional languages. Therefore, it appears that the morphological synthesis is more reflected at the word level while the

morphophonemic alternation is more related to the syllable level.

Table 3.3: Word complexity (WC_{TYPE} and WC_{TOKEN}) and difference between WC_{TYPE} and WC_{TOKEN} (ΔW), Syllable complexity (SC_{TYPE} and SC_{TOKEN}) and difference between SC_{TYPE} and SC_{TOKEN} (ΔS). The maximum and minimum values are marked in green and blue.

Group	**AI**			**AF**	**SA**					**SF**				
Language	cmn	vie	yue	eng	eus	hun	jpn	kor	tur	cat	deu	fra	ita	spa
WC_{TYPE}	1.98	1.06	1.90	2.17	3.74	3.07	3.06	3.15	3.24	3.19	2.86	2.32	3.38	3.13
WC_{TOKEN}	1.48	1.00	1.22	1.40	2.76	2.06	1.97	2.53	2.55	1.90	1.74	1.40	2.09	1.92
ΔW	0.5	0.06	0.68	0.78	0.98	1.02	1.09	0.62	0.69	1.30	1.12	0.92	1.29	1.21
SC_{TYPE}	3.97	3.99	4.00	3.46	2.92	3.07	2.83	2.95	3.00	3.20	3.38	3.29	3.09	3.11
SC_{TOKEN}	3.69	3.89	3.70	2.50	2.06	2.33	2.04	2.39	2.35	2.25	2.59	2.14	2.23	2.29
ΔS	0.28	0.10	0.30	0.96	0.85	0.75	0.79	0.56	0.65	0.96	0.79	1.15	0.87	0.82

Table 3.4: Correlations among *SR*, *WID* (word information density), *SID* (syllable information density), and linguistic complexity (*MC* denotes morphological complexity)

Parameter	**SR (N=14)**	**WID (N=14)**	**SID (N=14)**
WC_{TYPE}	r= 0.767**; p= 0.001 ρ= 0.754**; p= 0.002	r= 0.534*; p= 0.049 ρ= 0.486; p= 0.078	r= −0.870**; p< 0.001 ρ= −0.709**; p= 0.004
WC_{TOKEN}	r= 0.649*; p= 0.012 ρ= 0.660*; p=0.010	r= 0.619*; p= 0.018 ρ= 0.768**; p= 0.001	r= −0.773**; p= 0.001 ρ= −0.692**; p= 0.006
SC_{TYPE}	r= −0.828**; p< 0.001 ρ= −0.776**; p= 0.001	r= −0.626*; p= 0.017 ρ= −0.770**; p= 0.001	r= 0.934**; p< 0.001 ρ= 0.838**; p< 0.001
SC_{TOKEN}	r= −0.813**; p< 0.001 ρ= −0.851**; p<0.001	r= −0.446; p= 0.110 ρ= −0.361; p= 0.205	r= 0.849**; p< 0.001 ρ= 0.683**; p= 0.007
H(X)	r= −0.659*; p= 0.010 ρ= −0.532; p= 0.050	r= −0.629*; p= 0.016 ρ= −0.281; p= 0.331	r= 0.517; p= 0.059 ρ= 0.517; p= 0.058
$H(X_n|X_{n-1})$	r= −0.818**; p< 0.001 ρ= −0.859**; p< 0.001	r= −0.637*; p= 0.014 ρ= −0.620*; p= 0.018	r= 0.907**; p< 0.001 ρ= 0.796**; p= 0.001
$H(X_n|X_{n+1})$	r= −0.816**; p< 0.001 ρ= −0.847**; p< 0.001	r= −0.632*; p= 0.015 ρ= −0.654*; p= 0.011	r= 0.911**; p< 0.001 ρ= 0.808**; p< 0.001
MC	r= 0.655*; p= 0.011 ρ= 0.607*; p= 0.021	r= 0.254; p= 0.381 ρ= 0.203; p= 0.486	r= −0.731**; p= 0.003 ρ= −0.564*; p= 0.036

Table 3.4 recapitulates the correlations among *SR*, *ID*, and linguistic complexity. In terms of *SR*, a language with either more syllables per word or less segments (and tones if applicable) per syllable is assumed to be faster as *WC* and *SC* are in a negative relationship. A strong negative correlation is observed between *SR* and conditional entropy,

indicating that a language is spoken faster if the average amount of uncertainty obtained by means of its contextual information is lower.

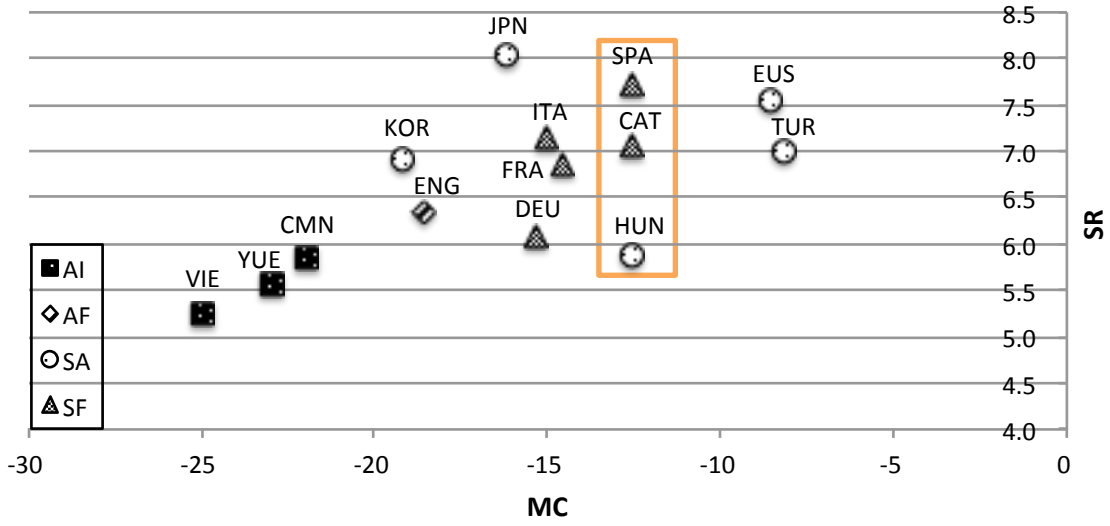

Figure 3.2: Morphological complexity (unitless) on x-axis and *SR* (average number of syllables uttered per second) on y-axis

A higher morphological complexity score (closer to 0) means that a language employs more inflectional morphological strategies whereas a lower score denotes that lexical strategies are preferred. A positive correlation exists between *SR* and morphological complexity (Pearson's $r = 0.655$*; p-value $= 0.011$; Spearman's $\rho = 0.607$*; p-value $= 0.021$; N $= 14$). However as displayed in Figure 3.2, if AI languages (Cantonese, Mandarin Chinese, and Vietnamese) are left aside, no correlation is found between them (Pearson's $r = 0.199$; p-value $= 0.557$; Spearman's $\rho = 0.183$; p-value $= 0.589$; N $= 11$). For instance, Spanish, Catalan, and Hungarian exhibit the same value of morphological complexity (-12.55) but in terms of *SR*, they vary from 5.87 (Hungarian) to 7.71 (Spanish). It is shown that analytic languages (AI and AF) favor lexical strategies over inflections compared to synthetic languages (SA and SF). Furthermore, the degree of inflection varies substantially within SA languages compared to the other types of languages. In the figure, languages can be divided into two types: SA and SF vs. AI and AF. The former, i.e. synthetic language, reveals high *SR* and high morphological complexity whereas the

latter, i.e. analytic language, exhibits low *SR* and low morphological complexity. Despite the general trend, German and Hungarian show relatively slow *SR* compared to the other synthetic languages. German displays the highest average number of segments per syllable (SC_{TYPE}= 3.38) among SF languages and Hungarian also reveals the highest average number of segments per syllable (SC_{TYPE} = 3.07) among SA languages in Figure 3.2, which may have influence on their relatively slow *SR*.

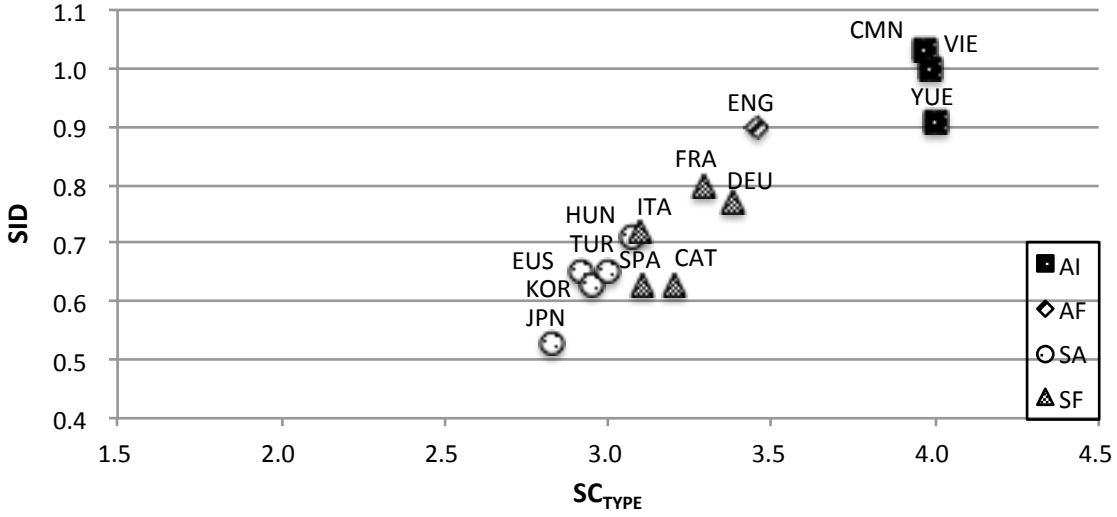

Figure 3.3: SC_{TYPE} (average number of segments (and tones, if applicable) per syllable) on x-axis & *SID* (average amount of information per syllable, unitless) on y-axis

Regarding *SID* and SC_{TYPE}, 4 groups display an almost linear behavior in Figure 3.3 (Pearson's $r = 0.934$**; p-value < 0.001; Spearman's $\rho = 0.838$**; p-value < 0.001; N = 14). Contrary to *SR* and *SID* where the ranges of SA and SF languages overlap, their ranges distinctively vary in terms of SC_{TYPE}, which reflects a distinction between agglutination and fusion (cf. 3.3.4). Furthermore, a clear distinction exists between analytic and synthetic languages, i.e. AI and AF vs. SA and SF. It is observed that analytic languages tend to encode more amount of information per syllable by means of more complex or longer syllables than synthetic languages.

Although a small set of languages analyzed in this study is not sufficient for drawing any typological generalization, the results may trigger further typological studies from

quantitative approaches. Moreover, the languages are classified based on the traditional morphological typology which has been criticized in modern theoretical linguistics since 20^{th} century. Nevertheless, the results presented in this study tries to demonstrate that such a classification is meaningful and can be applied to typological studies.

3.3.2 Information rate and linguistic complexity

IR is measured by a pairwise ratio between the mean duration of Vietnamese and a target language and denotes the amount of information conveyed per second. Thus the information related to the linguistic organization of the language (e.g. the number of words or syllables) is not considered in the calculation of *IR*, which distinguishes it from the other measures, i.e. *SR* and *ID*. This subsection investigates the relation between *IR* and linguistic complexity. While *SR* and *ID* display a wide range of variation (cf. Figure 3.2 for *SR* and Figure 3.3 for *ID*, respectively), *IR* exhibits a relatively narrow range of variation as shown in Figure 3.4. This relative "consistency" or "stability" of *IR* is viewed as the result of self-organization between *SR* and *ID* [Pellegrino, Coupé, & Marsico, 2011].

Table 3.5: Correlations between *IR* and linguistic complexity

Parameter	**IR (N=14)**
$\mathbf{WC_{TYPE}}$	$r = -0.581$*; $p = 0.029$, $\rho = -0.484$; $p = 0.079$
$\mathbf{WC_{TOKEN}}$	$r = -0.570$*; $p = 0.033$, $\rho = -0.586$*; $p = 0.028$
$\mathbf{SC_{TYPE}}$	$r = 0.659$*; $p = 0.010$, $\rho = 0.686$**; $p = 0.007$
$\mathbf{SC_{TOKEN}}$	$r = 0.494$; $p = 0.072$, $\rho = 0.350$; $p = 0.220$
$\mathbf{H(X)}$	$r = 0.238$; $p = 0.413$, $\rho = 0.259$; $p = 0.372$
$\mathbf{H(X_n \mid X_{n-1})}$	$r = 0.625$*; $p = 0.017$, $\rho = 0.579$*; $p = 0.030$
$\mathbf{H(X_n \mid X_{n+1})}$	$r = 0.636$*; $p = 0.015$, $\rho = 0.604$*; $p = 0.022$
MC	$r = -0.453$; $p = 0.103$, $\rho = -0.409$; $p = 0.147$

The correlations between *IR* and linguistic complexity are shown in Table 3.5 where no significant correlation is found between *IR* and morphological complexity. In terms of *IR*, there is no apparent tendency among the languages classified according to the morphological typology as displayed in Figure 3.4, although this observation should be confirmed

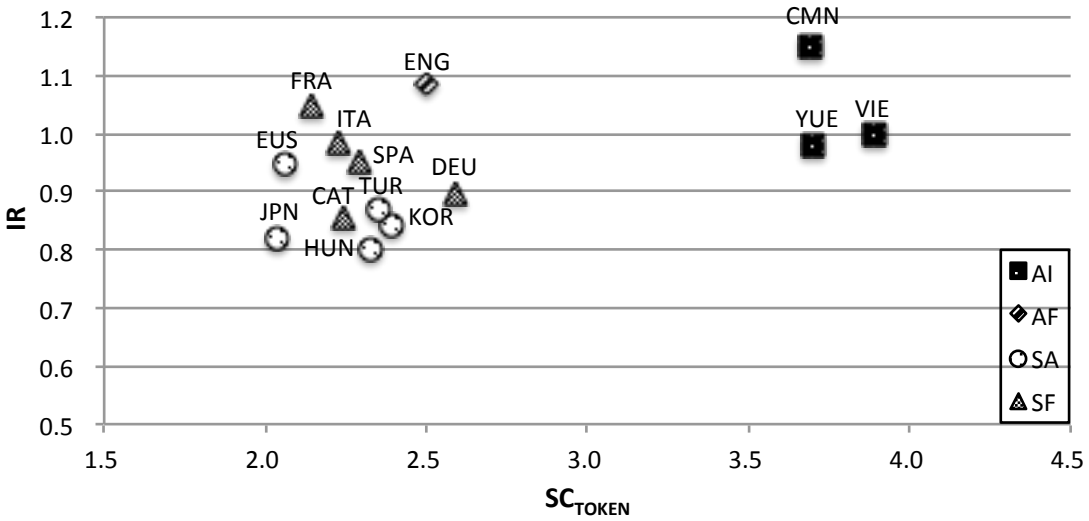

Figure 3.4: SC_{TOKEN} (average number of segments (and tones, if applicable) per syllable, weighted by relative frequency) on x-axis and IR (average amount of information per second, unitless) on y-axis

with more languages. In the same vein, with respect to phonological complexity, IR does not seem to be accounted for by SC_{TOKEN}, i.e. the average number of segments (and tones, if applicable) per syllable. It supports the assumption proposed by Pellegrino and his colleagues regarding the existence of "an optimal balance between social and cognitive constraints, taking also the characteristics of transmission along the audio channel into account" and their consideration of linguistic complexity as follows: "linguistic complexity merely defines the way each language encodes information, and says little about the actual rate of information transmitted during speech communication" [Pellegrino, Coupé, & Marsico, 2011].

3.3.3 Shannon entropy versus conditional entropy

The main goal of this present chapter is to assess a phenomenon of self-organization between morphological and phonological modules. Especially, two measures of phonological complexity are compared in terms of their relation with morphological complexity: entropy (or Shannon entropy) and conditional entropy. Shannon entropy measures the

average amount of uncertainty for using a syllable from a frequency distribution of syllables estimated from a large corpus. Conditional entropy, on the other hand, quantifies the average amount of unpredictability of syllable when its preceding or following context is known. Therefore, supposing that words may consist of more than one syllable, conditional entropy is lower than Shannon entropy since contextual information reduces such an unpredictability.

On the one hand, conditional entropy reflects the structure of words. For example, certain syllables, such as prefixes, may have tendency to appear more often at the initial position of words while the others, such as suffixes, occur more frequently at the final position of words. On the other hand, Shannon entropy is more concerned with the size of syllable inventory and the probability distribution of syllables. Let's say, if the syllables are all uniformly distributed in a language, its Shannon entropy reaches its maximum value. Moreover, if there are more syllables in the inventory, its Shannon entropy is higher as demonstrated by a positive correlation between Shannon entropy and the size of syllable inventory (Pearson's $r = 0.765^{**}$; p-value = 0.001; Spearman's $\rho = 0.793^{**}$; p-value= 0.001; N= 14).

Previous studies which adopted the entropy-based or conditional entropy-based measures were predominantly focused on the word-level analysis, investigating the relationship between context predictability and reduction of word: words with higher predictability are more likely to be reduced [Bell et al., 2009] [Jurafsky et al., 2001] [Pluymaekers, Ernestus, & Baayen, 2005], although there were also studies on the level of syllable [Aylett & Turk, 2004] and syntactic structure [Gahl & Garnsey, 2004].

Table 3.6 displays the results of morphological and phonological complexity of the 14 languages classified according to their morphological type. It is shown that morphological complexity range varies within the languages of the same morphological group, especially within SA languages ranging from -19.2 to -8.2. If the morphological complexity score is closer to 0, it is estimated that the language uses more inflectional strategies and if the

score is closer to −30, the language is assumed to employ more lexical strategies.

Table 3.6: Morphological and phonological complexity ($H(X)$, $H(X_n|X_{n-1})$, and $H(X_n|X_{n+1})$). The maximum and minimum values are marked in green and blue.

Group	L	MC	H(X)	$H(X_n\|X_{n-1})$	$H(X_n\|X_{n+1})$
AI	cmn	−21.95	8.69	6.96	6.99
	vie	−24.95	9.72	8.02	8.04
	yue	−22.95	7.97	6.53	6.59
AF	eng	−18.55	9.51	7.09	7.10
SA	eus	−8.55	8.32	4.83	5.05
	hun	−12.55	9.83	5.90	5.95
	jpn	−16.2	6.07	5.03	5.07
	kor	−19.2	8.05	5.56	5.53
	tur	−8.2	9.19	5.34	5.18
SF	cat	−12.55	8.10	5.49	5.53
	deu	−15.35	9.30	6.08	6.13
	fra	−14.55	8.39	6.68	6.76
	ita	−15.05	8.32	5.29	5.26
	spa	−12.55	8.32	5.43	5.41

Table 3.7: Correlations between morphological and phonological complexity

Parameter	Morphological complexity (N=14)			
WC_{TYPE}	r= 0.843**;	p< 0.001,	ρ= 0.784**;	p= 0.001
WC_{TOKEN}	r= 0.750**;	p= 0.002,	ρ= 0.674**;	p= 0.008
SC_{TYPE}	r= −0.791**;	p= 0.001,	ρ= −0.603*;	p= 0.023
SC_{TOKEN}	r= −0.816**;	p< 0.001,	ρ= −0.651*;	p= 0.012
H(X)	r= −0.035;	p= 0.905,	ρ= 0.060;	p= 0.839
$H(X_n\|X_{n-1})$	r= −0.761**;	p= 0.002,	ρ= −0.656*;	p= 0.011
$H(X_n\|X_{n+1})$	r= −0.759**;	p= 0.002,	ρ= −0.667**;	p= 0.009

The correlations between morphological and phonological complexity in the 14 languages are presented in Table 3.7. It is observed that WC and SC are significantly correlated with morphological complexity. In particular, the highest correlation is found between WC_{TYPE} and morphological complexity, indicating that languages with more syllables per word tend to have less complex inflectional morphology.

While there is no significant correlation between morphological complexity and entropy-

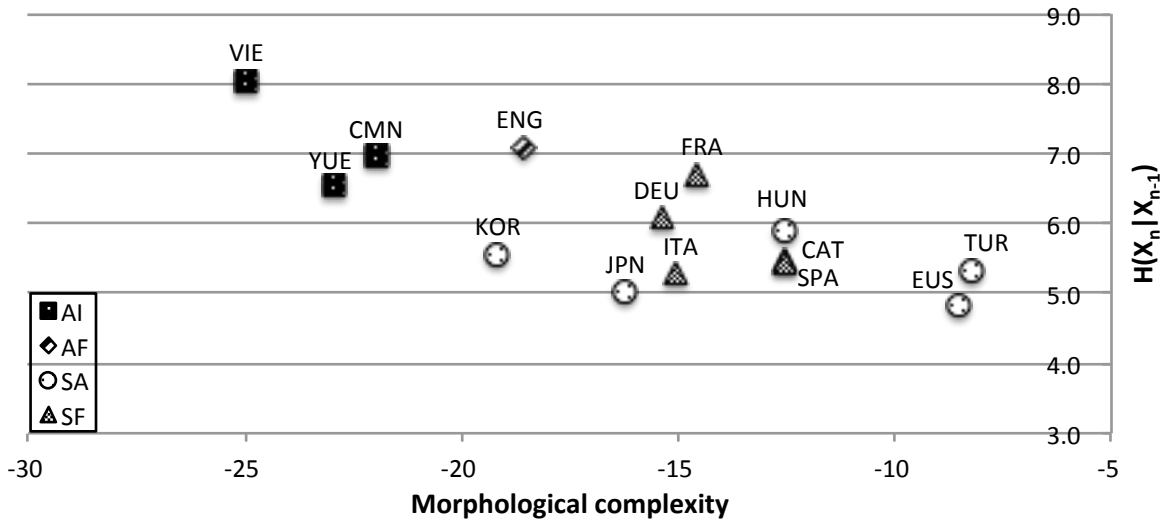

Figure 3.5: Morphological complexity (unitless) on x-axis and conditional entropy $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ (in bits) on y-axis

based measure of phonological complexity, the conditional entropy $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ with one preceding syllable and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$ with one following syllable as contextual information both display a significant negative correlation with morphological complexity. A high conditional entropy reflects a loose statistical relationship between each syllable and its environment while a low value reflects a tight statistical relationship, compatible with a larger morphological complexity. This correlation also denotes that languages with more complex inflectional morphology are likely to be more predictable in their phonological contextual information and thus, reveal less complex phonological complexity. In particular, it is shown that two agglutinative languages, Turkish and Basque, exhibit the highest score of morphological complexity.

Agglutinative languages are characterized by vowel harmony and strong affixation, which provide more contextual information for syllables. Therefore, they exhibit a lower conditional entropy than fusional languages as shown in Figure 3.6 (cf. Subsection 3.3.4). As previously mentioned at the beginning of this section, conditional entropy is expected to be more connected to morphological complexity than Shannon entropy since it reflects the structure of words, i.e. *structural complexity*.

3.3.4 Agglutination versus fusion

The traditional morphological distinction between agglutination and fusion has often been criticized in mainstream theoretical linguistics since 20$^{\text{th}}$ century.[46] This subsection aims to examine how this distinction is reflected by the quantifying measures of morphological and phonological complexity. The main reason that such a distinction is refuted by many linguists is based on the prevalent view that agglutination and fusion are considered as dichotomous opposition to each other in the traditional morphological typology. However, these two notions are still frequently employed when describing a language. In general, a language is classified as "agglutinative" if it prefers agglutination to fusion for its strategy of synthesis and as "fusional" if it prefers fusion. According to the degree of homogeneity, some languages are considered as "partially" or "strongly" agglutinative or fusional. For example, Basque and Turkish are typically regarded as strongly agglutinative languages and Finnish and Hungarian are considered as partially agglutinative languages in literature.

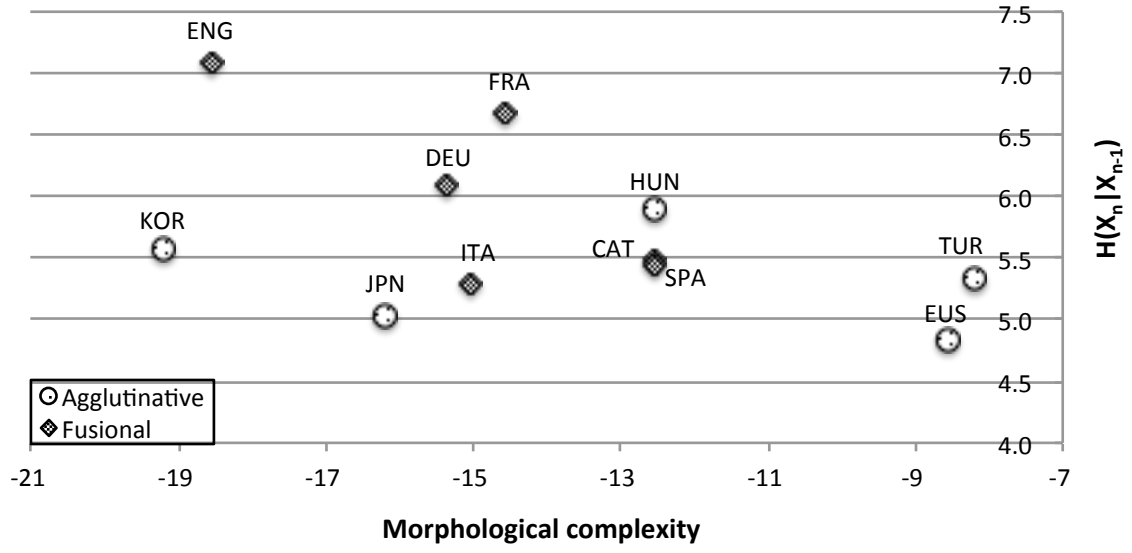

Figure 3.6: Agglutination vs. fusion: Morphological complexity score on x-axis and conditional entropy $H(X_{\text{n}}|X_{\text{n}-1})$ (in bits) on y-axis

Among the 14 languages considered, once the 3 isolating languages are left aside, we

[46] See [Plank, 1999] for a noteworthy exception among others.

have 5 agglutinative (Basque, Hungarian, Japanese, Korean, and Turkish) and 6 fusional languages (Catalan, English, French, German, Italian, and Spanish). In Figure 3.6, it is observed that agglutinative languages tend to exhibit a lower conditional entropy than fusional languages. As conditional entropy $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ is positively correlated to the size of syllable inventory in the 11 languages (Pearson's r=0.543**; p-value = 0.009; Spearman's ρ = 0.718*; p-value = 0.013; N = 11), agglutinative languages are assumed to exhibit a smaller size of inventory than fusional languages, although this tendency should be confirmed with numbers of typologically diverse languages. A low conditional entropy may result from the phenomenon of vowel harmony since the uncertainty of contextual information decreases in the languages with vowel harmony. Furthermore, Dressler asserted that "languages with vowel harmony are always (somewhat) agglutinating", reasoning that vowel harmony glues affixes and roots together [Dressler, 1985] [Moravcsik, 2003]. The effect of vowel harmony is revealed in some agglutinative languages (i.e. Hungarian, Korean, and Turkish) showing lower conditional entropy than fusional languages on average. However, there is no clear evidence of vowel harmony in the other two agglutinative languages.[47] In case of fusional languages, vowel harmony does not exist as a regular phenomenon.

Regarding morphological complexity, agglutinative languages are spread more widely ranging from −19.2 to −8.2 compared to fusional languages ranging from −18.55 to −12.55. Since morphological complexity score indicates the degree of inflection which encompasses both agglutination and fusion, the languages with remarkably high scores of morphological complexity, i.e. Turkish and Basque, can be regarded as languages with more "complex" inflection compared to the others.

Apart from comparing the degree of inflection, morphological complexity scores do not exhibit any difference between agglutination and fusion. The distinction between agglutination and fusion is better represented by the traditional linguistic measure of complexity,

[47]There are some arguments in favor of vowel harmony in many Basque dialects [Bereicua, 2013] and old Japanese [Ōno, 1970].

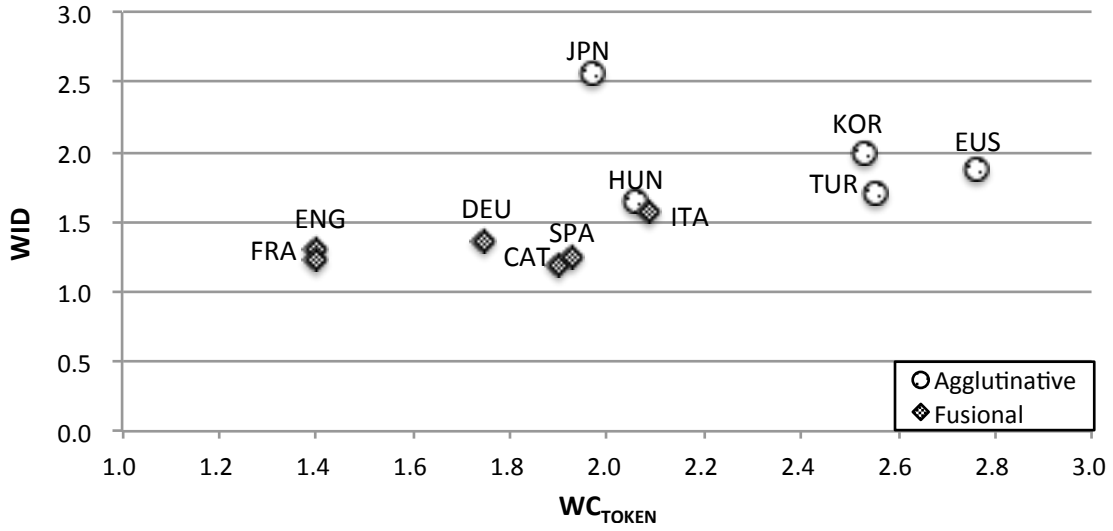

Figure 3.7: WC_{TOKEN} (average number of syllables per word weighted by relative frequency) on x-axis and *WID* (average amount of information per word, unitless) on y-axis

WC_{TOKEN}. In Figure 3.7, there is a clear division between two groups: agglutinative languages employ more number of syllables per word and convey more information per word than fusional languages. The differences between two types of languages are apparent and provide a hopeful evidence in supporting the distinction between agglutination and fusion. However, the limitation of this study is a small number of languages. Enlarging the data by adding more languages from several distinct language families may provide more convincing results.

