

An abstract graphic featuring a dense network of thin, curved lines in various colors (yellow, green, blue, purple) that converge and diverge. Interspersed among these lines are numerous spheres of different sizes and colors, including yellow, blue, purple, green, and red. The overall effect is a complex, multi-colored web of connections.

Scholars'
Press

Gerardo L. Febres

Quantifying the Complexity of Languages

A Novel Scope for the Comparative Analysis of
Communication Systems

Quantifying the Complexity of Languages. A novel Scope for the Comparative Analysis of Communication Systems. Gerardo L. Febres. 2016. Scholars' Press. Saarbrücken, Germany. ISBN: 978-3-639-85969-0.

Quantifying the Complexity of Languages

An innovative way to treat languages and systems, describing them as information structures, which allows for a comparative study of their properties, despite the obvious differences among their natures. To unveil some of the mysteries underlying all sort of languages—from music to natural languages and from bi-dimensional graphical representations to the synthesis of mathematics—a wide range of representations and information transfer phenomena are described at several levels of detail. Languages, considered as the expression of complex systems, are analyzed by means of ingeniously developed algorithms and methods. A vast field for applications opens in front of these universal analysis tools, promising a dense learning time to come.



Gerardo Luis Febres Añez is mechanical engineer from Universidad Simón Bolívar (USB), Venezuela. He holds a Master of Operations Research from Cornell University. He earned his PhD from USB, where he teaches systems' modeling and optimization. Among his research interests are the modeling of systems at diverse scales and quantitative linguistics.



978-3-639-85969-0

Quantifying the Complexity of Languages

A novel scope for the comparative analysis of communication systems

Gerardo Luis Febres Áñez

Science is the precise description of our observations...
Technology is the precise application of science.

Accurate observation motivates understanding. That's My Father.
Never end persevering. Never end learning. That's My Mother.
... for Them.

Preface

This book is based on my doctoral thesis dissertation, presented 9 months ago. Here I include the studies, those published and unpublished at the time, which integrated the doctoral work. However, and fortunately, it was only the beginning of my activity as a researcher in a rich field, full of interesting questions and mysteries waiting to be addressed by the scientific community. This research has continued. Interesting and meaningful results have come up to complement the results obtained prior to the presentation of the thesis. Always considering that completeness should prevail, I included in the now numbered Chapter VIII, the content of several studies still in progress, thus moving the Conclusion to Chapter IX, which now presents two additional sections.

With the new material, the scope of the book extends. Now, in addition to the original focus in developing tools and methods to recognize the identity of descriptions and to establish a comparative analysis between expressions of different languages, the book, towards its final chapters, points to a deeper understanding of the structure of information.

Three components of information are recognized: symbolic, spatial and semantic. A relationship of these three components leads to a method to estimate the semantic information; a quantity that has evaded most, if not all, tries for its quantification. I think this justifies including these results in this edition.

At this time, all works supporting the thesis are either, published or in press. I list these works in their current condition, and I add those works to be promptly submitted:

[63] G. Febres, K. Jaffé, C. Gershenson, Complexity measurement of natural and artificial languages, *Complexity*. 20 (2015) 429–453.
doi:10.1002/cplx.21529.

[71] G. Febres, K. Jaffé, Quantifying literature quality using complexity criteria, *J. Quant. Linguist.* (in press, 2016).

- [75] G. Febres, K. Jaffé, A Fundamental Scale of Descriptions for Analyzing Information Content of Communication Systems, *Entropy*. 17 (2015) 1606–1633. doi:10.3390/e17041606.
- [76] G. Febres, K. Jaffe, Calculating entropy at different scales among diverse communication systems, *Complexity*. early view (2015). doi:10.1002/cplx.21746
- [] G. Febres, K. Jaffe, Music viewed by its entropy content: A novel window for comparative analysis, *Advances in Complex Systems*. (in Press. 2016).
- [] G. Febres, Where is the information? (to be submitted. 2016).

Gerardo Febres

Caracas, Venezuela. February, 2016.

Acknowledgements

This is the result of the effort of many people. Here I want to thank them for their valuable help to achieve the completion of this project.

I thank Professor Klaus Jaffé for his continuous guidance and optimism. His attitude towards research has been an essential example to follow for the successful completion of this Thesis. Our recurrent technical discussions about complexity and other fields of research, constitute an important guideline in my future works as researcher. Equally important is the friendship and support he has always offered.

I have not yet personally met Professor Carlos Gershenson. But by means of technology, I have had the opportunity to receive his guidance about complexity. I want to thank for his generosity in sharing his deep knowledge with me.

The actual start of this study occurred years ago, when I devised a computerized platform to model the Venezuelan registry system in real-time fashion. The resulting network of systems became a logical representation of the relationships among the participants of the real system. As the new procedures were difficult and risky, support and confidence were as important as our technical capabilities. In this regard, I acknowledge the lawyer Beatriz Perez-Perazzo, who has been present since the primary phases of this study —the pre-academic stages—, and has constantly cared about the development of this work.

I feel grateful to Professors Alfredo Rios and Fernando Torre, for their fruitful conversations and encouraging comments about the progress of my work. I also thank my mother Amparo, my sister Dianora, and my brother Gabriel, who were always supportive and enthusiastic about the completion of this Thesis. Finally, I am deeply thankful to my friend Libia Álvarez Espín and my daughter Gabriela Febres Cueto, for the dedication they devoted to the review of some of the texts, graphs and tables of this study.

Acknowledgments

Quantifying the Complexity of Languages

Abstract

An overview of the various meanings of the term complexity is presented. Languages are then explored as to their capacity as vehicles for the description of complex systems. Measures were designed to compare different languages regarding some characteristics of the structure of the information they manage, such as entropy, diversity and complexity. These measures allowed to visualize regions in the diversity-entropy space which were specific to English, Spanish, computer programming language and music.

The analysis of various different types of languages with very diverse syntax, grammar and structure, was made possible by introducing the concept of a "Fundamental Scale" for the description of a system. This scale was defined as the one that produced the most parsimonious probability distribution when fitting various sizes of symbols representing descriptions in a given language.

A set of computer programs is presented that allows the design, implementation and control of the analysis of meta-data required for a quantitative description of complex systems. The program represents the system at different scales, each describing aspects related to its inner nature. The program uses modules developed to perform arithmetic operations that evaluate characteristic functions describing as tables and trees, multidimensional objects with regular and non-regular structures. These tools allowed to demonstrate that by using the degree of diversity of symbols and descriptive entropy measures, we can identify and quantify subtle differences between the languages studied. These differences are related to the form and style of each language. The tools allow the comparison of different kinds of complexity among very different types of languages. There are two areas where these findings may have important impact: (a) while the field of traditional compression techniques point towards the economy of transferring information, the study of languages by means of their structure and 'genome', is directed to what we could call *the compression of the interpretation*. (b) The computer implementation of a collective evaluation of texts, opens the possibility for massive or distant evaluation of system descriptions and natural language texts.

Key words: complexity, entropy, diversity, languages, scale, information.

Contents

Preface	v
Acknowledgements	vii
Abstract	ix
Contents	xi
List of figures	xix
List of tables	xxiii
Nomenclature	xxvii
Chapter I. Introduction	1
Chapter II. Visions of complexity	3
2.1 Interpretations of complexity	3
2.1.1 Symbolic complexity	5
2.1.2 Algorithmic complexity	6
2.1.3 Computational complexity	7
2.1.4 Effective complexity	8
2.1.5 Networks	9
2.2 Common paths of complexity	9
2.3 Languages as complex systems	10
Chapter III. Complexity measurement of natural and artificial languages	13
3.1 Methods	14

3.1.1 Text length L and symbolic diversity d	15
3.1.2 Entropy h	15
3.1.3 Emergence e	16
3.1.4 Self-organization s	16
3.1.5 Complexity c	16
3.1.6 Symbol frequency distribution f	17
3.1.7 Zipf's deviation J for a ranked distribution	18
3.1.8 Message selection	19
3.1.9 Symbol treatment	19
3.1.10 Software	21
3.2 Results	22
3.2.1 Diversity for natural and artificial languages	22
3.2.2 Entropy for natural and artificial languages	23
3.2.3 Emergence, self-organization and complexity	25
3.2.4 Symbol Frequency distributions	27
3.2.4.1 Zipf's deviation $J_{1,D}$ for ranked distribution	31
3.2.4.2 Tail Zipf's deviation $J_{\theta,D}$ for ranked tail distributions	31
3.3 Discussions	32
3.3.1 Diversity for natural and artificial languages	32
3.3.2 Entropy for natural and artificial languages	33
3.3.3 Symbol frequency distributions	35
3.4 Conclusions	36
Chapter IV. The representation of writing styles as symbolic diversity and entropy	39
4.1 Methods	41
4.1.1 Text length L and symbolic diversity d	42
4.1.2 Entropy h	42
4.1.3 Symbol frequency distribution f	42
4.1.4 Zipf's Deviation J	43
4.1.5 Relative deviations of texts properties	43
4.1.6 Writing Quality Scale WQS	44

4.1.7	Readability formulas <i>RES</i> and <i>IPSZ</i>	44
4.1.8	Message selection and groups	46
4.2	Results	47
4.2.1	Diversity for literature Nobel laureates and general writers	47
4.2.2	Entropy for literature Nobel laureates and general writers	48
4.2.3	Zipf's deviation $J_{1,D}$ for ranked distribution	50
4.2.4	Writing quality evaluation	53
4.2.5	Writing quality scales and readability indexes	56
4.2.6	Writing style change in time	57
4.3	Discussions	59
4.3.1	Diversity and entropy	59
4.3.2	Symbol frequency distribution profile	59
4.3.3	Writing Quality Scale versus Readability index	59
4.3.4	Tendencies of the writing style	60
4.4	Conclusions	61
Chapter V. The fundamental scale of descriptions		63
5.1	A quantitative description of a language	65
5.1.1	Quantity of information for a D 'nary language	65
5.1.2	Scale and resolution	65
5.1.3	The minimum length description scale	66
5.1.4	Language recognition	69
5.2	The Fundamental Scale Algorithm	70
5.2.1	Base language construction	70
5.2.2	Prospective symbol detection	71
5.2.3	Symbol birth process	72
5.2.4	Conservation of symbolic quantity	73
5.2.5	Symbol Survival Process	73
5.2.6	Controlling computational complexity	73
5.3	Tests and results	74
5.4	Discussions	78
CHAPTER VI. Several communication systems viewed at different scales		83

6.1 Methods	84
6.1.1 Diversity and entropy	84
6.1.2 Language scale	85
6.1.2.1 The character's scale	87
6.1.2.2 The word's scale	87
6.1.2.3 The fundamental scale	88
6.1.3 Scale downgrading	88
6.1.4 Message selection	90
6.1.4.1 Natural languages	90
6.1.4.2 Computer programming code	90
6.1.4.3 MIDI music	90
6.2 Results	93
6.2.1 Diversity	94
6.2.2 Entropy	95
6.2.3 Symbol frequency profiles	96
6.2.4 Stabilization length	98
6.3 Discussions	101
6.3.1 Diversity and entropy	102
6.3.2 Symbol frequency profiles	103
6.3.3 Description length	103
6.3.4 About the forces shaping languages	104
6.4 Conclusions	105
Chapter VII. Music entropy models	107
7.1 Methods	109
7.1.1 Language recognition	110
7.1.2 Specific diversity and entropy	111
7.1.3 The fundamental scale of a description	111
7.1.4 Scale downloading	112
7.1.5 Higher order entropy	112
7.1.6 Music selection	114
7.2 Results	116
7.2.1 Diversity and entropy	196
7.2.2 Information profiles	117

7.2.3 Symbol frequency profiles	119
7.2.4 Clusters and tendencies	121
7.3 Discussions	124
7.3.1 Diversity and entropy	125
7.3.2 Frequency profiles	126
7.3.3 About the evolution of music	127
7.4 Conclusions	128
Chapter VIII. Where is the information?	131
8.1 Properties of descriptions: Resolution, Scale and Scope	133
8.1.1 Resolution R	133
8.1.2 Scale D	133
8.1.3 Scope L	135
8.2 Balance of information content	135
8.3 An information flow model	139
8.4 Finding the fundamental scale	140
8.5 Comparing languages at different scales	141
8.6 Some test with different language expressions	142
8.6.1 Natural languages	142
8.6.2 Same symbolic structure. Different perceptions	143
8.6.3 Partial changes of resolution and scope	144
8.6.4 The impact of reorganizing	146
8.6.5 Music	147
8.6.6 Mathematics as a language	149
8.7 Information component fractions	149
8.8 Discussion	151
8.8.1 Implications of scale, scope and resolution	151
8.8.2 The balance of information	152
8.9 Summary	154
Chapter IX. Conclusion	157
9.1 Main results and contributions	159
9.1.1 Language quantitative analysis	159

9.1.2 The notion of scale as a numerical property	159
9.1.3 The fundamental scale	160
9.1.4 Notions of spatial and sematic information	160
9.1.5 An information flow model	161
9.1.6 A complex-experiment software platform	161
9.2 Possible future works	161
9.2.1 The concept of fundamental scale at multidimensional languages	161
9.2.2 Applying the method to other fields	162
Bibliography	163
Appendix A. MoNef: complex experiment modeling platform	169
A.1 Overview	170
A.2 Major components. Architecture	170
A.2.1 Environment	170
A.2.2 Data storage. File-object types	171
A.2.2.1 The .NPD extension	171
A.2.2.2 The .NPM extension	171
A.3 Object nature types	172
A.4 User interface	172
A.5 Object description	174
A.6 Model description and data input	174
A.7 Internal languages and syntaxes	175
A.7.1 The autonomous multidimensional object representation	175
A.7.2 The Localizer pseudo-language syntax	177
A.7.3 Functions and complex operations	178
Appendix B. Properties of natural languages and programing language texts	185
Appendix C. Literature Nobel laureates and non-laureates text properties	195
Appendix D. The Fundamental Scale Algorithm	205
Appendix E. Symbols of two descriptions at the fundamental scale	209

Appendix F. Language properties at different scales	219
Appendix G. MIDI music properties. Musicnet	227
Appendix H. Numerical data of the symbol frequency profiles for MIDI music	229
Appendix I. Symbol probability profiles for composers	233
Appendix J. Music styles by composer, in the space (<i>specific diversity, entropy, 2nd order entropy</i>)	237

List of Figures

Figure	Description	Page
2.1	Hierarchy of problem classes for Computational Complexity	7
3.1	Typical symbol ranked profile.	18
3.2	Diversity for messages expressed in English, Spanish and Computer Code.	22
3.3	Messages entropy vs. specific diversity for English, Spanish and computer code.	23
3.4	Messages entropy vs. specific diversity for English, Spanish and Computer Code.	25
3.5	Emergence, self-organization and complexity for English, Spanish and Computer Code.	26
3.6	Emergence, self-organization and complexity for English, Spanish and Computer Code.	26
3.7	Ranked symbol frequency distribution for English, Spanish and Computer Code.	27
3.8	Ranked symbol frequency distribution for English, Spanish and Computer Code.	29
3.9	Cumulative distribution function (CDF) of symbols ranked by frequency.	30
3.10	Zipf's deviation $J_{1,D}$ of symbol ranked frequency distributions depending on text length L .	30
3.11	Tail Zipf's deviation $J_{\theta,D}$ for symbol ranked frequency distributions vs. text tail length L .	32

List of Figures

4.1	Diversity D as a function of message length L for messages expressed in English and Spanish by non-Nobel and Literature Nobel laureates.	47
4.2	Entropy h vs. specific diversity d for messages expressed in English and Spanish by non-Nobel and Literature Nobel laureates.	49
4.3	Ranked symbol frequency distribution profiles.	51
4.4	Zipf's deviation $J1, D$ vs. message length L for messages expressed English and Spanish by non-Nobel and Literature Nobel laureates.	52
4.5	Writing quality evaluation for English and Spanish texts.	55
4.6	Text readability vs. Writing Quality Scale QWS for English texts and Spanish texts.	56
4.7	Average sentence length [words] vs. year when the speech was written for English and Spanish.	58
4.8	Writing Quality Scale WQS vs. year when the speech was written for English and Spanish.	58
5.1	Major components of the Fundamental Scale Algorithm.	71
5.2	Examples of reading a text to recognize prospective symbols with a sliding window.	72
5.3	Symbol profiles for an English text and a MIDI music text at different scales of observation.	77
5.4	Bertrand Russell's 1950 Nobel ceremony speech behavior according symbol length.	77
5.5	Beethoven's 9 th symphony 4 th movement MIDI music language behavior according symbol length.	78
6.1	Graphic representation of a language scale downgrading from scale D to scale S.	89
6.2	Diversity of as a function of description length measured in symbols.	92
6.3	Symbol entropy as a function of specific diversity.	95
6.4	Probability profiles for several communication systems.	97
6.5	Entropy vs. Length in symbols for different types of communication systems at their fundamental scale.	99

6.6	Model of entropy vs. description length in symbols.	100
7.1	Typical .symbol ranked probability profile with examples of 2nd order symbol bands.	112
7.2	Diversity as a function of music piece length measured in symbols.	116
7.3	Entropy as a function of specific diversity.	116
7.4	Variation of frequency profiles for several degraded scales and Information profiles calculated for three musical pieces.	118
7.5	Symbol Ranked Frequency profiles for 12 different types of western academic music.	120
7.6	2nd order Symbol Ranked Frequency profiles for 12 different types of western academic music.	121
7.7	Three views of the representation of music pieces in the space specific diversity, entropy, 2nd order entropy($d, h_D, h_D^{[2]}$).	123
7.8	Three views of the representation of music period/style groups in the space specific diversity, entropy, 2nd order entropy($d, h_D, h_D^{[2]}$).	124
7.9	Variation of 2nd order entropy over time for several types of music	125
8.1	Information flow graphical model.	140
8.2	Two perceptions of a 2D mosaic with a resolution 60 x 60 pixels.	143
8.3	Effects of changes of resolution and scope.	145
8.4	Four views of the same distribution of 30 squares.	146
8.5	A tiny fraction of the text which constitutes the Beethoven's 5th symphony 1st movement.	148
8.6	Five examples of mathematical descriptions.	149
8.7	Fraction of semantic information vs. fraction of spatial information.	153
9.1	The message's interpretation obeys the aspect the observer is interested in.	157
9.2	Languages are self-organizing sets of symbols.	158
A.1	MoNet's general architecture.	171

List of Figures

A.2	<i>MonNet's</i> hypothetical model file structure showing the relationships of files and their logical connections.	172
A.3	<i>MoNet's</i> graphic interface.	173
B.1	Diversity of words used in English speeches and novel segments vs. text length in words.	193
B.2	Diversity of words used in Spanish speeches and novel segments vs. text length in words.	193
C.1	Writing style for English speeches.	202
C.2	Writing style for Spanish speeches.	203
I.1	Symbol probability profiles of music by composer.	233-236
J.1	A view of Composers' style represented in the space: specific diversity, entropy, 2 nd order entropy.	238
J.2	A view of Composers' style represented in the space: specific diversity, entropy, 2 nd order entropy.	239
J.3	A view of Composers' style represented in the space: specific diversity, entropy, 2 nd order entropy.	240