3.3.5 Word order and linguistic complexity

This section presents a preliminary assessment of the relationship between word order and morphological and phonological modules in 12 languages. The most common word order in the languages of world is SOV (i.e. Subject-Object-Verb). In WALS, there are 566 SOV languages (41%) and 488 SVO languages (35%) among the 1 377 languages classified according to their word order. They are followed by the other types of word order: VSO (7%), VOS (2%), OVS (0.8%), and OSV (0.3%)[48]. Hence, three-quarters of languages fall

[48]It should be noted that 189 languages are considered as "no dominant order" in WALS.

into these two types of word order.

From an evolutionary perspective, some researchers claimed that "The earliest human language had rigid SOV word order" [Gell-Mann & Ruhlen, 2011] [Newmeyer, 2000] and that there's "an initial bias for SOV order" [Gibson et al., 2013]. In particular, Gibson and his colleagues explained a shift from SOV to SVO based on the noisy-channel hypothesis. The result of their study revealed that the SOV-SVO variation was triggered when there was a potential ambiguity such as reversing semantical roles between subject and object in SOV languages. Furthermore, it was shown that the languages with case-marking tend to maintain SOV order while SVO languages mostly lack of case-marking.

Table 3.8: Comparison between SOV and SVO: Number of case markers, morphological complexity, $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$, and WC_{TYPE}(word complexity) values are compared. MC scores are rounded off to the nearest whole number.

Word order	**SOV**				**SVO**								**No fixed order**	
Language	eus	jpn	kor	tur	cmn	vie	yue	eng	cat	fra	ita	spa	deu	hun
#Case marker	+10	8-9	6-7	6-7	×	×	×	2	×	×	×	×	4	+10
MC	−9	−16	−19	−8	−22	−25	−23	−19	−13	−15	−15	−13	−15	−13
$\mathbf{H(X_n|X_{n-1})}$	4.83	5.03	5.56	5.34	6.96	8.02	6.53	7.09	5.49	6.68	5.29	5.43	6.08	5.90
$\mathbf{H(X_n|X_{n+1})}$	5.05	5.07	5.53	5.18	6.99	8.04	6.59	7.10	5.53	6.76	5.26	5.41	6.13	5.95
WC	3.74	3.06	3.15	3.24	1.98	1.06	1.90	2.17	3.19	2.32	3.38	3.13	2.86	3.07

As shown in Table 3.8 and Figure 3.8, the 12 languages are divided into 2 types of word order, SOV and SVO, based on the information obtained online from WALS. German and Hungarian are not included in the figure, since they are classified as "no fixed word order". The explanation above regarding the relationship between case-marking and word order holds true with the 12 languages analyzed in this study. All SOV languages have a varying number of case markers whereas SVO languages do not have any case marker except for English. In terms of morphological classification presented in Section 3.2.4, SA languages correspond to SOV and the rest, i.e. AI, AF, and SF languages, concerns SVO.

A strong negative correlation between average word length and conditional entropy

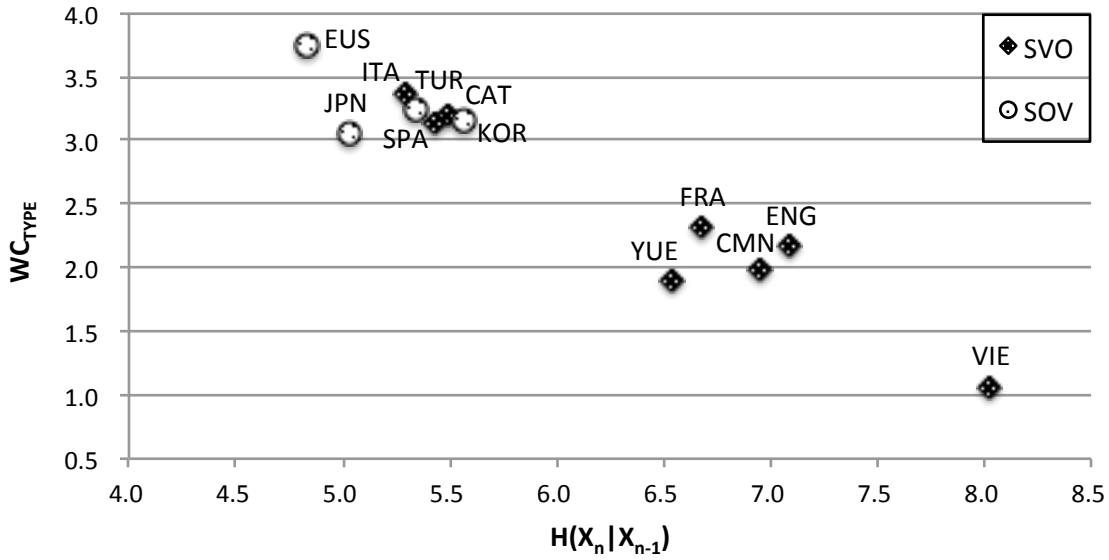

Figure 3.8: Assessing the relationship between conditional entropy $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ (in bits) on x-axis and WC_{TYPE} (average number of syllables per word) on y-axis regarding word order

$H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ is displayed (Pearson's $r = -0.964^{**}$; p-value < 0.001; Spearman's $\rho = -0.825^{**}$; p-value $= 0.001$; N $= 12$) in Figure 3.8. It is observed that i) SOV languages tend to have more syllables per word on average than SVO languages, and ii) SVO languages generally exhibit higher level of uncertainty in terms of the preceding phonological context than SOV languages. Moreover, SOV languages use postpositions which come after the object while SVO languages employ prepositions which precede the object. Therefore, it seems plausible to link word order, i.e. syntactic structure of language, to morphological and phonological complexity in an impressionistic manner.

On the contrary, comparing morphological and phonological complexity ($H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$) with respect to word order does not reveal any distinctive pattern. SOV and SVO languages both exhibit a varied range of inflectional morphology (11 points for SOV vs. 12 points for SVO). Morphological complexity score was calculated from the 29 linguistic features chosen in WALS (cf. Table 3.1) which measure the degree to which a language employs inflectional morphological strategies. Since the distinction between SOV and SVO is better reflected by the traditional linguistic measure of complexity rather than

a global measure of inflectional morphology, it can be assumed that word order is not directly related to the global complexity of inflectional morphology.

However, extending the study with an enlarged language sample with various types of word order can better account for the relationship between word order and the linguistic complexity. Furthermore, instead of considering word order as a criteria for comparison, taking more specific syntactic features (such as complex predicates or conjunctions) may reveal more convincing evidence that mesosystemic relation denotes compensatory interaction between linguistic subsystems such as morphology, phonology, and syntax.

3.4 Discussion

3.4.1 Equal overall complexity hypothesis: oversimplification or optimization?

This chapter investigates the validity of the equal overall complexity hypothesis by assessing the mesosystemic relationship between morphological and phonological complexity. This hypothesis has been criticized for the absence of null hypothesis [Fenk-Oczlon & Fenk, 2014] and its falsifiability in favor of the diversity of languages [Shosted, 2006]. In order to present supporting evidence for the validity of the hypothesis, different quantifying measures of linguistic complexity were chosen in this study as a function of linguistic module in question. Regarding phonological complexity, both information-theoretic and traditional grammar-based measures for quantifying linguistic complexity were used while usage-based measures, Shannon entropy and conditional entropy, were especially considered and compared. Regarding morphological complexity, grammar-based measure was used to calculate the global complexity of inflectional morphology. In addition, the 14 languages were classified into 4 groups based on two morphological criteria, degree of synthesis and morphophonemic alternation.

It is observed that the traditional classification of morphological typology (aggluti-

nation vs. fusion) and the distinction of word order (SOV vs. SVO) are better reflected by traditional measures of linguistic complexity such as *WC* and *SC*. The former clearly distinguishes different degrees of morphological synthesis and the latter reflects different levels of morphophonemic alternation. Furthermore, analytic and synthetic languages are also clearly distinguished by morphological complexity except for Korean.

Some general tendencies are found across the 14 languages classified according to the traditional morphological typology.

i) *SR* and morphological complexity: Synthetic languages tend to exhibit higher *SR* and more complex inflectional morphology than analytic languages.

ii) *SID* and SC_{TYPE}: Analytic languages are likely to encode more amount of information per syllable by means of more complex or longer syllables than synthetic languages.

iii) $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and morphological complexity: Analytic languages show a tendency toward higher phonological complexity and a preference toward lexical strategies over inflectional morphology while synthetic languages exhibit lower phonological complexity and favor inflectional strategies.

iv) WC_{TOKEN} and *WID*: Agglutinative languages contain more number of syllables per word and encodes more information per word than fusional languages.

v) $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and WC_{TYPE}: SOV languages can be characterized by more syllables per word and lower level of phonological complexity than SVO languages.

Although these results may need to be confirmed with data obtained from a wide range of languages, they provide a reasonably hopeful evidence for supporting the traditional morphological classification which has been criticized for "oversimplification" and "lack of evidence" in modern theoretical linguistics. Holistic typology and the equal complexity hypothesis were developed in the same vein, due to the popularity of biological taxonomy such as Darwinian classification [Darwin, 1859], considering language as a "(natural) organism possessing an *inner form*" [Robins, 1967] [Song, 2014]. However, as Comrie pointed out in *Language universals & linguistic typology*, there is a lack of empirical evidence for

holistic typology and moreover, it seems crucial to define parameters which describe language from a systemic perspective: "while we can state often wide-ranging correlations among logically independent parameters, these correlations are not sufficiently strong or sufficiently wide-ranging to give holistic types rather than cross-classification of languages on different parameters" [Comrie, 1989].

This study suggests that usage-based and information-theoretic measures provide empirical evidence with "sufficiently strong correlations" among linguistic complexity. Especially, while there was no correlation between morphological complexity and Shannon entropy, it was shown that two measures of conditional entropy, $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$, were negatively correlated with morphological complexity, distinguishing analytic and synthetic languages. This result can be interpreted that the effect of frequency better reflects the phonological complexity when if it is taken into account together with contextual information. Furthermore, the result is compatible with holistic typological distinction between analytic and synthetic languages. In addition, languages which differ in word order (SOV vs. SVO) exhibit distinct patterns: SOV languages tend to use postpositions and case markers and have more syllables per word on average and lower conditional entropy $H(X_{\mathrm{n}}|X_{\mathrm{n-1}})$ and $H(X_{\mathrm{n}}|X_{\mathrm{n+1}})$ than SVO languages which employ prepositions and lack case marker. Therefore, it is estimated that there is a mesosystemic interaction between linguistic modules, which enables a holistic typological distinction among the 14 languages.

A study by Gibson and his colleagues suggested that the variation from SOV to SVO results from the speaker's effort to reduce potential ambiguity of reversing semantical roles between subject and object in SOV languages [Gibson et al., 2013]. In cognitive and evolutionary linguistics, language has been considered as *a complex adaptive system* (CAS, henceforth) of which the structure emerges from the social interactions between speaker and hearer and their cognitive mechanisms [Beckner et al., 2009]. This idea was proposed before Darwin in the 18^{th} century [Christiansen & Chater, 2008]. From the CAS

perspective, language is shaped by the "interpersonal communicative and cognitive processes" [Slobin, 1997] and is not considered as the result of the adaptation of brain to the grammar of language [Christiansen & Chater, 2008]. This view on language as an emergent adaptive system leads to the inference that the equal overall complexity hypothesis and holistic typology may not be the consequence of theoretical oversimplification but result from the optimal balance between the social interactions and cognitive constraints which will be discussed in the next section.

3.4.2 Sociolinguistic and neurocognitive constraints on complexity

In this study, it is shown that *IR* was not accounted for by morphological complexity and exhibited a limited narrow range of variation among the 14 languages, confirming the result of previous study [Pellegrino, Coupé, & Marsico, 2011]. Despite the variation among linguistic complexity, the languages do not differ in terms of their capacity of transmitting information. Hence, they can be considered equally complex from functional and cognitive approaches. Moreover, the phenomenon of negative correlation has been used to account for this equal overall complexity hypothesis [Fenk & Fenk-Oczlon, 2006] [Shosted, 2006]. As it seems crucial to acknowledge that such a self-organization or trade-off among linguistic modules cannot wholly account for the equal overall complexity, the role taken by sociolinguistic factors and neurocognitive constraints in optimally balancing linguistic complexity should be highlighted and deserves further investigation.

In sociolinguistics and psycholinguistics, the relationship between linguistic structure and sociocultural constraints has been investigated by several researchers [Lupyan & Dale, 2010] [McWhorter, 2001] [Nettle, 2012] [Trudgill, 2011] [Wray & Grace, 2007] among many others. The result of those studies predominantly suggested that language is shaped by its social and cultural structures "just as biological organisms are shaped by ecological niche". Likewise, *Linguistic Niche Hypothesis* was proposed by [Lupyan & Dale, 2010], asserting

that "morphological complexity varies as a function of the learning population" due to "a greater pressure to be learnable by adult leaners". In *Sociolinguistic typology: social determinants of linguistic complexity*, Trudgill enumerated 4 types of social factors which influence linguistic structure as follows: (i) degree of linguistic contact (vs. isolation) of the community with the other communities speaking different languages, (ii) degree of social stability, (iii) size of community, and (iv) density of social network [Trudgill, 2011]. The results of the two studies mentioned above can be summarized as follows: that languages (a) spoken by a large number of population, (b) in an unstable linguistic community, (c) spread in a wide geographical range, (d) with loose social network and (e) high degree of contact with other linguistic communities and (f) being acquired by a large number of adult learners have a tendency towards *linguistic simplification* while the languages exhibiting the opposite tendencies are likely to go toward *linguistic complexification*.

In his recent book, *The language hoax*, McWhorter refuted the Sapir-Whorf hypothesis which claimed that language has a strong influence on the way people think and he argued that language is shaped by its culture and not the other way around [McWhorter, 2014]. Furthermore, within the framework of CAS, the structure of language evolves by the social interaction between speaker and listener creating "a conflict of interest" between them (i.e. conciseness vs. explicitness): speakers tend to reduce their effort of articulation and control their speech as a function of the needs of listeners, word frequency, and contextual or mutual information while listeners are likely to economize their effort of perception and reduce the probability of confusion [Beckner et al., 2009] [Bell et al., 2009] [Christiansen & Chater, 2008] [Gregory et al., 1999] [Jurafsky et al., 2001] [Lindblom, 1990]. Therefore, it appears that language is constructed by adapting itself not only to the sociocultural factors but also to the neurocognitive constraints on the interaction between speakers and listeners.

From functional and cognitive perspectives, the results of this chapter provide clues suggesting that the equal overall complexity hypothesis and holistic typology may be

regarded as convincing, especially by displaying a negative correlation between morphological and phonological modules at the mesosystemic level of analysis. Nonetheless, the equal overall complexity hypothesis cannot be wholly explained by this relationship for two reasons: (i) the optimizing socio-cognitive mechanisms underlying the communication between speakers and listeners do not seem to differ among languages, (ii) the rate of information transmission remains relatively stable among the 14 languages analyzed contrary to the variation in linguistic complexity. Thus, it is proposed for further study to combine both linguistic complexity and socio-cognitive factors to assess the equal overall complexity hypothesis and holistic typology.

3.4.3 Conclusion

In conclusion, some general tendencies were found among the languages classified according to holistic morphological typology by means of information-theoretic and grammar-based measures proposed in this chapter. In particular, it was observed that among different information-theoretic measures of phonological complexity, the values of conditional entropy, $H(X_n|X_{n-1})$ and $H(X_n|X_{n+1})$, were negatively correlated with morphological complexity, which demonstrates the effect of context information (vs. Shannon entropy). The results provide convincing evidence for supporting the validity of the morphological classification based on holistic typology and highlight a need to investigate the sociolinguistic and neurocognitive factors influencing the language structure along with linguistic complexity in further study.

Chapter 4

Functional load: microsystemic organization of phonological system

The function of a phonemic system is to keep the utterances of a language apart. Some contrasts between the phonemes in a system apparently do more of this job than others. [Hockett, 1966].

This chapter of thesis consists of an article entitled *Bridging phonological system and lexicon: insights from a corpus study of functional load* [Oh et al., forthcoming], which will be published in the special issue of Journal of Phonetics on *Speech sound systems*.

The first study in the previous chapters revealed that a cross-language tendency in terms of information transmission exists among the 17 languages in speech communication, at the macrosystemic level and the results confirmed the initial hypothesis that a relatively stable average information rate results from the phenomenon of self-organization between speech rate and information density. In the second study, the relationship between linguistic modules was assessed at the mesosystemic level and it was revealed that a negative correlation exists between phonological and morphological modules, which provides a supporting evidence for the equal complexity hypothesis.

The two previous studies confirmed that a phenomenon of self-organization exists both

at the macrosystemic and mesosystemic levels from a quantitative and typological perspective, by means of information-theoretic measures. In connection with the previous studies, in the present chapter, the phenomenon of self-organization is assessed at the microsystemic level from a quantitative and typological approach, using an information-theoretic measure, functional load (*FL*). *FL* has been used for measuring the relative importance carried by phoneme contrasts, based on the quantification method proposed by [Hockett, 1966] and it corresponds to the change of Shannon entropy of the phonological system if the contrasting pair is merged into one phoneme.

Two studies are conducted in this chapter. In the first study, the relative importance of phonological subsystems (e.g. vowels, consonants, stress, and tones) is examined in 9 languages (2 tonal and 7 non-tonal languages), taking morphological strategies (Lemma vs. Inflected) and usage frequency (Token vs. Type) into account. The second study consists of comparing the internal organization of phonological subsystems (vowels and consonants) in the 9 languages.

Since *FL* measures the relative importance of phonological subsystems and units, its value depends on the size of phoneme inventory and cannot be compared directly among languages exhibiting different phoneme inventory sizes. Thus, the goal of these studies is to observe general cross-language tendencies and language-specificities of the *organization* of phonological subsystems among the 9 languages, within the complex systems framework in which language is defined as a complex adaptive system adjusting itself to its environments by means of self-organization. The results confirm the following two hypotheses that (i) consonants play a more important role in lexical access than vowels, and that (ii) only a few phoneme contrasts play an important role in lexical access due to cognitive efficiency and robustness in speech communication, regardless language-specific differences.

4.1 Introduction

4.1.1 The concept of functional load

As stated by Hockett, "The function of a phonemic system is to keep the utterances of a language apart" [Hockett, 1966, p.1]. Phonemes are thus considered the elementary bricks on which contrasts between words are built. The most obvious procedure to identify them is by listing minimal pairs (when they exist): two sound sequences associated with two different meanings and differing by only one element. The set of such 'distinctive' elements constitutes the phonemic system of a particular language. For decades, studying phoneme inventories has been the gateway for understanding how languages work. This traditional approach to phonemes and relations between them has yielded highly significant insights into the organization of phonological systems [Crothers, 1978] [Hall, 2011] [Hyman, 2008] [Liljencrants & Lindblom, 1972] [Lindblom, 1986] [Lindblom & Maddieson, 1988] [Maddieson, 1984] [Marsico et al., 2003] [Schwartz et al., 1997] [Vallée, 1994]. However, a side-effect of this paradigm is that, because all phonemes in an inventory are given the same importance, disregarding their frequency and their role in contrasts[49], certain key phenomena remain underappreciated. To illustrate, consider asking a British English (RP: Received Pronunciation) speaker to provide an example of a minimal pair based on a consonantal contrast. Her answer is likely to include word pairs that exhibit a "high frequency" contrast such as /t-d/ (as in "*tip*" vs. "*dip*"), as opposed to word pairs that exhibit a "low frequency" contrast such as /ʒ-v/, (as in "*closure*" /ˈkləʊʒə/ vs. "*clover*" /ˈkləʊvə/). The point is that some phonemic contrasts in English, differentiate hundreds of word pairs (e.g. /t-d/) while others may only be involved in a handful of word pairs

[49] Vowels and consonants (as well as their natural subsets: stops, fricatives, etc.) are not considered identical, in terms of production [Ladefoged & Maddieson, 1996], acoustics ([Fogerty & Humes, 2012] [Ladefoged, 2001] [Stevens, 2002], among others), and perception ([Fry et al., 1962] [Kronrod, Coppess, & Feldman, 2012] [Liberman et al., 1975]). These differences have recently been mirrored by neurophysiological findings ([Caramazza et al., 2000] [Mesgarani et al., 2014] [Obleser et al., 2010] [Scharinger, Idsardi, & Poe, 2011]). Vowels and consonants are not identical in terms of functional role either ([Nespor, Peña, & Mehler, 2003] [New, Araújo, & Nazzi, 2008] [Toro et al., 2008]), should it be defined by usage frequency or *FL*, for instance.

(e.g. /ʒ-v/). This fact accords with Hockett's addendum to his characterization of the functional role of phonemes: i.e. that "Some contrasts between the phonemes in a system apparently do more [keeping apart of words] than others" [Hockett, 1966, p.1]. Moreover, this observation appears to hold true for other languages as well, with the work done by particular contrasts potentially varying across languages. Indeed, the Prague School thought that specific contrasts may differ from one language to another and that this "rendement fonctionnel" or "charge fonctionelle" (Functional Load, henceforth *FL*) should be taken into consideration when reasoning about phonological systems [Cercle Linguistique de Prague, 1931] [Jakobson, 1931].

4.1.2 Some landmarks on functional load

Despite a general agreement on what it covers, it should be noted that the concept of *FL* has often been considered in an impressionistic way (for a review, see [Surendran & Niyogi, 2003]). As a consequence, *FL* is generally described by a circumlocutions and no precise theoretical definition exists, beyond general statements such as "The term FUNCTIONAL LOAD is customarily used in linguistics to describe the extent and degree of contrast between linguistic units, usually phonemes" [King, 1967]. To be fair, one should also note that formal mathematical definitions arose as early as the mid-fifties [Hockett, 1955] and provided enough ground to address *FL*-related issues. Before this quantitative characterization, advocates of *FL* heavily relied on intuitions and extensions of the notion of phonological contrast. As stated in the previous section, phonological contrast and opposition were central concepts within the Prague School. Trubetzkoy later mentioned that an "economical" language would very often distinguish words by only one phoneme while "prodigal" languages would make usage of several phonological elements to keep words distinct [Trubetzkoy, 1939, p.240]. Kučera compared phonemic and syllabic inventory entropies, as well as some derived *FL* measures, in Russian and Czech [Kučera, 1963]. Yet, references to *FL* have remained sporadic for decades, probably because of

the difficulty to process large corpora, which were moreover hardly available. This state lasted until Surendran and Niyogi breathed new life into the concept at the beginning of this century. They compared *FL* of tones, stress, phonemes, and phonetic features in four languages (Dutch, English, German, and Mandarin) and highlighted the importance of the tonal system in Mandarin [Surendran & Niyogi, 2003]. This result was confirmed in a follow-up study [Surendran & Levow, 2004] and recently extended to Cantonese [Oh et al., 2013]. On and colleagues also compared the relative functional weight of consonantal, vocalic (and tonal, if any) systems in five languages (Cantonese, English, Japanese, Korean, and Mandarin). Their results suggest that the distributions of *FL* in a phonological system are very uneven, with only a few prominent contrasts. These differences in relative prominence may be useful to take into consideration for foreign language acquisition (following [Brown, 1988] [Munro & Derwing, 2006]).

Besides typology-oriented studies, the main topic for which *FL* was considered relevant was historical linguistics. Upon its inception, Martinet promoted the notion of *FL*, suggesting that it may play a role in language change [Martinet, 1938, 1955]. According to his hypothesis, also adopted later by [Hockett, 1966], phonemes involved in high-*FL* contrasts would be less prone to merging than those involved in low-*FL* contrasts. Corpus-based studies have failed to confirm this hypothesis for decades [King, 1967] [Surendran & Niyogi, 2003] [Surendran & Niyogi, 2006], but a recent cross-language study brought some support to it [Wedel, Kaplan, & Jackson, 2013]. Such conflicting results may be due to differences in corpora or to the small number of sound changes considered so far. It is also possible that, even if *FL* plays a role in phonetic change, its magnitude is limited, for example with regard to social factors [Labov, 2001]. As a consequence, even if *FL* does determine a pool of potential changes, their actual implementation in a language or a dialect probably depends on further aspects.

From a different angle, the availability of corpora in the field of child language acquisition also stimulated interest in the notion of *FL*. Its impact on the order of phoneme

acquisition by children was demonstrated [Pye, Ingram, & List, 1987] [Van Severen et al., 2012], in conjunction with language-specific properties [Stokes & Surendran, 2005]. Again, *FL* is not the only factor at play in the course of phonological acquisition, but converging cues indicate that the phonemes involved in high-*FL* oppositions have a tendency to be acquired earlier than the others [Van Severen et al., 2012]. Stokes and Surendran showed nevertheless that the effect of *FL* should be considered with caution since *FL* was not a significant predictor of consonant order of acquisition in Cantonese-speaking children, in contrast with what they observed in English-speaking children [Stokes & Surendran, 2005].

This re-emergence of the concept of *FL* can be seen as part of a general movement for promoting statistical and information-theoretic quantitative approaches (see [Goldsmith, 2000]). Today for instance, the relevance of usage frequency is well acknowledged, and many studies in psycholinguistics, phonology, and phonetics have proven that it significantly impacts cognitive processes, such as access to mental representations [Bybee, 2003] [Cholin, Levelt, &Schiller, 2006] [Jescheniak & Levelt, 1994] [Johnson, 1996] [Levelt, Roelofs, & Meyer, 1999] [Pierrehumbert, 2001] [Schilling, Rayner, & Chumbley, 1998] [Walsh et al., 2010]. It has nevertheless been less often mentioned in the study of phonological systems per se. However, we think that taking this functional approach into consideration can notably change our vision of phonological systems and can enrich our knowledge of speech cognitive processing. The goal of this paper is consequently to shed new light on phonological systems from the perspective of *FL*. The emphasis is placed on both their internal functional organization and their importance in building the lexicon. Results are then discussed on communicative and cognitive grounds, in connection with the main focus of this Special Issue.

For almost one century, *FL* has thus been suggested as a factor involved in the *acquisition* and the *evolution* of phonological units and systems as well as a *systemic* property rooted in lexical strategies. These three dimensions have in common the fact that they

deal with the dynamics of structural and functional relationships among the phonological units which define a phonological system. *FL* especially provides an additional approach to investigate the nature and dynamics of phonological units in the context of their systemic relations. The COSMO model introduced by [Moulin-Frier et al., forthcoming] provides a unifying framework able to address the nature of the cognitive architecture of communicating agents, in light of such systemic relations. From an epistemological viewpoint, Moulin-Frier and his colleagues advocate the implementation of alternative theories of speech communication in COSMO multi-agent simulations, and their testing against properties observed in real phonological systems. In their paper, this procedure is applied to regularities observed in phonological inventories (vocalic and consonantal systems) and syllable inventories through multi-agent deictic games. They also mention that their work can be extended to address compositionality, thus requiring more elaborate stimnuli for their communicating agents. We consider that *FL* may bring a new set of cross-linguistic regularities that would be especially relevant for testing extensions of the COSMO framework to lexically-based simulations. We suggest that the *FL* properties extracted from artificial corpora yielded by multi-agent naming games of similar setting [Steels & McIntyre, 1998] should be compared to properties observed in real human lexicons, beyond what has already been explored at the segmental level.

4.1.3 Paper outline

Section 4.2 introduces the methodology implemented in this paper. In Section 4.3 and 4.4, two directions are proposed to illustrate the potential of *FL* studies. In the first study, we investigate the structure of a phonological system as it is revealed by the *FL* of vowels, consonants, stress, and tones as whole subsystems. Morphological information available for five languages (British English, French, German, Italian, and Swahili) further leads to evaluate *FL* sensitivity to several factors. Considering token or type frequencies, wordforms or lemmas may reveal or confirm trends on the function of specific phonological

categories. More precisely, it has been shown, at least in some languages, that consonants and vowels tend to be preferentially involved in lexical access − for consonants − or rhythmic and syntactic information − for vowels [Bonatti et al., 2005] [Cutler et al., 2000] [Delle Luche et al., 2014] [Havy & Nazzi, 2009] [Nazzi & New, 2007] [Nazzi et al., 2009] [Nespor, Peña, & Mehler, 2003] [New, Araújo, & Nazzi, 2008] [Toro et al., 2008]. What has been coined Consonant Bias, potentially reflected by *FL*, will thus be the main issue at stake. In Section 4.4, the second study focuses on distribution of *FL* at the level of segmental units rather than phonological subsets. It thus investigates general trends or specificities regarding the internal functional organization of phonological systems in the world's languages. The quantitative measures of *FL* yielded by the framework suggest that representation of phonological (sub)systems based on frequency/usage (Figure 4.1, right) may be as useful as the more traditional, time-tested representations (Figure 4.1, left). Indeed, by directly encoding the different functional roles of vowels in terms of number

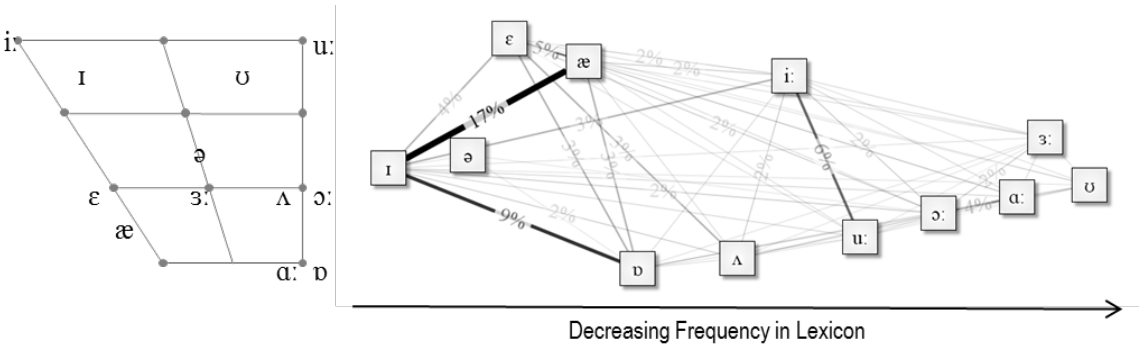

Figure 4.1: Illustrations of English (RP) vowel system. Left: Standard IPA chart. Right: Functional network-based representation. Vowels are ranked, from left to right, according to decreasing usage frequency. Edges (thickness and opacity) reflect the functional load associated with each vowel pair. Vertical positions of the vowel labels are arbitrary and chosen for legibility (data computed from WebCelex using the methodology described in Section 4.2 of the paper).

of contrasts, Figure 4.1 (right) reveals salient differences among vowels. For instance, the near-close vowels /ɪ/ and /ʊ/ behave very differently: /ɪ/ being frequent and engaged in a lot of lexical oppositions while the opposite is observed for /ʊ/. Moreover it gives a view

of the *system* as a set of intricate oppositions among its constituents, rather than a set of apparently independent segments, as in the left chart.

Finally in Section 4.5, results are discussed in terms of phonological units and features, relative weights of vowel vs. consonant, and general trends in the *FL* distribution within the phonological systems (see also [Oh et al., 2013]).

4.2 Rationale and methodology

4.2.1 Computing functional load

Several algorithmic approaches have been proposed to quantify FL [Hockett, 1955, 1966] [Ingram, 1989] [King, 1967] [Kučera, 1963] [Surendran & Niyogi, 2003] [Wang, 1967]. Following [Hockett, 1955], these approaches are grounded in information-theoretic methods [Shannon, 1948] and use entropy computed at various levels as the essential metrics. One noteworthy exception is the simple counting of the number of lexical minimal pairs based on each contrast [Ingram, 1989].

[Surendran & Niyogi, 2003] and [Van Severen et al., 2012] thoroughly discussed several of these metrics and the latter showed that Ingram's approach and an entropy-based metric implemented by [Surendran & Niyogi, 2003] are almost equivalent predictors of the age of acquisition of word-initial consonants in Dutch. However, they differ in the information they encompass and we chose to implement both metrics, referring to them as number of Minimal Pairs ($\#MP$) and Entropy FL (FL_E) respectively.

For each language studied, the material consists of a large set of word-forms associated with token frequencies drawn from a large, phonemically-transcribed, corpus. This dataset can optionally be pre-processed in order to filter out specific items (according to their token and lemma frequency, their grammatical category, etc., see Section 4.2.3). In this paper, the phonological inventory is defined as the pool of phonemes required to transcribe the corpus considered.

For each pair of phonemes in the inventory, *#MP* is the number of distinct word-forms that are discriminated by this specific pair. Because perceptual confusions (in language acquisition) and diachronic mergers (in language change) are more likely to occur between similar phonemes, several studies have limited the inspected contrasts to phoneme pairs that differ only by one phonological feature: place of articulation, manner of articulation or voicing for consonants [Van Severen et al., 2012] [Wedel, Kaplan, & Jackson, 2013]. However, since our goal was to study the global utilization of the phonological inventory for lexical purposes, no such limitation was implemented and all contrasts were considered. For example, in British English, the lexical items *hit*, *bit*, *pit*, and *sit* contributed to the contrasts /h-b/, /h-p/, /h-s/, /b-p/, /b-s/, and /p-s/. However, lexical differentiations involving an insertion did not contribute to *FL*; for instance, the lexical pair *hit-it* did not form a minimal pair.

Besides the Minimal Pair approach, we also implemented the information-theoretic approach proposed by [Hockett, 1966] and further elaborated by [Surendran & Niyogi, 2003]. Here, a language L is considered as a source of sequences made of word-forms w taken from a finite set of size N_L and composed of Vowels (possibly including diphthongs), Consonants (possibly including glides) and possibly Stresses and Tones taken from the phonological inventory $\mathbf{P} = \mathbf{V} \cup \mathbf{C} \cup \mathbf{S} \cup \mathbf{T}$. The amount of information of source L is estimated in terms of Shannon entropy *H(L)* [Shannon, 1948]:

$$H(L) = -\sum_{i=1}^{N_L} p_{w_i} \cdot log_2(p_{w_i}) \tag{4.1}$$

where p_{w_i} is the probability of word-form w_i, approximated by its relative token count estimated from the corpus.

Following [Surendran & Niyogi, 2003], we implemented the definition of *FL* given by [Carter, 1987] and derived from Hockett's initial proposal [Hockett, 1966]. The FL of a contrast between two phonemes φ and ψ, $FL_E(\varphi, \psi)$, is defined as the relative difference of entropy between two states of language L: the observed state L and a fictional state

$L^*_{\varphi\psi}$ in which the contrast is neutralized (or coalesced, in Hockett's terminology). $FL_E(\varphi, \psi)$ therefore quantifies the perturbation induced by merging φ and ψ, in terms of increase of homophony and of changes in the distribution of word frequencies:

$$FL_E(\varphi, \psi) = \frac{H(L) - H(L^*_{\varphi\psi})}{H(L)} \tag{4.2}$$

$FL_E(\varphi, \psi)$ is hence defined at the level of phonemic *contrasts*, as a ratio theoretically ranging from 0% to 100%.

In addition, one can also focus on the level of *phonemes* themselves, by summing $FL_E(\varphi, \psi)$ over all the contrasts in which a phoneme φ is involved. $FL_E(\varphi)$ thus measures the importance of phoneme φ in the language lexical network:

$$FL_E(\varphi) = \frac{1}{2} \sum_{\psi} FL_E(\varphi, \psi) \tag{4.3}$$

With the normalization factor ½ applied to ensure that:

$$\sum_{\varphi} FL_E(\varphi) = \sum_{\varphi, \psi \neq \varphi} FL_E(\varphi, \psi) \tag{4.4}$$

It can also be used to give a more global quantification of the functional weight of subparts of the phonological system. We defined FL_V (resp. FL_C) as the overall loss of information induced by comparing language L with a fictional state L^*_V (resp. L^*_C) in which all vowels (resp. consonants) are merged into a unique symbol. As an illustration, in L^*_V, the three English words *pit*, *bit*, and *pot* coalesce into two forms pVt and bVt while they result in two other forms CɪC and CɒC in L^*_C. Syllabic boundaries are taken into account to distinguish between words – e.g. Xī'ān and xīan in Mandarin – and for the computation of *FL*. For instance, during the computation of FL_C for English, the two words *mattress* /mæ.trɪs/ and *maxim* /mæk.sɪm/ result in two distinct entries /Cæ.CCɪC/ and /CæC.CɪC/, while they would merge into a single entry /CæCCɪC/ if syllable boundaries were not considered.

In addition to FL_V and FL_C, a more drastic reduction was implemented by only keeping the skeleton of the word-forms, i.e. consonantal and vocalic slots as well as stress and syllable boundaries. This so-called segmental FL, FL_{VC} measures the cumulative information carried by the identity of the segments in the wordlist. In the resulting L^*_{VC} language, the three words mentioned above merge into a CVC form.

$$FL_V(L) = \frac{H(L) - H(L^*_V)}{H(L)} \tag{4.5}$$

$$FL_C(L) = \frac{H(L) - H(L^*_C)}{H(L)} \tag{4.6}$$

$$FL_{VC}(L) = \frac{H(L) - H(L^*_{VC})}{H(L)} \tag{4.7}$$

By extension, stresses and tones can also be considered the same way. For instance in Mandarin, the lexical pair 判 ("sentence", /p$^{\text{h}}$an4/) and 盘 ("plate", /p$^{\text{h}}$an2/) contributes to the computation of FL_E between tone2 and tone4, and the global functional weight FL_T of the tonal system can thus be quantified mutatis mutandis, and an overall infra-syllabic FL_{VCTS} is also defined. It is important to note that FL_{VCTS} is not the sum of FL_V and FL_C. Although a strict mathematical proof is difficult to formulate, the following explanation can be given. Coalescing at the same time all vowels together and all consonants together necessarily merges all the word-forms that are merged by coalescing vowels only, and all the word-forms that are merged by coalescing consonants only (whether some word-forms merge in both cases are not relevant). Additionally, more mergers may occur between word-forms of similar phonological pattern (eg. CV, CVC, CV CCVC, etc.) that were not merged either in L^*_C or in L^*_V. Conversely, for FL_{VC} to be equal to FL_V+ FL_C, no word-form that did not get merged in either L^*_V or L^*_C should get merged in L^*_{VC}. This imposes strict constraints on the structure of word-forms that natural languages are usually far from respecting. As an example, while the invented language {pi, bi, pa, ba}

(with frequencies all equal to 1) satisfy the constraint, the slightly different language {pip, bi, pa, ba} (again, all frequencies equal to 1) does not.

#MP and FL_E differ in several ways, though they yielded similar results in previous studies [Surendran & Niyogi, 2003], [Van Severen et al., 2012]. For a given contrast φ-ψ, *#MP* only requires a knowledge of the word-forms in which the two phonemes are involved in order to count the relevant minimal pairs. However, $\#MP(\varphi, \psi)$ is not influenced by the rest of the lexicon, i.e. word-forms where φ and ψ are absent. It does not rely on any probability estimation either, which leads Wedel and colleagues to consider it as a *local* measure [Wedel, Kaplan, & Jackson, 2013]. On the contrary, Entropy FL is a *global* measure. The entropy is computed on the whole lexicon and involves probability estimations. As a consequence, $FL_E(\varphi, \psi)$ both requires a global knowledge of the lexicon *and* measures the impact of the φ-ψ contrast on the whole lexicon. Beyond the local influences on lexical access (e.g. [Luce & Pisoni, 1998]), it has been very recently suggested that global properties of the mental lexicon may influence lexical cognitive processing [Vitevitch, Chan, & Goldstien, 2014] and further investigations on the relationship between local and global levels will be insightful, though beyond the scope of this paper.

We introduced in this section several indices aimed at assessing the importance of phonological components in the maintenance of lexical distinctions. These components are however complemented with other dimensions: number of segments or syllables, syllabic structures, phonotactic and syllabotactic information, and more generally word structure. In the rest of this paper, we refer to these dimensions as structural information.

4.2.2 Language description

Table 1 provides the description of the data and phonological system of the nine languages (Cantonese, English, French, German, Italian, Japanese, Korean, Mandarin, and Swahili) analyzed in this paper. For five languages (English, French, German, Italian, and Swahili), lemmatized forms were available.

The number of vowels (including diphthongs), consonants, tones (if any) and stresses (if any) are provided for each language. The size of the phonological system may not correspond exactly to traditional phonological descriptions since the corpora used here included some loanwords and newly coined words derived from other languages.[50] For instance, in the Swahili corpus, there are plenty of Arabic and English loanwords which consequently extended syllabic structures beyond traditional "open" syllables (see Appendix A.5). Following [Maddieson, 2013], syllable complexity is estimated by a syllable index, ranging from 1 to 8 among the world's languages. This index corresponds to the sum of the potentially maximal number of onset, nucleus, and coda elements. For this study, indices were retrieved from the LAPSyD website [Maddieson et al., 2013]. The four Indo-European languages (English, French, German, and Italian) have complex syllable structures. The two Sino-Tibetan languages, Cantonese and Mandarin, as well as Korean and Japanese, have moderately complex syllable structures. Swahili has simple syllable structures.

Table 4.1: Language and corpus description. For each language, the size of its phonological system (V: #vowels, incl. diphthongs; C: #consonants; T: #tones; S:#stresses, if applicable), syllable index (based on LAPSyD), and the size of syllable inventory (#distinct syllables) are provided, as well as morphological typology information.

Language	ISO 639-3 code	Phonological system		Syllable index	Size of syllable inventory	Morphological type	Corpus
Cantonese	YUE	C V T	19 13 6	3	1 303	Analytic/ Isolating	A linguistic corpus of mid-20th century Hong Kong Cantonese
English	ENG	C V S	25 24 2	8	6 469	Analytic/ Fusional	WebCelex
French	FRA	C V	22 15	7	5 530	Synthetic/ Fusional	Lexique 3.80

[50]The phonemic inventories of the nine languages (obtained from each corpus) are given in Appendix A.5.

Table 4.1: Language and corpus description. For each language, the size of its phonological system (V: #vowels, incl. diphthongs; C: #consonants; T: #tones; S:#stresses, if applicable), syllable index (based on LAPSyD), and the size of syllable inventory (#distinct syllables) are provided, as well as morphological typology information (continued).

Language	ISO 639-3 code	Phonological system		Syllable index	Size of syllable inventory	Morphological type	Corpus
German	DEU	C V S	25 32 1	8	6 867	Synthetic/ Fusional	WebCelex
Italian	ITA	C V S	25 8 1	6	1 970	Synthetic/ Fusional	The Corpus PAISÀ
Japanese	JPN	C V	16 10	4	484	Synthetic/ Agglutinative	The Corpus of Spontaneous Japanese (CSJ)
Korean	KOR	C V	22 8	4	2 319	Synthetic/ Agglutinative	Leipzig Corpora Collection (LCC)
Mandarin	CMN	C V T	25 7 5	4	1 378	Analytic/ Isolating	Chinese Internet Corpus (S. Sharoff)
Swahili	SWH	C V	30 5	2	1 447	Synthetic/ Agglutinative	[Gelas, Besacier, & Pellegrino, 2012]

The small sample considered here also provides some variation in terms of morphological type. Morphological typology deals with the internal word structures. Languages are usually categorized along two dimensions: i) the internal complexity of words in terms of number of morphemes and ii) the assembling strategy for these morphemes. These two dimensions give rise to several morphological language types [Aikhenvald, 2007].

Regarding the number of morphemes per word, linguists distinguish between analytic and synthetic languages[51]. Analytic languages tend to limit the number of morphemes they pack in each word, a one-to-one correspondence being the norm. Synthetic languages on the contrary, make frequent use of words consisting of several morphemes. This distinction should be seen as a continuum, ranging from strictly analytic languages (e.g.

[51]There is also a third category which encompasses languages that express in one word what the other languages would distribute over several lexemes. These languages, such as Algonquian languages in Northern America, are called polysynthetic

Vietnamese) to languages where most words consist of several morphemes (e.g. Korean). Between them, one finds languages that lean towards analytic behavior (e.g. English has a tendency to have a low number of morphemes per word) or towards synthetic word formation (e.g. French and Italian are moderately synthetic).

With regards to the assembling strategy, the strict analytical languages have only one morpheme per word and they are thus said to be isolating. Languages that allow or impose several morphemes per word fall into two categories: Agglutinative languages (such as Korean and Japanese) have a strong tendency to maintain clear boundaries between these morphemes. In agglutinative languages, a word typically consists of a sequence in which each morpheme is clearly identified and carries one semantic feature (e.g. number, case, gender). In fusional languages, on the contrary, several semantic features may be merged into one morpheme and it may be difficult to identify the morphemes from the word-form. Romance and Germanic languages are fusional to some degree.

These categories of word formation only provide an outline that cannot account for the richness of morphological processing, both in terms of verbal vs. nominal domains or derivational vs. inflectional dimensions. For instance, both French and German are classified as synthetic / fusional languages, but nominal morphology is more elaborated in German than in French because of the case-marking system. In the rest of this paper, we only scratched the surface of this richness by comparing the *FL* patterns obtained with corpora consisting of lemmas vs. inflected forms, in order to shed light on potential differences between lexical and grammatical (bound) morphemes.

4.2.3 Data and preprocessing

For each corpus, the first step consisted of discarding erroneous word-forms (including non-alphabetical characters). Then, a specific preprocessing was applied as a function of the corpus nature.

For Mandarin, the Chinese Internet Corpus [Sharoff, 2006] was retrieved online. For

Cantonese, the *Linguistic corpus of mid-20th century Hong Kong Cantonese* [Research Centre on Linguistics and Language Information Sciences, 2013] was also downloaded. For both languages, public domain dictionaries and software - the CC-CEDICT dictionary [CC-CEDICT, 2012] and NJStar Chinese Word Processor [NJStar Software Corp, 2013] for Mandarin and CantoDict [Sheik, 2013] and JyutDict [Learner, 2013] for Cantonese - were used to get the pinyin and jyutping transcriptions respectively. For Mandarin, the transcription software was used when an entry of the corpus was missing in the dictionary. For Cantonese, the transcriptions provided by the two dictionaries were compared and, when differences between transcriptions reflected on-going changes, the most traditional pronunciations were retained. With assistance from Pr. Feng Wang at Peking University, the entries of the corpus with no corresponding transcription in the dictionaries were discarded, which reduced the size of the wordlist from 8 531 to 5 713. The corpus of spontaneous Japanese [NINJAL, 2011] provided transcriptions in katakana, which were then converted into phonological transcriptions by using a list of phonemic entities corresponding with morae in katakana. The initial corpus for Korean was retrieved from the Leipzig Corpus Collection and was converted into IPA by using a Korean pronunciation dictionary [Kim et al., 1993]

The WebCelex corpora in English and German [MPI for Psycholinguistics, 2013, 2014] were retrieved online. They included an automatic transcription derived from grapheme-to-phoneme conversion as well as corresponding lemma and grammatical category for each entry of the corpus. For French, Lexique 3.80 [New et al., 2001] was used, which is very similar to WebCelex with transcription, lemma and grammatical category for each word-form of the data. In some French variants, the opposition between /e/ and /ɛ/ tends to be neutralized [Gess, Lyche, & Meisenburg, 2012] but we decided to keep those phonemes apart in the data transcription.

For Italian, the corpus PAISÀ [Lyding et al., 2014] was retrieved online and was transcribed into IPA by using the dictionary of Italian pronunciation [Canepari, 2009]. When

there were missing entries in the dictionary, an automatic phonemic converter [Carnevali, 2009] was used and resulting transcriptions were corrected by the first author in order to follow the transcription rules of the pronunciation dictionary. The initial corpus provided corresponding lemma and grammatical information. Swahili data were collected at the Dynamique Du Langage Laboratory [Gelas, Besacier, & Pellegrino, 2012] and lemmatized with TreeTagger [Schmid, 1995].

For *FL* calculation, the 20 000 most frequent word-forms and lemmas were taken into account respectively from inflected and lemmatized data in each language except for Italian with 14 629 inflected word-forms (corresponding to 8 028 lemmas) and Cantonese with 5 172 entries (due to the relatively small corpus). All phonological entries in each language were syllabified and syllabic boundaries were considered for the computation of *FL*. In Section 4.3, the influence of the following parameters was assessed: TOKEN vs. TYPE and INFlected vs. LEMmatized, which resulted in 4 potential configurations - INF/TOKEN, INF/TYPE, LEM/TOKEN, and LEM/TYPE. For each version, FL_E and $\#MP$ were computed for vowel and consonant contrasts. Appendix A.6 provides a toy example to illustrate these different configurations. In Section 4.4, the *FL* carried by each individual vowel and consonant was calculated and discussed.

Among the four potential configurations above, the three most interesting ones will be reported in the paper. LEM/TYPE is the most lexicon-oriented dataset as it is reduced to lemmas and can be considered as a kind of "core" lexicon. On the contrary, INF/TOKEN version of data was the most usage-oriented corpus. Finally, INF/TYPE data can be regarded as the extended version of the mental lexicon. These three configurations gave insights on the structure of the core lexicon (LEM/TYPE), the influence of the inflectional morphology (INF/TYPE), and finally, the impact of the actual usage (INF/TOKEN).

4.3 Distribution of *FL* for susbsystems of the phonological inventory

In this section, the relative *FL* of each phonological subsystem (vowels, consonants, stress, and tones) are first explored in nine languages (Cantonese, English, French, German, Italian, Japanese, Korean, Mandarin, and Swahili). Further investigations are then performed with five languages (English, French, German, Italian, and Swahili) for which distinctions in terms of TOKEN/TYPE and LEMmatized/INFlected forms could be made. First, the range of variation of segmental FL is explored in the various configurations. The weights assumed by vocalic and consonantal subsystems are then examined.

4.3.1 Contributions of phonological subsystems to *FL*

To compute the *FL* of the phonological subsystems, the INF/TOKEN configuration was considered, as it was the only one available for all languages. Table 4.2 represents the *FL* associated with each phonological subsystem – vowels (FL_V) and consonants (FL_C) – as well as tones (FL_T) in Cantonese and Mandarin and lexical stresses (FL_S) in English, German, and Italian. FL reflects the relative importance of subsystem within each language.

Although the difference between consonantal and vocalic weight may be limited (as in French), FL_C was higher than FL_V in all nine languages. This result might be expected because of a universal trend to have more consonants than vowels in most of the world's languages: In LAPSyD [Maddieson et al., 2013] 646 over 696 languages have strictly more consonants than vowels. However, in the case of German, there were more vowels than consonants in the phonological inventory (32 vowels vs. 25 consonants in the data description) and the gap between FL_V and FL_C did not remarkably differ from those in other languages. Furthermore, the FL_V of German was the median in the dataset while the size of its vowel inventory was the largest.

While further investigating the influence of inventory size, a positive significant correlation between the size of the consonant inventory and FL_C was revealed (Spearman's $\rho = 0.792$; p-value = 0.011; N = 9). There was however no correlation between FL_V and the size of vowel inventory (Spearman's $\rho = 0.519$; p-value = 0.152; N = 9). For instance, the FL_V of a 5-vowel language (Swahili) and that of a 32-vowel language (German) were very similar while the FL_C of Swahili with 30 consonants differed considerably from that of Japanese with 16 consonants.

The impact of lexical tone was visible, with FL_T close to FL_V in Cantonese and superior to FL_V in Mandarin. Lexical stress had also some impact in Italian ($FL_S = 0.24\%$), but almost no impact in English and German.[52]

Table 4.2: Functional loads carried by vowels, consonants, tones and stress and Infra-syllabic FL_{VCTS}.

Language	yue	eng	fra	deu	ita	jpn	kor	cmn	swh
FL_V	4.55	6.70	14.83	4.37	7.61	3.76	3.30	3.24	4.11
FL_C	10.64	20.82	19.41	15.45	11.12	9.39	11.50	13.09	20.0
FL_S/FL_T	4.48	0.005	-	0.01	0.24	-	-	4.13	-
FL_{VCTS}	62.50	52.30	55.35	47.95	44.74	44.08	45.32	58.08	53.97

Information gathered in Table 4.2 is illustrated in Figure 4.2. The individual contribution of each phonological subsystem is displayed by the bars and the infra-syllabic FL_{VCTS} is represented by diamonds. Several studies have examined the relative importance of tone within a phonological system [Hua & Dodd, 2000] [Oh et al., 2013] [Surendran & Levow, 2004]. [Hua & Dodd, 2000] highlighted that in early language acquisition, tones are acquired earlier than other elements of syllables and that their role in distinguishing lexical meaning is more crucial than phonemes. In a corpus-based study, [Surendran & Levow, 2004] showed that the amount of information carried by tones is as important as the amount carried by vowels in Mandarin. [Oh et al., 2013] later confirmed this result

[52]In English and in German, homophony induced by stress coalescence is rare because of the high redundancy between stress and vowel quality encoding in WebCelex. Moreover when homophony arises, it impacts low frequency items.

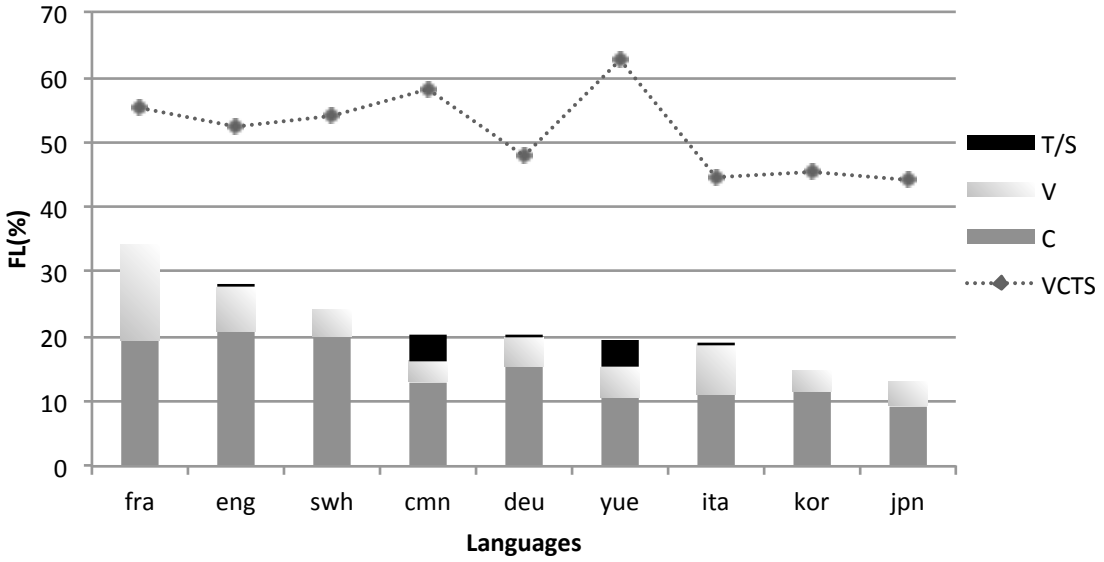

Figure 4.2: Functional loads carried by vowels (V), consonants (C), tones (T) and stress (S) and Infra-syllabic *FL* (FL_{VCTS}). X-axis shows languages by decreasing order of summed FL_V and FL_C.

with Cantonese data. Our results were in line with this and also suggested that there is no compensation between consonantal and tonal subsystems (see [Maddieson, 2007] and [Hombert, Ohala, & Ewan, 1979], for a diachronic perspective). We indeed found that both Cantonese and Mandarin relied on higher infra-syllabic FL_{VCTS} values than the other seven languages. However, the fact that the two tonal languages considered here are also isolating prevented from concluding on the origin of the heavy weight of the infra-syllabic information. More languages, with various tone systems, would be necessary to further assess this pattern.

4.3.2 Frequency, morphology, and *FL*

For English, French, German, Italian, and Swahili, lemmas corresponding to inflected forms were available, and INF/TOKEN, INF/TYPE and LEM/TYPE corpora could be extracted and investigated. None of these languages had tones, and lexical stress in English, German, and Italian was ignored given its very low *FL* with respect to consonants

and vowels.

The importance of the whole phonological inventory was assessed by examining FL_{VC} (Figure 4.3). Cross-language variations were visible, with a similar magnitude in the three corpus configurations. For LEM/TYPE corpora, the segmental *FL* varied from 37.9% in German to 57.6% in Swahili. In English, German, and Italian, segmental *FL* was lower than 50%, which implies that distinctions between lemmas mostly relied on the structural information in these three languages. Considering inflected forms rather than lemmas (LEM/TYPE vs. INF/TYPE comparison) had a limited impact on the load carried by segments, except in Italian. However, interpretations may differ across languages. In English, the identical FL_{VC} values reflected the limited productivity of the inflectional morphology. In German (and to a lesser extent in French and Swahili), the relative steadiness observed meant that the inflectional system is relatively neutral vis-à-vis the proportion of information based upon segments. In Italian, by contrast, word-forms were more distinguished via segmental differences in the INF/TYPE configuration than in the LEM/TYPE configuration (46.0% vs. 39.7% for FL_{VC}). This result is compatible with the regular inflectional system that produces a lot of (vowel) alternations in suffixes, both in verbal and nominal morphology.

FL consequently revealed that about one half of the words' "identity" was carried by other means than segmental distinctions in these five languages. This result may reflect a balance between time-localized (i.e. segmental) information and information spread along the whole word in speech communication. Such a syntagmatic organization may be more robust to noise and local degradation than a system where most of the information on word identity depends on a short-time window. Speakers tend to modulate their utterances during speech communication in order to optimize their transmission capacity. They are also likely to reduce words with less information (i.e. words with higher predictability) by employing both surface and structural information for estimating the predictability of words (see [Levy & Jaeger, 2007], among many others).

The importance of token frequency is abundantly described in psycholinguistics, where frequency effects are well documented, and it is also a corner stone in exemplar models in phonology [Johnson, 1996] [Pierrehumbert, 2001, 2003] [Walsh et al., 2010]. Here, we looked at the global changes induced in *FL* patterns when comparing type and token frequencies. Although the range of the cross-language variations was almost unchanged (FL_{VC} ranged from 40.3% in Italian to 55.4% in French), language-specific effects were visible.

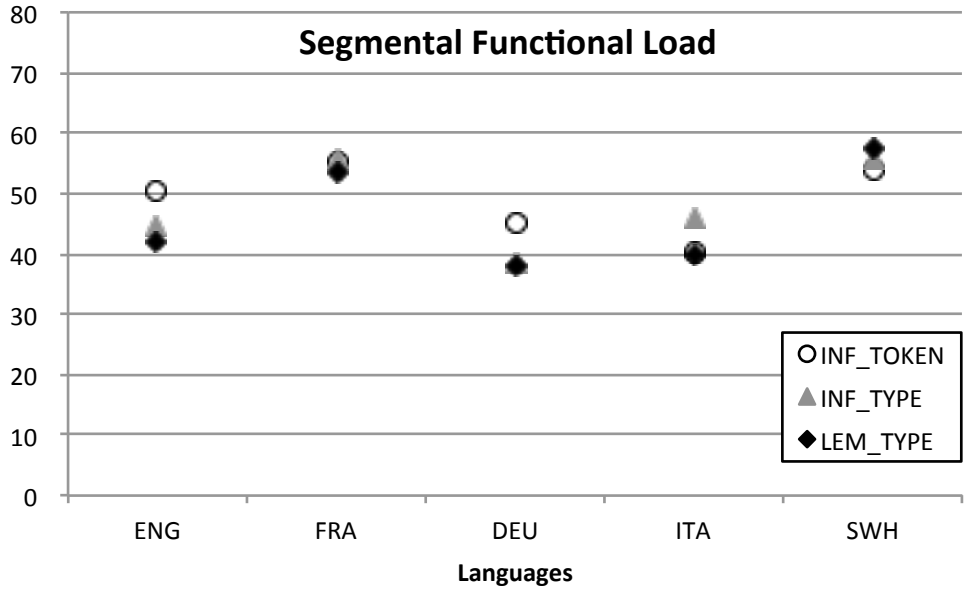

Figure 4.3: Segmental functional load (FL_{VC}) in five languages according to corpus configuration.