List of Tables

Table	Description	Page
2.1	Some important milestones in the fields of complexity and networks	4
3.1	Most frequently used symbols in English and Spanish.	28
3.2	Zipf's deviation $J_{\theta,D}$ and its correlation with length L for English, Spanish and artificial messages	31
3.3	Tail Zipf's deviation $J_{\theta,D}$ and its correlation with message tail length L_{θ} for English, Spanish and artificial messages	32
4.1	Comparing the relative specific diversity d_{rel} for English and Spanish messages by non-Nobel and Literature Nobel laureates	48
4.2	Comparing the relative entropy h_{rel} for English and Spanish messages by non-Nobel and Literature Nobel laureates	50
4.3	Comparing the relative Zipf's deviation $J_{1,D}$ for English and Spanish messages by non-Nobel and Literature Nobel laureates	52
4.4	Comparing the Writing Quality Scale WQS for English and Spanish messages by non-Nobel and Literature Nobel laureates	57
5.1	Results of the analysis of the Example Text at the three scales studied	75
5.2	Properties of two descriptions used to test the fundamental scale method	76
6.1	Number of messages processed for English, Spanish, computer programming code, and MIDI music	91

List of Tables

6.2	Properties of different communication systems considered as the union of all messages expressed in English, Spanish, computer programming code, and MIDI music	97
6.3	Average and standard deviation of the specific diversity and entropy for different types of communication systems, measured at the fundamental scale	100
7.1	Music classification tree and the data associated to the musical pieces considered within this study	115
7.2	Properties of western academic music	122
7.3	Properties of some traditional and popular music	122
8.1	Effects of different observation scales over the quantity of information of English texts	142
8.2	Properties of each interpretation of 2D patterns	144
8.3	Balance of information for the 2D example	145
8.4	Properties of each interpretation of 2D patterns	147
8.5	Effects of different observation scales over the quantity of information for two pieces of music	148
8.6	Properties of the mathematical descriptions	150
8.7	Relative weight of information type content for four information transmission media	151
A.1	<i>MoNet's</i> inherent attributes	174
A.2	Some examples of structured objects descriptions coded in the Autonomous Representation	176
A.3	List of transcendental function routines	178
A.4	List of matrix operation functions	179
A.5	List of probability distribution routines	179
A.6	List of file relative position functions	180
A.7	List of discrete functions and structures	181
A.8	List of language description functions	182
A.9	List of language discrete functions and structures	183

A.10	List of File system functions and other functions	184
B.1	Properties of artificial texts	186
B.2	Properties of English texts	187
B.3	Properties of Spanish texts	190
C.1	Properties of English texts by non-Nobel laureates	196
C.2	Properties of English texts by Literature Nobel laureates	198
C.3	Properties of Spanish texts by non-Nobel laureates	199
C.4	Properties of Spanish texts by Literature Nobel laureates	201
E.1	Word-scale profile of Bertrand Russell's speech given at the 1950 Nobel Award Ceremony	210
E.2	Fundamental-scale profile of Bertrand Russell's speech given at the 1950 Nobel Award Ceremony	213
E.3	Fundamental-scale profile for a MIDI version of Beethoven 9th Symphony, 4th movement	215
F.1	Properties of English language at different scales	220
F.2	Properties of Spanish language at different scales	223
F.3	Properties of computer programming code at different scales	225
F.4	Properties of MIDI Music at fundamental scale	226
G.1	MIDI Music properties. <i>MusicNet</i>	227
H.1	Numerical data for symbol probability for different types of music. Symbols determined by the fundamental scale method	230

Nomenclature

Notation

a	Some symbol rank limit.
b	Some symbol rank limit.
\mathbf{B}	Language regarded as a set of symbols.
d	Specific diversity.
d_{rel}	Specific diversity deviation respect to a diversity best fit curve.
C_i	Character indexed as i within a series of characters.
D	Diversity. Scale.
D_m	Diversity evaluated for the best fit entropy curve.
$D_{\mathbf{B}}$	Diversity of a message when observed using language \mathbf{B} .
L	Text length. Scope.
c	Complexity.
cs	Symbolic complexity.
C_{SY}	Number of characters in a syllable.
e	Emergency.
E, E_{Y_i}	Space required by a symbol, space required by symbol Y_i Measured in characters.
E	Distance between a symbol probability and the imaginary Zipf's distribution. Semantic information.
\mathbf{E}	Vector of distances E .
f	Symbol frequency.
f_r	Frequency of the symbol with rank r .
$F()$	Symbol's ranked frequency distribution.
g	Slope constant in ordered probability profiles
G	Degrees of freedom of a space.
h	Entropy.
h_B^s	Entropy of language B after adaptive stage s .

Nomenclature

h_m	Entropy evaluated for the best fit entropy curve.
h_{max}	Maximum entropy.
h_{rel}	Entropy deviation respect to an entropy best fit curve.
$h_{V,s}$	Entropy of a language after s adaptive stages of symbols of V chars.
h_{st}	Established entropy
$h^{[1]}, h^{[2]}$	First order entropy. Second order entropy.
$h_D^{[1]}$	First order entropy at observation scale D .
$h_D^{[2]}$	Second order entropy at observation scale D .
i	Index of an element within a vector, array or series.
j	Index of an element within an array or series.
J	Zipf's deviation ranked distribution.
$J_{\theta,D}$	Zipf's deviation J for a ranked distribution between θ and D .
k	Constant.
k	Index of an element.
$L_{a,b}$	Length between symbols ranked from a to b .
L_{SY}	Number of syllables a text.
L_W	Number of words a text.
M	Message. Total information of a message.
	Total message information.
N	Length of a text measured in symbols.
N_{st}	Stabilization Length of a text measured in symbols.
n_{st1}, n_{st2}	Number of descriptions with lengths N larger than the stabilization length N_{st} .
ns	Sub index to indicate the condition of <i>new symbol</i> .
os	Sub index to indicate the condition of <i>overlapped symbol</i> .
p_i	Probability of element indexed as i .
P	Some Fraction of nodes with highest degree in a network.
\mathbf{P}_c	Vector of probabilities of a symbol with condition c .
$P_{V,s}$	Vector of probabilities of a language after s adaptive stages of symbols of V characters.
$\mathbf{P}()$	Symbol's ranked frequency distribution.
q	Resolution of the classification of values of a distribution.
r	Rank.
	Density of resolution
r_i	Rank of element indexed as i .
R	Resolution
rs	Sub index to indicate the condition of <i>remaining symbol</i> .
s	Self-organization.
s	Scale.
s	As sub index: scale of observation of a language.
S	Spatial information.

$[u]$	As supra index: order of a property.
U	Element of the transformation matrix \mathbf{U} .
$\mathbf{U}, \mathbf{U}^{[so,fo]}$	Property order transformation matrix from order so to order fo
W	Some Fraction of edges in a network. Number of syllables per word.
Y	Symbol. Symbolic information.
Y_i	i 'th symbol in a sequence of symbols.
x	Generic variable.
x_i	Value of variable x for element indexed as i
$Z_{a,b}$	Zipf's reference value between symbols ranked from a to b .
α	A real number between 0 and 1.
λ	Entropy reduction parameter used as threshold for the birth of a fundamental symbol.
θ	Rank Zipf's profile tail start value.
μ, ν	Error minimization adjusting value.
Δ	Variation.
*	As sub index: product of an optimization process.
$\{C_1, C_2, \dots, C_N\}$	Set of elements C_i .
$(d_{rel}, h_{rel}, J_{rel})$	Coordinates of a point in a 3-dimensional space.

Fundamental Scale Algorithm notable variables and arrays

Phase	Number of characters skipped at the beginning of the reading scan of a text.
Symbol []	1 dimension array. Contains the symbols of used in description.
SymbolFrequency[i]	1 dimension array. Frequency of Symbol[i] .
SymbolPosition[i,j]	2 dimensions array. Position of the j 'th instance of Symbol[i]
ProspectiveSymbol []	1 dimension array. Contains the prospective symbols which may be used in a description.
ProspectSymbolFreq [i]	1 dimension array. Frequency of ProspectiveSymbol [i] .
ProspectSymbolPos[i, j]	2 dimensions array. Position of the j 'th instance of ProspectiveSymbol[i] .

Abbreviations

<i>ASCII</i>	American Standard Code for Information Interchange
<i>CFG</i>	Context-Free Grammar.
<i>DNA</i>	Deoxyribonucleic Acid.
<i>GTH</i>	Generative Theory of Tonal Harmony.
<i>GTM</i>	Generative Theory of Tonal Music.
<i>FSA</i>	Fundamental Scale Algorithm.
<i>IPSZ</i>	Szigritsz Perspicuity Index.
<i>MDL</i>	Minimal Description Length (Principle of).
<i>MIDI</i>	Musical Digital Interface.
<i>RES</i>	Flesch Readability Score.
<i>WQS</i>	Writing Quality Scale.

"Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains."
Steve Jobs

Chapter I

Introduction

In the literal sense, complexity means the difficulty to find a solution to a problem or situation. Until recently the classic management complexity was used as a way to signal some kind of limit on the ability to find solutions or analyze a problem. In recent decades, the development of computers, whose capacity has continued steadily to Moore's Law, has induced an interpretation of complexity that is essentially different from that which had dominated for centuries. The current capacity of computers have made it possible to see the complexity as an entity that can at least be studied and classified, and depending on the situation in which it occurs, even evaluated. Complexity is a subject of study where many areas of knowledge converge. Different conceptions of what complexity is, have appear and coexist to satisfy the interests of each field of study.

Complexity meanings could have started in the nineteen century with Auerbach's [1] observations about the structure of some nature facets. Fifty years later Zipf [2] found important similarities between the way human languages create and use words with these nature structures which Auerbach had observed formerly. Shannon [3] and Weaver found a method to quantify the information content of a description and Solomonoff [4]. Kolmogorov [5], and Chaitin [6], working independently, stablished the idea that complexity of an entity is strongly dependent on the quantity of information required to precisely depict it. But they referred to the shortest possible description without compromise of its exactitude, thus, since there is always the possibility to encounter a shorter way to describe something, the evaluation of the Kolmogorov's Complexity, as it is known, remains as an unreachable objective. Schrödinger's work 'What is Life' [7] started another conception of complexity

by pointing out that a complex system activity is devoted to organize itself and thus to keep its own entropy bounded to a certain limit. The idea that the entropy equilibrium point is an index of the system complexity has gained support among researchers [8–10]. Other perspectives of complexity have been suggested. Gell-Mann [11] for example, mentions that complexity could refer to the estimated capacity of a system to build up internal entropy in it pass into the future; it could be called 'Potential Complexity'. Bar-Yam [12] proposes to measure complexity by quantifying the degree of independence among several scales of observations of a system description.

Our ability to understand by means of synthesis and reductionism is growing at a slower pace than does our need to accurately interpret the complex systems developing around us. Describing systems and their behavior with more and more detail has been possible thanks to rapid growth of computers capacity as well as the evolution of the languages used to control them. The result is an explosion of possibilities to increase the knowledge about nature and about ourselves. The result is this new way of making science we call Complexity, present today in every corner of the unlimited field of the interdisciplinary sciences.

Languages are probably the most essential tool used to express, convey, and register information. Language may be depicted as a set of symbols and rules about their use, shared by the extremes of the communication process. The objective of this study is to encounter an appropriate ways to measure complexity in languages and descriptions. Even though complexity conceptions are focused in different aspects of the constitution of systems, most complexity indexes are heavily dependent on measures of entropy. Thus evaluating entropy becomes a necessary path to obtain a sense of complexity. The criteria used to select the symbols within a textual description, is a study parameter as well.

This thesis is organized in the same order the research have been addressed. It starts with an overview of the most important notions of complexity. Following, the possibility of analyzing English and Spanish descriptions was explored using the words as elementary symbol unit. In Chapter VI a method to find the most representative set of symbols in a description –the description Fundamental Scale– is developed. The thesis end with the use of the Fundamental Scale method to measure complexity at different scale in human natural languages, artificial languages and music.

"The aim of science is to seek the simplest explanations of complex facts. We are apt to fall into the error of thinking that the facts are simple because simplicity is the goal of our quest. The guiding motto in the life of every natural philosopher should be, 'Seek simplicity and distrust it.'"
Alfred North Whitehead

Chapter II

Visions of complexity

Many different conceptions of Complexity coexist today. Any description is subject to the relevance assigned to each aspect of the entity being described. We, therefore, do not describe the entity. We describe one or many aspects considered relevant of the entity. The subjectivity of the relevance has driven the conception of complexity to many objective, some of them even quantifiable, different concepts. I feel the term 'complexity' is still looking for its precise meaning. But it does not seem the term complexity will converge to a unique meaning. In this Chapter, a view of the most used conceptions of complexity is presented.

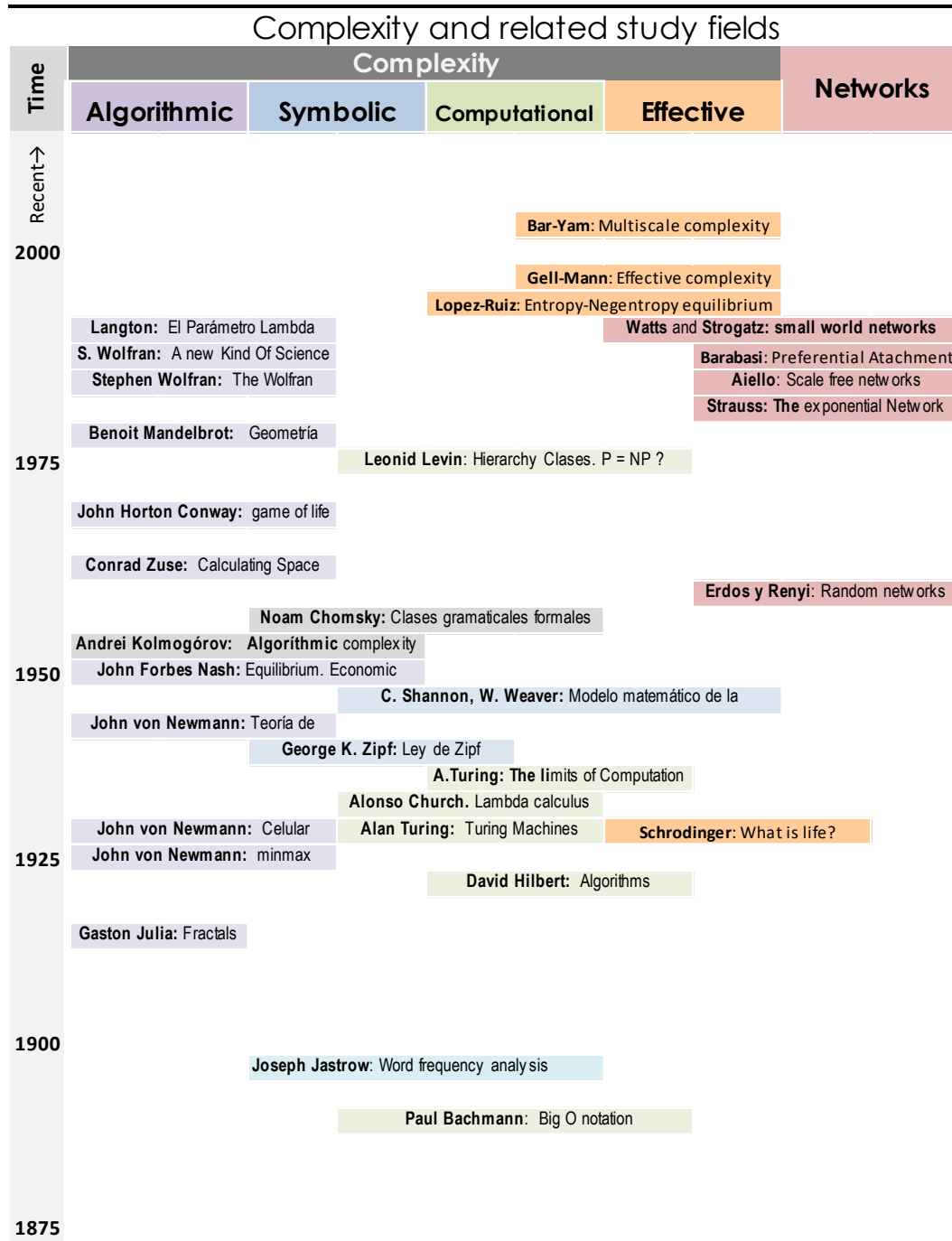
2.1 Interpretations of complexity

The approach to the study of complexity has been done from many perspectives, and traveling many avenues of research related to the concept of complexity. The analysis about the possibility of systems modeling by using machines, started with David Hilbert, Alan Turing and Alonso Church, who are forerunners of modern computing theory, and the concept of algorithm that we use today. They established the basis for modern computer science, and induced studies on the most central computational complexity question: How long does take a machine to perform a task?

Another approach to the implications of the complexity of systems was initiated by Claude Elwood Shannon and Warren Weaver, who created during the mid-40s, the field of communication theory, now called information theory, which

studies the efficiency with which information is handled during the communication process. This has led to the concept of complexity as a measure of the information needed to build description of a system.

Table 2.1: Some important milestones in the fields of complexity and networks.



There have appeared a conception of complexity as a measure of the internal activity of the system to self-organize its internal parts. This conception of complexity is at first glance, a depiction of the physical aspects of the system instead of a depiction of the information describing the system. However, any physical property comes in the form of information. So, leaving the concept of information outside our consideration, is not yet possible.

Whichever the conception of complexity is adopted, reducing the complexity of a system to a single number, challenges the very notion of complexity. It was not long ago, when the need for synthesis to produce a description a system aspect, was inescapable. There were no computers, no registration automated forms or automated data results. In many cases the study required rethinking the problems to reduce them to conditions of symmetry, uniformity, homogeneity and ideality, in order to implement the most capable synthesis-tool that we know: Mathematics. Today's computers are powerful enough to change that situation, and several ways of looking at complexity are now quantifiable.

Finally, the meaning of complexity have specialized versions, and nowadays it depends on the area in which it is used. Although closely related, these definitions carry themselves essential differences. Some ways to evaluate complexity are absolute. Some are relative. While some refer to the difficulty of synthesize or compress an information package, other refer to the effort required to expand a previously compacted message or description. The next paragraphs are devoted to go slightly deep under the surface of today's most recognized concepts of complexity.

2.1.1 Symbolic complexity

Symbolic Complexity, also called Shannon's complexity, is a function of the amount of information contained in a message according to its representation as a sequence of symbols. The amount of information contained in a message is estimated by calculating its entropy. The name of entropy as a property of a text was borrowed from thermodynamic concepts that express the 'way' in which it is or may be organized energy or temperature fields.

Naturally, the loan term is due to the parallelism that occurs between the physical and thermodynamic information field situations. In his work Shannon [3] showed that the minimal length of binary message is proportional to the entropy calculated on the basis of the frequency of appearance of the two symbols.

Being entropy a measure of the minimal length of a message and therefore an indication of the resources involved in the transmission of a message, some authors, especially those oriented toward the fields of information and

communications, consider entropy and complexity mutually proportional. Thus, representing the probability of encountering the *i*'th symbol as p_i , and using the positive constant k as the proportionality factor, the symbolic complexity cs based on Shannon's entropy can be determined as

$$cs = k \cdot h = -k \cdot \sum_{i=1}^2 p_i \cdot \log_2 p_i . \quad (2.1)$$

2.1.2 Algorithmic complexity

Algorithmic complexity, also known as Kolmogorov's complexity, is a concept resulting from the independent works of Ray Solomonoff [4] and Andrei Kolmogorov [5] and the contributions of Gregory Chaitin [6] in 1969. As the symbolic complexity, the algorithmic complexity is also a measure of the amount of information need to describe an object. But differently from the first, in algorithmic complexity the description may use the regularities of the description to code by any means, mechanisms capable of reproducing the developed object description. In other words, this complexity measures the length of the description of the algorithm that reproduces the object symbolic description. To illustrate the idea consider an object which can be described as the following series of numbers:

7, 3, 10, 13, 23, 36, 59, 95, 154, 249, 403, 652, 1055, 1707, 2762, 4469, 7231,
 11700, 18931, 30631, 49562, 80193, 129755, 209948, 339703, 549651,
 889354,1439005, 2328359, 3767364, 6095723, 9863087, 15958810, 25821897,
 41780707, 67602604, 109383311, 176985915, 286369226, 463355141,
 749724367.