In English and German, shifting from type to token increased the weight of segments in distinguishing among inflected forms (+5.6 points and +6.4 points respectively). This effect was probably a consequence of the predominance of shorter words, including many monosyllabic words[53] in the most frequent words [Bell et al., 2009], [Zipf, 1949]. These words have more phonological neighbors with high frequency and they more heavily rely on segmental contrasts than longer low-frequency words since they incorporate much less

[53]The English corpus includes more than 5 700 different monosyllabic word-forms, and the German corpus more than 1 600 ones.

structural information. An opposite trend was visible in Italian, since the segmental *FL* diminished from 46.0% to 40.3% from the type to the token-based corpus.

Compared to English and German, Italian has a lower syllabic complexity which clearly limits the number of monosyllabic word-forms (less than 500 are present in the corpus) and may explain this different behavior. In French and Swahili, changes induced by taking inflections and token frequencies into account were limited compared to other languages. Moreover, in the three corpus configurations, segmental loads were higher than in the other languages (values between 53.9% and 57.6% in Swahili, and between 53.7% and 55.7% in French). In Swahili, this preponderance shall be put in perspective with both the vastly predominant CV syllable structure (except in loanwords) and the strict morphological structure induced by Bantu case marking and verbal morphology. As a consequence, structural information is more limited in Swahili than in fusional languages which allow more variations, in frequent as well as infrequent word-forms. In French, the interpretation is different. On the one hand, a large variety of syllabic structures are present, allowing a large number of monosyllabic word-forms for instance (more than 3 600 are present in the corpus), in contrast to Italian and Swahili. On the other hand, the role of segments in lexical distinctions (as illustrated through the LEM/TYPE configuration) is much larger than in English and German.

An interim conclusion is that variations were visible in i) the relative weight of segmental vs. structural information in lexical distinctions and ii) the impact of token frequencies on this balance. The small language sample prevented from drawing any typological conclusions, but it suggested that the relative weight of segmental vs. structural information results from an interaction of factors that cannot be reduced to the basic size of the phonological system.

FL_V and FL_C values for each corpus configuration are presented in Table 4.3. $\#MP$ are not reported because of their similarity with FL_E estimated from types. FL_V ranged from 1.4% to 14.8%, whether accounting for frequency or morphology. FL_C ranged accord-

ingly from 9.5% to 24.4%. FL_E values for INF/TYPE and INF/TOKEN configurations were highly correlated (Spearman's $\rho = 0.952^{**}$; p-value < 0.001; V and C series pooled together; N = 10).

Table 4.3: Functional loads (in %) associated with vowel and consonant inventories, as a function of the corpus configuration in five languages (see text for details).

		Language	eng	fra	deu	ita	swh
TYPE	INF	FL_V	3.5	7.6	2.0	6.1	3.6
		FL_C	18.0	15.7	11.8	11.2	16.8
	LEM	FL_V	3.0	5.2	1.4	1.8	5.6
		FL_C	14.8	15.2	9.8	9.5	24.4
TOKEN	INF	FL_V	6.7	14.8	4.4	7.6	4.1
		FL_C	20.8	19.4	15.4	11.1	20.0

Reinforcing observations made in Section 4.3.1, FL_C was higher than FL_V in the five languages, for each corpus configuration. While there was a positive significant correlation between the size of the consonant inventory and FL_C for nine languages, there was none between the size of a phonological system (i.e. vowel or consonant subsystem) and its global FL neither in INF/TYPE (Spearman's $\rho = 0.215$; p-value = 0.551; V and C series pooled together; N = 10) nor in INF/TOKEN (Spearman's $\rho = 0.325$; p-value = 0.359; N = 10). These results indicated that the size of a phonological system was not a good predictor of the amount of lexical information its segmental contrasts accounted for.

4.3.3 Consonontal bias

In order to investigate more specifically the potential bias towards consonants vs. vowels, we defined the difference-over-sum of FL_C and FL_V, expressed as a percentage:

$$CBias = 100 * \frac{FL_C - FL_V}{FL_C + FL_V} \tag{4.8}$$

If the vocalic and consonantal subsystems have equal FL (unbiased system), *CBias* is equal to zero. The more a system is biased towards consonants, the higher *CBias* is, up

to a theoretical limit of 100%. On the contrary, a system biased towards vowels would yield negative values, with a theoretical limit of –100%. The difference-over-sum provides a normalized criterion to contrast languages with each other and it is more appropriate than the difference FL_C-FL_V since a significant range of variation exists for both FL_C and FL_V.

CBias indices are given in Figure 4.4. Three series, corresponding to each corpus configuration, are displayed. In LEM/TYPE configuration, a strong positive *CBias* was visible for each language. It ranged from 49.1% in French to 75.2% in German.

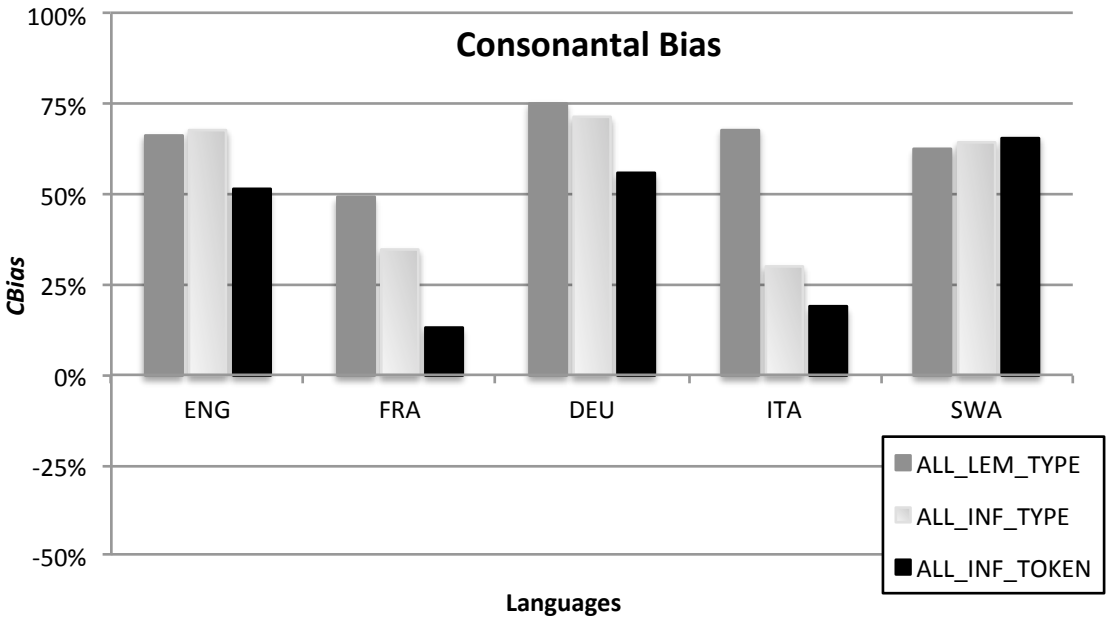

Figure 4.4: *CBias* according to corpus configuration

We then explored the influence of corpus configuration (in terms of TOKEN vs. TYPE, and LEMmatized vs. INFlected data) on *CBias*. Regarding the influence of inflectional morphology, several patterns were visible on the INF/TYPE series (Figure 4.4). Though German and English are quite distinct from each other in terms of richness of inflectional morphology (both verbal and nominal), they exhibited almost similar patterns, with a limited impact with regard to the lemmatized configuration. In French and Italian, on the contrary, changes were notable, with *CBias* dropping from 68.2% (LEM/TYPE) to 30.0%

(INF/TYPE) in Italian. In Swahili, changes between LEM and INF corpora were limited. These results suggested that this bias is not only a matter of morphological productivity.

Taking token frequency into account (INF/TOKEN series) led to decreasing *CBias*, except in Swahili. Even if it resulted in a low consonantal bias in French (13.4%), no language reached a situation biased toward vowels or even balanced. Cross-language differences were nevertheless much more visible in this configuration than in the LEM/TYPE configuration previously discussed, with *CBias* ranging from 13.4% in French to 65.9% in Swahili.

This approach revealed the existence of a large *CBias* in the core lexicon (LEM/TYPE configuration) in the five languages. The magnitude of this effect was not directly linked either to the absolute size of the vowel system (Swahili exhibited a large value with a 5-vowel system) or to its relative size compared to the number of consonants (German showed the highest *CBias* though it has more vowels than consonants). Moreover, *Cbias* seemed to be insensitive to syllabic complexity and syllable inventory size (English and Swahili reached similar magnitudes with very different syllabic complexity). The comparison of LEM/TYPE and INF/TYPE configurations provided a way to evaluate the impact of the inflectional morphology. Two profiles were shown. On the one hand, morphology had a limited impact on *CBias* in English, German, and Swahili, though these languages drastically differ in their morphological productivity. On the other hand, inflectional morphemes had a tendency to counter-balance the bias towards consonants in French and especially in Italian. Finally, when token frequency is considered, i.e. when we switched from a "flat" lexical representation of word-forms to a usage-based representation, the *CBias* range of variation became larger, even if this pattern was still present in the five languages.

Computing the *CBias* for Cantonese, Japanese, Korean, and Mandarin in INF/TOKEN configuration led to 40.1%, 42.8%, 55.4% and 60.3% respectively. These values were all positive, and fell within the range of previous values.

These results suggested that the consonantal bias may be a robust trend at the lexical level, beyond large typological differences among languages in terms of size of phonological system, syllabic complexity, and morphology. This *CBias* was nevertheless modulated by usage, with possible consequences on the cognitive representations of the speakers.

4.4 Distribution of *FL* within phonological subsystems

In this section, all nine languages are considered in INF/TOKEN configuration. The distributions of FL_E and $\#MP$ are investigated in the vowel and consonant subsystems, as well as their consequences in terms of system economy. The individual phonemes with the highest FL_E and $\#MP$ in each language are then discussed from a typological perspective. Like in Section 4.3, the 20 000 most frequent word-forms were employed, except in Cantonese and Italian where only 5 172 and 14 629 entries were present respectively, due to limitations in corpus size. Language data and preprocessing were previously described in detail in Subsection 4.2.3.

4.4.1 Patterns in *FL* distributions

Up to this point, we presented cumulative results, at the scale of each phonological subsystem or at the more global scale of infra-syllabic information as a whole. *FL* is also useful to rank contrasts within a language subsystem and to cross-linguistically compare their distributions. In Figures 4.5 and 4.6, such distributions are displayed for vowels and consonants respectively. Pairs are ranked by decreasing order of *FL* on the x-axis with FL_E on the left y-axis (grey triangles) and $\#MP$ on the right y-axis (black circles). Since the number of contrasts lawfully followed the number of vowels and consonants in each language according to a $n(n\text{-}1)/2$ relationship, x-axis ranges differ between languages. Accordingly, the y-axes depend on FL_E and $\#MP$ values but scales have been matched in order to ease comparison of the distribution shapes. The first striking observation is that

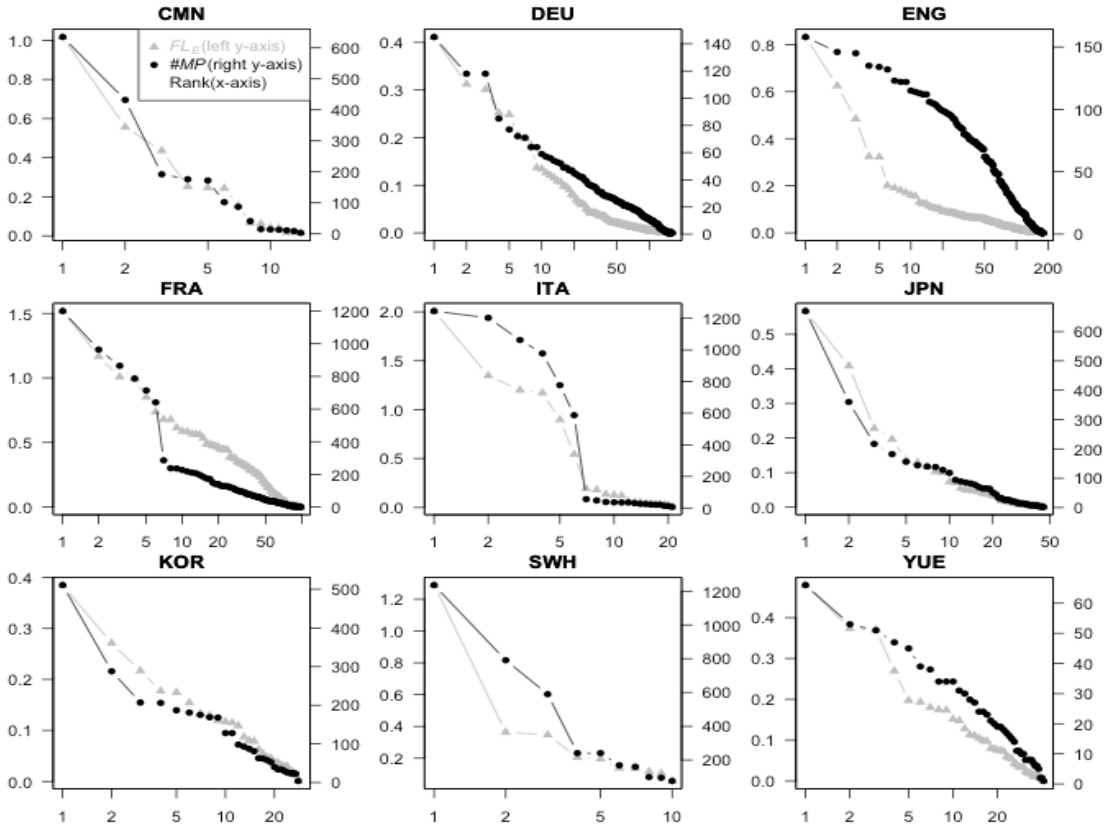

Figure 4.5: Distribution of vowel pairs: FL_E on the left y-axis (in gray) and $\#MP$ on the right y-axis (in black). Pairs are listed by their decreasing order of FL values using a logarithmic scale.

none of the nine languages evenly relied on its vowel or consonantal system to carry its FL. For both vocalic and consonantal contrasts (for $\#MP$ and FL_E), the general shape consisted of two sections: high-ranked contrasts, characterized by a rather abrupt decline, and low-ranked contrasts, with a slow decrease. The relative size of each section might be variable, but most of the time, it consisted of five pairs or less, which is a very small number of contrasts to rely on. Despite this common trend towards uneven distributions, language-specific differences were also visible. In some cases, the decline was regular, without any clear inflection point (e.g. distribution of vowel contrasts in German or Cantonese, or distribution of consonant contrasts in English). On the contrary, Italian for vowels and Japanese for consonants exhibited "S-shape" distributions. In Italian, the first two vocalic contrasts were involved in almost the same number of minimal pairs, and the same pat-

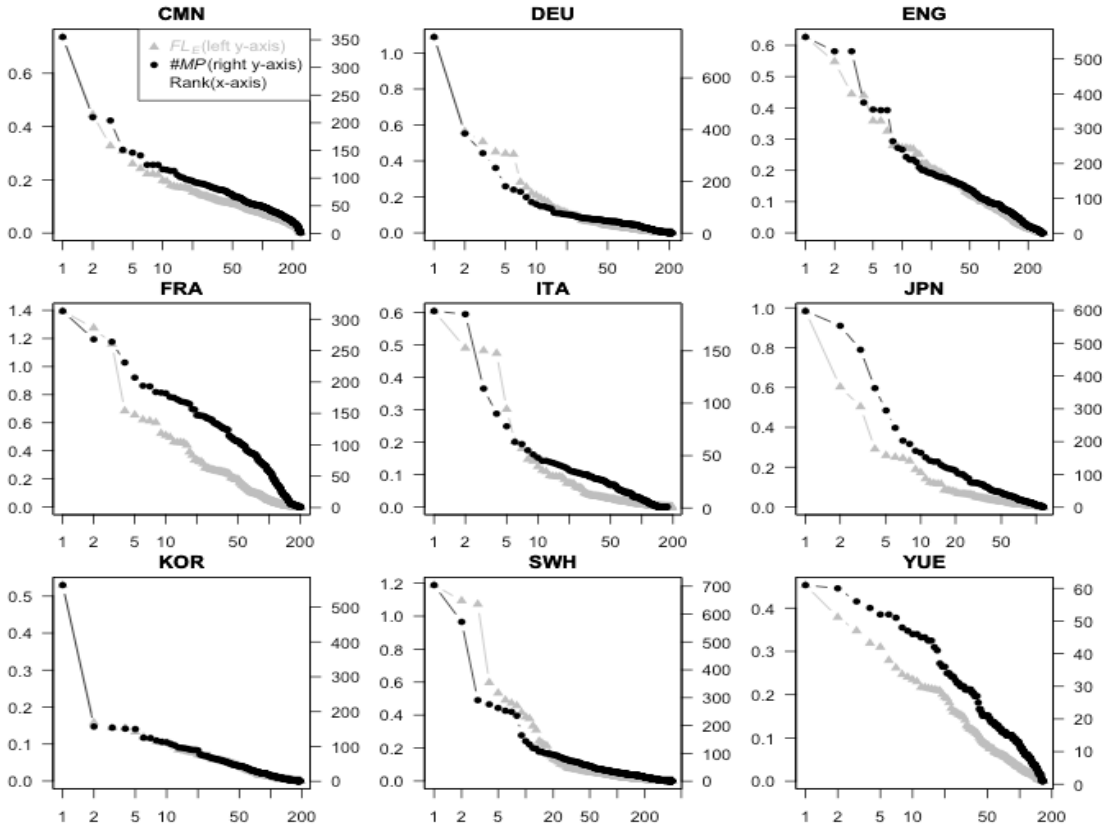

Figure 4.6: Distribution of consonant pairs: FL_E on the left y-axis (in gray) and $\#MP$ on the right y-axis (in black). Pairs are listed by their decreasing order of FL values using a logarithmic scale.

tern held for consonants. In other cases, the decrease in FL between the first and the second contrast was large (e.g. in Japanese, Korean, Swahili for vowels and in German and Korean for consonants). Cross-linguistically, phonological contrasts didn't follow a regular distribution, such as Zipf's law (observed for word-form frequencies) or another heavy-tailed distribution (such as Yule distribution, see [Martindale et al., 1996]).

Comparison between $\#MP$ and FL_E distributions may also be insightful since they point towards potentially different cognitive processes. $\#MP$ distribution is related to the whole set of word-forms in the language, and it thus corresponds to the organization of mental lexicon. In contrast, by including token frequency, FL_E is more related to frequent words and to online processing in situations of communication. In several cases, the two distributions were analogous (e.g. Korean for consonants, or Mandarin for both vowels and

consonants). In other cases, the different distributions observed meant that the structure of the basic lexicon (consisting of the frequent word-forms) differed from the structure of the extended lexicon. More precisely, two patterns were present. When FL_E distribution was partly above the $\#MP$ distribution, as for vowels in Korean or consonants in German or Swahili, a few contrasts were promoted by usage. On the contrary, having the FL_E distribution below the $\#MP$ distribution signified that for frequent word-forms, less information was conveyed by the infra-syllabic level. This pattern is common in our sample (in German, English, Italian, and Cantonese for vowels and in French, Japanese, and Cantonese for consonants). It may be related to the amount of other linguistic information available, which helps to understand words and consequently limits the burden carried by each word itself.

We showed in Figures 4.5 and 4.6 that a lot of contrasts were characterized by a very low FL and that they marginally contributed to the segmental FL. They conveyed consequently a very low amount of information and we performed a simulation in order to evaluate how the nine languages behave at the systemic level in this respect. The algorithmic principle was to reduce the phonological set, by iteratively eliminating the segment with the smallest FL until only one segment remained. For instance, in Swahili, we observed for the vowels: $FL(/e/) < FL(/o/) < FL(/u/) < FL(/i/) < FL(/a/)$. In the first iteration, /e/ was eliminated from the system, and coalesced with the vowel /a/ with which it was involved in the maximum number of minimal pairs. We computed the relative loss of entropy corresponding to the lexicon described by this new 4-vowel system. In the second iteration, /o/ underwent the coalescence process, resulting in a lexicon described by a 3-vowel system. The process was next applied to /u/, then to /i/, and resulted in a 1-vowel system (with entropy consequently equal to FL_V). The results of this simulation are displayed in Figures 4.7 and 4.8.[54] For legibility, the y-axis represents the proportion of initial entropy preserved in the altered system. It is thus the complement of FL on 100%.

[54]See Appendix A.7 for the list of the contrasting pairs of vowels & consonants.

The iteration step in the simulation is indicated on the x-axis (zero being the original system, with a *FL* of 100%).

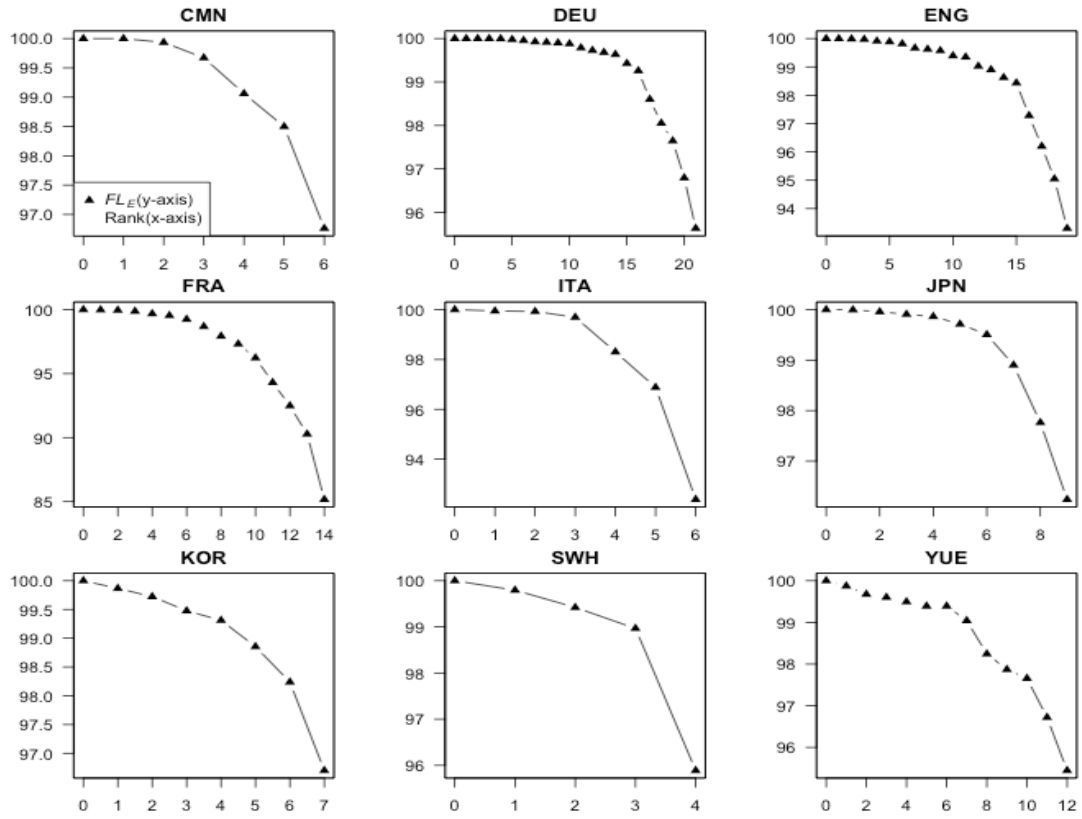

Figure 4.7: Simulation of the relative loss of entropy induced by reducing vowel system, % of FL_E on the y-axis (in black), phonemes are listed by their increasing order of *FL* (x-axis).

Two major patterns are visible in the graphs. The first configuration illustrated that some systems were more sensitive to changes induced by the reduction process. This pattern was present for instance in Korean and Swahili for vowels, and in Mandarin, Japanese, and Cantonese for consonants. In most cases, however, systems were very resilient to reducing the size of the phonological systems, and the loss in *FL* induced was barely noticeable at least at the beginning of the process. It was especially salient in German and English for vowels and for German, English, French, Italian, Korean, and Swahili for consonants. In German, for instance, the majority of the vowel system could

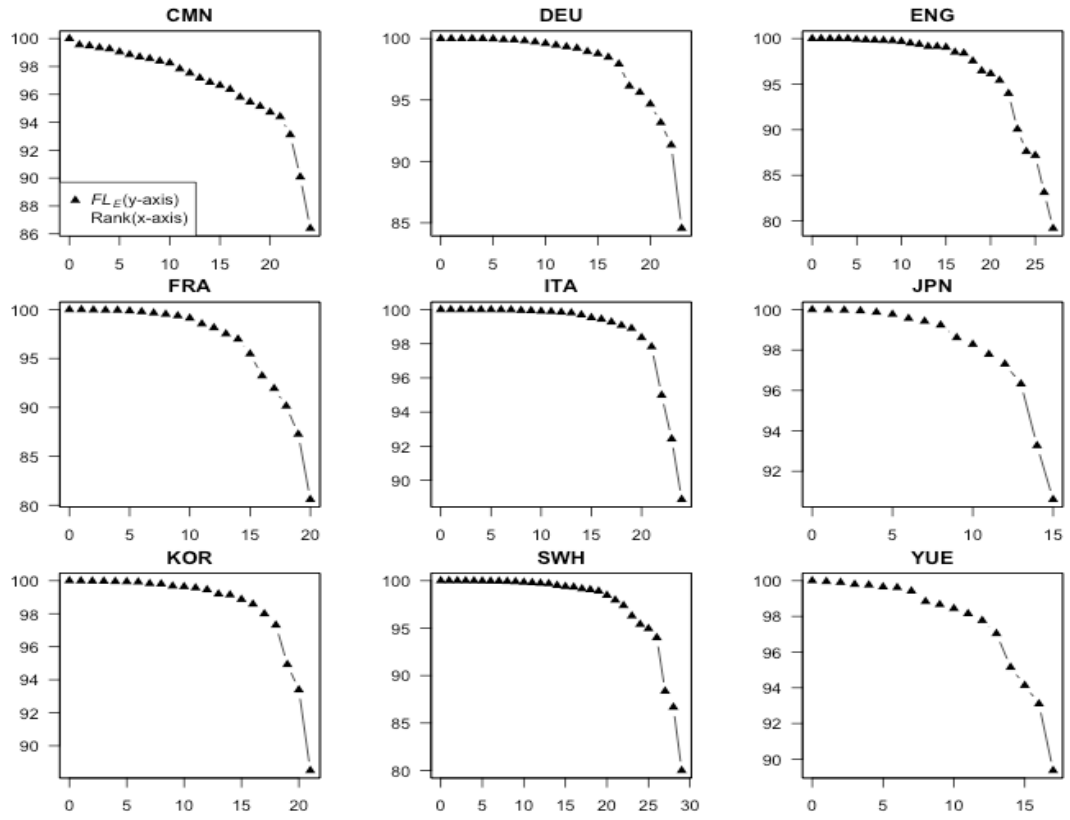

Figure 4.8: Simulation of the relative loss of entropy induced by reducing consonant system, % of FL_E on the y-axis (in black), phonemes are listed by their increasing order of FL (x-axis).

be coalesced with an information loss of less than 1%.

In general, the amount of information loss induced by a merger varied more widely in consonants than in vowels among the languages, which is consistent with the larger FL associated to the consonantal component. One major exception is Italian, for which a drastic reduction of the number of consonants would have minor consequences in terms of information loss. This result is coherent with Section 4.3 on *CBias* where the importance of consonants in structuring the lexicon was highlighted, but limited in Italian.

One could consider that keeping such contrasts distinct in a language is costly and provides no real advantage, especially if these contrasts rely on segments that do not participate in any high FL pair. However, [Vitevitch, 2008] described the self-organization

of phonological word-forms in the mental lexicon by employing the concepts of small-world topology and scale-free network. These networks are characterized by a small average path length, a high clustering coefficient (for both network patterns), a power-law degree-distribution and a preferential attachment (for the latter) in the growth theory of Barabási-Albert [Barabási & Albert, 1999]. In this approach, the structures of phonological system and mental lexicon can both be described as scale-free networks due to their preferential attachment - a small number of giant components with many other smaller components. Such properties facilitate language acquisition, production and perception with its robustness and resilience to errors and damages of components. From this perspective, the observed distribution of vowel and consonant systems shown in the figures above can be regarded as the consequences of cognitive efficiency and optimization for language acquisition and information retrieval, which is a robust property of natural languages. For instance, [Morales & Taylor, 2007] have shown that variable frequencies of language elements improve language acquisition compared to the elements with equal frequencies. Such characteristics of a natural language which self-organizes the structure of its systems result from the cognitive efficiency and optimization during language learning and speech communication.

4.4.2 Cross-language trends in preferred phonological features

Figures 4.5 and 4.6 pointed towards the high proportion of information coded by the five highest-ranked contrasts in the nine languages. Consequently, we further examined these specific contrasts in this subsection, as well as the highest-ranked segments themselves. Tables 4.4 and 4.5 display the five vowel pairs with the highest FL_E computed with the INF/TOKEN configuration of corpus and the five individual vowels with the highest FL_E respectively.

Among the five vowel pairs with the highest FL_E listed by their decreasing order of FL_E in Table 4.4, there was no pair which was present in all the nine language studied. In

	Language																	
	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	ɔː-aː	0.48	aɪ-eɪ	0.83	e-a	1.52	a-ɛ	0.41	e-a	2.01	e-a	0.57	i-e	0.39	ə-a	1.02	i-a	1.29
2	ɛː-ɔː	0.37	ɪ-æ	0.62	ø-e	1.17	ɪ-aɪ	0.31	i-e	1.35	o-a	0.41	o-i	0.27	u-i	0.56	u-i	0.36
3	o-ɐ	0.37	eɪ-iː	0.48	ø-a	1.01	a-ɪ	0.30	i-a	1.20	i-a	0.23	i-a	0.22	u-ə	0.44	u-a	0.35
4	aː-ɐ	0.27	aɪ-iː	0.32	ã-e	0.99	aː-iː	0.25	o-a	1.17	o-e	0.20	o-e	0.18	u-a	0.25	e-a	0.21
5	u-i	0.20	ɪ-ɒ	0.32	ɛ-e	0.85	a-aɪ	0.25	o-i	0.90	u-i	0.14	o-a	0.17	y-i	0.25	o-a	0.20

Table 4.4: 5 Vowel pairs with the highest FL_E

fact, we observed 28 different contrasts (the maximum possible being 45) composed of 18 different vowels. However, four contrasts appeared in four different languages: /i-a/, /i-u/, /e-a/ and /o-a/. Interestingly, they rely on /i, e, a, o, u/, the five most frequent vowels in the world's languages. Among those four contrasts, three involved the low vowel /a/, this vowel being implicated as well in eight of the nine most important contrasts found in our sample. This points towards a particular role of the maximally opened vowel. The only language without the vowel /a/ in its most salient contrast is Korean, with the pair /i-e/. This time it is the maximally closed vowel that is found. Again vowel height seems to be an important dimension for vowel oppositions as it operates in 16 out of the 28 different most salient contrasts, either maximally /i-a/ or minimally /i-e/ for example.

Although Swahili obeyed a kind of maximum contrast selection (with respectively /i-a/, /u-i/, and /u-a/ on the podium), the general trend was to prefer moderate to low acoustical distances in these contrast sets, as illustrated by /ɔː-aː/ in Cantonese or /a-ɛ/ in German. Redundant contrasts, defined as contrasts where more than one feature (frontness, aperture, and rounding) is involved, were also very common but they were rarely based on a secondary feature, with the exceptions of /aː-ɐ/ in Cantonese and /ɑ̃-e/ in French. In Italian, three of the five pairs with the highest FL_E, (/e-a/, /i-e/, and /i-a/) seem to reflect the inflectional morphology as they contain the thematic vowels /a/, /e/, and /i/, which is the marker of inflection class in verbal morphology [Da Tos, 2013].

Several remarks can be made at the level of the vowels themselves (Table 4.5). First,

	Language																	
	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	ɔː	0.71	eɪ	1.12	e	3.63	a	0.71	a	2.34	a	0.76	i	0.58	u	1.73	a	1.02
2	aː	0.66	aɪ	1.00	a	3.51	iː	0.68	e	2.14	e	0.50	a	0.48	i	1.71	i	0.95
3	ɐ	0.65	iː	0.99	ø	2.74	aɪ	0.57	i	1.87	o	0.48	o	0.48	ə	1.66	u	0.45
4	iː	0.45	ɪ	0.93	ã	2.72	ɪ	0.52	o	1.34	i	0.33	e	0.36	a	1.54	o	0.29
5	ɛː	0.39	æ	0.75	ɛ	2.36	ɛ	0.46	ɔ	0.29	oː	0.25	ʌ	0.27	y	0.54	e	0.24

Table 4.5: 5 Individual vowels with the highest FL_E

the differences among the five vowels with the highest FL_E were less important than the ones between the five most salient contrasts, this means that the load is more evenly divided at the level of the segments than what appears to be when looking directly at contrasts. Second, for almost all languages, the vowels with the highest FL_E were the ones implicated in the five most salient contrasts. When looking at the vowel qualities present in this set, we observed 24 different vowels (again maximum is 45). The low vowel (/a/-like) was not always the preferred attractor or hub, (four languages out of nine) but it was present in the table for each language, either as a monophthong or as the beginning of a diphthong. It is followed by /i/ or /iː/, present in eight out of nine languages. /e/ and /o/ or /oː/ were found in five languages. Surprisingly, the back high vowel /u/ is only present in two languages (Mandarin and Swahili), yet the five most frequent vowels are the most contrast bearer ones. In terms of features, among the 45 vowels and diphthongs of the table, 23 vowels are front, 10 are central (incl. /a/-beginning diphthongs) and 12 are back. Finally, we noticed that the larger the vowel inventories, the more likely the set of "preferred" vowels will be to include vowels other than /i, e, a, o, u/.

The first remark that can be made for consonants (Tables 4.6 and 4.7) is that they show more variability than vowels. We observed 37 different contrasts out of the 45 possible relying on 22 different consonants. Only six contrasts were present in more than one language: three in three different languages and three in two different languages. All six contain coronal consonants and four include a nasal. These trends can in fact be

	Language																	
	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	n-m	0.45	n-t	0.63	l-d	1.40	R,r-n	1.09	l-n	0.60	s-k	0.98	l-n	0.53	t-l	0.74	j-n	1.19
2	ts-t	0.38	z-t	0.55	l-s	1.28	R,r-m	0.57	s-d	0.49	w-g	0.60	g-t	0.16	ŋ-n	0.45	j-w	1.09
3	ts-k	0.35	h-ð	0.44	s-d	1.16	z-d	0.51	l-d	0.48	n-t	0.50	n-g	0.14	t-ʂ	0.33	w-n	0.17
4	ts-j	0.32	n-z	0.44	n-d	0.69	s-n	0.45	n-d	0.47	m-n	0.29	n-d	0.14	tʂ-k	0.31	z-j	0.60
5	ts-s	0.31	ð-b	0.36	l-n	0.66	v-d	0.44	k-l	0.30	m-k	0.26	n-m	0.13	tɕ-ɕ	0.26	j-l	0.53

Table 4.6: 5 Consonant pairs with the highest FL_E

generalized across the entire set of preferred contrasts. The first rank contrast involved at least one coronal consonant in all 9 languages. More generally, coronal consonants are present in 43 of the 45 contrasts listed in Table 4.6, with a prominence of the voiced nasal /n/ (in 18 contrasts), followed by the voiced stop /d/ and the lateral approximant /l/ (both in 9 contrasts). In terms of manner of articulation, oral and nasal stops, fricatives, affricates, and approximants are present, with a preference for nasals and stops, followed by fricatives and approximants.

	Language																	
	yue		eng		fra		deu		ita		jpn		kor		cmn		swh	
1	ts	1.36	t	1.74	s	3.40	n	1.49	d	1.07	k	1.26	n	0.79	t	3.44	n	2.18
2	k	1.28	n	1.57	l	3.25	R,r	1.17	l	0.96	s	0.86	g	0.61	l	2.86	j	2.08
3	s	1.08	m	1.35	d	3.14	m	1.03	n	0.81	t	0.79	l	0.51	ʂ	2.85	w	2.01
4	h	0.96	ð	1.28	m	2.01	d	0.85	s	0.76	n	0.74	s^h	0.46	tʂ	2.53	l	1.35
5	t	0.95	s	1.24	n	1.93	z	0.74	k	0.46	m	0.58	d	0.42	p	2.12	z	1.31

Table 4.7: 5 Individual consonants with the highest FL_E

Table 7 shows the five consonants with the highest FL_E. We found 19 different consonants out of the 45 possible, 13 of which were coronal. 8 out of 19 different consonants were found in more than one language, only two of them were not coronals (/m/ and /k/). Coronal consonants appeared with various manners in the first row in all languages except Japanese (/k/). Another general trend was a preference for voiced consonants,

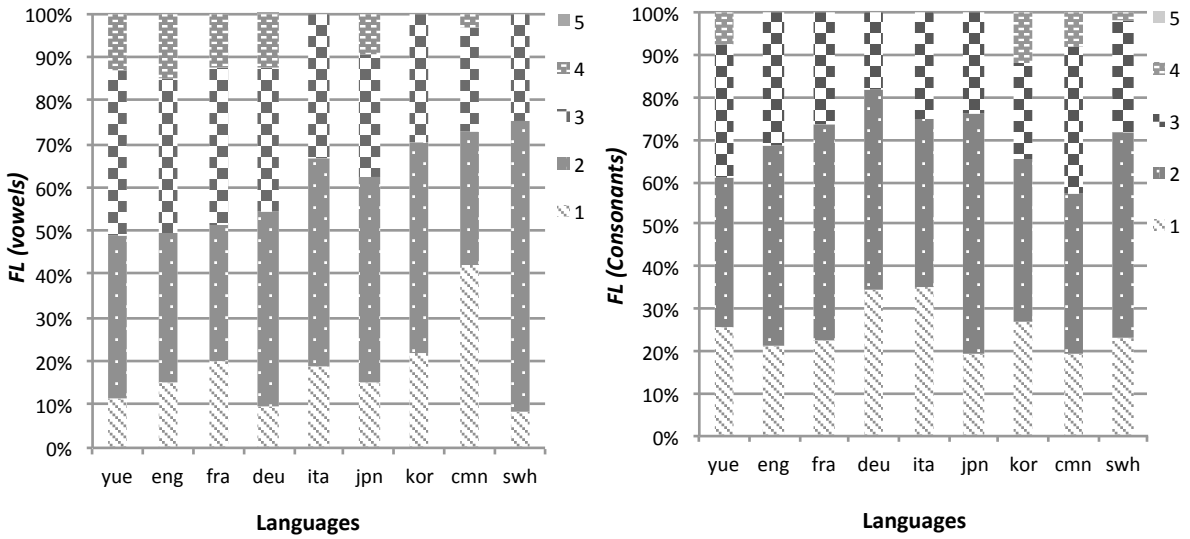

Figure 4.9: Distribution of FL_E as a function of feature distances of the contrasts. Left: vowels. Right: consonants.

which accounted for 27 consonants over all 45. This preference was nevertheless relative, since five over nine first-rank consonants were voiceless, and this reversed tendency even pervaded almost the entire table for Cantonese and Mandarin. To this regard, we can note that when the consonant inventory of the language includes voiced stops, the most frequent contrasts rely on sonorants, whereas if the inventory lacks voiced stops the most frequent contrasts involved obstruents.

Finally, we adopted a different perspective by investigating the *FL* distribution in terms of distance between the members of contrastive pairs. Figure 4.9 shows FL_E distributed according to a feature distance, for vowels (Left panel) and consonants (Right panel). The feature distance between two segments was computed on the basis of their segmental definitions in the UPSID database [Maddieson, 1984] [Maddieson & Precoda, 1990]. Specifically, features are compared within the natural classes they belong to (frontness, roundedness, manner, place, etc.), and the distance is equal to the number of classes in which segments differ. Secondary contrasts such as *nasalized* or *long* define distinct additional classes. For example, the distance between /i/ {*high*; *front*; *unrounded*} and /u/ {*high*; *back*; *rounded*} is 2. The distance between /oː/ {*long*; *lower-mid*; *back*; *rounded*}

and /õ/ {*nasalized*; *lower-mid*; *back*; *rounded*} is 2 also since the *nasalized* and *long* features belong to two distinct classes. The distance between /p/ {*unvoiced*; *labial*; *occlusive*} and /v/ {*voiced*; *labio-dental*; *fricative*} is 3. In the data set, most distances ranged from 1 (e.g. /n-m/) to 4, with six contrasts yielding a distance of 5 (/ãː-ʏ/, /ãː-ɔ/, /ɶ̃ː-au/ in German and /k^{hw}-j/, /w-k^{hw}/, /m-k^{hw}/ in Cantonese). For vowels in the nine languages, more than 50% of the *FL* was carried by distinctions of one or two features except in Cantonese and English. However 2-feature contrasts were favored over 1-feature or 3-feature contrasts, except in Cantonese, English, French, and Mandarin. It highlighted a trend to prefer redundant vocalic contrasts over the most economical ones (1-feature contrasts). In Mandarin, 1-feature contrasts almost accounted for one half of the total FL_V by themselves. On the contrary, in Cantonese, English and French, 3-feature contrasts were the most important. One can also mention that in French, 4-feature contrasts were the favored ones. They involved one nasal vowel and one oral vowel with qualities differing in their 3 dimensions and their importance may be related to the frequent use of grammatical words consisting only of one nasal vowel (such as on [õ], un [œ̃], en [ɑ̃]).

For consonants, "2-feature-or-more" contrasts were in majority, and similarly to vowels, a cross-linguistic tendency towards an economical system was illustrated by the predominance of 2-feature contrasts.

Comparing the results of vowel systems and consonant systems in the 9 languages leads us to assume that cognitive principles for organizing vowel and consonant systems are different in nature. In the case of vowel systems, languages employ the principle of maximal perceptual contrasts [Jakobson, 1941] for organizing phonological structures and lexicon. In the context of language acquisition, Rose, as cited in [Van Severen et al., 2012], mentioned that consonants with high *FL* tend to have the least articulatory complexity and the highest perceptual salience, which corresponds to the characteristics shared by the coronal consonants [Rose, 2009]. Presumably, different acoustic characteristics of vowels and consonants may also play an important role regarding the different organizations of

both vowel and consonant minimal pairs - the perception of consonants is more categorical and the perception of vowels is more continuous [Liberman et al., 1975] [Fry et al., 1962].

4.5 General discussion

The information-theoretic approach implemented in this paper directly bridged the level of the phonological components and the level of the lexicon. We thus proposed a shift from a common view of phonological systems as an inventory of components (segments, stress, tones) toward a functional perspective encompassing lexical relationship between these components. This approach relies on large corpora and facilitates cross-language comparison since the same methodology was applied to each language.[55]

4.5.1 *FL* at the level of phonological subsystems

The study presented in Section 4.3 gave support to the existence of a lexical consonantal bias in five languages (two Romance languages, two Germanic languages, and one Bantu language). Japanese, Korean and two Sinitic languages were further examples where FL_C was much larger than FL_V. The index we defined, *CBias*, ranged from 49% to 75%, reflecting a preference toward consonant-based distinctions rather than vowel-based distinctions when analyzing a corpus of lemmatized forms and leaving token frequency aside. However, this trend was modulated as soon as inflected word-forms and/or token frequency were considered. In the INF/TOKEN corpus configuration, for instance, *CBias* ranged from 13% in French to 66% in Swahili, with various tendencies among the languages. Consequently, this consonantal trend should not be seen as an absolute and monolithic phenomenon since it resulted from the interaction between several linguis-

[55] Additional studies will obviously be necessary to extend the report done in the following lines to a larger number of languages. We thus do not pretend reaching any typological conclusion given the small language set studied so far. Similarly, the robustness of our approach has to be more thoroughly investigated. Preliminary experiments showed that distributional patterns seem to be robust against the variation in the corpora size.

tic dimensions (phonological inventories, but also syllabic diversity, and morphological type, as well as differences between lemmas and affix structures). For instance, Italian and Swahili had somewhat similar *CBias* at the lemmatized level (respectively 68% and 63%) but their behavior drastically differed in the INF/TOKEN configuration (resp. 19% and 66%). These observed differences between type and token *FL* may be related to language-specific configurations in phonological representations and mental lexicon, with consequences on online processing as well as on the dynamics of language acquisition.[56]

In their seminal paper, Nespor, Peña, and Mehler advocated for a greater relevance of consonants to build the lexicon, and a greater relevance of vowels to carry grammatical information, and they mentioned linguistic and cognitive motivations [Nespor, Peña, & Mehler, 2003]. They indicated the facts that most languages have more consonants than vowels in their inventories, that the number of consonantal "slots" is larger or equal to the number of vocalic slots in syllables (except in the basic V syllable structure) and finally that consonants have a general tendency to disharmonize within words, while vowel harmony (as well as vowel reduction) is frequent in the world's languages. According to these authors, these factors converge towards a more salient role of consonants than vowels in word distinctiveness.[57] Further evidence comes from psycholinguistic experiments on word transformations [Cutler et al., 2000] and later confirmation in language acquisition [Nazzi, 2005] [Nazzi & New, 2007]. Nespor, Peña, and Mehler also mentioned that in the area of inflectional morphology, the "division of labor" between consonants and vowels has some "fuzzy boundaries", leaving a more thorough assessment to future investigation [Nespor, Peña, & Mehler, 2003, p.204]. Nazzi and New shed some light on this issue by showing that in French the whole lexicon (roots and inflected forms) relies less

[56]For instance, [Kissling, 2012] showed that phonological differences in two languages impact short-term memory processing. More precisely, she showed that English native speakers recall vowel series better than consonant series whereas the reverse is true for Arabic native speakers. In our opinion, corpus studies based on data collected during language acquisition would offer an interesting perspective to complement psycholinguistic experiments on vowel and consonant perception and representation.

[57]It has also been argued that speech consists more of consonantal than vocalic substance (in terms of duration), but [Easterday, Timm, & Maddieson, 2011] mitigated this assumption since in their corpus of 22 languages, the proportion of vocalic duration ranged from 43.3% to 60.1%, with an average of 53.8%.

heavily on consonantal contrasts than lexical roots only, when types are considered [Nazzi & New, 2007, p.277]. They thus endorsed the influence of morphology on the relative role of consonants in the lexicon. This statement was supported by the present study, as the *CBias* effects for French and Italian indicated that the inflectional system moderates consonantal bias to some degree, in contrast with the effects for German. More generally, comparing *CBias* between LEM/TYPE and INF/TYPE configurations may help refining the "fuzzy boundaries" for each language considered.

Moreover, recent studies show that the role of consonants to access the lexicon might not be as monolithic as supposed, and especially that there is an interaction between the information carried by consonants or vowels and their position in words. Estimating this information through conditional entropy, Tanaka-Ishii nicely established that in English, at the beginning and at the end of the words, information carried by consonants is much larger than information carried by vowels, while within words, this difference is reduced [Tanaka-Ishii, 2012]. Very recently, Delle Luche and colleagues also showed that consonantal bias is sensitive to the syllable and rhythm structure of the words in French and English [Delle Luche et al., 2014]. Finally, it is important to notice that the consonant advantage visible in the lexicon disappears in production and perception, and is even replaced by a vowel advantage, when whole sentences are considered [Fogerty & Humes, 2012] [Kewley-Port, Burkle, & Lee, 2007] [Owren & Cardillo, 2006]. Stilp and Kluender, in a radical acoustic approach that does not consider segments as primitives, also show a prevalence of vowels over consonants in speech intervals characterized by high values of their index of cochlea-scaled spectral entropy (and thus high information amount) [Stilp & Kluender, 2010]. The approach developed in this section did not address the balance between consonantal and vocalic information in sentences since it was based on lexical data. However, the differences observed between processing at word and sentence levels are consistent with the importance of temporal organization of information in speech. Under this view, the differences of lexical structures revealed in this section, for instance between

type frequency and token frequency, may reflect this prominence, since token frequency not only influences cognitive representations but also expectations (and thus information) in the processing of connected speech. The corpus-oriented study presented here, although limited, can complement other approaches, such as behavioral experiments in the search for explanations of the distinct role of consonants and vowels in language. Section 4.3 also aimed at studying the relative contribution of vowels, consonants, stress, and tones to lexical distinctions. The importance of tone system in Cantonese and Mandarin was first confirmed. Together with their isolating morphology which strikingly limits the structural information in the lexicon, it might explain the large infra-syllabic *FL* observed for these two languages (63% and 58% respectively). Among the nine languages on average, 51.7% of the lexical distinctions relied on infra-syllabic components. It pointed towards a balance between localized short-term information (measured by infra-syllabic *FL*) and longer term information. One has nevertheless to keep in mind that the phonemic transcriptions of word-forms only provide part of the picture. The speech phonetic substance is not in a one-to-one relationship with the phonemic "theoretical" sequence and continuous speech moreover involves predictability effects that alter the realization and perception of the words themselves (see [Aylett & Turk, 2004] [Levy & Jaeger, 2007] [Piantadosi, Tily, & Gibson, 2009] for discussion).

4.5.2 *FL* distribution within phonological subsystems

As developed in Section 4.4, uneven distributions of *FL* among the available contrasts were also present in the nine languages and suggested the existence of a cross-linguistic trend. Hockett's diagnostic quoted in the introduction was thus confirmed, and our quantitative approach also shed light on the concentration of *FL* on very few contrasts (Figures 4.5 & 4.6). In the case of vocalic contrasts, they were moreover built upon a small set of vowels, while, for consonants, these high-*FL* contrasts are more disseminated over the consonant system, yet it is important to note the strong presence of coronals and nasals

in the set of most salient consonants. Finally, we observed a small significant negative correlation between the *FL* of consonantal contrasts and the feature-distance of its constituents: the higher the *FL*, the closer the members of the pair (it was just a tendency for vowels).

Finally, a remarkable trend was illustrated in Figures 4.7 and 4.8 despite the differences in phonological inventories among the sample. *FL* concentration on a few contrasts also resulted in a kind of resilience of the lexicon vis-à-vis an alteration of its phonological inventory. For the nine languages, the simulations based on an iterative process of coalescence, yielded a two-phase pattern: removing step by step the majority of the phonemes led to a gradual and limited decrease of the *FL*. The second phase, characterized, on the contrary, by an abrupt slope, led to major changes in the information encoded by the phonological system. It would be interesting to reproduce the same methodology with a larger number of languages.

The existence of cross-language trends should not hide that language-specific patterns were also revealed. For instance, the differences between FL_E and $\#MP$ distributions (Figure 4.5 and 4.6) widely varied from one language to another, especially for vowels. In some cases, taking token frequency (as in FL_E) into account led to more continuous distributions while in other cases, considering only minimal pairs, without any usage-based count (as in $\#MP$) yielded the most regular distributions. Such differences might i) mirror structural differences in the language lexicon and ii) have consequences on the cognitive processing of the speakers' mental lexicon. Further studies, including a more comprehensive examination of each language distribution, will be necessary to go beyond this simple report.

4.5.3 Conclusion

We would like to highlight that the distributions studied here may be put in relation with graph representations of lexicons, phonological systems, etc. The methodology pre-

sented here makes the phonological system *emerge* from interactions between word-forms in a lexicon. These interactions are often represented as graphs, and their regularities are often viewed as mirroring the phenomena from which they develop (see [Arbesman, Strogatz, & Vitevitch, 2012] [Gerlach & Altmann, 2013] [Jäger, 2012] [Kello & Beltz, 2009] and [Kello et al., 2012], for discussion). When it comes to language, emergence can be considered at different levels. Moulin-Frier and colleagues emphasizes how phonological properties may emerge from a set of nonlinguistic (cognitive, motor, perceptual, communicative) abilities [Moulin-Frier et al., forthcoming]. Implementing language games additionally highlights how properties shared by a community of speakers may emerge from local interactions. These two perspectives are at work in the COSMO model. However, the linguistic structures manipulated in language games cannot yet approach the complexity of real word-forms, and *FL* is thus insightful to investigate how actual word-forms interact. Avoiding homophony arising from phonetic change, for example in the case of the loss of stop codas /p, t ,k/ between Late Middle Chinese and Standard Mandarin, may lead to the emergence of new phonemic contrasts. Moreover, other evolutions may take place, as it was the case in Chinese, at the morphological level with the disyllabification of words, which reduced homophony. Diachronic corpora of texts may therefore be useful to test evolutionary hypotheses, and move beyond synchronic analyses of *FL* as those performed in this paper.

Chapter 5

Conclusion

...at many levels and time-scales, language provides the necessary conditions to support spontaneous emergence of patterns through self-organizational pathways [Wedel, 2011].

The main goal of this thesis was to investigate general tendencies (i.e. statistical universals) among typologically diverse languages within the complex systems framework. In this framework, language is characterized as an emergent self-organizing system resulting from multi-constrained optimization and is structured for optimal and efficient communication. Thus, it is assumed that self-organization phenomena exist at several levels of linguistic analysis, due to cognitive optimization. To confirm this hypothesis, three studies were conducted from a typological and quantitative perspective, by means of information-theoretic measures respectively at the macrosystemic, mesosystemic, and microsystemic levels.

In the general introduction (Chapter 1), some basic notions used in the thesis and some of the relevant studies were presented:
(i) Regarding the general framework of the thesis, the notion of complex adaptive system was described and the classification of language universals proposed by Comrie and the two main contrasting approaches to language universals (Greenberg vs. Chomsky) were illustrated.

(ii) In terms of methodology, the two different approaches to quantifying linguistic complexity were described: traditional linguistic approach (grammar-based complexity) vs. information-theoretic approach (usage-based complexity).
(iii) With respect to language-external factors influencing information encoding and transmission, sociolinguistic factors (e.g. population size, geographic spread, and the degree of linguistic contact) and neurocognitive factors (e.g. "a conflict of interest" between speaker and hearer and the UID hypothesis) were discussed.
(iv) Previous relevant studies on information encoding were revisited: Zipf's law, the entropy rate constancy principle, the uniform information density hypothesis, and the smooth signal redundancy hypothesis.

The first study (Chapter 2) was focused on the assessment of cross-language tendencies of information encoding among the 18 languages, based on the initial hypothesis proposed in [Pellegrino, Coupé, & Marsico, 2011] that the average information rate (IR) does not differ significantly among languages due to a trade-off between speech rate (SR) and information density (ID). In addition to the extension of the previous study from 7 to 18 languages, an information-theoretic approach was added in order to examine whether there is a significant correlation between the IR computed by a pairwise comparison using Vietnamese as a reference and the IR obtained by information-theoretic measures. The Information theory was chosen as a methodological framework since it provides the mathematical formalization of information density. Our approach corresponds to a cross-language study on a general tendency of regulating the average information flux among the 18 languages and differs from the UID hypothesis which is more related to the cognitive aspects of "speakers' choices about structuring their utterances" using also information-theoretic measures.

In the results of the first study, among the information-theoretic measures used for computing ID, those taking account of context (i.e. conditional entropy $H(X_n/X_{n-1})$ and $H(X_n/X_{n+1})$) are more strongly correlated with the syntagmatic measure of ID than Shan-

non entropy and surprisal which were obtained from a unigram language model. Furthermore, conditional entropy is also strongly connected with the morphological strategies of languages (e.g. the patterns of affixation) and it distinguishes between synthetic and analytic languages by exhibiting a lower conditional entropy for synthetic languages than analytic languages, except for Mandarin Chinese.

Regarding the relationship among *IR* obtained by information-theoretic measures and by a pairwise comparison, *IR* computed by conditional entropy ($IR_{H(X_n|X_{n-1})}$ and $IR_{H(X_n|X_{n+1})}$) exhibits less cross-language variation and lower average value than $IR_{H(X)}$ and $IR_{S(X)}$. Such result suggests that when context is taken into account, the languages with different *IR* are leveled out and that conditional entropy could be matched to the effort of disambiguation (hearer's effort) whilst Shannon entropy is regarded as the memory and recognition (the effort for both speaker and hearer).

One of the major limitation of the information-theoretic approach is that it is strongly dependent on the size and the characteristics of corpus, as shown in Subsection 2.3.2. In order to better estimate the distribution of syllable frequencies, a text corpus large enough to build a robust language model is required. In addition, an oral corpus which is phonologically balanced and large enough is necessary to better predict the average *IR*. In this study, there are several languages with limited text corpora and the length of oral corpus is quite short, containing 3−5 sentences in each text (for a total of 15 texts). If the data size is relatively small, Shannon entropy is considered as a more appropriate measure of *ID* than surprisal which is more data-dependent since it takes the individual syllable into account on the local scale. For instance, contrary to Shannon entropy and conditional entropy, *IR* obtained from surprisal is not significantly correlated with the interaction between *SR* and *ID*, since surprisal is more dependent on corpus than Shannon entropy and conditional entropy.

In the second study (Chapter 3), the mesosystemic relationship between phonological complexity and morphological complexity was assessed, based on the equal complexity

hypothesis. Along with holistic typology, this hypothesis has been criticized for its falsifiability and absence of null hypothesis since the end of the 20th century in modern theoretical linguistics. The aim of this study was thus to investigate whether a phenomenon of self-organization exists between the complexity of linguistic subsystems such as phonology and morphology, which is assumed to be manifested by a negative correlation between phonological complexity and morphological complexity. The notion *complexity* allows us to quantify the *richness* of linguistic system and the regulations (i.e. the structure of expression) and compare typologically distinct languages. Phonological complexity was obtained by estimating the average amount of information (in bits) which is required to encode a random syllable, by means of Shannon entropy and conditional entropy.

Some general tendencies were found among the 14 languages classified according to holistic morphological typology. It was shown that in comparison with analytic languages, synthetic languages (i) exhibit higher *SR* and encode less information per syllable, (ii) display lower phonological complexity and more complex inflectional morphology. Although those tendencies should be confirmed with a wide range of languages, these results provide a hopeful evidence in favor of the traditional morphological classification and holistic typology.

Furthermore, it was suggested that morphological complexity was negatively correlated with phonological complexity: in particular, while there was no significant correlation between morphological complexity and Shannon entropy, two conditional entropy $H(X_n|X_{n-1})$ and $H(X_n|X_{n+1})$ were significantly and negatively correlated with morphological complexity, which led to the conclusion that conditional entropy reflects the structure of words whereas Shannon entropy is more related to the size of syllable inventory. However, in the framework of complex adaptive system, the role of non-linguistic factors, i.e. sociolinguistic factors and neurocognitive constraints, in optimally balancing linguistic complexity should be highlighted and deserves further studies.

In the third study (Chapter 4), the phenomenon of self-organization was investigated

at the microsystemic level, by using an information-theoretic measure, functional load (FL) which is a tool for measuring the relative importance carried by phonemic contrasts. The goal of this study was to find general cross-language tendencies of the organization of phonological subsystems among the 9 languages. This chapter consisted of the following two studies: (i) the relative importance of phonological subsystems (vowels, consonants, stress and tones) was examined and compared among the 2 tonal and 7 non-tonal languages, taking morphological strategies and usage frequency into account, (ii) the internal organization of each phonological subsystem (vowels and consonants) among the 9 languages were compared.

Regarding the relative importance of phonological subsystems, the results confirmed that among phonological subsystems, consonants play a more important role in lexical access than vowels and, in particular, variations were visible among the languages if morphological strategies and usage frequency are considered. In terms of the internal organization within a phonological subsystem, it was shown that only a few phoneme contrasts play an important role in lexical access among the 9 languages while high-FL contrasts are language-specific and there are no general tendency found among them. Such characteristic of phonological system is considered as the result of cognitive optimization, allowing the system to be robust and resilient to damage and errors, despite language specificities.