The amount of information description, taken as the literal sequence of digits, can be determined by writing down the sequence using only the symbols of a binary base, a then apply Equation (1.1) to obtain a number proportional to the amount of information. But if some property of the sequence is detected, such that it can be synthetized in a shorter string of characters, then it would be possible to wrap the whole sequence of numbers in a lighter information package. In fact those number x in the sequence obey the rules

$$x_i = x_{i-1} + x_{i-2} , x_0 = 7, x_1 = 3 , 0 \leq i \leq 40 .$$

These rules, which resemble the well-known Fibonacci series, clearly express the original sequence of numbers in a shorter string; more effective if the goal is to transmit the message over a media with charges or penalties over the amount of information transmitted.

The algorithmic complexity is then an evaluation of the capability of a language—or an algorithm—to synthetize a system symbolic description, into shorter

symbol-string. Kolmogorov's algorithmic complexity is an unreachable concept because it is not possible to be certain about the very best and most compact way of writing a description. Still some authors as Funes [13], regard it as a “*simple idea with practical and philosophical consequences*”.

2.1.3 Computational complexity

In the context of computing, complexity refers to the resources needed to reach the end of a set of tasks leading to the solution of a problem by way of computational media. Resources used as a measure of complexity can be expressed in time, memory space, information transmission capability, or some combination thereof.

This complexity measure is typically expressed in terms of the resources required by a computerized system, but the analysis deals with the algorithm that consists of logical elements, and not directly with the computer or device where really are the resources that will account for the complexity calculated.

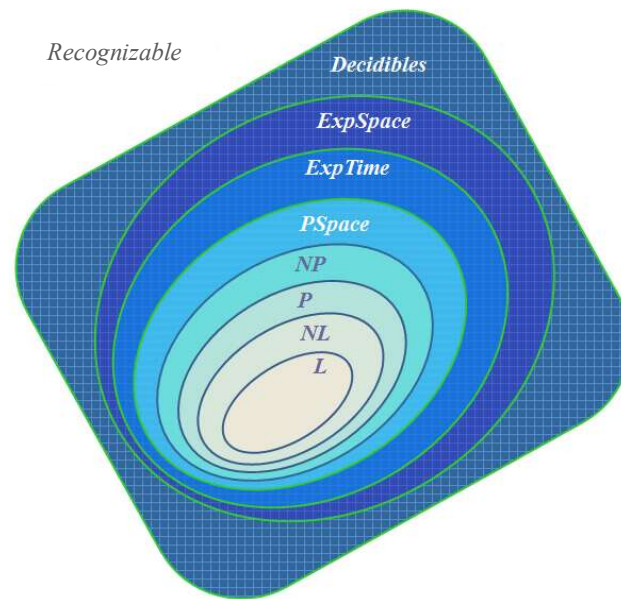


Figure 2.1: Hierarchy of problem classes for computational complexity

Each problem is described by values in one or more dimensions. Some of these dimensions are found to be dominant in the treatment of the problem. To these dimensions we refer to as key dimensions and magnitude as the characteristic value. In the development of methods to study the computational complexity, sets of problems that share some property with respect to the use of resources required for their solution, are grouped. Usually the required resources are calculated on the basis on their growth rate when the problem is large upon one

of its characteristic components. For example, if the problem consists of finding all prime numbers below a certain value n . Then n is the characteristic value of this problem. By monitoring the growing computational resources required by the computer for large values of n , this type of problem can be classified. The problem's resource growth rate is compared to the growth rate for other problems and set in a class where the problems share similar values for their required resources growth rate. Additionally, these problem classes are defined by using some reference functions. The functions commonly used to perform as complexity references are, ordered from lower to higher complexity: logarithmic, linear, polynomial y exponential. Then it is said that a problem is polynomial is its solution requires resources with growth rate equal or lower than that from some polynomials.

Figure 1.2 shows the relationship between different computational complexity classes. The result is a powerful map of sets that summarizes the universe of problems according to our ability to solve them with some effectiveness and economy of time and space.

2.1.4 Effective complexity

At some point researchers began to think of the term complexity as a measure of the internal activity among the system parts to reach the needed equilibrium, or perhaps a pseudo-equilibrium, in order to keep sustainability over some time. This conception of complexity disregards the use of any resource to convey information from a sender to a receiver. It does not pay attention to the communication process. It does, however, need an estimate of the system entropy since it represents an evaluation of the internal order condition.

As entropy calculations are applied to system descriptions, the resulting figure ends up being proportional to the quantity of information of the system's description. But this does not mean this conception of system complexity is directly related to the information needed to describe the system. It only means that symbolic entropy is used to compare the system internal activity with references placed at the extremes of total order —minimal entropy— and total disorder —maximum entropy. The fact entropy also serves measure of information should be taken as an interesting coincidence which explanation is beyond the purposes of this work.

A quantification of this complexity came with the work of Lopez-Ruiz, Mancini and Calbet [1]. They set the complexity value as proportional to the distance of the system entropy to the entropy of a total ordered system, and also proportional to the entropy of a totally disordered system. This leads to the product of disorder (entropy h) times the order (1 minus entropy h). Since values

of normalized entropy go from zero to one, the plain product of $h \cdot (1 - h)$ gets as high as 0.25 when $h = 0.5$. To normalize the complexity obtained after this product, Fernandez, Maldonado and Gershenson [14] proposed a factor of 4. The final estimate of complexity c is

$$c = 4 \cdot h \cdot (1 - h) . \tag{2.2}$$

2.1.5 Networks

Networks are the logical representation of related entities. They are not a measurable quantity. They are not even a quantity. Networks is the name given to a language suited for describing non-regular spaces as those found in the topology of complex systems. The Networks as a field, is closely related with complexity. Both disciplines feed each other and in fact, have grown with parallel rhythms. Networks can be useful to characterize systems. Network properties as node degree distribution, diameter, clustering coefficient, and other network associated indexes can be the basis for measures of complexity of systems.

Some specific branches of the field of networks look very likely to become relevant sources of knowledge when complexity steps from a stationary index to a dynamic index. We are not there yet, but it will arrive. Node preferential Attachment [15,16] and Small World Networks [17] are examples studied hypothesis that teach us about probable ways of complexity grow patterns.

2.2 Common paths of complexity

Different interpretations of complexity have been mentioned above. The difficulty to quantify them can differ in the arithmetic depth employed as well as the method used. For example, symbolic complexity is quantified by figuring out the entropy on the basis of the frequency of occurrence of symbols. Whether symbolic complexity is proportional to entropy is not an issue here. The point is that symbolic complexity is a function of entropy and therefore can be quantify. Computational complexity, on the other hand, is quantified by classifying a problem as belonging to a specific set within a hierarchical structure of sets. This hierarchic is itself is a language capable of quantifying the computational complexity of problem. Finally, algorithmic complexity evaluates the complexity of algorithms more than they do it for actual system descriptions.

Even more elaborated forms of complexity have appeared. They are born as sketchy ideas promising useful ways of understanding reality. I would classify within this category Murray Gell-Mann's '*potential complexity*' [11] and Bar-Yam's '*multi-scale complexity*' [12] based on what he calls complexity profiles –

other refer to this representation as information profiles-. But an exact method for quantifying them must wait until a language for their description organizes itself and settles down into a commonly accepted method.

Studies based on the frequency of use of certain words, according to the subject they belong to and their grammatical function, have been made since long time ago. It has always been a very tempting idea to discover through writings, the personality, abilities, talents, preferences, and even the intelligence of their authors. In psychology, for example, in 1896 and 1891 Joseph Jastrow [18,19] contrasted the habits of men and women on how often they use words according to their class. Jastrow's motivation was confined to the psychological analysis that could be drawn from the language of the people and not the information content of messages.

More than half a century after Jastrow's studies, George Kinsley Zipf's works [2] appeared. His results, expressed with elegant simplicity in a very short equation, still resonate today in our understanding of aspects of complexity whose scope goes beyond languages and texts. Zipf's Law states that for a sufficiently long text, the frequency with which a word is inversely proportional to its rank (position with respect to other words according to the frequency). Using p_r to denote the probability of finding a word from the rank r and k a constant that depends on the number of words that make up the language, Zipf's Law is written as:

$$p_r = \frac{k}{r} \quad . \quad (2.3a)$$

Zipf's Law is observed in many data collections. Distributions sorted by population, religions by number of parishioners, companies by number of employees, non-English languages and many other sorted data sets, closely respect the generalized Zipf law, which incorporates the exponent g for the ranking term. The expression becomes:

$$p_r = \frac{k}{r^g} \quad . \quad (2.3b)$$

2.3 Languages as complex systems

Whichever the complexity concept is adopted, quantities like Information, entropy, self-organization, diversity, equilibrium, excess entropy, are well defined and methods for their calculations are established. Despite differences among many Complexity concepts, they are actually not conflictive. They simply use the same term "complexity" to refer different -sometimes only slightly different- aspects of systems. We, therefore pay little attention to syntactical meaning of complexity. Instead, we prefer to study the connections between entropy and other diversity measures with the way we perceive messages. In this regard the

concept of language, as an entity that organizes information, is the common framework where all these ways of looking at complexity lays. The study is approached as the analysis of several experiments performed over expressions of different types of languages. Thus, languages are considered as instances of complex systems.

“‘Meow’ means ‘woof’ in cat.”
George Carlin

“The limits of my language means the limits of my world.”
Ludwig Wittgenstein

Chapter III

Complexity measurement of natural and artificial languages

The study of symbol frequency distribution for English was initially addressed by Zipf [2] in 1949 and Heaps during the 70s [20], giving rise to Zipf's and Herdan-Heaps' laws respectively (frequently referred to as Heaps' law). Zipf suggested that the scale free shape of the word frequency distribution, typically found for English long texts, derives from his *Principle of Least Effort*. As in many other large scale phenomena, the origin of the tendency of natural languages to organize around scale free structures, remains controversial [21] and a plentiful source of hypothesis and comparisons with other 'laws of nature' [1,22,23]. The relationship between both Laws has been studied [24] and their validity for various natural alphabetic languages tested [25–27]. Yet, a generally accepted mechanism to explain this behavior is still lacking, as Zipf's and Heaps' laws have been traditionally applied only to probabilistic consequences of grammar structure and language size.

Language grammar has been addressed in the study of basic grammar rules and the mechanisms to buildup English phrases, initiated by Chomsky [28] in the late 50's. Later Jackendoff [29] developed the X-bar theory, fostering the idea of underlying effects driving human communication processes to produce grammar properties common to all natural languages. Yet clear descriptions of the fundamental sources of such a behavior, remains a matter of discussion, perhaps because it is a problem too complex to be completely understood employing only theoretical methods. Important differences arise from the nature and content of message than is transmitted. Yet, languages viewed as describing tools, have their own capacity to deliver a message more effectively

or more efficiently. Therefore languages are susceptible of being evaluated. As George Markowsky [30] expressed:

“An important point to stress here, ... , is that the algorithmic complexity¹ of an object depends very much on the language in which the object is described! We can make the complexity of any particular object as small or as large as we choose by picking the appropriate language or by modifying an existing language.”

In this Chapter, we treat languages as complex systems made of large sets of symbols, and following other authors suggestion [2, 3]. We compare messages expressed in natural and artificial languages using metrics developed to quantify complexity. Our comparison is based on measurements of message symbol diversity, entropy and symbol frequency distributions. Zipf's distribution profiles and Heaps' functions are identified for different messages samples. We evaluate the impact of these measures over emergence, self-organization and complexity of messages expressed in natural and artificial languages.

Our strategy is to evaluate a wide range of texts for each language studied, including text pieces from a variety of writers distributed over a timespan of more than 200 years. All texts were recorded in a computer file directory and analyzed with purposely developed software called *MoNet* [31] (see section 3.1.10), as explained in Sections 3.1.1 to 3.1.6.

3.1 Methods

We compared three aspects of English, Spanish and artificial languages: symbol diversity D , entropy h , and the symbol frequency distribution f . For the available measures of diversity and information, we follow Gershenson and Fernandez [10] to evaluate emergence and self-organization for natural and artificial languages. For complexity, we use the definition of Lopez-Ruiz et al. [1] which sees complexity as a balance between chaotic and stable regimes. All computations are directed to the symbolic analysis. We have made an effort to recognize slight differences in the way words or punctuation signs are presented in a text. Nevertheless our analysis disregards any syntactical meaning.

¹ The concept of Algorithmic Complexity is not rigorously the same concept of Complexity, Emergence or Information we apply in this study. Still, Markowsky's point of view justifies perfectly our study.

3.1.1 Text length L and symbolic diversity d

The length of a text L is measured as the total number of symbols or words used and the diversity D as the number of different symbols that appear in the text. We define the specific diversity d as the ratio of diversity D and length L , that is

$$d = \text{specific diversity} = D/L. \quad (3.1)$$

In this study symbols are considered at the scale of words. Here a word is a considered as a sequence of characters delimited by some specific characters such as a blank space (see section 3.1.9). Most recognized symbols were natural and artificial language words. Nevertheless some single character symbols, such as periods and commas, appeared by themselves with complete meaning and function and therefore playing a role comparable to that of normal words.

3.1.2 Entropy h

Entropy calculations are based on Shannon's information [3], which is equivalent to Boltzmann-Gibbs entropy. Message information is estimated by the entropy equation is based on the probability of appearance of symbols within the message. Symbols (words) are treated all with the same weight, ignoring any information that might be associated to meanings, length or context. Shannon's entropy expression for a text with a symbol probability distribution $P(p_i)$ is:

$$h(p_i) = - \sum_{i=1}^2 p_i \log_2 p_i. \quad (3.2)$$

Shannon was interested in evaluating the amount of information and its transmission processes; therefore his entropy expression was presented for a binary alphabet formed by the symbols '0' and '1'. Entropy measurement in this study is at the scale of words, where each word is a symbol, extending the original Shannon's expression for a D -symbol alphabet:

$$h(f_r) = - \sum_{r=1}^D \frac{f_r}{L} \log_D \frac{f_r}{L}, \quad (3.3)$$

where the term p_i have been replaced with its equivalent in terms of the symbol frequency distribution $F(f_r)$ and the text length L measured as the total number of symbols. The values for the symbol frequency distribution $F(f_r)$ are ordered on r , the symbol rank place ordered by their number of appearances in the text. Since there are D different symbols, r takes integer values from 1 to D . Notice the base of the logarithm is the diversity D and hence h is bounded between zero and one.

3.1.3 Emergence e

As a system description is based on different scales —the number of different symbols used— the quantity of information of the description varies. Emergence measures the variation of the quantity of information needed to describe a system as the scale of the description varies, thus, emergence can be seen as a profile of quantity of information for a range of system scales. Therefore we express emergence e as a function of the quantity of information respect to the description length L (total number of symbols) and the specific symbol diversity d . This is given Shannon's information (3.3), so we have:

$$e(F(f_r)) = h(F(f_r)). \quad (3.4)$$

3.1.4 Self-organization s

The self-organization of a system can be seen as the capacity to spontaneously limit the tendency of its components to fill the system space —symbols in our case— in a homogenous, totally random distributed fashion. Since entropy reaches a maximum when the system components are homogeneously randomly dispersed, self-organization s is measured as the difference of the maximum entropy level $h_{max} = 1$, and the actual system entropy [14].

$$s(F(f_r)) = h_{max} - h(F(f_r)) = 1 - e(F(f_r)). \quad (3.5)$$

3.1.5 Complexity c

Message entropy calculations are based on Shannon's expression [4]. Message information is estimated by the entropy equation based on the probability of appearance of symbols within the message. Symbols (words) have all the same weight here, ignoring putative differences in information associated to the word's meanings, length or context. We used the complexity definition proposed by López-Ruiz et al. [9], and its quantifying expression proposed by Fernández et al. [14].

$$c(F(f_r)) = 4 \cdot e(F(f_r)) \cdot s(F(f_r)) = 4 \cdot h(F(f_r)) \cdot [1 - h(F(f_r))]. \quad (3.6)$$

In this definition, complexity is high when there is a balance between emergence (entropy, chaos) and self-organization (order). If either is maximal, then complexity is minimal. Equations (3.4-3.6) depend on Shannon's information and can be reduced to it. Still, it is explanatory to study each of these separately, as it will be seen in our results below, emergency e is a measure of “disorder”, entropy s measures order and complexity c their balance.

3.1.6 Symbol frequency distribution f

For any message or text the number of words in a rank segment $[a, b]$ may computed as:

$$L_{a,b} = \sum_{r=a}^b f_r, \quad (3.7)$$

where a and b are respectively the start and the end of the segment where symbol were ranked. For any segment, $a = 1$ and $b = D$.

Zipf's law states that any sufficiently long English text will behave according to the following rule [1] [5]:

$$f(r) = \frac{f_a}{(r - a)^g}, \quad (3.8)$$

where r is the ranking by number of appearances of a symbol, $f(r)$ a function that retrieves the numbers of appearances of word ranked as r , f_a the number of appearances of the first ranked word within the segment considered, and g a positive real exponent.

For any message, we define Zipf's reference $Z_{a,b}$ as the total number of symbol appearances in the ranking segment $[a, b]$ assuming that it follows Zipf's Law. Therefore $Z_{a,b}$ is

$$Z_{a,b} = \sum_{r=a}^b f_r = \sum_{r=a}^b \frac{f_a}{r^g}. \quad (3.9)$$

Equation (3.8) allows us to determine the Zipf's reference Z for any segment within the symbol rank dominion. We computed versions of Zipf's reference Z for the complete message, specifically named $Z_{1,D}$, and for the tail of the message frequency distribution (see Figure 3.1), named $Z_{\theta,D}$. The sub index θ is used to indicate the ranking position r_θ where the head-tail transition occurs.

Head-tail transition location can be a difficult parameter to set and is often considered to be among a range of possibilities. We used the following definition: For a discrete symbol ranked frequency or probability distribution, the region of the lowest frequency of ranked symbols starts where the symbols with a unique frequency (or probability = 1) end. Figure 3.1 illustrates an example of symbol frequency profile. The point signaled with the arrow corresponds to the 20th rank position and has 7 occurrences, and no other symbol shares the same number of appearances. At that point we define the start of the tail which includes the distribution domain shadowed in yellow in the figure.