This thesis aims to provide a multi-level study from a typological and quantitative approach at the macrosystemic, mesosystemic, and microsystemic levels of linguistic analysis, by analyzing the written and spoken linguistic data. The results of the three studies have suggested or provided supports for the following arguments:

(i) Languages have been structured by their usage to optimally encode and transmit information in human communication.

(ii) Within the framework of complex adaptive systems, language is defined as a self-organizing system which is characterized by the phenomenon of emergence and self-organization.

(iii) Due to the characteristics of language as a complex adaptive system, some general tendencies (not absolute universals) are observed among the typologically distinct languages.

(iv) Those cross-language tendencies are found at the three different levels: first, in terms of the average information rate in speech communication at the macrosystemic level, second, in terms of the trade-off between linguistic complexity in phonology and morphology at the mesosystemic level, and third, in terms of the internal structure of linguistic subsystem in phonology at the microsystemic level.

As for the perspective for further studies, the following points can be developed:

(i) Enlarging the language sample and adding more languages from various language families with simple syllable structures (e.g. Hawaiian and Navajo) or with polysynthetic morphology (e.g. Algonquian languages) in further studies may yield more compelling arguments toward general trends of information rate in typological perspective.

(ii) Non-linguistic aspects are necessary in order to confirm the hypothesis of complex adaptive systems theory that the phenomenon of self-organization results from the interaction between linguistic and extra-linguistic factors. Especially, the relationship between information rate and socio-cognitive factors can be studied by assessing extra-individual factors (i.e. social environments such as speaker population size, geographic spread, and the degree of linguistic contact) [Trudgill, 2011] and intra-individual factors (i.e. sociolinguistic profile of individual speakers such as age, sex and lifestyle). Based on the previous study on the comparison between the *SR*s of Basque/Spanish and Catalan/Spanish bilinguals in Spain [Oh et al., 2013], further studies on the comparison between the *IR* of the bilinguals and the monolinguals, for instance, in Spain, can provide some insight into the effects of sociolinguistic and cognitive factors, while controlling linguistic factors.

(iii) The slope and distributions of *FL* can be further examined based on the graph theory and network science [Barabási & Albert, 1999], which was previously adopted by Vitevitch and colleagues who described the structures of mental lexicon as scale-free networks using

preferential attachment [Arbesman, Strogatz, & Vitevitch, 2012] [Vitevitch, 2008] [Vitevitch, Chan, & Goldstien, 2014].

(iv) As the smooth signal redundancy hypothesis previously suggested that predictability is inversely related with syllable duration and prosodic prominence [Aylett & Turk, 2004], the relationship between IR/ID and language structures can be further assessed at the interface between phonetics and phonology by taking phonetic features into account.

Appendix A

A.1 Information about oral data

Table A.1: Speaker description. For each language, the information regarding speakers (code speaker, #texts, sex, total # speakers, and age) are provided.

Language	Speaker	# Texts	Sex	# Speakers	Age
Basque	F1(Am)	15	F	10	28
	F2(Ux)	15	F		19
	F3(En)	15	F		30
	F4(Ux2)	15	F		29
	F5(Am2)	15	F		31
	M1(Bo)	15	M		26
	M2(Ai)	15	M		36
	M3(An)	15	M		27
	M4(In)	15	M		22
	M5(Ax)	15	M		32
British English	fc	5	F	10	-
	ff	9	F		-
	fg	7	F		-
	fh	1	F		-
	fj	5	F		-
	fa	10	M		-
	fb	2	M		-
	fd	7	M		-
	fe	4	M		-
	fi	10	M		-

Table A.1: Speaker description. For each language, the information regarding speakers (code speaker, #texts, sex, total # speakers, and age) are provided.

Language	Speaker	# Texts	Sex	# Speakers	Age
Cantonese	F1(Vi)	15	F	10	20
	F2(Ka)	15	F		21
	F3(Ce)	15	F		23
	F5(Ye)	15	F		24
	F6(Tra)	15	F		21
	M2(ka)	15	M		20
	M3(Bra)	15	M		22
	M4(Al)	15	M		24
	M5(Ed)	15	M		22
	M6(Hu)	15	M		23
Catalan	F1(Su)	15	F	10	42
	F2(De)	15	F		50
	F3(Mo)	15	F		21
	F4(Mi)	15	F		28
	F5(An)	15	F		39
	M1(Xa)	15	M		28
	M2(Al)	15	M		29
	M3(Da)	15	M		31
	M4(Ma)	15	M		44
	M5(Jo)	15	M		42
Finnish	F1(Ul)	15	F	10	30
	F2(Re)	15	F		35
	F3(Ki)	15	F		41
	F4(Pri)	15	F		16
	F5(Ma)	15	F		22
	M1(Ee)	15	F		52
	M2(He)	15	F		28
	M3(Mi)	15	F		45
	M4(Ma)	15	F		37
	M5(Ma)	15	F		26
French	F1(Je)	15	F	10	25
	F2(Be)	15	F		41
	F3(Ma)	15	F		28
	F4(Lu)	15	F		24
	F5(Na)	15	F		46
	M1(Se)	15	M		37
	M2(Ar)	15	M		36
	M3(Pi)	15	M		27
	M4(No)	15	M		25
	M5(Chr)	15	M		36

Table A.1: Speaker description. For each language, the information regarding speakers (code speaker, #texts, sex, total # speakers, and age) are provided.

Language	Speaker	# Texts	Sex	# Speakers	Age
German	aj	6	F	10	-
	ga	6	F		-
	jm	9	F		-
	mi	9	F		-
	ss	9	F		-
	bg	9	M		-
	hm	9	M		-
	mj	6	M		-
	qk	6	M		-
	sm	6	M		-
Hungarian	F1(An)	15	F	10	39
	F2(Ga)	15	F		33
	F3(Il)	15	F		51
	F4(As)	15	F		57
	F5(Ju)	15	F		31
	M1(Ar)	15	M		42
	M2(Ma)	15	M		27
	M3(Er)	15	M		27
	M4(Ga)	15	M		69
	M5(Mi)	15	M		17
Italian	a0	8	F	10	-
	b6	6	F		-
	ba	4	F		-
	bf	3	F		-
	bl	4	F		-
	ag	6	M		-
	au	3	M		-
	b4	6	M		-
	b7	6	M		-
	bk	8	M		-
Japanese	F1(Ma)	15	F	10	20
	F2(Hi)	15	F		20
	F3(Ju)	15	F		53
	F4(Ay)	15	F		29
	F5(Mi)	15	F		22
	M1(Ni)	15	M		51
	M2(Shi)	15	M		40
	M3(Ke)	15	M		22
	M4(Da)	15	M		21
	M5(Yo)	15	M		28

Table A.1: Speaker description. For each language, the information regarding speakers (code speaker, #texts, sex, total # speakers, and age) are provided.

Language	Speaker	# Texts	Sex	# Speakers	Age
Korean	F1(My)	15	F	10	28
	F2(Ji)	15	F		31
	F3(Eu)	15	F		33
	F4(Hy)	15	F		35
	F6(Jw)	15	F		19
	M1(Sa)	15	M		36
	M2(Do)	15	M		16
	M3(Sh)	15	M		19
	M4(Ju)	15	M		50
	M5(Jh)	15	M		19
Mandarin Chinese	F1(Hu)	15	F	10	19
	F2(Yu)	15	F		19
	F3(Ma)	15	F		25
	F4(Fe)	15	F		28
	F6(Xu)	15	F		-
	M1(Cha)	15	M		19
	M2(Ye)	15	M		24
	M4(Yi)	15	M		31
	M5(Qi)	15	M		24
	M6(Na)	15	M		19
Serbian	F1(Li)	15	F	10	30
	F2(Le)	15	F		34
	F3(Je)	15	F		32
	F4(So)	15	F		31
	F5(Ol)	15	F		38
	M1(Go)	15	M		44
	M2(Iv)	15	M		34
	M3(Pe)	15	M		19
	M4(Vo)	15	M		21
	M5(Ste)	15	M		23
Spanish	F1(Am)	15	F	10	28
	F1(Su)	15	F		42
	F2(De)	15	F		50
	F3(En)	15	F		30
	F3(Mo)	15	F		21
	M1(Bo)	15	M		26
	M4(In)	15	M		22
	M4(Ma)	15	M		44
	M5(Ax)	15	M		32
	M5(Jo)	15	M		42

Table A.1: Speaker description. For each language, the information regarding speakers (code speaker, #texts, sex, total # speakers, and age) are provided.

Language	Speaker	# Texts	Sex	# Speakers	Age
Thai	F1(Ja)	15	F	10	33
	F2(Si)	15	F		28
	F3(Fa)	15	F		23
	F5(Ki)	15	F		32
	F7(Pi)	15	F		43
	M1(Pa)	15	M		27
	M2(Ra)	15	M		23
	M3(Shi)	15	M		31
	M4(Su)	15	M		31
	M5(Ik)	15	M		30
Turkish	F1(Em)	15	F	10	-
	F2(Fe)	15	F		-
	F3(Be)	15	F		25
	F4(Ra)	15	F		31
	F5(Nu)	15	F		37
	M1(Ta)	15	M		-
	M2(Al)	15	M		30
	M3(En)	14	M		37
	M4(An)	15	M		24
	M5(Me)	15	M		44
Vietnamese	F1(DTND)	15	F	10	-
	F2(DTNH)	15	F		-
	F3(Ly)	15	F		25
	F4(Ma)	15	F		26
	F5(Vu)	15	F		21
	M1(NCP1)	15	M		-
	M2(NVS1)	15	M		-
	M3(Van)	15	M		28
	M4(Qua)	15	M		31
	M5(Ti)	15	M		32
Wolof	F1(An)	15	F	10	43
	F2(Ad)	15	F		42
	F3(So)	15	F		35
	F4(To)	15	F		52
	F5(Sa)	15	F		29
	M1(Jl)	15	M		67
	M2(Sa)	15	M		36
	M3(El)	15	M		40
	M4(Da)	15	M		35
	M5(Di)	15	M		55

A.2 Translations of oral script (text Q1)

CAT: En aquell turó hi ha una drecera cap a casa meva. Alguns veïns diuen que el turó està embruixat, de fet, a ningú li fa gràcia passar per aquella zona quan s'ha fet fosc. És clar, jo no em crec aquestes supersticions, i de fet, m'agrada passar per aquella ruta ja que la trobo molt pintoresca.

CMN: 到我家有条翻山的小路。有的本地人说山上有鬼。天黑以后就没有人敢从那走。当然，我是不相信这些迷信的东西。我就是这条风景如画的路。

DEU: Von hier aus gibt es zu meinem haus auch eine abkürzung über den hügel. Die meisten leute erzählen daß es dort spukt. Im dunklen würde keiner von ihnen da lang gehen. Natürlich glaube ich nicht an so einen übernatürlichen blödsinn. Ich gehe aber trotzdem lieber den schönen weg um den hügel herum auch wenn es ein bißchen länger dauert.

FIN: Tuon mäen yli on oikotie kotiini. Joidenkin paikallisten mukaan mäellä kummittelee. Kukaan ei halua mennä noiden peltojen läpi pimeällä. Minä en tietenkään usko mitään tuollaista taikauskoista hölynpölyä. Se on vain viehättävä reitti maatilan läpi.

FRA: Il y a un raccourci par cette colline jusqu'à chez moi. Des gens du coin disent qu'elle est hantée. Personne n'aime traverser ces champs la nuit tombée. Bien sûr, je ne crois pas à ces bêtises superstitieuses. C'est juste que j'aime la promenade pittoresque à travers la propriété.

ENG: There's a short cut over that hill to my house. Some local people say the hill is haunted. No-one likes to pass through those fields after dark. Of course, I don't believe in any of that superstitious nonsense. I just like the picturesque route through the estate.

EUS: Bada bidezidor bat muino hartatik nire etxeraino. Herriko batzuek diote muinoa sorginduta dagoela. Inork ez du hortik igaro nahi gauez. Jakina, nik ez dut horrelako zentzugabekeriarik sinesten. Gogoko dut hango bide bitxi hura.

HUN: Egy kis ösvény vezet a dombon keresztül a házamig. A helybéliek azt mondják, hogy a dombot kísértetek lakják. Senki nem szeret keresztülmenni azokon a mezőkön sötétedés után. Természetesen nem hiszek semmi ilyen babonás szamárságban. Csak szeretem a birtokon keresztül vezető festői szépségű utat.

ITA: C'e' una scorciatoia in collina per arrivare alla mia casa. Alcune persone del luogo dicono che la collina e' abitata dai fantasmi. Nessuno passa di la' dopo il tramonto. Naturalmente io non credo a queste stupide superstizioni. Mi piace molto la strada pittoresca che attraversa la mia tenuta.

JPN: わが家へ至る近道があるんです。その道はあの丘を越えて行くもので、地元の人たちは「幽霊の出る丘」と呼び、暗くなると誰もその道を通りたがりません。もちろん、私はそのような迷信は信じていません。うちの所有地を通るその道は、絵画のようで本当に美しいんですよ。

KOR: 저 언덕 너머에 우리 집으로 가는 지름길이 있어요. 동네 사람들 몇몇은 그 언덕에서 귀신이 나온다고 하죠. 그래서 어두워지고 나면 아무도 그 곳을 지나가려하지 않아요. 전 그런 말도 안되는 미신 따위는 당연히 믿지 않아요. 전 단지 그 곳을 지나갈 때 보이는 그림같은 풍경을 좋아할 뿐이에요.

SPA: A mi casa se puede ir por un atajo cruzando el bosque. Los vecinos dicen que el bosque está embrujado. Nadie quiere pasar por allí cuando oscurece. Yo, por supuesto, no soy supersticioso. A mí me encanta ir por ese camino tan pintoresco.

SRP: Ima jedna prečica preko tog brda do moje kuće. Neki lokalci kažu da na brdu ima duhova. Niko ne voli da prelazi preko tih polja kad padne mrak. Naravno, uopšte ne verujem u ta glupa sujeverja. Jednostavno uživam u šetnji slikovitim pejzažem preko imanja.

THA: มีทางลัดข้ามเขามาที่บ้านฉัน คนแถวนี้บอกว่าทางลัดมีผีคอยหลอกหลอนอยู่ และไม่มีใครกล้าเดินผ่านทางนี้ตอนกลางคืน แต่ฉันก็ไม่เชื่อเรื่องพวกนี้ ฉันแค่ชอบที่จะใช้เส้นทางสวยๆ น่าเดิน.

TUR: Evime giden yol üzerinde kestirme bir patika var. O civardakiler bu yolun büyülü olduğunu söylüyorlar. Hava karardıktan sonra bu yoldan geçmeyi kimse sevmez. Bu tür batıl inançları hep saçma bulmuşumdur. Ben bu manzaralı yoldan yürümeyi çok seviyo- rum.

VIE: Để đến nhà tôi anh có thể đi đường tắt qua quả đồi Người làng tôi không thích đi đường đấy vì họ cho rằng quả đồi có ma nhất là vào ban đêm Tôi không tin vào những lời đồn đại đó Tôi lại rất thích đi đường đó vì nó rất thơ mộng.

WOL: Ngir ñów sama kër, mënees naa jél am mbartal ci tund wi. Waa réew mi taa-muwuñu lool jaare fa nee ñu ndax bërëb boobu dafa am rab, rawati-na bu timis fàddoo. Gëmuma lenn ci wax yooyu. Man, moom, foofu laa tàmma jaar : Aka wuteek yeneen yoon yi !

YUE: 經過嗰座山去我屋企會有一條捷徑，但係有啲當地人話嗰度成日鬧鬼。無人願意係天黑之後行嗰度。當然，我唔信呢啲鬼鬼怪怪嘅野啦。其實我真係好鍾意嗰條風景如畫嘅山路架。

A.3 Comparison of translations

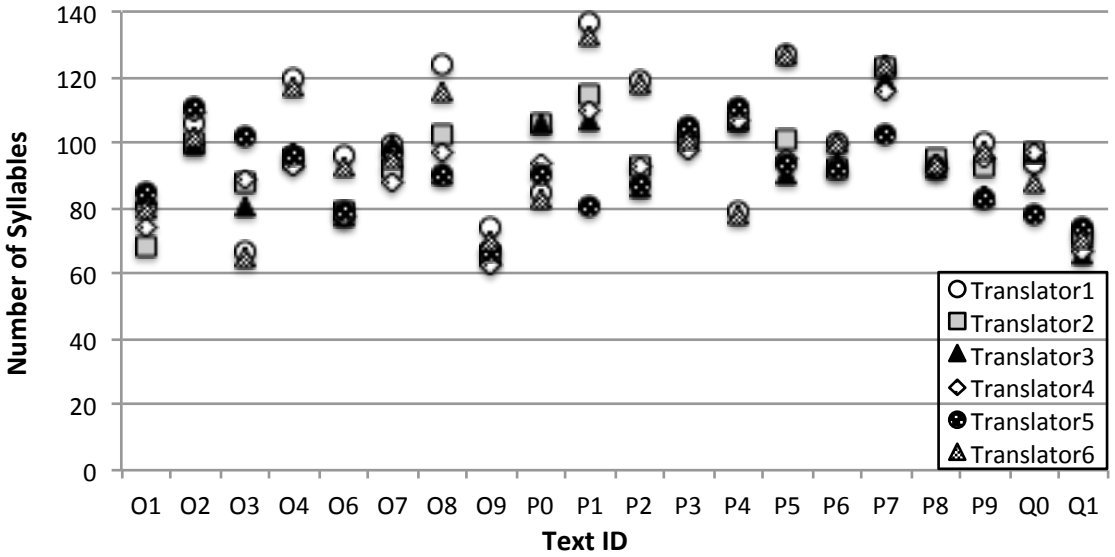

Figure A.1: Comparison of translations in French: Translator6 was used in this study.

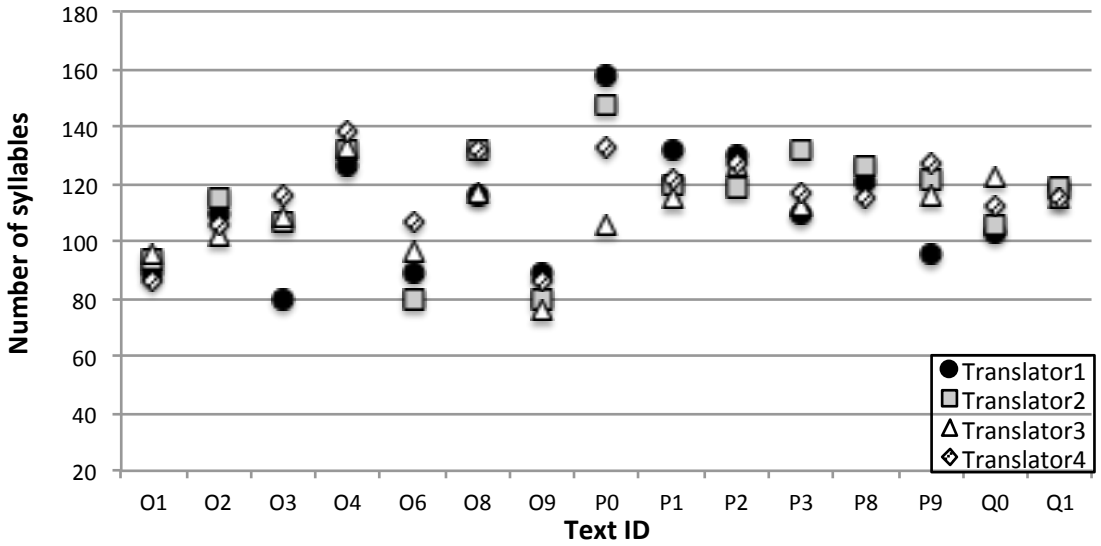

Figure A.2: Comparison of translations in Korean: Translator4 was used in this study.

A.4 20 most frequent words in 18 languages

Figure A.3: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
CAT	1	de	də	75290.53
	2	la	lə	45514.26
	3	a	ə	32028.86
	4	i	i	31990.98
	5	que	kə	26477.24
	6	el	əl	21743.01
	7	un	un	19533.05
	8	en	ən	18316.12
	9	del	dəl	16223.43
	10	per	pər	15926.75
	11	les	ləs	12156.68
	12	els	əls	10928.91
	13	al	əl	10002.71
	14	amb	əm	9567.99
	15	no	nˈo	8313.85
	16	es	əs	6781.37
	17	el	əl	6244.13
	18	ha	a	5256.04
	19	aquest	ə_kˈɛt	5220.11
	20	dels	dəls	5154.30
CMN	1	的	tə5	45402.97
	2	了	lə5	14216.88
	3	我	wə3	12935.94
	4	一	i1	11879.45
	5	在	tsai4	9854.95
	6	是	ʂi4	9574.07
	7	他	tʰa1	6117.56
	8	个	kə4	5910.71
	9	你	ni3	5043.01
	10	和	xə2	4595.85
	11	不	pu4	4594.67
	12	有	jəu3	4503.92
	13	就	tɕiəu4	4194.9
	14	这	tʂə4	4104.01

Figure A.4: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
CMN	15	也	jə3	3876.87
	16	说	ʂuə1	3740.92
	17	人	ʐən2	3738.57
	18	她	$t^{h}a1$	3646.28
	19	着	tʂuə2	3594.69
	20	都	təu1	3361.52
DEU	1	und	ˈʊnt	30086.71
	2	in	ˈɪn	22054.43
	3	der	ˈdeːɐ	14581.28
	4	der	ˈdeːɐ	14581.28
	5	der	ˈdeːɐ	14581.28
	6	die	ˈdiː	13651.73
	7	die	ˈdiː	13651.73
	8	die	ˈdiː	13651.73
	9	nicht	ˈnɪçt	9544.34
	10	ist	ˈɪst	8867.87
	11	es	ˈɛs	7647.66
	12	dass	ˈdas	7034.16
	13	ich	ˈɪç	6968.20
	14	er	ˈeːɐ	6905.03
	15	zu	ˈt͡su	6162.79
	16	zu	ˈt͡su	6162.79
	17	auch	ˈaux	5764.38
	18	von	ˈfɔn	5584.27
	19	von	ˈfɔn	5584.27
	20	mit	ˈmɪt	5121.10
ENG	1	the	ˈðiː	62326.06
	2	of	ˈɒv	31410.03
	3	and	ˈænd	29948.00
	4	a	ˈeɪ	24562.03
	5	a	ˈeɪ	24070.66
	6	in	ˈɪn	19263.82
	7	it	ˈɪt	11626.86
	8	i	ˈaɪ	11523.34
	9	i	ˈaɪ	11523.28
	10	is	ˈɪz	9843.92
	11	he	ˈhiː	9210.18
	12	for	ˈfɔːr	8282.86

Figure A.5: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
ENG	13	that	ˈðæt	7838.07
	14	you	ˈju	7484.18
	15	to	ˈtu	7309.88
	16	with	ˈwɪð	7267.66
	17	be	ˈbi:	6482.88
	18	on	ˈɒn	6346.68
	19	at	ˈæt	5825.58
	20	his	ˈhɪz	5805.29
EUS	1	eta	e_tˈa	44713.38
	2	ez	ˈes	16712.48
	3	da	dˈa	15416.06
	4	ere	e_ɾˈe	9508.37
	5	izan	i_sˈan	6801.69
	6	dira	di_ɾˈa	6407.49
	7	zen	sˈen	6192.70
	8	zuen	su_ˈen	5873.13
	9	egin	e_ɣˈin	5738.16
	10	du	dˈu	5721.98
	11	bere	be_ɾˈe	5435.59
	12	edo	e_ðˈo	5262.77
	13	baina	baɪ_nˈa	4867.20
	14	behar	be_ˈar	4533.91
	15	beste	beʂ_tˈe	3950.44
	16	egiten	e_ɣˈi_cen	3536.77
	17	den	dˈen	3382.33
	18	hau	ˈa_ʊ	3087.71
	19	esan	e_ʂˈan	2962.62
	20	dute	du_tˈe	2799.13
FIN	1	ja	jˈa	36906.95
	2	on	ˈon	28609.60
	3	ei	ˈei	10968.44
	4	että	ˈet_tæ	10314.87
	5	oli	ˈo_lɪ	8485.71
	6	se	sˈe	5980.18
	7	hän	hˈæn	5795.83
	8	mutta	mˈut_ta	5218.66
	9	ovat	ˈo_vat	5152.16
	10	kuin	kwˈin	4733.06

Figure A.6: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
FIN	11	myös	mˈyøs	4716.89
	12	kun	kˈun	4252.55
	13	ole	ˈo_le	4083.71
	14	sen	sˈen	3872.76
	15	tai	tˈai	3434.48
	16	joka	jˈo_ka	3299.60
	17	niin	nˈiːn	3020.47
	18	mukaan	mˈu_kaːn	2936.78
	19	jo	jˈo	2859.18
	20	vain	vˈain	2547.70
FRA	1	de	dø	38928.92
	2	la	la	23633.92
	3	et	e	20879.73
	4	à	a	19209.05
	5	le	lø	18310.95
	6	il	il	15832.09
	7	les	le	14662.3
	8	un	œ̃	13550.68
	9	lˈ	l	12746.76
	10	dˈ	d	11876.35
	11	je	ʒø	10862.77
	12	des	de	10624.93
	13	une	yn	9587.97
	14	pas	pa	8795.14
	15	en	ɑ̃	8732.57
	16	dans	dɑ̃	8296.08
	17	qui	ki	7897.91
	18	ne	nø	7752.09
	19	elle	ɛl	6991.49
	20	du	dy	6882.16
HUN	1	a	ˈɑː	80023.70
	2	az	ˈɑz	26332.29
	3	és	ˈeːʃ	18461.96
	4	hogy	hˈoɟ	15411.31
	5	a	ˈɑː	14729.23
	6	is	ˈiʃ	12402.43
	7	nem	nˈɛm	12294.83
	8	egy	ˈɛɟ	6516.87

Figure A.7: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
HUN	9	az	ˈɑz	5044.78
	10	meg	mˈɛg	4315.30
	11	volt	vˈolt	3659.94
	12	csak	tʃˈɑk	3608.37
	13	de	dˈɛ	3459.65
	14	már	mˈaːr	3422.05
	15	azt	ˈɑst	3017.99
	16	még	mˈeːg	2996.61
	17	ha	hˈɑ	2950.30
	18	van	vˈɑn	2915.45
	19	mint	mˈint	2863.86
	20	az	ˈɑz	2827.28
ITA	1	di	di	48038.46
	2	e	ˈe	30666.64
	3	il	il	26353.83
	4	la	la	25510.43
	5	in	in	19263.61
	6	a	ˈa	17160.09
	7	del	del	14614.32
	8	un	ˈun	13447.63
	9	per	per	13403.74
	10	che	ke	12704.05
	11	si	si	11546.50
	12	della	ˈdel_la	11293.78
	13	lˈ	l	11202.28
	14	i	ˈi	9824.64
	15	con	kon	9731.02
	16	una	ˈu_na	9548.89
	17	nel	nel	8852.02
	18	da	da	8840.67
	19	è	ɛ	8779.13
	20	le	le	8356.32
JPN	1	の	no	41309.58
	2	に	nji	23746.63
	3	は	ha	22227.66
	4	て	te	20965.96
	5	を	wo	20326.59
	6	が	ga	20112.91

Figure A.8: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
JPN	7	で	de	16705.79
	8	た	ta	16554.98
	9	と	to	15096.28
	10	し	sji	10165.53
	11	も	mo	7902.62
	12	な	na	6864.84
	13	ない	na_i	5901.97
	14	ます	ma_su	5692.12
	15	こと	ko_to	5451.05
	16	か	ka	4750.67
	17	です	de_su	4619.6
	18	いる	i_ru	4428.88
	19	する	su_ru	4140.14
	20	から	ka_ra	4059.67
KOR	1	있다	id_ta	17941.42
	2	등	dɯŋ	12253.47
	3	이	i	11110.94
	4	있는	in_nɯn	10601.81
	5	수	s^hu	10106.96
	6	고	go	7070.32
	7	것이다	gʌ_s^hi_da	5738.60
	8	또	to	5024.69
	9	한	han	4827.27
	10	위해	wi_hɛ	4261.85
	11	말했다	mal_hɛd_ta	4168.77
	12	했다	hɛd_ta	3906.41
	13	것	gʌd	3863.11
	14	이날	i_nal	3831.51
	15	경우	gjʌŋ_u	3682.14
	16	밝혔다	bal_k^hjʌd_ta	3657.47
	17	및	mid	3599.89
	18	한다	han_da	3436.24
	19	것이	gʌ_s^hi	3294.23
	20	할	hal	3138.81
SPA	1	de	de	86901.25
	2	la	la	50428.12
	3	en	en	31423.60
	4	y	i	30006.74

Figure A.9: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
SPA	5	el	el	28971.85
	6	que	ke	24101.39
	7	a	a	22060.33
	8	los	los	15440.46
	9	del	ˈdel	13512.46
	10	un	un	10288.30
	11	por	ˈpor	10271.09
	12	se	se	10115.66
	13	con	ˈkon	10015.93
	14	par	ˈpar	9900.74
	15	e	e	8627.58
	16	no	no	7983.24
	17	una	ˈu_na	7726.73
	18	al	al	5936.49
	19	su	su	5472.77
	20	o	o	4774.15
SRP	1	i	ˈi	46281.52
	2	u	ˈu	35845.22
	3	je	jˈe	31940.08
	4	da	dˈa	27109.77
	5	na	nˈa	19666.20
	6	se	sˈɛ	19197.23
	7	za	zˈa	18343.67
	8	od	ˈod	10300.27
	9	su	sˈʊ	9602.94
	10	sa	sˈa	9145.61
	11	a	ˈa	7193.27
	12	ne	nˈɛ	6780.50
	13	koji	kˈo_jɪ	5902.71
	14	o	ˈo	5618.92
	15	to	tˈo	5236.41
	16	iz	ˈiz	4479.33
	17	kao	kˈa_o	4389.20
	18	do	dˈo	3586.31
	19	ili	ˈɪ_lɪ	3480.95
	20	ali	ˈa_lɪ	3316.53
THA	1	ที่	tʰîː	25249.91
	2	เป็น	pen	17017.25