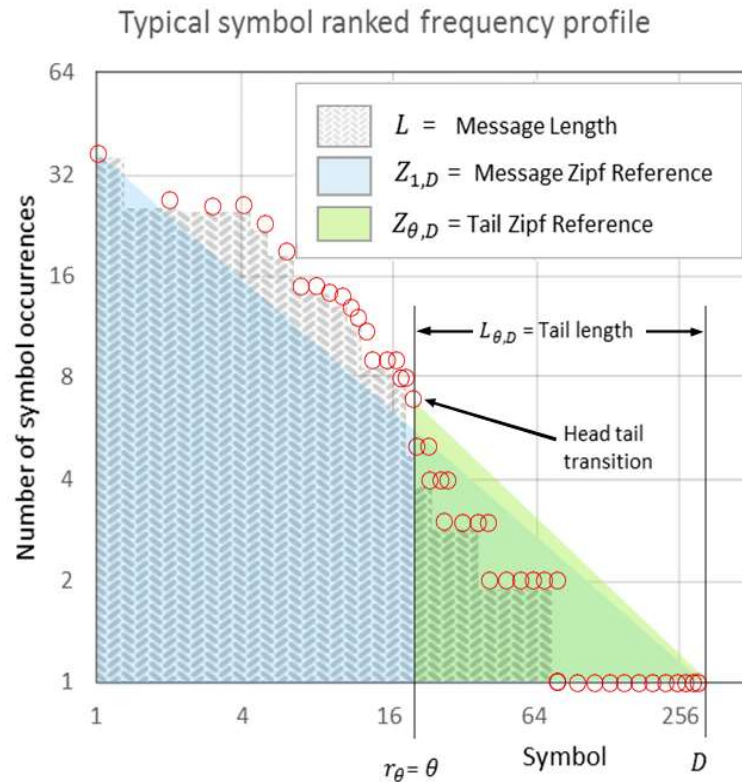


Figure 3.1: Typical symbol ranked profile. Red dots indicate the number of occurrences and the ranking position of the symbols of a given text. Message Zipf's and tail Zipf's references are the blue and yellow shadowed areas respectively.

3.1.7 Zipf's deviation J for a ranked distribution

The complete message Zipf's reference is determined by Expression (3.8). The corresponding Zipf's deviations $J_{1,D}$ from a Zipfian distribution and the deviation of its tail $J_{\theta,D}$ are

$$J_{1,D} = (L - Z_{1,D}) / Z_{1,D} \quad (3.10a)$$

$$J_{\theta,D} = (L_{\theta,D} - Z_{\theta,D}) / Z_{\theta,D} \quad (3.10b)$$

Identifying the starting point for the tail of each message or code profile is a search intensive task. We included in the software *MoNet*, the capability of locating within a frequency profile the points with properties characterizing the start of the tail, and to split messages and codes in heads and tails. Once the tail starting rank r_{θ} is determined, Zipf's tail deviation was obtained by applying Equations (3.10a) and (3.10b).

3.1.8 Message selection

We built text libraries containing consisting of large text fragments, obtained from English and Spanish speeches, segments of stories and novels, and computer codes written in high level programming languages (C, C#, Basic, Matlab, Java, HTML and PHP). The program then produced descriptive indices and attributes for each of these. Each message could be analyzed as an individual object or as a part of a collective group of objects.

Natural language message selection

Natural language messages were collected from historic speeches available in on the web as texts expressed in English or Spanish. Natural language texts include speeches from politicians, human rights defenders and literature Nobel Laureates. The language used to write the original speech was not a selection criterion. There are speeches in our selection originally written in English, Spanish, French, Russian, Italian, German, Arabic, Portuguese, Chinese and Japanese. Translated speeches and texts are indicated as such, providing data for studying translations. Novel fragments were authored in English or Spanish by popular writers and by some Nobel laureates in literature. We collected 156 texts in English and 158 in Spanish. The shortest speech was 87 words long, whereas the longest speech contained more than 20000 words.

Artificial language message selection

We included 49 computer codes devoted to perform recognizable tasks. Artificial text lengths go from a C# code which generates Fibonacci numbers with just 62 symbols, to computer logs with more than 160000 symbols. This selection of artificial texts include codes written in C, C#, Basic, Java, *MatLab*, *HTML* and *PHP*. The Table in Appendix B gives details of codes and their fragments used here.

3.1.9 Symbol treatment

Special treatment of certain character strings or symbols were considered as follows:

Word: A word is any character string isolated by the characters 'space' or 'line return'. The word is the symbolic unit.

Space: The space works as a delimiter for symbols or words.

Line return or line feed: Is a delimiter for paragraphs.

Punctuation signs: Any sign is considered as a complete independent symbol. In natural languages, the punctuation signs have specific meaning that, with very few exceptions, are not sensitive to other surrounding characters. When located next to numeric characters, if a punctuation sign appeared attached to another symbol, the sign was handled as being separated by the space character to keep it as a single symbol. This rule provides a coherent solution to the very frequent case where words appear attached to punctuation symbols.

Numbers: For natural languages, a digitally written number might be a unique sequence of characters. Numbers express quantities and work as adjectives or modifiers of an idea. All numbers in a natural language message are then considered as different symbols.

Capital letters: Words are case sensitive. In English and Spanish a word with its first letter written with a capital letter, refers to a specific name. Therefore, a name appropriately written with a first capital letter is different from the same character sequence written with all letters in lower case. But when the word starting with capital letter comes after a period sign, we assume it is a common lower case word, unless other appearances of the same word indicates it certainly is a proper name that should keep its first capital letter.

For Spanish messages.

Accents: in Spanish, vowels are sometimes marked with an accent over it to indicate where the sound stress or emphasis should be. Rules to indicate when the accent mark should be present and when it shouldn't, are easy to apply and are part of what any Spanish speaker should know from elementary school. Forgetting accent marks when they should appear is associated with poor writing abilities; it is unacceptable in any serious literary work. We consider that any accented word is different, and has some different meaning, from the same character sequence without accents.

For artificial languages (computer code).

Comments: in artificial languages comments do not affect any action of the interpreter or compiler. Additionally, comments are intended to convey ideas to the human programmer, administrator or maintenance personnel, hence most comments are written in phrases dominated by natural languages. Comments were thus excluded from any code analyzed.

Computer messages: Most computer codes rely on the possibility of informing the user or operator about execution parameters. This information is normally expressed in different languages to that of the code. Computer message contained in a code were converted to a single word by extracting all spaces.

Numbers: Differently from natural languages, in artificial languages sequences of digits may represent variable names or memory addresses, which are objects with different meaning. In artificial languages, any difference in a digit is considered to result in a different word.

Capital letters: We considered artificial language symbols as case sensitive.

Variables: When in different parts of the code, two or more variable names were presented as the same symbol or characters string, but we know that sometimes they could have a totally different meaning since they could be pointing to a different memory address. This may introduce some deviation in the results.

3.1.10 Software

Two software programs were developed to analyze the texts. First, we built a file directory structure containing, and classifying the messages each with its inherent and invariant text-object properties. We refer to the file directory as the library. The second software program, called *MoNet* [6], manage the library and produced the data for our study.

Library: The library holds descriptions of each existing text-object with its attribute values. The scope of each object description can be adjusted adding attributes or even modifying their data representation nature and dimensionality. We built a text library containing hundreds of these text-objects. Libraries can be updated by deleting or adding text-objects.

MoNet: Is a bundle of scripts, interpretations, programs and visual interfaces designed to analyze complex systems descriptions at different scales of observation. *MoNet* describes a system as a collection of objects and object families connected by hierarchical and functional relationships. *MoNet* can treat every text included in a library as well as the library itself, offering results for text-objects as independent elements or as groups. For every component of the system modeled, descriptions at different scales can co-exist. Individual objects can be selected combining logical conditions based on properties or attribute values.

3.2 Results

3.2.1 Diversity for natural and artificial languages

Figure 3.2 shows how diversity varies with the message length in texts written in English, Spanish and Computer Code. Diversity increases as messages grow in length, but there seems to be an upper bound of diversity for each message length. For English this upper bound is slightly lower than for Spanish. As message length increases, English also shows a wider dispersion toward lower diversities of words. Artificial messages represented by computer code showed a much lower diversity than the natural languages.

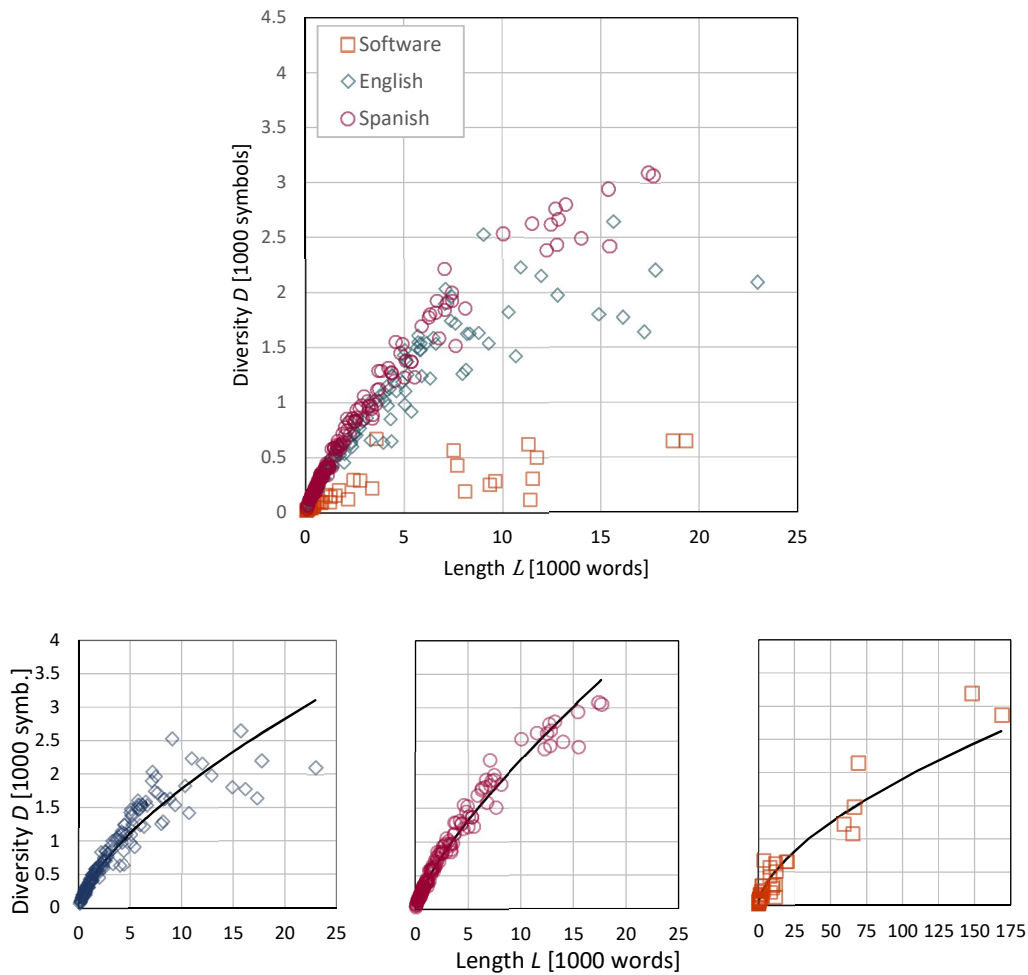


Figure 3.2: Diversity for messages expressed in English, Spanish and Computer Code. Lower row presents fit dots (black) for messages expressed in English (left), Spanish (center) and Software (right).

The regression models of Heaps' Law [7] for message diversities and message length are:

$$\text{English:} \quad D = 3.766 \cdot L^{0.67} \quad (3.11a)$$

$$\text{Spanish:} \quad D = 2.3 \cdot L^{0.75} \quad (3.11b)$$

$$\text{Software:} \quad D = 2.252 \cdot L^{0.61} \quad (3.11c)$$

3.2.2 Entropy for natural and artificial languages

Figure 3.3 shows entropy h values for texts expressed in natural languages and computer code programs as a function of specific diversity d (see section 3.1.1). Extreme values of entropy are the same for messages expressed in all languages; entropy drops down to zero when diversity decreases to zero and tends to a maximum value of 1 as specific diversity approaches 1. For artificial messages entropy is dispersed over a wider range of values, perhaps as a consequence of the many different computer languages included in this work's sample.

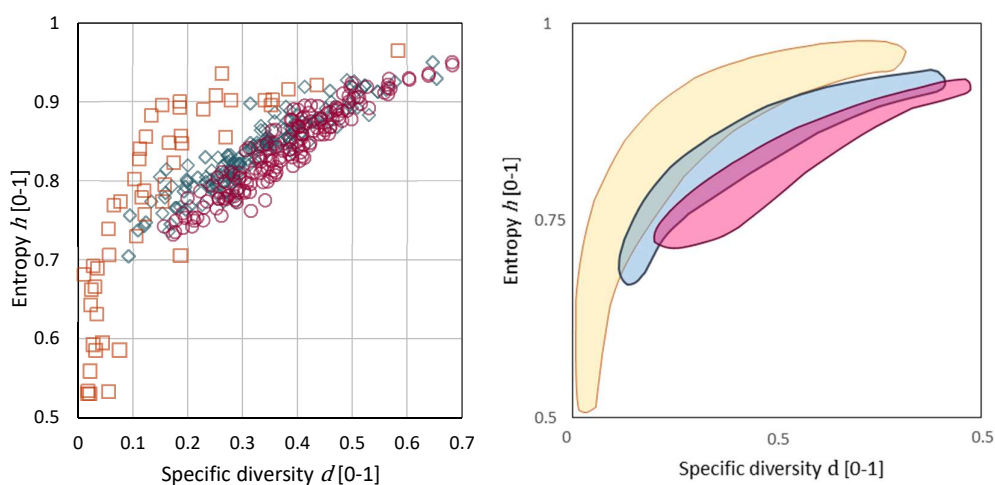


Figure 3.3: Messages entropy vs. specific diversity for English (blue rhomboids), Spanish (red circles) and Computer Code (orange squares). On the left graph each dot represents a message. The right graph shows the area where most messages lie upon its corresponding language.

Natural languages show less dispersion in entropy levels, nevertheless differences among languages show up in the areas they cover over the plane of entropy-diversity with few overlapping shared areas over that space. See Figure 3.3.

The entropy expression shown in Eq. (3.3) is a function with $D - 1$ degrees of freedom; there are $D - 1$ different ways of varying the variable F that affect the resulting value of entropy h . Nevertheless, when specific diversity is at extreme

values $d = 0$ and $d = 1$, the distribution F becomes homogenous and function $h(F)$ adopts the following predictable behavior.

$$1: \quad h(F | d \rightarrow 0) = 0 \quad . \quad (3.12a)$$

$$2: \quad h(F | d \rightarrow 1) = 1 \quad . \quad (3.12b)$$

Having these extreme conditions for $h(F)$, we propose a real function $h(d)$ to characterize the entropy distribution of a language over the range of specific diversity. The dispersion of the points is due to the fact that none of the texts obeys perfectly a Zipf's law, yet each language tends to fill a particular area of the space entropy-specific diversity.

To model the curves along the core of these clusters of dots, that is entropy as a function of specific diversity, we refer to the so called Lorenz curves [32] which can be used to describe the fraction of edges W of a scale-free network with one or two ends connected to a node which belongs to the fraction P of the nodes with highest degree [23]. The family of Lorenz curves is expressed by

$$W = P^{(\alpha-2)/(\alpha-1)} \quad . \quad (3.13)$$

Now consider the network associated to a text where the nodes represent words or symbols and the edges represent the relation between consecutive words. In a network like this, all nodes, except those corresponding to the first and the last words, will have a degree of connectivity that doubles the number of appearances of the represented word.

Thus, the resulting ranked node degree distribution will be analogous to a Zipf's distribution and therefore, the network as defined, will have a scale-free structure. On the other hand, entropy can be interpreted as the cumulative uncertainties that every symbol adds or subtracts from the total uncertainty or entropy. Viewing entropy h of a ranked frequency distribution as the cumulative uncertainty after adding up the contributions of the D most frequent symbols, we should expect this entropy h to have a scale-free behavior with respect to changes D . After the analogies between these conditions and those needed to expect a behavior like the Lorenz curves dictate, we propose the use of the one-parameter Expression (3.13) to describe any language's entropy as a function of d and the parameter α . So that:

$$h = \left(\frac{D}{L}\right)^{(\alpha-2)/(\alpha-1)} = d^{(\alpha-2)/(\alpha-1)} \quad . \quad (3.14)$$

Figure 3.4 compares the data using the entropy model for the languages studied. Values of α were obtained to minimize square errors between the entropy model and the experimental results obtained from each text of the library. Numerical results were $\alpha = 2.123$ for English, $\alpha = 2.178$ for Spanish and $\alpha = 2.1$ for artificial. The figure shows a much wider range of entropy values for artificial languages compared to the natural languages studied.

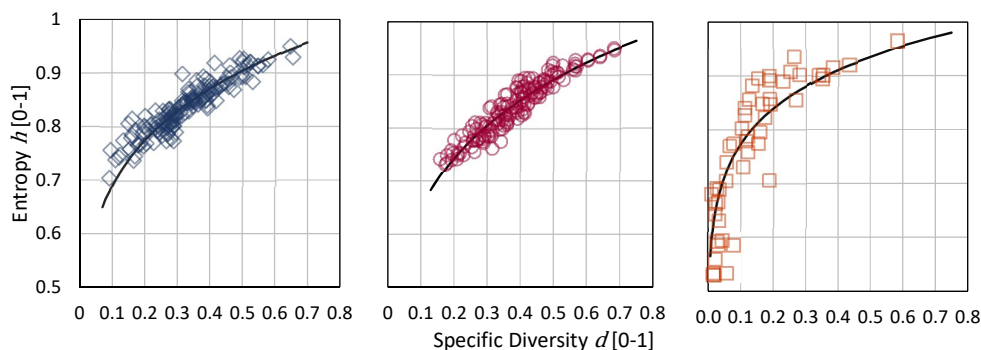


Figure 3.4: Messages entropy vs. specific diversity for English (left), Spanish (Center) and Computer Code (right).

Equations (3.15a), (3.15b), and (3.15c) present specific cases of function $h(d)$ for each language studied:

$$\text{English:} \quad h = d^{0.1511} . \quad (3.15a)$$

$$\text{Spanish:} \quad h = d^{0.1756} . \quad (3.15b)$$

$$\text{Software:} \quad h = d^{0.09091} . \quad (3.15c)$$

3.2.3 Emergence, self-organization and complexity

Starting with functions for entropy, obtaining expressions for emergence, self-organization and complexity is straightforward using results of Equations (3.15a), (3.15b) and (3.15c) with Equations (3.4), (3.5) and (3.6). Figure 3.5 illustrates these results. To obtain expressions of emergence, self-organization as functions of the message length L , we combined Equations (3.15a), (3.15b) and (3.15c) with (3.11a), (3.11b) and (3.11c) respectively. See the results in Figure 3.6.

For all languages, emergence increases with specific diversity and decreases with length. Self-organization follows opposite tendencies, decreasing with specific diversity and increasing with length. Complexity is maximal for low specific diversities and then decreases, although much less for natural languages. Complexity increases with length for all languages. The most conspicuous result here is that artificial languages show a different pattern in complexity depending on specific diversity, as the maximum complexity for

III. Complexity measurement of natural and artificial languages

artificial languages is close to zero and then decreases faster than natural languages. This might reflect fundamental differences in organizing the symbols (grammar) between both types of languages.

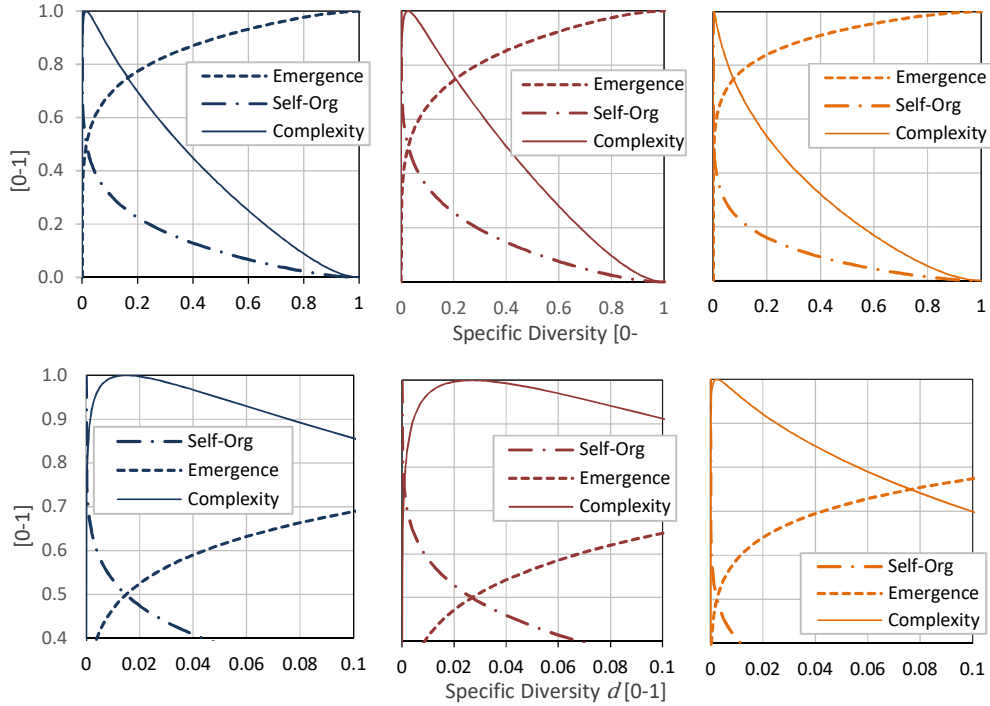


Figure 3.5: Emergence, self-organization and complexity for English (left), Spanish (Center) and Computer Code (right). Vertical axis is dimensionless [0-1]. Graphs placed on the lower row correspond to the detail very near the value zero for horizontal axis. These plots are based on Equations (3.4), (3.5) and (3.6) combined with Equations (3.15a), (3.15b) and 3.15c).

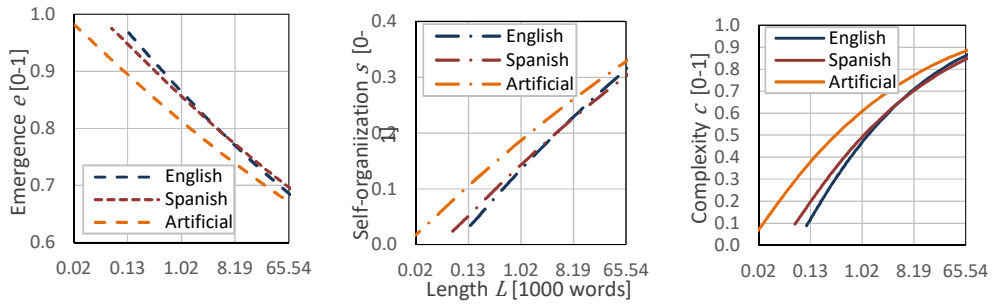
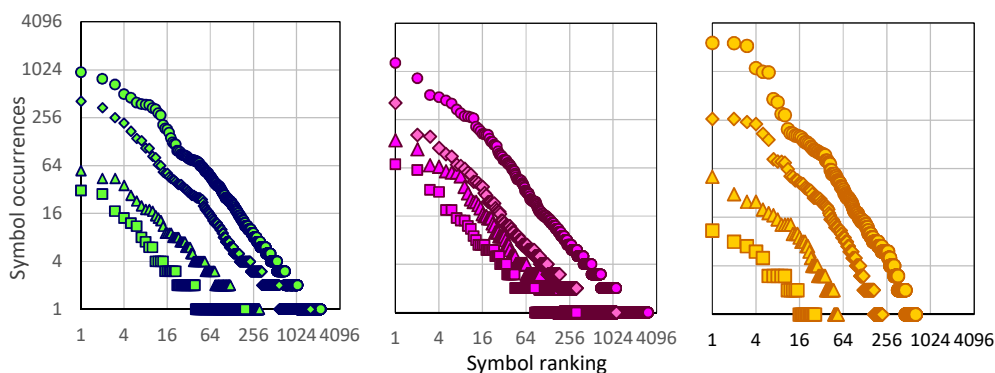


Figure 3.6: Emergence (left), self-organization (center) and complexity (right) for English, Spanish and Computer Code. Vertical axis are dimensionless [0-1]. These plots are based on Equations (3.4), (3.5) and (3.6) combined with Equations (3.11a), (3.11b) and (3.11c).

3.2.4 Symbol frequency distributions

Profile of symbol frequency distributions were inspected in two ways: first by a qualitative analysis of their shapes and second by characterizing each profile with its area deviation J with respect to a Zipfs distributed profile.

A sample of symbol frequency distributions profiles for the considered languages is represented in Figure 3.7. Each sequence of markers belongs to a message and each marker corresponds to a word or symbol within the message. While no important differences are observed among messages profiles expressed in the same language, a noticeable tendency to a faster decreasing frequency profile appears for messages expressed in artificial languages, perhaps a consequence of the limited number of symbols these types of languages have.



English:	square - 1945.BS.Eng.GabrielaMistral	rhombus - 1950.NL.Eng.BertrandRussell	triangle - 1921.MarieCurie	circle - 1890.RusselConwell
Spanish:	square - 1936.DoloresIbarruri	rhombus - JoseSaramago.Valencia,	triangle - 1982.Gabriel G. Márquez	circle - C.J.Cela.LaColmena.Cap1
Artificial:	square - FibonacciNumbers.CSharp	rhombus - Sociodynamica.Module3	triangle - QuickSort.CSharp	circle - WebSite.Inmogal.php

Figure 3.7: Ranked symbol frequency distribution for English (left), Spanish (center) and Computer Code (right). A sample of three or four messages for each language is shown.

By building these frequency profiles, we could obtain a list of the most used words in English and Spanish. An equivalent list for artificial languages is also obtainable; however it is difficult to interpret due to the diversity of programming languages used in our artificial text sample.

III. Complexity measurement of natural and artificial languages

Table 3.1: Most frequently used symbols in English and Spanish. Open-class words are shown with italic characters. Closed-class word are shown with normal characters. Top ranked open-class words are shown with italic-bold letters.

Natural languages symbol frequency					
English. Total symbols = 23398			Spanish. Total Symbols = 33249		
Rank	Word (Symbol)	Use [%]	Rank	Word (Symbol)	Use [%]
1	the	5.51921	1	,	5.7697
2	,	4.96449	2	de	5.0643
3	.	4.58479	3	.	3.8664
4	of	2.96836	4	la	3.5446
5	and	2.89258	5	que	3.0410
6	to	2.39816	6	y	2.8992
7	a	1.71795	7	el	2.3789
8	in	1.63451	8	en	2.0957
9	that	1.42234	9	a	1.9270
10	i	1.33711	10	los	1.5953
11	is	1.29327	11	no	1.1690
12	it	1.09772	12	las	0.9659
13	we	1.09103	13	un	0.9562
14	not	0.79216	14	se	0.9486
15	"	0.78874	15	con	0.8530
16	for	0.73284	16	del	0.8395
17	he	0.70253	17	por	0.7923
18	have	0.70204	18	una	0.7836
19	was	0.63881	19	para	0.6962
20	be	0.62708	20	es	0.6939
21	this	0.55440	21	-	0.6241
22	as	0.54185	22	lo	0.6229
23	you	0.53549	23	su	0.5637
24	are	0.53370	24	al	0.4811
25	with	0.52637	25	más	0.4503
26	they	0.50694	26	como	0.4330
...
58	<i>man</i>	0.24761	58	<i>pueblo</i>	0.1435
...	59	<i>mundo</i>	0.1408
62	<i>people</i>	0.23883	60	sobre	0.1344
...
71	<i>world</i>	0.17423	67	<i>vida</i>	0.1256
...
500...	<i>indeed...</i>	0.01867...	500...	<i>poeta...</i>	0.01749...
...8000	<i>...yard</i>	...0.000732	...7339	<i>...flujo</i>	...0.000843
8002 - 9920	<i>adapt - vitiated</i>	0.00055	7340 - 8841	<i>funda...insurgimos</i>	0.000843
9923 - 13505	<i>actress - Zemindars</i>	0.00037	8842 - 11736	<i>adictos...zumbido</i>	0.000632
13506 - 23398	<i>Aaron-Zulu</i>	0.00018	11737 - 15622	<i>abastecimientos... Zelli</i>	0.000419
			15783 - 33249	<i>abanderado ... Xavier</i>	0.000209

Table 3.1 shows statistics about the use of symbols for English and Spanish. Table 3.1 was constructed overlapping symbol frequency profiles of English and Spanish messages contained in our working library. After these calculations, two frequency profiles (probability distributions) were obtained: one for English, the other for Spanish. The first 25 rows of Table 3.1 correspond to the 25 most used

symbols. After this high ranked symbols, rows in Table 3.1 show groups of symbols sharing ranges with the same or approximate percentage of use. In accordance with our definition of tail form this study, head-tail transition occurs at rankings 40 and 35 for English and Spanish respectively.

Joining the text messages in three sets, according to the language they are written with, we obtained an approximation of the symbol frequency profiles for the 'active' fraction of the languages studied (see discussion). Figure 3.7 shows these profiles. Natural languages exhibit a wide range of ranks where the symbol frequency decays with an approximately constant slope g , sustaining Zipf's law for English and extending its validity to Spanish, at least up to certain range of the symbol rank dominion.

Even though we included many programming languages and artificial code as if they were all part of a unique language, which they are not, artificial languages do not show a range where we can consider slope g a constant, evidencing the fact that artificial languages are much smaller than natural ones. The values of exponent g were calculated for the three profile tails and included in Figure 3.7; profile slopes are all negative but g values are shown positive to be consistent with Equation (3.8).

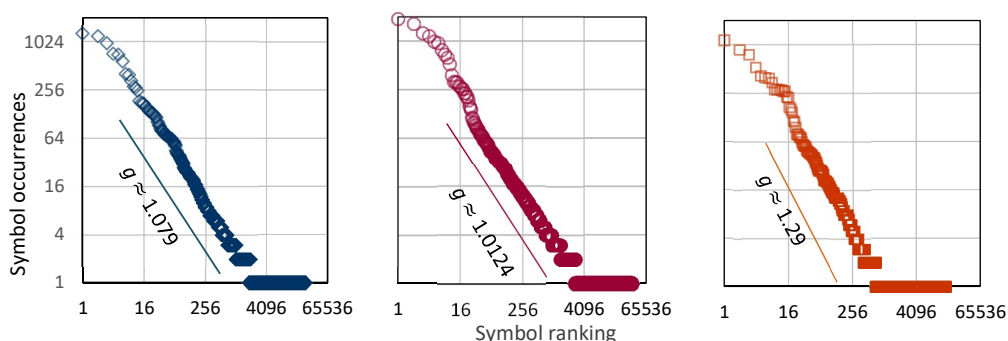


Figure 3.8: Ranked symbol frequency distribution for English (left), Spanish (center) and artificial languages (right).

Notice that Spanish has, among the languages studied here, the smallest tail slope, meaning the heaviest tail; an indication of the variety of words included in all the Spanish messages. At the other end of our sample, artificial languages present the fastest decaying slope and the most limited number of symbols. Direct measurement of differences between profile shapes is not straight forward. We converted the symbol frequency distributions into probability distributions and graph their corresponding CDF (cumulative function distribution) shown in Figure 3.8.

III. Complexity measurement of natural and artificial languages

As expected, artificial languages' CDF grow faster than the others; the five hundred most frequently used symbols are enough to comprise almost 90% of all symbols included in our list of more than 13000 artificial symbols. The first 500 words cover 74% of the 23398 English words included in our library and 70% for the 33249-word Spanish library. The profile heads also reflect some differences between languages. In spite of the general faster growing English's CDF as compared with Spanish, the latter's CDF is higher up to symbol ranked about 56, where the two curves cross. This Spanish faster growing CDF within the head region implies a more intensive use of the close-words group and consequently the tendency of a more structured use of this particular language.



Figure 3.9: Cumulative distribution function (CDF) of symbols ranked by frequency. Horizontal axis is scaled to show the curves for the 4096 most frequently used words for English, Spanish and Artificial language. Note the logarithmic scale in horizontal axis.

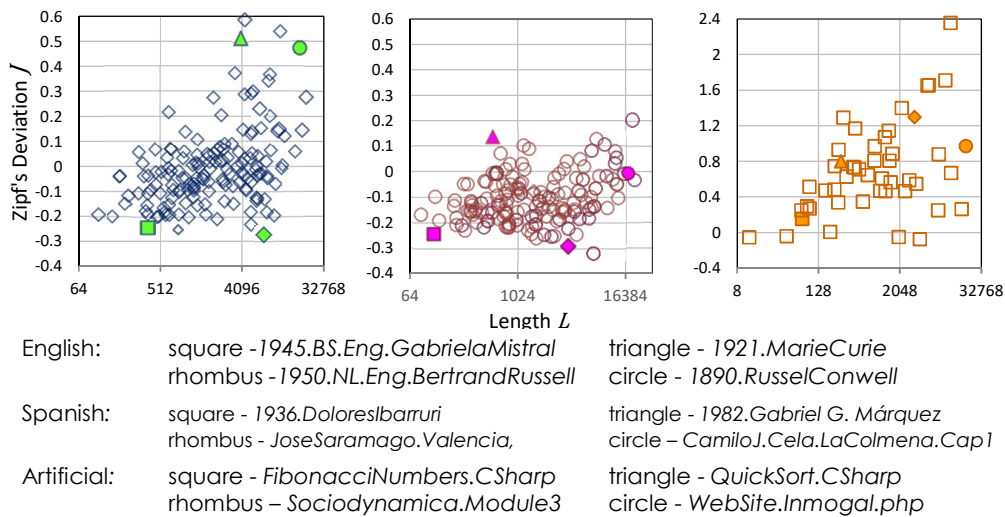


Figure 3.10: Zipf's deviation $J_{1,D}$ of symbol ranked frequency distributions depending on text length L . English (left), Spanish (center) and Computer Code (right).

Table 3.2: Zipf's Deviation $J_{1,D}$ and its correlation with length L for English, Spanish and artificial messages.

Zipfs' deviation $J_{1,D}$ for natural and artificial languages				
	n	J 1, D average	J 1, D Std. Dev.	Correlation J _{1,D} : L
English	156	0.0045	0.1719	0.560
Spanish	158	-0.1074	0.0943	0.351
Computer Code	49	0.6944	0.4961	0.102
t-test	n1 - n2	p-value		
English - Spanish	156 - 158	6.58E-12		
Natural - Software	314 - 49	9.47E-64		

3.2.4.1 Zipf's deviation $J_{1,D}$ for ranked distribution

We computed Zipf's deviations $J_{1,D}$ for natural and artificial languages. Figure 3.10 shows the result of these calculation on the plane Zipf's deviation $J_{1,D}$ vs. Length L . Dependence between Zipf's deviation $J_{1,D}$ and Length L was evaluated with standard deviation and correlations.

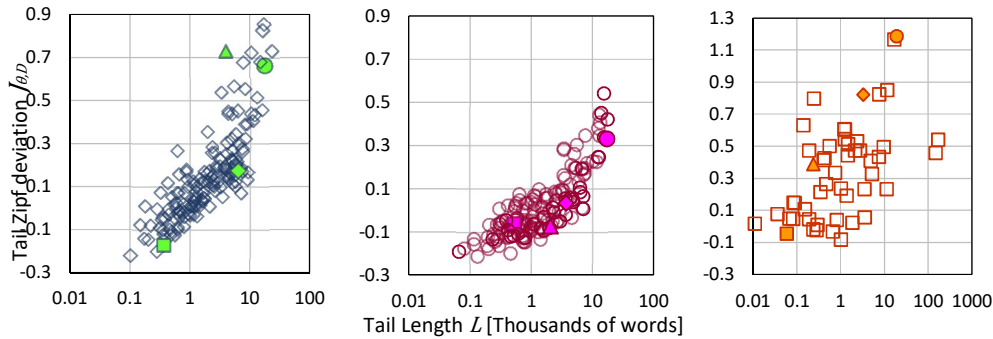
Two additional tests were performed with Student-t distributions to compare the Zipf's deviations $J_{1,D}$. The first tests the hypothesis of English and Spanish Zipf's distribution being the same. The second tests the hypothesis for natural and artificial languages to be the same. Results for all tests show that p-values are very small indicating that Zipfs deviation differed statistically in very significant ways between the three different languages studied. Table 3.2 summarizes these results.

3.2.4.2 Tail Zipf's deviation $J_{\theta,D}$ for ranked tail distributions

Zipf's deviation was also inspected for the tails of the ranked frequency distributions as described in Section 3.1.6. This evaluation provides some further understanding of the tails shapes and relates some tendencies to other variables associated to the messages and the languages.

Figure 3.11 shows the Zipf's deviation $J_{\theta,D}$ based on the messages tails for the three languages included in this study. The incidence of language and different group of writers over the tail of ranked frequency distributions was evaluated by performing a Student-t test which results are included in Table 3.3. Student-t tests to compare the distributions of the texts tail Zipf's deviations $J_{\theta,D}$ show very small

p-values, indicating that tail Zipfs deviation differed statistically in very significant ways between the three different languages studied.



English: square - 1945.BS.Eng.GabrielaMistral triangle - 1921.MarieCurie
 rhombus - 1950.NL.Eng.BertrandRussell circle - 1890.RusselConwell
 Spanish: square - 1936.Doloresbarruri triangle - 1982.Gabriel García Márquez
 rhombus - JoseSaramago.Valencia, circle - CamiloJ.Cela.LaColmena.Cap1
 Artificial: square - FibonacciNumbers.CSharp triangle - QuickSort.CSharp
 rhombus - Sociodynamica.Module3 circle - WebSite.Inmogal.php

Figure 3.11: Tail Zipf's deviation $J_{\theta,D}$ for symbol ranked frequency distributions vs. text tail length L . English (left), Spanish (center) and Computer Code (right). Reference texts are highlighted with filled markers.

Table 3.3: Tail Zipf's deviation $J_{\theta,D}$ and its correlation with message tail length L_{θ} for English, Spanish and artificial messages.

Tail Zipf's deviation $J_{\theta,D}$ for natural and artificial languages				
	n	$J_{\theta,D}$ average	$J_{\theta,D}$ Std. Dev.	Correlation $J_{\theta} : L_{\theta}$
English	156	0.1502	0.2108	0.809
Spanish	158	0.0235	0.1493	0.856
Computer Code	49	0.3528	0.3062	0.640
t-test	n1 - n2	p-value		
English - Spanish	156 - 158	2.34E-09		
Natural - Software	314 - 49	2.79E-15		

3.3 Discussions

3.3.1 Diversity for natural and artificial languages

Setting a precise number for the total number of words of a natural language is impossible, as words appear and disappear constantly. However it has been estimated that English contains more words than Spanish [10,33,34]. Living

languages evolve over time and structural differences make it difficult to compare figures of language size measure. Nevertheless the numbers of lemmas in dictionaries provide us a reference to compare language sizes. The dictionary of the Real Academia Española contains 87.718 Spanish lemmas [35] while the Oxford English dictionary includes about 600.000 words [36]. Despite the larger size of English dictionaries, Spanish texts showed higher and less dispersed symbol diversity than English.