Figure A.10: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
THA	3	จะ	tɕàʔ	16207.27
	4	การ	kaːn	16034.59
	5	ไม่	mâj	15679.18
	6	มี	miː	15595.65
	7	ใน	naj	15404.86
	8	ของ	kʰɔ˘ːŋ	15219.00
	9	และ	lɛ́ʔ	13903.67
	10	ได้	dâj	12717.72
	11	ไป	paj	12321.87
	12	ให้	hâj	11839.58
	13	ว่า	wâː	11802.52
	14	มา	maː	11036.27
	15	ก็	kɔ̂ː	10796.95
	16	คน	kʰon	7565.44
	17	แล้ว	lɛ́ːw	6922.52
	18	ความ	kʰwaːm	6747.53
	19	กับ	kàp	6385.27
	20	อยู่	jùː	6258.86
TUR	1	ve	vˈɛ	41659.66
	2	bir	bˈɪr	24454.82
	3	bu	bˈʊ	10005.71
	4	ile	ˈi_lɛ	9502.31
	5	için	i_tʃˈɪn	8857.75
	6	bu	bˈʊ	8776.79
	7	da	dˈa	8669.88
	8	de	dˈɛ	7834.35
	9	olarak	o_la_rˈak	7424.36
	10	olan	o_lˈan	4436.14
	11	çok	tʃˈɔk	4275.26
	12	daha	da_hˈa	4099.85
	13	veya	ve_jˈa	3683.64
	14	en	ˈɛn	3648.35
	15	gibi	ɟi_bˈɪ	3413.77
	16	her	hˈɛr	3053.61
	17	kadar	ka_dˈar	2916.60
	18	ise	ˈi_se	2813.85
	19	sonra	sɔn_rˈa	2784.78
	20	göre	ɟœ_rˈɛ	2305.26

Figure A.11: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
VIE	1	không	xoŋ͡m1	16963.44
	2	một	mot6	16697.85
	3	của	kuə4	14648.37
	4	và	va2	14125.75
	5	hai	haj1	13652.29
	6	mươi	mɯəj1	11212.45
	7	các	kak5	10711.40
	8	là	la2	10340.21
	9	đã	da3	9491.44
	10	được	dɯək6	9288.53
	11	có	kɔ5	8940.67
	12	trong	cɔŋ͡m1	8762.85
	13	trăm	căm1	7594.13
	14	người	ŋɯəj2	7355.85
	15	năm	năm1	7300.85
	16	cho	cɔ1	6948.78
	17	với	vɤ̆j5	6801.19
	18	ba	ba1	6766.24
	19	những	ɲɯŋ3	6437.22
	20	này	năj2	6342.32
WOL	1	ko	ko	27392.40
	2	ci	ci	26216.04
	3	mu	mu	20166.18
	4	ma	ma	15908.88
	5	nga	nga	15498.09
	6	ba	ba	15404.72
	7	la	la	14135.00
	8	ne	ne	13649.52
	9	bi	bi	13145.36
	10	ñu	ñu	13014.66
	11	di	di	12865.28
	12	na	na	11670.25
	13	xam	xam	10661.94
	14	yi	yi	9896.37
	15	bu	bu	9186.82
	16	né	né	8290.54
	17	am	am	7991.78
	18	wax	wax	7805.06

Figure A.12: 20 most frequent words in 18 languages

Language	Rank	Word	Transcription	Frequency
WOL	19	rekk	rekk	7058.16
	20	lu	lu	6796.75
YUE	1	你	nei5	52266.88
	2	我	ŋɔ:5	46001.19
	3	呀	a:1	37684.63
	4	唔	m4	29420.86
	5	係	hɐi6	22944.05
	6	呢	nɛ:1	17424.81
	7	嗰	kɐm2	16331.52
	8	佢	k^{h}oy5	15916.82
	9	嘞	la:k3	15698.16
	10	嘅	kɛ:3	14891.39
	11	個	kɔ:3	14258.03
	12	好	hou2	14092.15
	13	嚟	lɐi4	13677.46
	14	都	tou1	12599.24
	15	喇	la:1	12335.35
	16	就	tsɐu6	11709.53
	17	咩	mɛ:1	10178.92
	18	去	hoy3	9839.63
	19	啲	ti:1	9130.87
	20	得	tɐk1	8693.55

A.5 Phonemic inventories of 9 languages

Figure A.13: Vowel inventories of 9 languages (obtained from each corpus analyzed and may contain some phonemes from the transcription of loanwords)

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
V	i	iː	iː	i	i	i	i	i	i
	y	yː	uː	y	u	iː	ɯ	u	iː
	u	uː	ɪ	u	e	ɯ	u	e	y
	ə	ɪ	ʊ	e	ø	ɯː	e	o	yː
	o	ʏ	ə	ø	o	e	o	a	u
	ɚ	eː	ɜː	o	ɛ	eː	ɛ		uː
	a	øː	ɛ	ə	ɔ	o	ʌ		e
		ʊ	ʌ	ɛ	a	oː	a		o
		oː	ɔː	œ		a			ɛː
		ə	æ	ɔ		aː			œː
		ɜː	æ̃	ɛ̃					ɔː
		ɛ	æː	œ̃					ɐ
		ɛː	ɒ	ɔ̃					aː
		œ	ɑː	a					
		œ̃ː	ɑ̃ː	ɑ̃					
		ʌ	ɒ̃ː						
		ɔː	eɪ						
		ɔ	aɪ						
		æ	ɔɪ						
		æ̃	əʊ						
		æ̃ː	aʊ						
		a	Iə						
		aə	ɛə						
		ɒ̃ː	ʊə						
		ã:							
		eɪ							
		aɪ							
		ɔɪ							
		aʊ							
		ai							
		au							
		ʊy							

Figure A.14: Consonant inventories of 9 languages (obtained from each corpus analyzed and may contain some phonemes from the transcription of loanwords)

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
C	p	p	p	p	p	p	p	p	p
	t	t	t	t	t	t	t	t	t
	k	k	k	k	k	k	c	c	k
	pʰ	b	b	b	b	b	k	k	kʷ
	tʰ	d	d	d	d	d	pʰ	b	pʰ
	kʰ	g	g	g	g	c	tʰ	d	tʰ
	ts	p͡f	f	f	t͡s	g	cʰ	ɟ	kʰ
	tʂ	t͡s	v	v	d͡z	f	kʰ	g	kʷʰ
	tɕ	t͡ʃ	θ	s	t͡ʃ	s	b	m	t͡s
	tsʰ	d͡ʒ	ð	z	d͡ʒ	z	d	n	t͡sʰ
	tʂʰ	m	s	ʃ	f	h	g	mv	f
	tɕʰ	n	z	ʒ	v	m	d͡ʑ	nd	s
	f	ŋ	ʃ	ʁ	θ	n	m	ɲɟ	h
	s	f	ʒ	m	s	r	n	ŋg	m
	ʂ	v	x, ç	n	z	w	ŋ	mb	n
	ʐ	s	h	ɲ	ʃ	j	s	nz	ŋ
	ɕ	z	t͡ʃ	ŋ	ʒ		sʰ	f	l
	x	ʃ	d͡ʒ	l	m		h	v	w
	w	ʒ	m	R	n		l	θ	j
	ɥ	X,ç	n	w	ɲ		w	ð	
	j	h	ŋ	ɥ	l		ɯɥ	s	
	l	l	l	j	r		j	z	
	m	R,r	r, R		ʎ			ʃ	
	n	w	w		w			x	
	ŋ	j	j		j			ɣ	
								h	
								l	
								r	
								w	
								j	

A.6 Illustration of different configurations

Below is a toy example that illustrates the differences between the configurations INF/TOKEN, INF/TYPE, LEM/TOKEN and LEM/TYPE. The starting point is a fictitious corpus based on an extraction of entries of the WebCelex English corpus:

Table A.2: Fictitious corpus

Inflected form	Lemma	Phonetic form	Grammatical category	Frequency
beautiful	beautiful	ˈbjuː-tə-fʊl	Adjective	2075
beautifully	beautifully	ˈbjuː-tə-flɪ	Adverb	278
drink	drink	ˈdrɪŋk	Verb	728
drinks	drink	ˈdrɪŋks	Verb	111
drink	drink	ˈdrɪŋk	Noun	1414
drinks	drink	ˈdrɪŋks	Noun	440
drinker	drinker	ˈdrɪŋ-kəR	Noun	30
drinkers	drinker	ˈdrɪŋ-kəRs	Noun	44
drank	drink	ˈdræŋk	Verb	620

For each corpus, entries are merged on the basis of similar phonetic forms, regardless of grammatical categories. For a set of entries with an identical phonetic form, the frequency of the resulting entry is equal to the sum of the frequencies of the merged entries.

To build the INF/TOKEN corpus, one therefore only needs to merge identical phonetic forms, more precisely here i) /ˈdrɪŋk/ as a verb and as a noun, ii) /ˈdrɪŋks/ as a verb and as a noun:

Table A.3: INF/TOKEN corpus

Inflected form	Phonetic form	Frequency
beautiful	ˈbjuː-tə-fʊl	2075
beautifully	ˈbjuː-tə-flɪ	278
drink	ˈdrɪŋk	2142 (728+1414)
drinks	ˈdrɪŋks	551 (111+440)
drinker	ˈdrɪŋ-kəR	30
drinkers	ˈdrɪŋ-kəRs	44
drank	ˈdræŋk	620

To obtain the LEM/TOKEN corpus, we first merge the entries of the initial set according to their lemmas. The frequency of a lemma form is equal to the sum of the frequencies of the corresponding inflected forms:

Table A.4: Intermediate corpus while building the LEM/TOKEN corpus

Lemma	Phonetic form	Grammatical category	Frequency
beautiful	ˈbjuː-tə-fʊl	Adjective	2075
beautifully	ˈbjuː-tə-flɪ	Adverb	278
drink	ˈdrɪŋk	Verb	1459 (728+111+620)
drink	ˈdrɪŋk	Noun	1854 (1414+440)
drinker	ˈdrɪŋ-kəR	Noun	74 (30+44)

The second step is to merge entries according to their phonetic forms, as done previously for the INF/TOKEN corpus:

Table A.5: LEM/TOKEN corpus

Lemma	Phonetic form	Frequency
beautiful	ˈbjuː-tə-fʊl	2075
beautifully	ˈbjuː-tə-flɪ	278
drink	ˈdrɪŋk	3313 (1459+1854)
drinker	ˈdrɪŋ-kəR	74 (30+44)

Considering types rather than tokens amounts to equating all frequencies to 1. We can therefore easily derive the INF/TYPE corpus from the previous INF/TOKEN corpus. Note that equating the frequencies to 1 should take place *after* extracting the 20 000 most frequent entries, as mentioned in section 4.2.3 (this is not relevant for our small toy corpus). The LEM/TYPE corpus is obtained from the LEM/TOKEN corpus the way that the INF/TYPE corpus is derived from the INF/TOKEN corpus:

Table A.6: INF/TYPE corpus

Inflected form	Phonetic form	Frequency
beautiful	ˈbjuː-tə-fʊl	1
beautifully	ˈbjuː-tə-flɪ	1
drink	ˈdrɪŋk	1
drinks	ˈdrɪŋks	1
drinker	ˈdrɪŋ-kəR	1
drinkers	ˈdrɪŋ-kəRs	1
drank	ˈdræŋk	1

A.7 Contrasting pairs of vowels & consonants

List of the contrasting pairs of vowels & consonants (ranked by increasing *FL*) for the nine languages under study, used in the simulation presented in Section 4.4.1 to estimate the relative loss of entropy when gradually coalescing lower-*FL* segments with higher-*FL* segments.

Figure A.15: Contrasting pairs of vowels (ranked by increasing *FL*)

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
V	o→a	æ̃:→i:	ɒ̃:→ɒ	ə→a	ø→a	a:→ɯ	ɯ→a	e→a	y→u
	ɘ→y	ɒ̃:→u:	ɑ̃:→ɒ	œ→ɛ	ɛ→a	ɯ:→i	ɛ→a	o→a	i→u
	y→i	ã:→Y	ʊə→ɔ:	ɔ→a	u→o	i:→i	u→a	u→i	œ:→e
	i→u	ə→ɪ	ə→I	ɛ̃→a	ɔ→a	e:→a	ʌ→i	i→a	u:→a:
	u→ə	ø:→o:	ɔI→eI	œ̃→ɛ	o→a	ɯ→a	e→i		y:→i:
	a→ə	œ→ɔ	Iə→ɔ:	y→ɛ	i→e	o:→a	o→i		u→ɐ
		ɛ:→a:	ʊ→æ	o→a	e→a	i→a	a→i		e→o
		ʊy→au	ɛə→ɔ:	u→a		o→a			o→ɐ
		Y→ɛ	ɜ:→ɔ:	ɔ̃→ɛ		e→a			ɛ:→ɔ:
		ʊ→a	aʊ→eI	i→ɛ					i:→a:
		y:→a:	ɑ:→eI	ɛ→e					ɐ→a:
		u:→a:	u:→i:	ɑ̃→e					a:→ɔ:
		au→ai	ʌ→æ	ø→e					
		ɔ→ɛ	ɒ→I	a→e					
		o:→i:	ɔ:→I						
		e:→a:	ɛ→I						
		a:→i:	əʊ→eI						
		ɛ→a	æ→I						
		ɪ→a	I→eI						
		ai→a	i:→eI						
		i:→a	aI→eI						

Figure A.16: Contrasting pairs of consonants (ranked by increasing *FL*)

Language	CMN	DEU	ENG	FRA	ITA	JPN	KOR	SWH	YUE
C	ŋ→n	ʒ→b	x, ç→p	ŋ→t	θ→n	f→k	tʰ→n	x→h	kʷʰ→t͡sʰ
	ɥ→j	t͡ʃ→p	ʒ→s	ɥ→ʁ	ʒ→f	p→k	kʰ→g	mv→v	kʰ→l
	f→ʂ	d͡ʒ→b	ŋ→d	ɲ→ʁ	d͡z→b	b→k	c→b	θ→ɟ	pʰ→k
	pʰ→m	p͡f→t	θ→d	w→ʁ	w→r	c→k	p→d	ɣ→w	ŋ→k
	ʐ→ʂ	ŋ→s	ʤ→t	z→ʁ	ɲ→b	z→k	w→j	ð→k	kʷ→t͡s
	tɕʰ→tɕ	j→v	t͡ʃ→k	g→t	j→r	r→k	s→d	nz→c	f→k
	kʰ→ʂ	ʃ→k	j→l	j→ʁ	ʃ→v	j→n	ŋ→n	ɲɟ→t	w→m
	tsʰ→l	p→n	g→k	ʃ→ʁ	t͡s→t	h→k	pʰ→g	g→t	tʰ→s
	s→ʂ	t͡s→t	v→d	b→k	g→d	w→g	k→d	f→t	n→m
	w→m	g→b	p→k	f→v	z→n	d→k	ɯɥ→g	ʃ→k	p→t
	ɕ→tɕ	h→v	r, R→z	ʁ→l	b→t	g→k	j→sʰ	r→t	l→t
	tʂʰ→ʂ	b→R,r	ʃ→t	ʒ→s	d͡ʒ→t	m→n	cʰ→sʰ	d→k	j→s
	tɕ→ʂ	k→n	k→t	v→t	p→m	n→t	d͡ʑ→sʰ	nd→k	t͡sʰ→s
	ts→ʂ	l→R,r	f→b	k→p	f→m	t→k	t→b	v→k	m→t
	x→t	X,ç→s	l→t	p→t	t͡ʃ→s	s→k	h→sʰ	ŋg→t	t→k
	m→l	f→s	b→t	t→s	r→t		b→g	mb→n	h→k
	j→ʂ	t→n	h→w	n→s	v→t		m→g	b→k	s→ts
	k→tʂ	s→n	z→d	m→s	ʎ→d		d→g	ɟ→w	k→ts
	tʰ→l	v→z	w→t	d→s	m→k		sʰ→g	p→k	
	n→l	z→R,r	d→t	l→s	t→n		l→n	h→k	
	p→t	d→R,r	s→t		k→s		g→n	s→k	
	tʂ→ʂ	m→R,r	ð→m		s→n			t→n	
	ʂ→l	R,r→n	m→t		n→d			m→k	
	l→t		n→t		l→d			k→l	
								c→w	
								z→l	
								l→n	
								w→j	
								j→n	

Bibliography

[Ackerman & Malouf, 2013] Ackerman, F. & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language.* 89:429-464.

[Aikhenvald, 2007] Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. In T. Shopen (Ed.) *Language Typology and Syntactic Description, Vol. 3.* Cambridge: Cambridge University Press, 1-65.

[Akmajian et al., 2001] Akmajian, A., Demer, R. A., Farmer, A. K., & Harnish, R. M. (2001). *Linguistics: An Introduction to Language and Communication.* MIT Press.

[Altmann, 1980] Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika, 2,* 1–10.

[Arbesman, Strogatz, & Vitevitch, 2012] Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos, 20(03),* 679-685.

[Aroonmanakun, Tansiri, & Nittayanuparp, 2009] Aroonmanakun, W., Tansiri, K., & Nittayanuparp, P. (2009). Thai National Corpus: a progress report. In *Proceedings of the 7th Workshop on Asian Language Resources.* Association for Computational Linguistics, 153-158.

[Ashby, 1947] Ashby, W. R. (1947). Principles of the Self-Organizing Dynamic System. *Journal of General Psychology.* volume 37, 125–128.

[Ashby, 1962] Ashby, W. R. (1962). Principles of the self-organizing system, In Von Foerster, H. & Zopf, Jr. G. W. (Eds.), *Principles of Self-Organization:Transactions of the University of Illinois Symposium,* Pergamon Press: London, UK, 255-278.

[Atkinson, 2011] Atkinson Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science 332,* 346–349.

[Aylett & Turk, 2004] Aylett, M. P. & Turk A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech, 47:1,* 31-56.

[Bichel & Nichols, 2013] Bickel, B. & Nichols, J. (2013). Fusion of Selected Inflectional Formatives. In Dryer, M. S. & Haspelmath, M. (Eds.) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

[Bane, 2008] Bane, M. (2008.) Quantifying and measuring morphological complexity. In *Proc. of the 26th West Coast Conference on Formal Linguistics,* 69-76.

[Barabási & Albert, 1999] Barabási, A. L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science, 286(5439),* 509–512.

[Bates et al., 2015] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-8, http://CRAN.R-project.org/package=lme4.

[Beckner et al., 2009] Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language learning,* 59(s1), 1-26.

[Bereicua, 2013] Bereicua, R. M. S. (2013). A survey of linguistic variables in the central zone of Deva river valley. *Anuario del Seminario de Filología Vasca "Julio de Urquijo",* 6, 20-28.

[Bell et al., 2009] Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language, 60(1),* 92–111.

[Bickel & Nichols, 2005] Bickel, B. & Nichols, J. (2005). Inflectional synthesis of the verb. In Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (Eds.), *The World Atlas of Language Structures,* 94– 97. Oxford: Oxford University Press.

[Biemann et al., 2007] Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig Corpora Collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic.*

[Blache & Meunier, 2004] Blache, P., & Meunier, C. (2004). Language as a complex system: the case of phonetic variability. In *Congreso de Lingüística General.* pp. 192-195.

[Blevins, 2004] Blevins, J. (2004). *Evolutionary Phonology: The emergence of sound patterns.* Cambridge:Cambridge University Press.

[Blevins, 2013] Blevins, J. (2013). The information-theoretic turn. *Psihologija,* 46(4), 355-375.

[Bonatti et al., 2005] Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science, 16(6),* 451-459.

[Bronfenbrenner, 1979] Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design.* Cambridge, MA: Harvard University Press.

[Brown, 1988] Brown, A. (1988). Functional load and the teaching of pronunciation. *Tesol Quarterly, 22(4),* 593-606.

[Bybee, 2003] Bybee, J. (2003). *Phonology and Language Use.* Cambridge University Press.

[Byrd, 1994] Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication,* 15(1), 39-54.

[Campione & Véronis, 1998] Campione, E. & Véronis, J. (1998). A multilingual prosodic database, *Proc. of the 5th International Conference on Spoken Language Processing (ICSLP'98),* Sydney, Australia, 3163-3166.

[Canepari, 2009] Canepari, L. (2009). *Dizionario di pronuncia italiana.* Bologna: Zanichelli Editore.

[Caramazza et al., 2000] Caramazza, A., Chialant, D., Capasso, R., & Miceli, G. (2000). Separable processing of consonants and vowels. *Nature, 403(6768),* 428-430.

[Carnevali, 2009] Carnevali, S. (2009). Fonetica, downloaded from http://www.webalice.it/sandro.carnevali2011/

[Carter, 1987] Carter, D. M. (1987). An information-theoretic analysis of phonetic dictionary access. *Computer Speech & Language, 2(1),* 1-11.

[CC-CEDICT, 2012] CC-CEDICT Dictionary, downloaded on 30 Nov 2012 from http://cc-cedict.org/wiki/.

[Cercle Linguistique de Prague, 1931] Cercle Linguistique de Prague (1931). Réunion phonologique internationale : 18-21 / XII 1930, tenue à Prague. *Travaux du Cercle linguistique de Prague, 4.* Prague: Jednota ceskosloven-skych matematiku a fysiku.

[Chen & Zechner, 2011] Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1,* pp. 722-731.

[Cholin, Levelt, &Schiller, 2006] Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006) Effects of syllable frequency in speech production. *Cognition, 99(2),* 205-235.

[Chomsky, 1959] Chomsky, N. (1959). A Review of B. F. Skinner's Verbal Behavior In Leon, A. J. & Murray S. M. (Eds.), *Readings in the Psychology of Language,* Prentice-Hall, 1967, pp. 142-143.

[Chomsky, 1965] Chomsky, N. (1965). *Aspects of the Theory of Syntax,* Cambridge, Massachusetts: MIT Press.

[Christiansen & Chater, 2008] Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and brain sciences,* 31(05), 489-509.

[Cohen Priva, 2008] Priva, U. C. (2008). Using information content to predict phone deletion. In *Proceedings of the 27th west coast conference on formal linguistics.* pp. 90-98.

[Comrie, 1989] Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology.* University of Chicago press.

[Coupé, Oh, Pellegrino, & Marsico, 2014] Coupé, C., Oh, Y. M., Pellegrino, F., & Marsico, E. (2014). Cross-linguistic investigations of oral and silent reading. In *Fifteenth Annual Conference of the International Speech Communication Association.*

[Cover & Thomas, 2012] Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory.* John Wiley & Sons.

[Croft, 1996] Croft, W. (1996). *Typology and Universals.* Cambridge: Cambridge University Press.

[Croft, 2002] Croft, W. (2002). *Typology and universals.* Cambridge University Press.

[Crothers, 1978] Crothers, J. (1978). Typology and universals of vowel systems in phonology. In Greenberg, J. H., Ferguson, C. A., & Moravcsik, E. A. (Eds.) *Universals of human language, vol. 2.* Stanford, CA: Stanford University Press, 93-152.

[Curran & Osborne, 2002] Curran, J. R., & Osborne, M. (2002). A very very large corpus doesn't always yield reliable estimates. In *Proceedings of the 6th conference on Natural language learning.* Volume 20, 1-6. Association for Computational Linguistics.

[Cutler et al., 2000] Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition, 28(5),* 746-755.

[Dabrowska, 2010] Dabrowska, E. (2010). Native v expert intuitions: An empirical study of acceptability judgements. *The Linguistic Review,* 27, 1-23.

[Dahl, 2004] Dahl, Ö. (2004). *The Growth and Maintenance of Linguistic Complexity.* Philadelphia: John Benjamins.

[Darwin, 1859] Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* London: Murray. [1st ed.]

[Da Tos, 2013] Da Tos, M. (2013). The Italian FINIRE-type verbs: a case of morphomic attraction. *The Boundaries of Pure Morphology: Diachronic and Synchronic Perspectives, 4,* 45.

[Davis & Zajdo, 2010] Davis, B. L. & Zajdo, K. (2010). *The Syllable in Speech Production: Perspectives on the Frame Content Theory.* Taylor & Francis.

[de Boer, 2000] de Boer, B. (2000). Self-organization in vowel systems. *Journal of Phonetics 28.* 441–465.

[de Boer, 2012] de Boer, B. (2012). Self organization and language evolution. In Tallerman, M. (Ed.) *The Oxford handbook of language evolution.* Oxford University Press, pp. 612-620.

[DeGraff, 2001] DeGraff, M. (2001). On the origin of Creoles: A Cartesian critique of Neo-Darwinian linguistics. *Linguistic Typology,* 5(2/3), 213-310.

[Delle Luche et al., 2014] Delle Luche, C., Poltrock, S., Goslin, J., New, B., Floccia, C., & Nazzi, T. (2014). Differential processing of consonants and vowels in the auditory modality: A cross-linguistic study. *Journal of Memory and Language, 72,* 1-15.

[Demberg & Keller, 2008] Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition,* 109(2), 193-210.

[Demberg et al., 2012] Demberg, V., Sayeed, A. B., Gorinski, P. J., & Engonopoulos, N. (2012). Syntactic surprisal affects spoken word duration in conversational contexts. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning.* Association for Computational Linguistics. pp. 356-367.

[Dressler, 1985] Dressler, W. (1985). Typological aspects of Natural Morphology. *Acta Linguistica Academiae Scientiarum Hungaricae* 35 (1-2), 51-70.

[Dryer & Haspelmath, 2013] Dryer, M. S. & Haspelmath, M. (Eds.) (2013). *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

[Duanmu, 1990] Duanmu, S. (1990). *A formal study of syllable, tone, stress and domain in Chinese languages.* PhD Thesis. Massachusetts Institute of Technology.

[Easterday, Timm, & Maddieson, 2011] Easterday, S., Timm, J., & Maddieson, I. (2011). The effects of phonological structure on the acoustic correlates of rhythm. *ICPhS XVII,* 623-626.

[Evans & Levinson, 2009] Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences,* 32(05), 429-448.

[Fenk-Oczlon & Fenk, 1999] Fenk-Oczlon G. & Fenk, A. (1999). Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology 3:2.* 151-177.

[Fenk-Oczlon & Fenk, 2004] Fenk-Oczlon, G. & Fenk, A. (2004). Systemic typology and crosslinguistic regularities. *Text processing and cognitive technologies,* 229-234.

[Fenk-Oczlon & Fenk, 2005] Fenk-Oczlon, G. & Fenk, A. (2005). Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In Fenk-Oczlon, G. & Winkler, C. (Eds.) *Sprache und Natürlichkeit, Gedenkband für Willi Mayerthaler,* Tübingen: Gunther Narr.

[Fenk, Fenk-Oczlon, & Fenk, 2006] Fenk, A., Fenk-Oczlon, G. & 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 & Fenk-Oczlon, 2006] Fenk, A. & Fenk-Oczlon, G. (2006). Crosslinguistic computation and a rhythm-based classification of languages. In *From Data and Information Analysis to Knowledge Engineering.* Springer Berlin Heidelberg.

[Fenk-Oczlon, 2013] Relationships between semantic complexity, structural complexity and markedness: Frequency matters. *International Congress of Linguists.*

[Fenk-Oczlon & Fenk, 2014] Fenk-Oczlon, G. & Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznan Studies in Contemporary Linguistics.* Volume 50, Issue 2, Pages 145–155.

[Ferragne, Flavier, & Fressard, 2013] Ferragne, E., Flavier, S., & Fressard, C. (2013). ROCme! software for the recording and management of speech corpora. In *Proceedings of Interspeech,* pp. 1864-1865.

[Ferrer i Cancho & Solé, 2003] Ferrer i Cancho, R. & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Science, 100(3),* 788-791.

[Ferrer i Cancho, 2006] Ferrer-i -Cancho, R. (2006). On the universality of Zipf's law for word frequencies. In Grzybek, P. & Köhler, R. (eds.) *Exact Methods in the Study of Language and Text. To Honor Gabriel Altmann* Berlin: Gruyter, pp. 131–40.

[Ferrer i Cancho & Díaz-Guilera, 2007] Ferrer i Cancho, R. & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment.*

[Fogerty & Humes, 2012] Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America, 131(2),* 1490-1501.

[Frank & Jaeger, 2008] Frank, A. & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society* Washington, DC: Cognitive Science Society. pp. 933-938.

[Fry et al., 1962] Fry D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech, 5,* 171-189.

[Gahl & Garnsey, 2004] Gahl, S. & Garnsey, S. (2004). Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. *Language.* 80: 748–774(2004).

[Gahl, 2008] Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language, 84(3),* 474-496.

[Gale & Sampson, 1995] Gale, W. & Geoffrey Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics,* vol. 2, 217-37.

[Gelas, Besacier, & Pellegrino, 2012] Gelas, H., Besacier, L., & Pellegrino, F. (2012). Developments of Swahili resources for an automatic speech recognition system. In *Proceedings of the Third International Workshop on Spoken Languages Technologies for Under-resourced Languages,* 94-101.

[Gell-Mann & Ruhlen, 2011] Gell-Mann, M., & Ruhlen, M. (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences,* 108(42), 17290-17295.

[Genzel & Charniak, 2002] Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. *Proceedings of the 40th annual meeting on association for computational linguistics.* pp. 199–206.

[Genzel & Charniak, 2003] Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. *Proceedings of the 2003 conference on Empirical methods in natural language processing.* Association for Computational Linguistics. pp. 65-72.

[Gerlach & Altmann, 2013] Gerlach, M., & Altmann, E. G. (2013). Stochastic model for the vocabulary growth in natural languages. *Physical Review X, 3(2).*

[Gess, Lyche, & Meisenburg, 2012] Gess, R., Lyche, C., & Meisenburg, T. (Eds.). (2012). *Phonological Variation in French: Illustrations from three continents.* John Benjamins Publishing.

[Gibson et al., 2013] Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological science.*

[Goldsmith, 2000] Goldsmith, J. A. (2000). On information theory, entropy, and phonology in the 20th century. *Folia Linguistica 34:1-2.* 85-100.

[Goldsmith, 2001] Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics,* 27(2), 153-198.

[Goldsmith, 2002] Goldsmith, J. (2002). Probabilistic models of grammar: phonology as information minimization, *Phonological Studies,* Vol. 5, pp. 21-46.

[Greenberg, 1960] Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics, 26(3),* 178-194.