The higher word diversity of Spanish may thus be due to factors such as syntactical rules or grammar which affect both languages differently. Verb tenses and conjugations, for example, are all considered as one word when included in a dictionary, but each of them was recognized as a different symbol here.

For Spanish, most articles, pronouns and subject genres vary from masculine to feminine while for English this only happens for particular cases like *his/her*. These grammar characteristics may increase the number of different symbols used in any Spanish texts, but considering the relative size of closed and open word groups, this effect should be marginal with regard to general text symbol diversity. On the other hand, verbs, which belong to the open group of words, have more tenses and conjugations for Spanish and therefore increase Spanish word diversity in ways not accounted for in dictionaries. Grammar is then one feature that explains greater Spanish word diversity compared to English.

These differences might explain only parts of the results shown here. A wider use of words in Spanish, compared to English, despite a larger number of words in English dictionaries, cannot be excluded.

3.3.2 Entropy for natural and artificial languages

There is no qualitative difference for this property between English and Spanish, perhaps a consequence of the similar structure and functionality both natural languages share. Nevertheless entropy appears slightly higher for messages expressed in English than for those in Spanish; being English a larger language in terms of words, this result might be explained as consequence of a more elaborated grammar in Spanish allowing for lower entropy levels. The topic also has an impact over the properties we measured. For example, religious speeches in English and political speeches in Spanish show a lower symbol diversity than those texts influenced by other topics. Clearly, the semantic speech content has an incidence over the text properties as the symbolic diversity and entropy. In addition to these theme-associated differences, there are however, overlapping differences between the languages themselves. We think the number of messages considered and the wide range of natural

language themes and computer code functions included in our library of study, suffice to avoid any important bias in our comparison between natural and artificial languages caused by the differences in the semantic content of texts.

Natural languages have developed to express concepts and complex ideas. Natural languages can express many different types of messages such as information, persuasion, inspiration, instruction, distraction and joy. Artificial languages, in contrast, are designed to give precise instructions; they are more formal than natural ones [37] as they must convey precise and unequivocal information to machines. Artificial languages are represented by computer programs; collections of instructions having extensive number of symbols and commands. The number of symbols that an artificial language usually contains is very small when compared to natural ones. Connecting and auxiliary words like prepositions and articles are limited to conditional and logical expressions. Adjectives are replaced by numeric variables which may quantify some aspects modeled. With these limitations, computer languages have little room for style compared to natural languages. Computer code is valued for its effectiveness rather than its beauty. The limited structure to form sentences in artificial languages leads to a relatively flatter frequency distribution and therefore higher entropy levels.

Since emergence is defined as equivalent to Shannon's information (entropy), the higher emergence for artificial languages implies that less symbols are used to produce 'more meaning'. In other words, there is less redundancy in artificial than in natural languages. Redundancy can lead to robustness [38], which is desirable in natural languages where communication may be noisy. However, artificial languages are created for formal, deterministic compilers or interpreters, so there is no pressure to develop robustness.

Self-organization, as opposed to emergence, is higher in artificial than in natural languages. This is because of the same reason explained above: artificial languages require more structure to be more precise, which fulfills their purpose. Natural languages are less organized because they require flexibility and adaptability for their purpose, which includes the ability of having different words with the same meaning (synonymy) and words with different meanings (polysemy).

For the same specific diversity d , complexity is higher for natural languages (Figure 3.5). However, for the same length L , complexity is higher for artificial languages, as emergence dominates the properties of all languages ($e > 0.5$) (Figure 3.6). Artificial languages are slightly more regular, but all languages have a relatively high entropy and thus emergence.

3.3.3 Symbol frequency distributions

Intuition may suggest that the symbol frequency profile of a symbol limited language will decay faster than a richer language in terms of number of available symbols. Figure 8 illustrates how, for the natural languages considered here, the points of each message rank distribution profile lay close to a straight line connecting the first with the last ranked word. This indicates that g values for natural languages are approximately constant over the range of symbol ranking. For artificial texts, on the contrary, symbol-frequency vs. symbol-ranking does not show a constant decay value. The slope of the graph is low for most used symbols and increases its decay rate as the symbols considered approach the least used ones, giving the rank symbol profile of artificial language the concave downward shape characteristic of an approximation to the cut-off region [17]. This increasing slope g that artificial messages exhibit over ranges of the ranking dominion indicate these languages are close to the physical limit of their total number of symbols. For natural languages g values are not only lower but also closer to a constant, denoting that natural language profiles are within the scale-free region and therefore far from the physical limit [17] imposed by the number of symbols they are constituted with. Natural languages are significantly larger than the artificial languages all together.

There is a qualitative difference of the symbol frequency distributions for natural and artificial languages; texts written in natural languages correlate with a power law distribution for all the Symbol Ranking ranges while artificial texts show an increasing decay slope for ranges of least used symbols. This difference may be related to the fact that for natural languages any message uses only a tiny fraction of the whole set of words of the language, while any reasonable long computer code will use a large fraction of the whole set of symbols available in the computer language.

The most conspicuous difference between natural and artificial languages was revealed using Zlpf's deviation $J_{1,D}$. Statistical analysis revealed highly significant differences between natural and artificial languages in this variable. Tail Zipf's deviation $J_{\theta,D}$, confirmed these differences, focusing only on the tails of these distributions. No loss of information was evidenced when focusing our analysis only on the tails, compared with analysis using the complete frequency profile of the Zlpf's deviation $J_{1,D}$.

Another interesting aspect of this list of symbols is where the words of open and close classes lay according to their frequency of use; close and open word classes are also known as core and non-core word types. As Andrew Moore explained [39], English grew by adding new words to its open-word class

consisting of nouns, verbs and qualifiers, (adjectives and adverbs). The close-word class contains determiners, pronouns, prepositions and conjunctions; words that establish functionality and language structure. The dynamic process of word creation and the 'flow' of words from one class to the other have been recently modeled [40]. Changes over time are slow, thus for our purpose of this study, we considered the open and close classes as invariant groups. Being the open-class the sustained faster growing type of words of natural languages, it is reasonable to expect the open words class to be much larger than the group of closed words. The smaller size of the closed-word class and the highly restricted character of its components (most of them do not even have synonyms), explain the high frequency of their use and their tendency to be placed near the top of the ranked list shown in Table 3.1, letting the open-class words to sink down to lower ranked positions of the list. There are formal indications of this tendency of close words to group near the top of frequency ranked list in a study by Montemurro et al. [41], where pronouns are presented as the most frequently used word-function in Shakespeare's Hamlet.

Besides being necessary to understand the structure of English and Spanish, the classification of words as members of the open and closed groups is important because analyzing the ranking among the open-class words may lead to some practical uses as the recognition of message subject or theme. The highest ranked open-class words are represented using italic-bold characters. For the messages included in this study, the most used open-class words were '*man*', '*people*' and '*world*' for English, and '*pueblo*', '*mundo*' and '*vida*' for Spanish; all of them are terms with strong connection to government, religion, and human rights as the main theme treated by the majority of the messages.

3.4 Conclusions

Diversity is higher for Spanish messages than for English ones, suggesting that there is influence of cultural constraints over message diversity. Being more restricted to very specific uses and less dependent on writing style, artificial languages showed a considerably lower diversity than natural languages.

Entropy measures for natural languages are higher than those for artificial. The larger symbolic diversity for natural languages dominates the resulting text entropies, leaving frequency profiles to a more subtle influence. When comparing English and Spanish however, symbolic diversities are closer to each other while entropy differences become relevant. Future work could include sets of legal, clinical or technical documents. Since these seem to be more specific, they should have properties in between the natural and artificial sets studied here.

We have shown that important differences among languages become evident by experimentally measuring symbolic diversity, emergence and complexity in collections of texts. The differences detected are the result of the combination of the current status of their respective evolution as well as cultural aspects that affect the style of communicating and writing. These differences among languages are evidenced measuring symbolic diversity, emergence and complexity in collections of texts. Yet the most reliable measure was the symbolic diversity. Applying this procedure over the basis of a 'grammar scale complexity' would provide a deeper sense of languages nature and behavior.

From a wider scope, the results obtained constitute a strong indication that languages can be regarded beyond a large set of words and grammar rules, and as a collections of interacting organisms to which the concepts of complexity, emergence and self-organization apply.

We believe that the present study showed that complexity analysis can add to our understanding of features of natural languages. For example, automatic devises to differentiate text written by computers from text produced by real persons might be feasible using this knowledge. Yet our study also revealed that Complexity Science is in a very incipient state regarding its capacity to extract meaning from the analysis of texts. Much interesting work lies ahead.

"Brevity is the soul of wit."
William Shakespeare, *Hamlet*

"I have only made this letter longer because I have not had the time to make it
shorter."

Blaise Pascal, *The Provincial Letters*

Chapter IV

The representation of writing styles as symbolic diversity and entropy

In 1880 Lucius Sherman [42,43] studied the structure of the English language from a statistical point of view, finding that the average number of words in English sentences had diminished from 50, before the times of Queen Elizabeth, to 23 during the time Sherman lived. A second result showed that writers are consistent in the average number of words per sentence [8]. Efforts to construct methods to evaluate text readability have continued since then. During the early twentieth century, teachers evaluated texts relying on the *Teacher's Word Book* [44] by Thorndike; a collection of the 10,000 most frequently used words in English published in 1921 and extended to 20,000 words in 1932 [45] and 30,000 in 1944 [46]. These word-frequency lists were mostly used to evaluate the appropriateness of reading material for children at elementary schools. The evaluation of quality of writing consisted, basically, in counting the number of different words in a text as a measurement of the author's size of vocabulary.

The vocabulary lists became the basis for describing an underlying structure, as is the English language word frequency distribution, known today as the Zipf's law [21]. Due to George Kingsley Zipf's renowned work, *Human Behavior and The Principle of Least Effort* [2]. The evaluation of quality of writing consisted, basically, in counting the number of different words in a text as a measurement of the author's size of vocabulary.

Starting with his PhD thesis [47], Rudolf Flesch published a series of books studying English texts [48–52]. These efforts led to the Reading Ease Score, usually referred to as RES, a formula based on the weighted combination of vocabulary,

average word character-length and average sentence word-length, useful to evaluate the ease, or difficulty, to read and understand the content of texts. After Flesch's original work, other researchers built formulas based on to Flesch's *RES* formula. Adaptations for specific uses such as the evaluation of applicants to enter the US navy [53] and institutions in charge of assessing reading and comprehension of prospective students of American universities [54], as well as for analyzing the suitability of basic school texts appeared and became the theme of much research and experimentation. Within the fields where readability formulas have been a useful tool, health occupies an important place [55–57], but the field of education resulted best suited for the application of these readability formulas versions specially made by Chall [58], Kincaid and others.

In 1959 Fernandez-Huerta [59] adjusted the original Flesch's readability formula and produced the '*Formula de lecturabilidad de Fernández-Huerta*' (Readability Formula) for Spanish. Another adaptation of the Flesch's formula, presented by Szigriszt-Pazos [60] named '*Formula de Perspicuidad*' (Perspicuity Formula) or *IPSZ*, as we will refer to it, has become the current standard to evaluate the readability of Spanish texts.

In recent years, a different approach to measure readability as appeared. Relying on today's computing capacity Tanaka-Ishii K., Tezuka S. and Terada H. [61], proposed looking at readability as a relative property of texts instead of an absolute assessment. Theirs is a method based on Support Vector Machines and sorting algorithms. Yet, traditional readability formulas are widely accepted, and remain as the most used method to evaluate the appropriateness of texts in accordance with the audience they are intended for.

The relationship between readability measures and word frequency profiles is the focus during the 1960's by Klare [62]. Klare added to the 'Human Behavior and The Principle of Least Effort', the mechanisms that explain the high frequency of appearance short words in natural language texts. Klare stated that the size of words is an underlying 'learning' factor which makes the communication process more effective, since shorter words are faster and better understood by both interacting parts, the emitter and the receiver, highlighting the fact that any communication process is not only less laborious, but also more effective when shorter symbols are used.

The possibility of measuring the quality of writing became a need with the emergence of qualifying tests for schools and colleges. The writing skills, in spite of being a neat reflex of intellectual capabilities, are an elusive property to measure. On the other hand, complexity measurements of text messages, address only the evaluation of the quantity of information needed to specify

and transmit a message, compressibility and other aspects of the information, focusing on the mere descriptive process and disregarding the idea content, beauty or any other form of valuable characteristic of the message. As an alternative, we suggest evaluating quality of writing by formulas based on characterizations of texts Zipf's profiles. The particular language's grammar rules establish restrictions over some degrees of freedom of the symbol frequency distribution profile [63], but there is still enough space for the text's symbol frequency profile, to be sensitive to some properties of a text as for example: organization, coherence, vocabulary richness, length of sentences and word difficulty, which have influence over readability.

Entropy was suggested as an index sensitive to the writing style by Kontoyiannis in 1997 [64]. In his study, Kontoyiannis computed entropy at the scale of characters; in other words, entropy was estimated considering a fixed symbolic diversity determined by the 26 characters of the English alphabet. Despite the obvious weak connection between the letter frequency distribution and the style of writing in any natural language, Kontoyiannis was able to conjecture the existence of some correlation between entropy and the style of writing. In another path of research, Jackes Savoy [65,66] presents evidence of the influence of the time period and the political affiliation of the authors and the frequency of use of specific type-of-words as verbs, pronouns, and adverbs. Savoy used a sample of a few hundred speeches pronounced by American presidents.

In this Chapter we investigate the impact of quality of writing over word diversity, entropy and ranked frequency profiles. To perform our experiments, we built a library with 138 English and 136 Spanish texts. The authors of the texts include politicians, military, Literature Nobel laureates, writers, scientists, artists, and other public figures. To overcome the bias introduced by the variety of text lengths, we evaluated differences in symbol diversity and entropy indices that might be related to writing quality. Finally, we propose an evaluation scale for English and Spanish that, we claim, is related to the quality of writing. Representing the set of speeches in the plane specific diversity-entropy, we visually highlight that relationship.

4.1 Methods

We based this work on a library containing English texts and Spanish texts. Texts were grouped in two categories: one integrated by those texts originated by authors who were laureate with the Literature Nobel Prize, the other formed by texts produced by renowned writers, politicians, military and social personalities. Combining writers and Nobel laureates for English and Spanish, we obtained four groups for our analysis.

Each text was characterized by its symbolic diversity D , entropy h , and distribution of symbolic frequency f in accordance with the definitions shown below. We built a mathematical model with these properties for the four groups created. Mean values and dispersion were studied by statistical methods. Finally we produced quality of writing scales for English and Spanish.

4.1.1 Text length L and symbolic diversity d

The length of a text L is measured as the total number of symbols used, and the diversity D as the number of different symbols that appear in the text. We define the specific diversity d as the ratio of diversity D and length L , that is

$$d = \text{specific diversity} = D/L. \quad (4.1)$$

As symbols we consider words as well as punctuation signs, therefore the number of symbols is obtained adding the count of both types of symbols.

4.1.2 Entropy h

Shannon's entropy expression [4] is used to measure texts information. Symbols (words) are treated as information units, disregarding any differential information weight that may be associated to the word meanings, length or context. The entropy h for texts is evaluated following definition:

$$h = - \sum_{r=1}^D \frac{f_r}{L} \log_D \frac{f_r}{L}. \quad (4.2)$$

where f_r is the number of appearances of the symbol occupying the place r within the ranking of symbols' frequency. Notice the base of the logarithm is the diversity D and hence $h(L, D)$ is bounded between zero and one. Setting the base of the logarithm to 2, expression (2) becomes

$$h = - \frac{1}{\log_2 D} \sum_{r=1}^D \frac{f_r}{L} \log_2 \frac{f_r}{L}. \quad (4.3)$$

4.1.3 Symbol frequency distribution f

When the symbols of a message are arranged according to the number of their appearances, from the most frequently found symbol to the least, we obtain the ranked symbol profile. For any symbol profile, the number of words in a rank segment $[a, b]$ may be computed as:

$$L_{a,b} = \sum_{r=a}^b f_r. \quad (4.4)$$

where r is a word frequency rank position while a and b are the start and the end of the considered symbol rank segment respectively. For any segment, $a = 1$ and $b = D$.

4.1.4 Zipf's deviation J

Zipf's law states that any sufficiently long English text will behave according to the following rule [9] [5]:

$$f(r) = \frac{f_a}{(r - a)^g}, \quad (4.5)$$

where r is the ranking by number of appearances of a symbol, $f(r)$ a function that retrieves the numbers of appearances of word ranked as r , f_a the number of appearances of the first ranked word within the segment considered, and g a positive real exponent.

For any message, we define Zipf's reference $Z_{a,b}$ as the total number of symbol appearances in the ranking segment $[a, b]$ assuming that it follows Zipf's Law. Therefore $Z_{a,b}$ is

$$Z_{a,b} = \sum_{r=a}^b f_r = \sum_{r=a}^b \frac{f_a}{r^g}. \quad (4.6)$$

The complete message Zipf's reference $Z_{1,D}$ is determined by expression (4.6) and the corresponding Zipf's deviations for the whole distribution $J_{1,D}$ is

$$J_{1,D} = \frac{(L_{1,D} - Z_{1,D})}{Z_{1,D}}. \quad (4.7)$$

4.1.5 Relative deviations of text properties

As Grabchak et al. [67] explain, statistics of specific diversity and entropy for natural languages texts have a bias upon the text length. This bias is due to the language structure and the definition of these properties and have been estimated by Febres, Jaffe and Gershenson for English, Spanish and some computer programming languages [63]. To compensate for the bias introduced by the diversity on text lengths of our library, we used the model presented by Febres, Jaffe and Gershenson, which consists of a minimal square error regression to determine the expected values of the specific diversity and the

entropy, as functions of the text length. The difference between the properties from data and the regression model is referred to as *Relative Deviation*. Applied cases to diversity relative deviation d_{rel} and entropy relative deviation h_{rel} are included in Eqs. (4.8) and (4.9).

$$d_{rel} = \frac{D - D_m}{D_m} . \quad (4.8)$$

$$h_{rel} = h - h_m . \quad (4.9)$$

Notice that Zipf's deviation, calculated as Expression (4.7) indicates, also works expresses a relative deviation.

4.1.6 Writing Quality Scale *WQS*

We did not find any computerized method to evaluate quality of writing. Thus, we designed a method for evaluating the quality of writing which results in a value we called Writing Quality Scale (*WQS*). Our method is based on evaluations of Equations (4.7), (4.8) and (4.9) for several hundred texts organized in groups as will be explain in Section 4.1.8.

4.1.7 Readability formulas *RES* and *IPSZ*

Readability formulas are available for many languages. They do not measure quality of writing but the appropriateness of a text for certain group of readers, like for example, children belonging to a school grade. Thus we used some readability formulas as a reference to compare the *WQS* with them. For English we used the *Reading Easy Score (RES)* by Flesch [50]:

$$RES = 206.835 - 84.6 W - 1.015 S , \quad (4.10)$$

where W is the average of the word length measured in syllables and S the average of the phrase length measured in words. For Spanish we used the adaptation that Szigriszt [60] made to the *RES* formula, named the *Perspicuity Index (IPSZ)*:

$$IPSZ = 206.835 - 84.6 W - S , \quad (4.11)$$

In Equation (3.11) W and S represent the same as in *RES* formula. Values of S were obtained as $S = L_w / L_{ph}$ where L_w is the text length measured in words and L_{ph} is the text length measured in phrases. In English as well as in Spanish, a phrase ends every time a period, colons, semicolons, question mark, exclamation sign or ellipsis appears. Thus L_{ph} equal the addition of the

appearances of the mentioned punctuation signs. The average number of syllables per word W is calculated as

$$W = L_{SY}/L_W, \quad (4.12)$$

where L_{SY} is the number of syllables of the whole text. Determining the number of syllables L_{SY} , is more difficult than counting words or punctuation signs; syllables are the textual representation of single sounds, whose start and end may be difficult to recognize, and additionally, the rules to extract syllables from a text have many exceptions, and vary from language to language. In fact, some authors [10] refer to the deviation from a regular correspondence between a written symbol and the associated phoneme, as *letter-phoneme complexity*, or *orthographic depth*; a completely different notion of complexity from the one we are dealing with in the present study. Thus, recognizing syllables from graphemes with an automated process is not a straight forward task. Especially for English, which is considered an *orthographically deep languages*², strict correspondence between writing and pronunciation, and vice versa, rarely exists. For Spanish, there is correspondence from writing to pronunciation, meaning that starting from a written word, we know its sound; but there may be many ways of writing down a sound we hear. This ambiguous correspondence between writing and pronunciation appears in English, Spanish and up to some degree in most alphabetic natural languages³. To prevent the writing of software codes to count syllables in English and Spanish texts, we decided to estimate L_{SY} by computing

$$L_{SY} = L_{CH}/C_{SY}, \quad (4.13)$$

where L_{CH} is the number of characters not including punctuation signs in the text and C_{SY} is the average number of characters contained in a syllable. Looking for other researcher's indicators of the number of characters per syllable, we found three pairs of L_{SY} values for English and Spanish. First: in her PhD thesis Barrio Cantalejo [55] explains how Szigriszt used Eaton's dictionary [68] to estimate values $C_{SY} = 1.69$ and $C_{SY} = 2.67$ for English and Spanish respectively. Second: in their study Trauzettel-Klosinski et al [56], measured the number of words, syllables and characters for 17 languages. They obtained values of $C_{SY} = 3.15$ and $C_{SY} =$

² In psycholinguistics a language is considered *orthographically deep* when there is little consistency between its written and spoken form. Some deep languages are Hebrew, English and French. Serbo-Croatian and Italian are examples of shallow languages [101].

³ Natural alphabetic languages are those whose writing consists of words build up with syllables represented by characters of an alphabet. Most known alphabets are the Latin, Cyrillic, Greek and Arabic. Natural syllabic languages represent phonograms by a single symbol or morpheme, therefore a set of letters (graphemes) to represent single sounds is not needed. Natural syllabic languages include Chinese and Japanese.

1.9 for English and Spanish. Third: Gualda Gil [69] compares the density of information conveyed by English and Spanish texts. As part of his analysis, Gualda Gil reports values of $C_{SY} = 3.57$ and $C_{SY} = 2.94$ for English and Spanish. Observing the lack of coherence among these values, we did our own count of syllables over a sample texts and calculated values of C_{SY} . Our results were within a 5% difference from those reported by Gualda Gil and therefore, we used the values he reported into Equation (4.13).

4.1.8 Message selection and groups

This study is based on written texts from historic famous speeches available in the web. The texts were originally written in different languages including English, Spanish, Portuguese, French, Italian, German, Japanese, Arabic, Russian, Chinese and Swedish. Since the analysis was done for English and Spanish, many of the texts used are translations from the original versions. Most texts are from politicians, human rights defenders and Literature Nobel laureates. We selected speeches to keep our library texts, as close as possible to the genuine writing ability of the author. Some other writing genres as the novel, have expressions from personages who distort the writing capabilities of the author. Thus, we have restricted the texts to analyze, to speeches.

We created groups of speeches and novel segments for English and Spanish: one group of texts with undoubtedly good language users, those who received Nobel Prizes for Literature, and another group by authors for which we have no special reason to assume an out-of-average use of the language. Texts classified as written by a Nobel laureate are all in their original language.

Some speeches written by Literature Nobel laureates, and translated from their original languages to English or Spanish are included in the graphs of Figures 1 to 8. These texts are clearly signaled by different markers, and are used only to obtain some sense of the effects of translations over texts authored by Nobel laureates and then subject to a translation process, but these translated Nobel laureate texts are not considered in anyone of the computation or our comparison.

To compute word frequencies, we considered punctuation signs as words. A detailed explanation about especial symbol considerations can be found in [63]. Text libraries, computations and results registering were administered by *MoNet*, a complex-system analysis framework we have developed to elaborate and combine results from the network of experiments which constitute this and previous works.

4.2 Results

4.2.1 Diversity for Literature Nobel laureates and general writers

Figure 4.1 shows how diversity varies with the message length. The parameters of models expressed in Equations (4.14a) and (4.14b), which express Heaps' Law [7], were adjusted to minimize the summation of squared errors between the data and each model. The result is represented by the black lines in each graph of Figure 4.1.

$$\text{English:} \quad D_m = 3.766 \cdot L^{0.67}, \quad (4.14a)$$

$$\text{Spanish:} \quad D_m = 2.3 \cdot L^{0.75}. \quad (4.14b)$$

Notice that messages coming from Nobel laureate writers appear in the higher-diversity side of the regression line defined by Equations (4.14a) and (4.14b), suggesting the possibility of grading quality of writing around diversity values. The exceptions correspond to novels, as they include long texts reflecting popular discourse of the characters of the novel rather than language normally used by the writer.

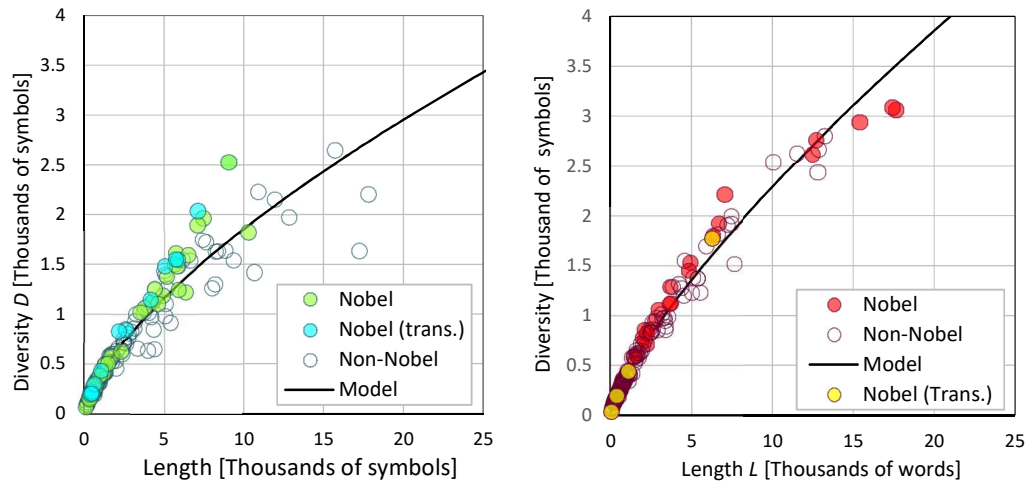


Figure 4.1: Diversity D as a function of message length L for messages expressed in English (left) and Spanish (right) by non-Nobel and Literature Nobel laureates. Texts authored by Literature Nobel laureates are highlighted with filled markers.

The differences between messages diversity D and the diversity model expressed in Equations (4.14a) and (4.14b) was evaluated statistically for English and Spanish. Comparisons of these differences for non-Nobel laureates and Literature Nobel laureates, are shown in Table 4.1. The upper sector of Table 4.1 shows a comparison of diversities for texts written by Nobel and non-Nobel laureates. While the row for writers shows negative values for relative diversity

deviations, the counterpart row for Nobel laureates, show positive values, confirming the tendency of Nobel laureate writers to use a richer vocabulary for both, English and Spanish.

Table 4.1: Comparing the relative specific diversity d_{rel} for English and Spanish messages by non-Nobel and Literature Nobel laureates. Upper section of the table shows the average relative specific diversity and its standard deviation. Lower section shows the p-values for Student t-tests applied to different group combinations.

		Relative specific diversity d_{rel}		
		n	d_{rel} average	d_{rel} std.dev.
English:	Nobel laureates	37	0.02690	0.152
English:	Non Nobel	101	-0.05741	0.133
Spanish:	Nobel laureates	19	0.07296	0.097
Spanish:	Non Nobel	117	-0.02339	0.085
t-test		n1 - n2	p-value	
English:	Nobel - Non Nobel	37 - 101	0.00186	
Spanish:	Nobel - Non Nobel	19 - 117	0.00001	
Nobel:	English - Spanish	37 - 19	0.23604	
Non Nobel:	English - Spanish	101 - 117	0.02340	

When comparing English and Spanish for categories non-Nobel and Nobel laureate, the p-values are very low (especially for Spanish), meaning that the null-hypothesis should be rejected. This indicates that in English and Spanish there is a relevant difference between the relative deviation of the specific diversity d_{rel} , in the texts written by Nobel and non-Nobel writers.

On the other hand, p-values for comparisons between non-Nobel and Nobel laureates indicate values sufficiently low to reject the null-hypothesis for English and Spanish. According to this, the relative deviations of the specific diversity d_{rel} , behave differently and offer information useful to recognize whether or not a text was written by a Literature Nobel laureate. Results show that Spanish Nobel laureates differ from other Spanish writers more than the English colleagues. Non-Nobel laureates did not differ between Spanish and English writers.

4.2.2 Entropy for Literature Nobel laureates and general writers

Figure 4.2 shows entropy h values for speeches expressed in natural languages versus specific diversity d . Blue rhomboidal dots represent English messages and red circular ones represent Spanish.

Entropy must drop down to zero when diversity decreases to zero. It also tends to a maximum value of 1 as specific diversity approaches 1. Therefore the entropy of any message can be modelled as a function of its specific diversity [63], according to

$$h = \left(\frac{D}{L}\right)^{(\alpha-2)/(\alpha-1)} = d^{(\alpha-2)/(\alpha-1)}, \quad (4.15)$$

where α is a real number. Expressions (4.16a) and (4.16b) were obtained after adjusting parameter α to fit experimental data.

English:
$$h = d^{0.1523}, \quad (4.16a)$$

Spanish:
$$h = d^{0.1763}. \quad (4.16b)$$

Figure 4.2 also differentiates between writers and Nobel laureates. Color filled dots represent speeches written by Literature Nobel laureate. Messages originated by other writers are represented by empty dots. It is visually noticeable that dots representing texts from Nobel laureates tend to lie at a lower entropy level than that indicated by the lines representing models (4.16a) and (4.16b). Nobel laureate texts show less entropy than the average for non laureates in both Spanish and English. The difference between the two categories was analyzed statistically and results are shown in Table 4.2.

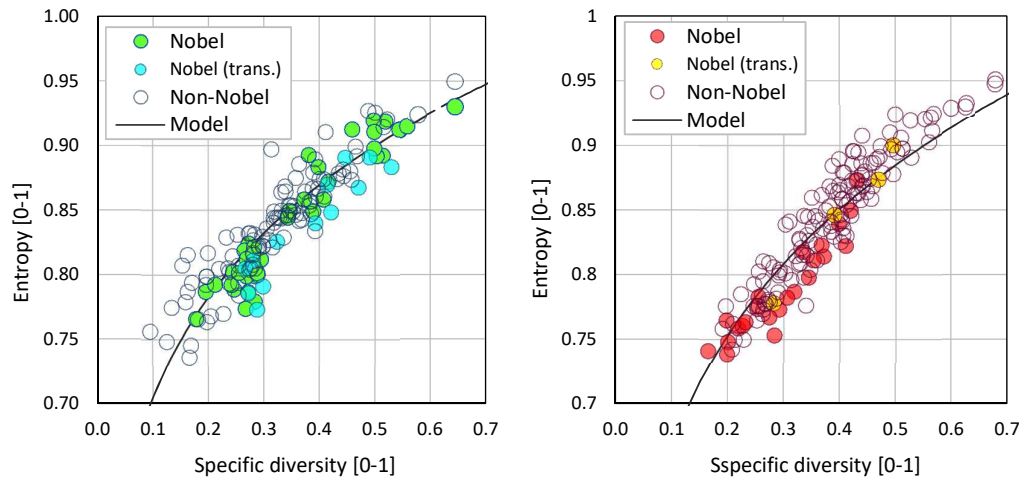


Figure 4.2: Entropy h vs. specific diversity d for messages expressed in English (left) and Spanish (right) by non-Nobel and Literature Nobel laureates. Texts authored by Literature Nobel laureates are highlighted with filled markers.

Relative entropies h_{rel} for non-Nobel writers and Nobel writers show opposite signed values. This difference in the distribution of relative entropy for writers and Nobel laureates is confirmed by the Student t-test; p-values printed in bold numbers are very low and therefore the hypothesis is rejected for English and Spanish.

Table 4.2: Comparing the relative entropy h_{rel} for English and Spanish messages by non-Nobel and Literature Nobel laureates. Upper section of the table shows the average relative entropy and its standard deviation. Lower section shows the p-values for Student t-tests applied to different group combinations.

Relative entropy h_{rel}				
		n	h_{rel} average	h_{rel} std.dev.
English:	Nobel laureates	37	-0.00567	0.0192
English:	Non Nobel	101	0.00318	0.0184
Spanish:	Nobel laureates	19	-0.01168	0.0192
Spanish:	Non Nobel	117	0.00579	0.0183
t-test		n1 - n2	p-value	
English:	Nobel - Non Nobel	37 - 101	0.00659	
Spanish:	Nobel - Non Nobel	19 - 117	0.00005	
Nobel:	English - Spanish	37 - 19	0.2067	
Non Nobel:	English - Spanish	101 - 117	0.2852	

4.2.3. Zipf's deviation $J_{1,D}$ for ranked distribution

Profile of symbol frequency distributions were inspected in two ways: first by a qualitative analysis of their shapes, and second by characterizing each profile with its area deviation J with respect to a Zipf distributed profile.

A sample of symbol frequency distributions profiles for the considered languages, is represented in Figure 4.3. Each sequence of markers belongs to a message and each marker corresponds to a word or symbol within the message. The size of the sample included in Figure 4.3 is limited to avoid excessive overlapping of markers which would keep from appreciating the shape of each profile. No important differences are observed among messages profiles expressed in the same language, however.

IV. The representation of writing styles as symbolic diversity and entropy

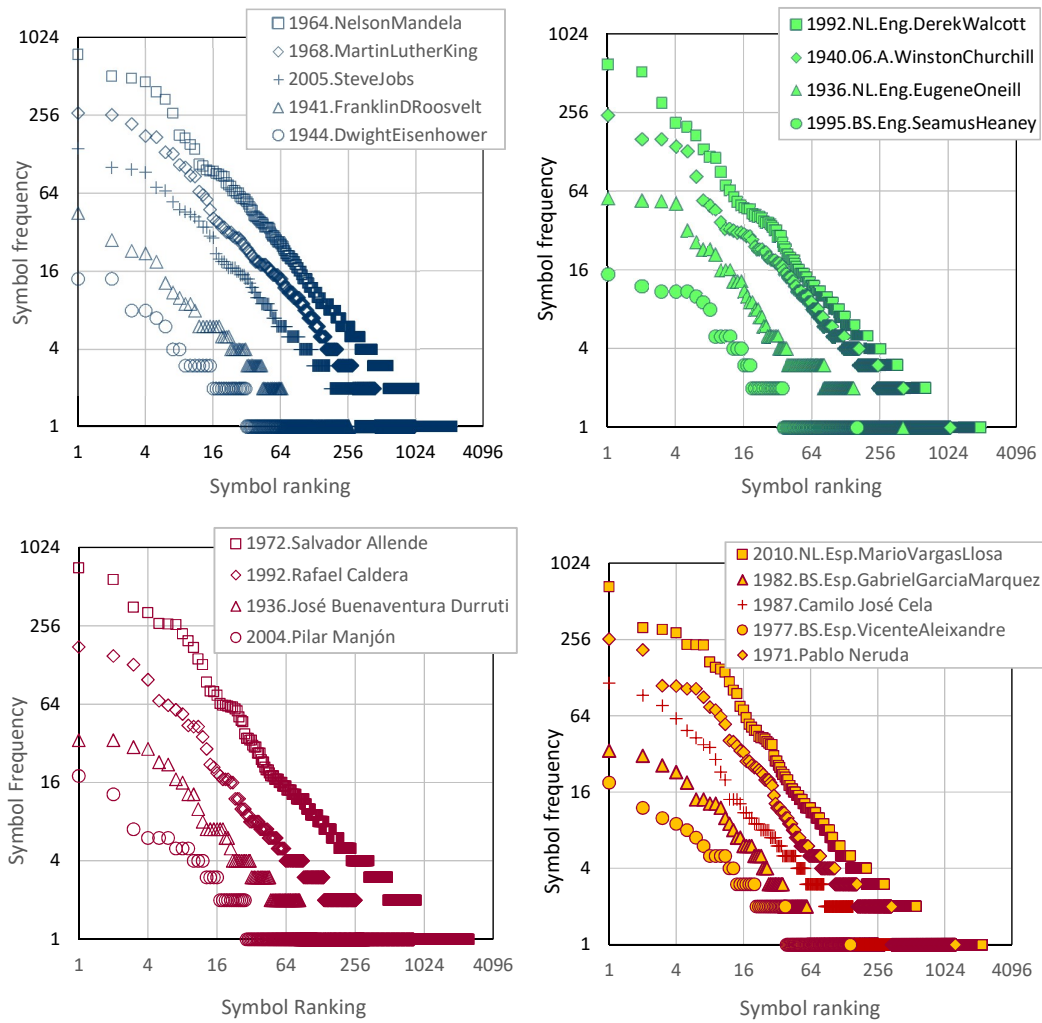


Figure 4.3: Ranked symbol frequency distribution profiles. Sample of three profiles for each category. Upper row shows English message profiles and lower row Spanish message profiles. Left column graphs show the profiles for writer originated texts and right column Nobel laureate texts.

Zipf's deviations $J_{1,D}$ for messages written by writers and Literature Nobel laureates are illustrated in Figure 4.4. Messages written by Literature Nobel laureates exhibit lower values of Zipf's deviations $J_{1,D}$ in comparison to Zipf's deviation of texts coming from non-laureate writers. To measure this difference, we computed Zipf's deviations $J_{1,D}$ for different groups of data: languages and writers class. Table 4.3 summarizes these results.