[Greenberg, 1966] Greenberg, J. H. (1966). *Universals of language,* second edition. The Cambridge, Mass.: MIT Press.

[Greenberg, 1978] Greenberg, J. H. (1978). Diachrony, synchrony and language universals. Universals of Human Language. In Greenberg, J. H., Charles A. Ferguson, C. A. & Edith, A. M. (Eds.) *Vol. III: Word Structure* 47-82. Stanford: Stanford University Press.

[Greenberg, 1999] Greenberg, S. (1999). Speaking in a shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication,* 29. 159-176.

[Gregory et al., 1999] Gregory, M. L., Raymond, W. D., Bell, A., Fosler-Lussier, E., & Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In *CLS-99.* University of Chicago.

[Hale, 2001] Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies.* Association for Computational Linguistics, pp. 1-8.

[Hale, 2003] Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research,* 32(2), 101-123.

[Hall, 2011] Hall, D. C. (2011). Phonological contrast and its phonetic enhancement: dispersedness without dispersion. *Phonology, 28(01),* 1-54.

[Hammarström et al., 2015] Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. (2015). *Glottolog 2.4.* Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://glottolog.org)

[Havy & Nazzi, 2009] Havy, M. & Nazzi, T. (2009). Better processing of consonantal over vocalic information in word learning at 16 months of age. *Infancy, 14(4),* 439-456.

[Hawkins, 2003] Hawkins, J. A. (2003). Efficiency and complexity in grammars: Three general principles. *The nature of explanation in linguistic theory,* 121-152.

[Hay & Bauer, 2007] Hay, J. & Bauer, L. (2007). Phoneme inventory size and population size. *Language 83,* 388–400.

[Hockett, 1955] Hockett, C. F. (1955). *A manual of phonology.* Waverly Press: Baltimore.

[Hockett, 1958] Hockett, C. F. (1958). *A Course in Modern Linguistics.* The Macmillan Company: New York.

[Hockett, 1966] Hockett, C. F. (1966). The quantification of functional load: A linguistic problem. *Report Number RM-5168-PR,* Rand Corp. Santa Monica.

[Hombert, Ohala, & Ewan, 1979] Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language,* 37-58.

[Hua & Dodd, 2000] Hua, Z., & Dodd, B. (2000). The phonological acquisition of Putonghua (modern standard Chinese). *Journal of Child Language, 27(01),* 3-42.

[Hyman, 2008] Hyman, L. M. (2008). Universals in phonology. *The Linguistic Review, 25,* 83-137.

[Ingram, 1989] Ingram, D. (1989). *First language acquisition: Method, description and explanation.* Cambridge University Press.

[Institute for the Languages of Finland, 1996-1998] Department of General Linguistics, University of Helsinki and Institute for the Languages of Finland, (1996-1998). *Finnish Parole Corpus,* available through http://kaino.kotus.fi/sanat/taajuuslista/parole.php

[Jakobson, 1931] Jakobson, R. (1931). Principes de phonologie historique. In Troubetzkoy, N. S. *Principes de phonologie.* Paris, Klincksieck, 1976, 315-336.

[Jakobson, 1941] Jakobson, R. (1941; 1962). Kindersprache, Aphasie und allgemeine Lautgesetze. Reprinted in *Selected Writings I.* Mouton, The Hague, 328-401.

[Jaeger, 2010] Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology,* 61(1), 23-62.

[Jäger, 2012] Jäger, G. (2012). Power laws and other heavy-tailed distributions in linguistic typology. *Advances in Complex Systems, 15.*

[Jaynes, 19] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review,* 106(4), 620.

[Jescheniak & Levelt, 1994] Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(4),* 824-843.

[Johnson, 1996] Johnson, K. (1996). Speech perception without speaker normalization. In Johnson, K. & Mullennix (Eds.) *Talker Variability in Speech Processing.* San Diego. Academic Press.

[Johnson, 2004] Johnson, K. (2004). Massive reduction in conversational American English. In Yoneyama. K. & Maekawa, K. (Eds.)*Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium.* The National International Institute for Japanese Language: Tokyo. 29-54.

[Joseph, 2000] Joseph, J. E. (2000). *Limiting the Arbitrary: Linguistic Naturalism and Its Opposites in Plato's Cratylus and the Modern Theories of Language (Vol. 96).* John Benjamins Publishing.

[Joseph & Newmeyer, 2012] Joseph, J. E. & Newmeyer, F. J. (2012). 'All Languages Are Equally Complex': The rise and fall of a consensus. *Historiographia Linguistica, vol 39,* no. 2-3, pp. 341-368.

[Juola, 1998] Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics, 5(3),* 206–213.

[Jurafsky et al., 2001] Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language,* 45, 229-254.

[Keller, 2004] Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. *Proceedings of the conference on empirical methods in natural language processing.* Vol. 317, No. 324.

[Kello & Beltz, 2009] Kello, C. T., & Beltz, B. C. (2009). Scale-free networks in phonological and orthographic wordform lexicons. In Pellegrino, F., Marsico, E., Chitoran, I. & Coupé, C. (Eds.) *Approaches to Phonological Complexity.* Phonology & Phonetics Series vol. 16, Berlin, New York, Mouton de Gruyter, 171-190.

[Kello et al., 2012] Kello, C. T., Brown, G. D., Ferrer i Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in cognitive sciences, 14(5),* 223-232.

[Kewley-Port, Burkle, & Lee, 2007] Kewley-Port, D., Burkle ,T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing impaired listeners. *The Journal of the Acoustical Society of America, 122,* 2365–2375.

[Kim et al., 1993] Kim, S., Yi, H., Yu, C., & Han'guk Pangsong Kongsa. (1993). *Py'ojun Han'gugŏ parŭm taesajŏn =: A Korean pronunciation dictionary.* Sŏul T'ŭkpyŏlsi: Ŏmun'gak.

[King, 1967] King, R. D. (1967). Functional Load and Sound Change, *Language, 43:4,* 831-852.

[Kirby, 2008] Kirby, J. (2008). vPhon: a Vietnamese phonetizer (version 0.2.4). Retrieved from http://lel.ed.ac.uk/j̃kirby/vphon.html.

[Kissling, 2012] Kissling, E. M. (2012). Cross-linguistic differences in the immediate serial recall of consonants versus vowels. *Applied Psycholinguistics, 33(03),* 605–621.

[Kostić, 1991] Kostić, A. (1991). Informational approach to processing inflectional morphology: Standard data reconsidered. *Psychological research,* 53, 62–70.

[Kronrod, Coppess, & Feldman, 2012] Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A unified model of categorical effects in consonant and vowel perception. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society,* 629-634.

[Kučera, 1963] Kučera, H. (1963). Entropy, redundancy and functional load in Russian and Czech. *American Contributions to the Fifth International Congress of Slavists.* Mouton & Company: The Hague, 191-219.

[Kusters, 2003] Kusters, W. (2003). *Linguistic complexity: the influence of social change on verbal inflection.* Utrecht: LOT.

[Ladefoged & Maddieson, 1996] Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages.* Blackwells, Cambridge.

[Ladefoged, 2001] Ladefoged, P. (2001). *Vowels and consonants: An introduction to the sounds of languages.* Oxford: Blackwells.

[Ladefoged & Johnson, 2014] Ladefoged, P. & Johnson, K. (2014). *A Course in Phonetics.* Cengage Learning.

[Labov, 2001] Labov, W. (2001). *Principles of linguistic change, Vol. II: Social factors.* Oxford: Blackwell.

[Le et al., 2004] Le, V. B., Tran, D. D., Castelli, E., Besacier, L., & Serignat, J. F. (2004). Spoken and Written Language Resources for Vietnamese. In *LREC.* 4, pp. 599-602.

[Learner, 2013] Learner, Z. (2013). JyutDict, downloaded on March 3, 2013 from http://zhongwenlearner.com/down-loads/jyutdict/.

[Levelt, Roelofs, & Meyer, 1999] Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22(01),* 1-38.

[Levinson, 2000] Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT Press.

[Levy & Jaeger, 2007] Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J. & Hoffman, T. (Eds.) *Advances in Neural Information Processing System.* Cambridge, MA: MIT Press, 849-856.

[Liberman et al., 1975] Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1975). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358-368.

[Liljencrants & Lindblom, 1972] Liljencrants, J. & Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language, 48,* 839-862.

[Lindblom, MacNeilage, & Studdert-Kennedy, 1984] Lindblom, B., MacNeilage, P. & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of language universals. In Butterworth, B., Comrie, B., & Dahl, Ö. (Eds.) *Explanations for language universals,* Berlin: Mouton, pp.181-203.

[Lindblom, 1986] Lindblom, B. (1986). Phonetic universals in vowel systems. In Ohala, J. J. & Jaeger, J. J. (Eds.), *Experimental phonology.* Orlando, FL: Academic Press, 13.

[Lindblom & Maddieson, 1988] Lindblom, B. & Maddieson, I. (1988). Phonetic universals in consonant systems. In Hyman, L. M. & Li, C. N. (Eds.) *Language, speech and mind.* London: Routledge, 62-78.

[Lindblom, 1990] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modeling.* Springer Netherlands. pp. 403-439.

[Lloret, 2007] Lloret, M. R. (2007). On the nature of vowel harmony: Spreading with a purpose. *Stefano Canalis,* 15.

[Lloyd, 2001] Lloyd, S. (2001). Measures of complexity: A nonexhaustive list, *IEEE Contr. Syst. Mag. 21 (4),* 7-8, 2001.

[Luce & Pisoni, 1998] Luce, P. A. & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing, 19(1),* 1-36.

[Lupyan & Dale, 2010] Lupyan, G. & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE 5(1): e8559.*

[Lyding et al., 2014] Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., Pirrelli, V. (2014). The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9).* Association for Computational Linguistics, Gothenburg, Sweden, 36-43.

[Maddieson, 1984] Maddieson, I. (1984). *Patterns of sounds.* Cambridge, MA: Cambridge University Press.

[Maddieson & Precoda, 1990] Maddieson, I., & Precoda, K. (1990). Updating UPSID. *Working Papers in Phonetics 74.* Department of Linguistics, UCLA, UC Los Angeles.

[Maddieson, 1991] Maddieson, I. (1991). Testing the universality of phonological generalizations with a phonetically specified segment database: Results and limitations. *Phonetica 48.* 193-206.

[Maddieson, 2006] Maddieson, I. (2006). Correlating phonological complexity: data and validation. *Linguistic Typology.* 10:1, 106-123.

[Maddieson, 2007] Maddieson, I. (2007). Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. In Solé, M. J., Beddor, P., & Ohala, M. (Eds.) *Experimental Approaches to Phonology.* Oxford University Press, Oxford and New York: 93-103.

[Maddieson, 2013] Maddieson, I. (2013). Syllable Structure. In: Dryer, M. S. & Haspelmath, M. (Eds.) *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.(Available online at http://wals.info/chapter/12)

[Maddieson et al., 2013] Maddieson, I., Flavier, S., Marsico, E., Coupé, C., & Pellegrino, F. (2013). LAPSyD: Lyon-Albuquerque phonological systems database. In *Proceedings of Interspeech 2013.* 3022-3026.

[Mahowald et al., 2013] Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition,* 126(2), 313-318.

[Marsico et al., 2003] Marsico, E., Maddieson, I., Coupé, C., & Pellegrino, F. (2003). Investigating the "hidden" structure of phonological systems. In *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society,* 256-267.

[Martindale et al., 1996] Martindale, C., Gusein-Zade, S. M., McKenzie, D., & Borodovsky, M. Y. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics, 3(2),* 106–112.

[Martinet, 1938] Martinet, A. (1938). La phonologie. *Le français moderne, 6,* 131-146.

[Martinet, 1955] Martinet, A. (1955). Économie des changements phonétiques. *Traité de phonologie diachronique.* Francke: Berne.

[MATLAB (R2011a)] *MATLAB and Statistics Toolbox Release R2011a,* The MathWorks, Inc., Natick, Massachusetts, United States.

[McWhorter, 2001] McWhorter, J. (2001) The world's simplest grammars are creole grammars. *Linguistic Typology, 5,* pp.125-166.

[McWhorter, 2014] McWhorter, J. H. (2014). *The language hoax: Why the world looks the same in any language.* Oxford University Press.

[Meillet, 1915] Meillet, A. (1915). *Introduction à l'étude comparative des langues indo-européennes,* Paris, Hachette, 4e édit.

[Mitchell, 2009] Mitchell, M. (2009). *Complexity: A guided tour.* Oxford University Press.

[MPI for Psycholinguistics, 2013] Max Planck Institute for Psycholinguistics, *WebCelex,* retrieved on March 18, 2013 and on August 6, 2014 from http://celex.mpi.nl.

[Mesgarani et al., 2014] Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343(6174),* 1006-1010.

[Mithun, 1984] Mithun, M. (1984). Levels of linguistic structure and the rate of change. In Fisiak, J. (Ed.) *Historical Syntax.* Berlin: Mouton.

[Morales & Taylor, 2007] Morales, F. & Taylor, J. R. (2007). *Learning from relative frequency.* Available as LAUD (Linguistic Agency, University of Duisburg) preprint, paper no. 690.

[Moravcsik, 2003] Moravcsik, E. (2003). Inflectional morphology in the Hungarian noun phrase: a typological assessment. In Plank, F. (ed.) *Noun phrase structure in the languags of Europe.* Berlin: Mouton de Gruyter.

[López, 2004] López, X. (2004) Transcriptor fonético automático del español retrieved online on 11 December 2014 from http://www.aucel.com/pln/sobre.html

[Moscoso del Prado, Kostić, & Baayen, 2004] Moscoso del Prado, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition,* 94(1), 1-18.

[Moscoso del Prado, 2011] Moscoso del Prado, F. (2011). The Mirage of morphological complexity, In *Proc. of the 33rd Annual Conference of the Cognitive Science Society,* 3524-3529.

[Moulin-Frier et al., forthcoming] Moulin-Frier, C., Diard, J., Schwartz, J.-L., & Bessière, P. (forthcoming). COSMO ("Communicating about Objects using Sensory-Motor Operations"): a Bayesian modeling framework for studying speech communication and the emergence of phonological systems, *Journal of Phonetics.*

[Mueller et al., 2003] Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity and phonological complexity in verbal working memory. *Journal of Experimental Psychology, Learning, Memory, and Cognition,* 29:6, 1353-1380.

[Munro & Derwing, 2006] Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34(4),* 520-531.

[Nichols, 2007] Nichols, J. (2007). The distribution of complexity in the world's languages. *81st Annual Meeting of the Linguistic Society of America.*

[NINJAL, 2011] National Institute for Japanese Language and Linguistics and National Institute of Information and Communications Technology, *The corpus of spontaneous Japanese (CSJ),* Third printing, 2011.

[Nazzi, 2005] Use of phonetic specificity during the acquisition of new words: Differences between consonants and vowels. *Cognition, 98,* 13-30.

[Nazzi & New, 2007] Nazzi, T., & New, B. (2007). Beyond stop consonants: Consonantal specificity in early lexical acquisition. *Cognitive Development, 22(2),* 271-279.

[Nazzi et al., 2009] Nazzi, T., Floccia, C., Moquet, B., & Butler, J. (2009). Bias for consonantal information over vocalic information in 30-month-olds: Cross-linguistic evidence from French and English. *Journal of Experimental Child Psychology, 102(4),* 522-537.

[Nespor, Peña, & Mehler, 2003] Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio, 2(2),* 203-230.

[Nettle, 2012] Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 367(1597).* The Royal Society, pp. 1829–1836.

[New et al., 2001] New B., Pallier C., Ferrand L., & Matos R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE 3.80, *L'Année Psychologique, 101,* 447-462. http://www.lexique.org.

[New, Araújo, & Nazzi, 2008] New, B., Araújo, V., & Nazzi, T. (2008). Differential processing of consonants and vowels in lexical access through reading. *Psychological Science, 19(12),* 1223-1227.

[Newmeyer, 2000] Newmeyer, F. J. (2000). On the Reconstruction of 'Proto-World' Word Order. In Knight, C., Studdert-Kennedy, M., & Hurford, J. (Eds.) *The Evolutionary Emergence of Language,* Cambridge University Press, 372-388.

[Nichols, 1995] Nichols, J. (1995). Diachronically stable structural features. Historical Linguistics 1993. In Anderson, H. (Ed.) *Selected Papers from the 11th International Conference on Historical Linguistics,* Los Angeles 16-20 August 1993, 337-355. Amsterdam/Philadelphia: John Benjamins Publishing Company.

[Nicolis & Prigogine, 1977] Nicolis, G. & Prigogine, I. (1977). *Self-organization in non-equilibrium systems.* Wiley, New York.

[NJStar Software Corp, 2013] NJStar Software Corp (2013). Chinese Word Processor v.5.30, downloaded from http://www.njstar.com/cms/njstar-chinese-word-processor/.

[Obleser et al., 2010] Obleser, J., Leaver, A., VanMeter, J., & Rauschecker, J. P. (2010). Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization. *Auditory Cognitive Neuroscience, 1: 232.*

[Oh et al., 2013] Oh, Y. M., Pellegrino, F., Coupé, C., & Marsico, E. (2013). Cross-language comparison of functional load for vowels, consonants, and tones. In *Proceedings of Interspeech 2013.* Lyon, France, 3032-3036.

[Oh et al., 2013] Oh, Y. M., Pellegrino, F., Marsico, E., & Coupé, C. (2013). A quantitative and typological approach to correlating linguistic complexity. In *Proceedings of the 5th Conference on Quantitative Investigations in Theoretical Linguistics,* Leuven, Belgium.

[Oh et al., 2013] Oh, Y. M., Coupé, C., & Pellegrino, F. (2013). Effect of bilingualism on speech sate: the case of Catalan and Basque bilinguals in spain. In *Proceedings of The International Congress of Linguists,* Geneva.

[Oh et al., forthcoming] Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (forthcoming). Bridging phonological system and lexicon: insights from a corpus study of functional load, *Journal of phonetics.*

[Ōno, 1970] Ōno, S. (1970). *The origin of the Japanese language.* Kokusai Bunka Shinkokai.

[Ortega, 2003] Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics,* 24(4), 492-518.

[Oudeyer, 2006] Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech.* Oxford: Oxford University Press.

[Owren & Cardillo, 2006] Owren, M. J, Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of the Acoustical Society of America, 119,* 1727–1739.

[Pachès et al., 2000] Pachès, P., de la Mota, C., Riera, M., Perea, M. P., Febrer, A., Estruch, M., ... & Nadeu, C. (2000). Segre: An automatic tool for grapheme-to-allophone transcription in Catalan. In *Proc. of Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC.* pp. 52-61.

[Paninski, 2003] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation,* 15(6), 1191-1253.

[Pellegrino, Coupé, & Marsico, 2007] Pellegrino, F., Coupé, C., & Marsico, E. (2007). An Information theory-based approach to the balance of complexity between phonetics, phonology and morphosyntax, *81st Annual Meeting of the Linguistic Society of America,* Anaheim, CA, USA, 4-7 January 2007.

[Pellegrino, Coupé, & Marsico, 2011] Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language, 87(3),* 539–558.

[Perea et al., 2006] Perea, M., Urkia, M., Davis, C. J., Agirre, A., Laseka, E., & Carreiras, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods, 38(4),* 610-615.

[Piantadosi, Tily, & Gibson, 2009] Piantadosi, S. T., Tily, H. J. and Gibson, E. (2009). The communicative lexicon hypothesis. *Proceedings of the 31st annual meeting of the Cognitive Science Society (CogSci09)*, 2582-2587.

[Piantadosi, Tily, & Gibson, 2011] Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America,* 108(9):3526–3529.

[Piantadosi, Tily, & Gibson, 2012] Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition,* 122(3), 280-291.

[Pierrehumbert, 2001] Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. & Hopper, P. (Eds.) *Frequency Effects and Emergent Grammar.* Amsterdam: John Benjamins Publishing, 137-157.

[Pierrehumbert, 2003] Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In Bod, R., Hay, J., & Jannedy, S. (Eds.) *Probability Theory in Linguistics.* The MIT Press. Cambridge, MA, 177-228.

[Plank, 1998] Plank, F. (1998). The co-variation of phonology with morphology and syntax: A hopeful history. *Linguistic Typology. Volume 2, Issue 2,* Pages 195–230.

[Plank, 1999] Plank, F. (1999). Split morphology: How agglutination and flexion mix. *Linguistic Typology.* Volume 3, Issue 3, Pages 279–340.

[Pluymaekers, Ernestus, & Baayen, 2005] Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica,* 62(2-4):146– 159.

[Prigogine & Nicolis, 1985] Prigogine, I. & Nicolis, G. (1985). Self-organisation in nonequilibrium systems: towards a dynamics of complexity. *Bifurcation Analysis.* Springer Netherlands, pp.3-12.

[Pye, Ingram, & List, 1987] Pye, C., Ingram, D., & List, H. (1987). A comparison of initial consonant acquisition in English and Quiché. In Nelson, K. & Van Kleek, A. (Eds.), *Children's Language, 6,* 175-190. Hillsdale: Erlbaum.

[Qian & Jaeger, 2009] Qian, T., & Jaeger, T. F. (2009). Evidence for efficient language production in Chinese. *CogSci09,* 851-856.

[Qian & Jaeger, 2012] Qian, T. & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive science,* 36(7), 1312-1336.

[R Core Team, 2013] R Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[Research Centre on Linguistics and Language Information Sciences, 2013] Research Centre on Linguistics and Language Information Sciences, The Hong Kong Institute of Education (2013). *A linguistic corpus of mid-20th century Hong Kong Cantonese.* Retrieved on March 1, 2013 from http://hkcc.livac.org.

[Rissanen, 1984] Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory IT-30:4,* pp. 629-36.

[Roberts & Winters, 2013] Roberts, S., & Winters, J. (2013) Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits. *PLoS ONE 8(8): e70902.*

[Robins, 1967] Robins, R. H. (1967). *A short history of linguistics.* Indiana University Press, Bloomington and London.

[Rose, 2009] Rose, Y. (2009). Internal and external influences on child language productions. In Pellegrino, F., Marsico, E., Chitoran, I., & Coupé, C. (Eds.) *Approaches to phonological complexity.* Berlin: Mouton de Gruyter, 329-351.

[Rose, 2014] Rose, Y. (2014). Corpus-based Investigations of Child Phonological Development: Formal and Practical Considerations. In Durand, J., Gut, U., & Kristoffersen, G. (Eds.) *The Oxford Handbook of Corpus Phonology.* Oxford: Oxford University Press. 265-285.

[Sapir, 1970] Sapir, E. (1970). *Language: An Introduction to the Study of Speech.* London: Rupert Hart-Davis (first published 1921).

[Schmid, 1995] Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland.

[Scobbie & Stuart-Smith, 2008] Scobbie, J. M., & Stuart-Smith, J. (2008). Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. In *Contrast: Perception and Acquisition: Selected Papers from the Second International Conference on Contrast in Phonology.* Berlin: Mouton de Gruyter, 87-113.

[Scharinger, Idsardi, & Poe, 2011] Scharinger, M., Idsardi, W. J., & Poe, S. (2011). A comprehensive three-dimensional cortical map of vowel space. *Journal of cognitive neuroscience, 23(12),* 3972-3982.

[Schilling, Rayner, & Chumbley, 1998] Schilling, H. H., Rayner, K., & Chumbley, J. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26(6),* 1270-1281.

[Schwartz et al., 1997] Schwartz, J. L., Boë, L. J., Vallée, N., & Abry, C. (1997). Major trends in vowel system inventories. *Journal of Phonetics, 25(3),* 233-253.

[Seyfarth, 2014] Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition, 133(1),* 140-155.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379-423, 623-656.

[Sharoff, 2006] Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S. (Eds.) *WaCky! Working papers on the web as corpus,* Gedit, Bologna, http://corpus.leeds.ac.uk/query-zh.html.

[Sheik, 2013] Sheik, A. (2013). CantoDict, http://www.cantonese.sheik.co.uk/.

[Sherard, 1985] Sherard, M. (1985). Morphological structure of the pronominal and verb systems in two pronominalized Himalayan languages. In McCoy, J. & Light, T. (Eds.) *Contributions to Sino-Tibetan Studies.* 101-134. Leiden: Brill.

[Shosted, 2006] Shosted, R. (2006). Correlating complexity: A typological approach. *Linguistic Typology 10,* pp.1-40.

[Slobin, 1997] Slobin, D. I. (1997). The origins of grammaticizable notions: Beyond the individual mind. In Slobin, D. I. (Ed.), *The crosslinguistic study of language acquisition.* (Vol. 5, pp. 265–323). Mahwah, NJ: Lawrence Erlbaum.

[Song, 2014] Song, J. J. (2014). *Linguistic Typology: Morphology and Syntax.* Longman Linguistics Library, Taylor & Francis.

[Steels & McIntyre, 1998] Steels, L., & McIntyre, A. (1998). Spatially distributed naming games. *Advances in complex systems, 1(04),* 301-323.

[Stevens, 2002] Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111(4),* 1872-1891.

[Stilp & Kluender, 2010] Stilp, C. E., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility, *Proceedings of the National Academy of Sciences, 107,* 12387–12392.

[Stokes & Surendran, 2005] Stokes, S. & Surendran, D. (2005). Articulatory complexity, ambient frequency and functional load as predictors of consonant development in children. *Journal of Speech and Hearing Research, 48(3).*

[Surendran & Niyogi, 2003] Surendran, D. & Niyogi, P. (2003). Measuring the usefulness (functional load) of phonological contrasts. *Technical Report TR-2003-12.* Department of computer science, University of Chicago.

[Surendran & Levow, 2004] Surendran, D. & Levow, G. A. (2004). The functional load of tone in Mandarin is as high as that of vowels. In *Proceedings of Speech Prosody 2004* Japan.

[Surendran & Niyogi, 2006] Surendran, D. & Niyogi, P. (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In Thomsen, O. N., (Ed.), *Competing Models of Linguistic Change: Evolution and Beyond.* Amsterdam and Philadelphia: John Benjamins, 43-58.

[Tanaka-Ishii, 2012] Tanaka-Ishii, K. (2012). Information Bias Inside English Words. *Journal of Quantitative Linguistics, 19(1),* 77-94.

[Tily et al., 2009] Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition,* 1(2), 147-165.

[Toro et al., 2008] Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding words and rules in a speech stream functional differences between vowels and consonants. *Psychological Science, 19(2),* 137-144.

[Trubetzkoy, 1939] Trubetzkoy, N. S. (1939). *Grundzüge der Phonologie.* Göttingen: Vandenhoek and Ruprecht.

[Trudgill, 2011] Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity.* Oxford University Press.

[Vallée, 1994] Vallée, N. (1994). *Systèmes vocaliques : de la typologie aux prédictions.* PhD dissertation, Université Stendhal, Grenoble, France.

[Van Severen et al., 2012] Van Severen, L., Gillis, J. J., Molemans, I., Van Den Berg, R., De Maeyer, S., & Gillis, S. (2012). The relation between order of acquisition, segmental frequency and function: the case of word-initial consonants in Dutch. *Journal of child language, 40(04),* 703-740.

[van Son & Pols, 2003] van Son, R. J. J. H. & Pols, L. C. W. (2003). How efficient is speech? *Proceedings Institute of Phonetic Sciences,* 25, 171-184.

[Váradi, 2002] Váradi, T. (2002). The Hungarian National Corpus. In *LREC.*

[Villasenor et al., 2012] Villasenor, J., Han, Y., Wen, D., Gonzalez, E., & Chen, J. (2012). The Information Rate of Modern Speech and It's Implications for Language Evolution, *Proceedings of The Ninth International Conference on the Evolution of Language (Evolang 9),* Kyoto, Japan, pp. 376-383.

[Virpioja et al., 2013] Virpioja, S., Smit., P., Grönroos, S. A., & Kurimo, M. (2013). *Morfessor 2.0: Python implementation and extensions for morfessor baseline.*

[Vitevitch, 2008] Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research: JSLHR, 51(2),* 408–422.

[Vitevitch, Chan, & Goldstien, 2014] Vitevitch, M. S., Chan, K. Y., Goldstein, R. (2014). Insights into failed lexical retrieval from network science. *Cognitive Psychology, 68,* 1-32.

[von Foerster, 1960] von Foerster, H. (1960). On self-organising systems and their environments. In Yovits, M. C. & Cameron, S. (Eds.) *Self-Organising Systems.* Pergamon Press, London, pp. 30- 50.

[Walsh et al., 2010] Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive Science, 34(4),* 537-582.

[Wang, 1967] Wang, W. Y. (1967). The measurement of functional load. *Phonetica, 16(1),* 36-54.

[Weaver, 1953] Weaver, W. (1953). Recent contributions to the mathematical theory of communication. *ETC: A Review of General Semantics,* 261-281.

[Wedel, 2011] Wedel, A. (2011). *Self-organization in phonology.* The Blackwell companion to phonology, 1, 130-147.

[Wedel, 2012] Wedel, A. (2012). Self-organization and categorical behavior in phonology. In *Annual Meeting of the Berkeley Linguistics Society.* Vol. 29, No. 1.

[Wedel, Kaplan, & Jackson, 2013] Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition, 128(2),* 179-186.

[Wichmann & Holman, 2009] Wichmann, S. & Holman, E. W. (2009). *Assessing Temporal Stability for Linguistic Typological Features.* München: LINCOM Europa.

[Wichmann, Rama, & Holman, 2011] Wichmann S., Rama T., Holman E. W. (2011). Phonological diversity, word length and population sizes across languages: the ASJP evidence. *Linguistic typolology 15,* 177–197.

[Wray & Grace, 2007] Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua,* 117(3), 543-578.

[Zipf, 1949] Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley Press.

[Zoubir, A. M. & Iskander, 2007] Zoubir, A. M. & Iskander, D. R. (2007). Bootstrap Methods and Applications : A Tutorial for the Signal Processing Practitioner. *IEEE - Signal Processing Magazine 24(4).* pp. 10-19.

[Zséder et al., 2012] Zséder, A., Recski, G., Varga, D., & Kornai, A. (2012). Rapid creation of large-scale corpora and frequency dictionaries. In *Proceedings to LREC 2012.*