IV. The representation of writing styles as symbolic diversity and entropy

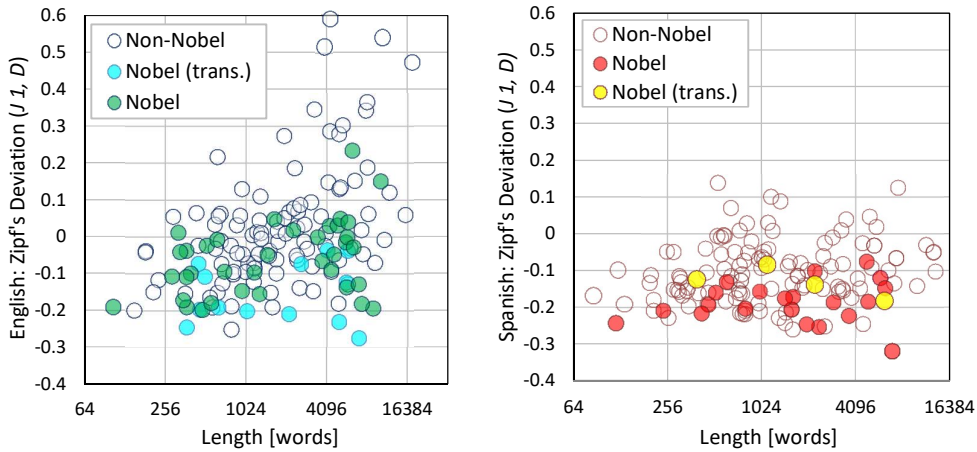


Figure 4.4: Relative Zipf's deviation $J_{1,D}$ vs. message length L for messages expressed in English (left) and Spanish (right) by non-Nobel and Literature Nobel laureates. Texts authored by Literature Nobel laureates are highlighted with filled markers.

Table 4.3: Comparing the relative Zipf's deviation $J_{1,D}$ for English and Spanish messages by non-Nobel and Literature Nobel laureates. Upper section of the table shows the average relative Zipf's deviation and its standard deviation. Lower section shows the p-values for Student t-tests applied to different group combinations.

		Relative Zipf's deviation $J_{1,D}$		
		n	$J_{1,D}$ average	$J_{1,D}$ s td.dev.
English:	Nobel laureates	37	-0.05779	0.0994
English:	Non Nobel	101	0.03232	0.1768
Spanish:	Nobel laureates	19	-0.19167	0.0561
Spanish:	Non Nobel	117	-0.10382	0.0856
t-test		n1 - n2	p-value	
English:	Nobel - Non Nobel	37 - 101	0.00396	
Spanish:	Nobel - Non Nobel	19 - 117	0.00003	
Nobel:	English - Spanish	37 - 19	< 0.00001	
Non Nobel:	English - Spanish	101 - 117	< 0.00001	

Spanish texts from Nobel-laureates show different Zipf's deviations when compared with texts from non-Nobel writers. For English texts, this difference is more subtle than the difference when the language is Spanish. Comparing the non-Nobel with Nobel writers, the p-value for Spanish is less than 0.00003, low enough to reject the null hypothesis, meaning that for Spanish the deviation of the Zipf's distribution is different for the two writer categories considered. For English the p-value of 0.00396, is also sufficiently low to reject the null hypothesis between these two categories. In fact, average values $J_{1,D}$ for English-non-Nobel

writers (0.03232) and English Nobel (-0.05779) are relatively far from each other. For Spanish this statistic is different; values $J_{1,D}$ for non-Nobel (-0.10382) and Spanish Nobel (-0.19167) are sensibly different.

4.2.4 Writing quality evaluation

Not being a Literature Nobel laureate does not mean poor writing capabilities. But winning a Literature Nobel Prize is guarantee of being gifted for excellent writing as well as master knowledge and control over a natural language. Some measurable statistical difference should emerge from classifying writers by those who were recognized with a Nobel Prize, and those who were not.

Figures 4.1, 4.2 and 4.4 present a clear evidence of the tendency of speeches from Nobel laureates to differ from the average style of writing of public figures. When comparing Nobel and non-Nobel laureate messages, the average of the former group tends to show higher specific diversity d and lower entropy h . This is interesting because the higher specific diversity of Nobel laureate texts should promote a higher entropy due to the larger scale D of the language used implied by the larger vocabulary. See Equations (4.2) and (4.15) to observe how D affects the resulting entropy h . Nonetheless, in spite of the larger vocabulary exhibited by Literature Nobel laureates in their texts, the associated entropies h are lower. Thus h_{rel} is a second variable to include in a writing quality evaluation scale.

Our data shows that Zipf's deviation $J_{1,D}$ is a third variable to have influence over a writing quality evaluation scale.

As some clustering is observed for the Nobel laureate class, we estimated the coordinates of the centers, and a direction vector pointing from the non-laureate class center to the Nobel laureate class center. These directions provide a sense for creating a scale that is sensitive to the quality of writing for English and Spanish. The clusters centers coordinates are:

$$\text{English writers class: } (d_{rel}, h_{rel}, J_{rel}) = (-0.05741, 0.00318, -0.03232) \quad (4.17a)$$

$$\text{English Nobel class: } (d_{rel}, h_{rel}, J_{rel}) = (0.0269, -0.00567, -0.05779) \quad (4.17b)$$

$$\text{Spanish writers class: } (d_{rel}, h_{rel}, J_{rel}) = (-0.02339, 0.00579, -0.08785) \quad (4.17c)$$

$$\text{Spanish Nobel class: } (d_{rel}, h_{rel}, J_{rel}) = (0.07296, -0.01168, -0.19167) \quad (4.17d)$$

The direction vectors are:

$$\begin{aligned} \text{English direction vector } (d_{rel}, h_{rel}, J_{rel}) & & (4.18a) \\ & = (0.68147, -0.07153, -0.72835) \end{aligned}$$

$$\begin{aligned} \text{Spanish direction vector } (d_{rel}, h_{rel}, J_{rel}) & & (4.18b) \\ & = (0.73241, -0.13280, -0.66779) \end{aligned}$$

Based on the director vectors and the non-Nobel writer's class center coordinates, we suggest the following a *Writing Quality Scale (WQS)* which we claim is sensible to the quality of writing.

$$\begin{aligned} \text{English: } WQS & = 5.5082 (d_{rel} + 0.02690) - 0.5782 (h_{rel} - 0.00318) & (4.19a) \\ & - 5.8871 (J_{1,D} - 0.03232) \end{aligned}$$

$$\begin{aligned} \text{Spanish: } WQS & = 5.5674 (d_{rel} + 0.02339) - 1.0095 (h_{rel} - 0.00579) & (4.19b) \\ & - 5.0762 (J_{1,D} - 0.10382) \end{aligned}$$

We computed *WQS* for each text as Equations (4.19a) and (4.19b) indicate. Values of relative specific diversity d_{rel} , relative entropy h_{rel} , Zipf's deviation $J_{1,D}$ are the same included in Appendix C, and graphically shown in Figure 4.5.

Whether or not the language of a speech or a novel, is the author's native language, may be a factor with some influence over the evaluation of the *WQS*. For the case of all statistics in this study, a speech is considered to be authored by a Nobel Prize winner only in the version the text is presented in the author's native language. That choice assumes that the difference, which can be subtle, between the style of writing of a Nobel laureate and a non-Nobel writer, could vanish in the process of translation.

The selected criterion for considering a text written by a Literature Nobel laureate, evades possible effects of the translation, when it is performed, over the statistics presented and also over the models we call *WQS*. However, in Figure 4.5, those texts originally written by Nobel laureates and thereafter translated into English or Spanish, are included in the left graphs together with the texts authored by Nobel laureate writers. In the graphs on the left of Figure 4.5, the bubbles representing texts by Nobel laureates, are bordered with a thick dark ring, while translated texts have a thin border.

IV. The representation of writing styles as symbolic diversity and entropy

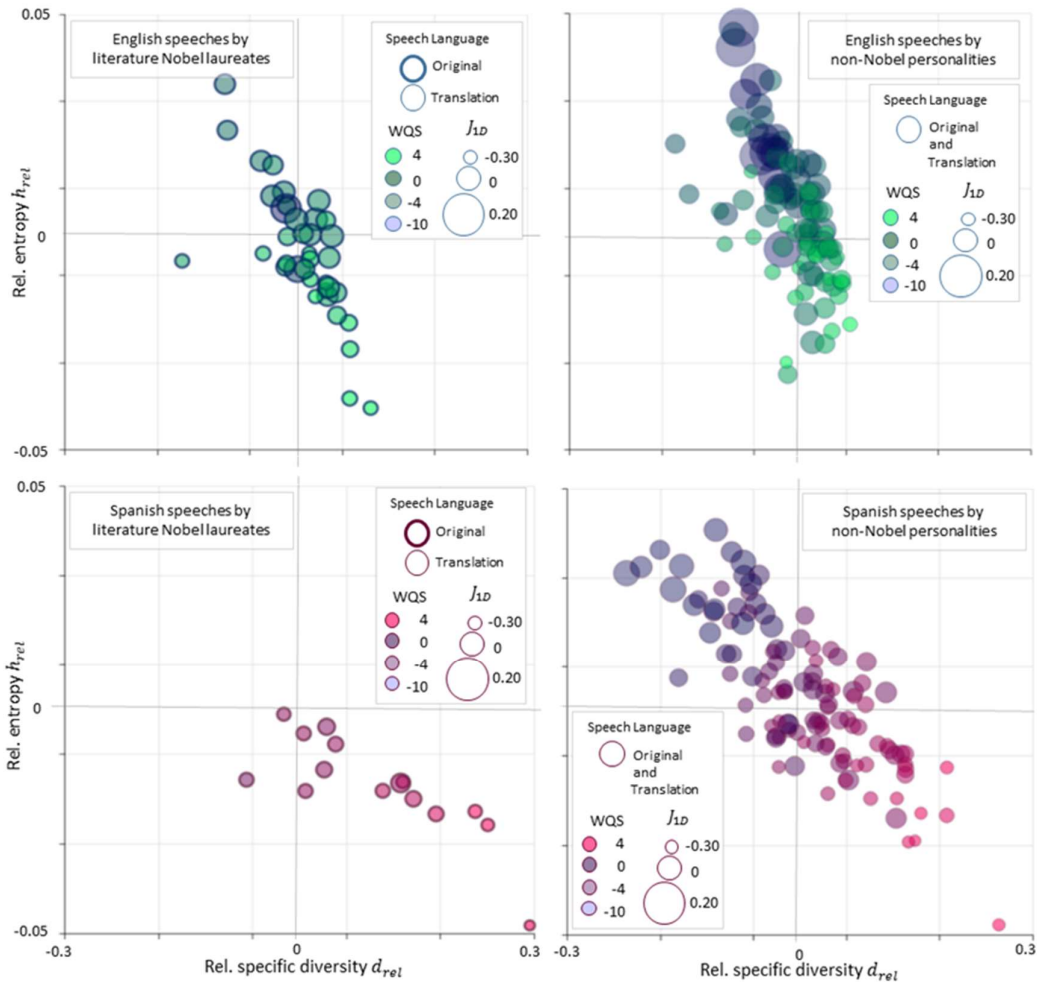


Figure 4.5: Writing quality evaluation for English (top) and Spanish (bottom) texts. Left graphs correspond to Literature Nobel laureate texts and right graphs to non-Nobel texts. Horizontal axes represent relative specific diversity d_{rel} . Vertical axes represent relative entropy h_{rel} . The Zipf's deviation $J_{1,D}$ is represented as proportional to the radius of the bubbles. The writing quality scale WQS value is represented by the bubble color.

Representing Nobel laureate texts and translated texts from original Nobel laureate texts, allows for visually appraise the impact over the WQS value of texts translations. The left graphs in Figure 4.5 suggest that there is no important difference among the WQS values for original and translated Nobel laureate texts. Comparing left graphs with right graphs, there is a noticeable tendency for the Spanish texts written by Literature Nobel laureates, to cluster around the point signaled by Expression (4.17d). English texts do not show as much clustering as Spanish texts do, which is consistent with the p-values of the Student t-test shown in Table 4.3.

4.2.5 Writing quality scales and readability indexes

The *Writing Quality Scales (WQS)* developed in section 4.2.4 were compared with the readability indexes from Flesch and Szigriszt. Figure 4.6 shows graphs of readability indexes versus the *WQS* obtained for each text in the library. In the graphs, each dot represents a text. To enable the graphs to visually show the difference between text categories, filled dots correspond to texts written by Literature Nobel laureates and empty dots show texts by non-Nobel writers. For Spanish, there is a higher density of dots representing texts by Nobel laureates towards the high *WQS* region, placed to the right of horizontal axis. For English, texts written by Nobel laureates and non-laureates do not show any important difference in their dispersion over the space of any axis.

Numerical comparisons between these different texts evaluations, are included in Table 4.4, confirming the visual appreciation mentioned above. Even though small, there is a difference between the averages of the distributions of Spanish readability indexes *IPSZ* for texts authored by Nobel writers and non-Nobel writers. At the same time, the small p-value obtained from Student-t tests for these distributions, indicates they are different, and that Literature Nobel laureates tend to produce more readable texts than others writers. The Student-t test performed between the distributions of English readability indexes *RES* for Nobel and non-Nobel texts, resulted in a high p-value indicating that there is not any important difference between these distributions of the readability index.

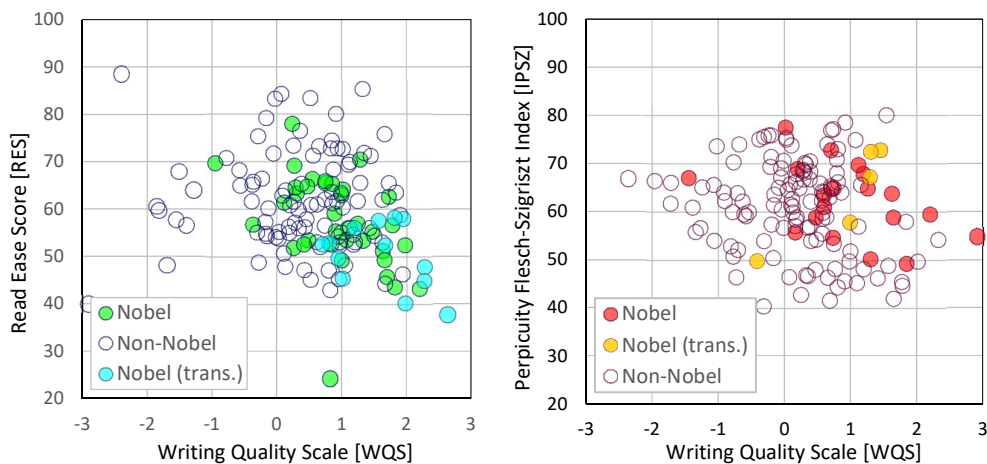


Figure 4.6 Text readability vs. Writing Quality Scale *WQS* for English texts (left) and Spanish texts (right). English readability is measured as Flesch *RES* (Reading Ease Score). Spanish readability is measured as Szigriszt *IPSZ* (Perspicuity Index). Filled dots (green for English texts and orange for Spanish texts) correspond to texts written by Literature Nobel laureates. Empty dots correspond to non-Nobel texts.

For English and Spanish texts, values of *WQS* for Nobel laureates showed higher when compared with values for texts coming from non-Nobel writers. Probable differences between distributions of the *WQS* of texts written in English and Spanish, were evaluated by Student t-tests. The results of these tests, included in Table 4.4, indicate that for English, Nobel and non-Nobel *WQS* values are likely to come from different distributions, while for Spanish these distributions are definitively different.

Table 4.4: Comparing the Writing Quality Scale *WQS* for English and Spanish messages by non-Nobel and Literature Nobel laureates. Upper section of the table shows the average Writing Quality Scale and its standard deviation. Lower section shows the p-values for Student t-tests applied to different group combinations.

Writing Quality Scale and Readability							
		n	Readability*				
			<i>WQS</i> average	<i>WQS</i> std.dev.	average [0-100]	Readability std.dev.	Correlation <i>WQS-Ready.</i>
English:	Nobel and non-Nobel	138	0.43	1.09	59.90	10.52	-0.34
	Nobel laureates	37	0.90	0.68	56.91	10.21	-0.38
	Non -Nobel	101	0.25	1.16	61.00	10.47	-0.31
Spanish:	Nobel and non-Nobel	136	0.25	0.97	61.03	9.33	-0.15
	Nobel laureates	19	1.16	0.75	63.43	7.45	-0.39
	Non -Nobel	117	0.11	0.92	60.64	9.57	-0.19
t-test		n1 - n2	p-value	t-test	n1 - n2	p-value	
English <i>WQS</i> : Nobel - Non Nobel		37 - 101	0.002	Spanish <i>WQS</i> : Nobel-Non Nobel	19 - 117	0.000	
English RES: Nobel - Non Nobel		37 - 101	0.043	Spanish IPSZ: Nobel-Non Nobel	19 - 117	0.229	

* Readability is measured as *REF* for English and *IPSZ* for Spanish.

4.2.6 Writing style change in time

Even though our main objective is to compare differences in writing style, we take advantage of the data fed, in order to investigate the change that the average sentence length, as measured in words, and the writing quality scale *WQS*, may have had over the last couple of centuries.

Figure 4.7 presents the average number of word per sentence for each speech of our texts library. Figure 4.8 shows the Writing Quality Scale *WQS* values. In both Figures, four graphs are presented for the combinations of speeches expressed in English and Spanish, and authored by non-Nobel writers and Literature Nobel laureate speeches. The speeches authored by Nobel laureates, but resulting from translation from another language, are included in the graph dedicated to Nobel laureates, but they are distinguished with a different marker from the one used for Nobel laureate texts expressed in their original language.

IV. The representation of writing styles as symbolic diversity and entropy

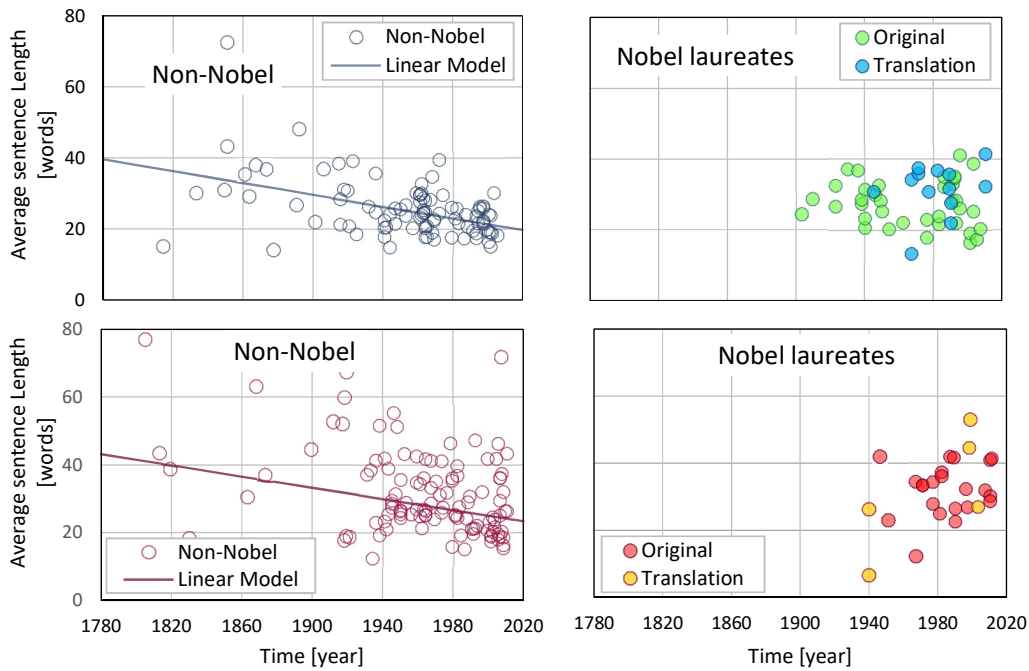


Figure 4.7: Average sentence length [words] vs. year when the speech was written for English (top row) and Spanish (bottom row). Left graphs show texts written by non-Nobel writers. A continuous line represent an error minimum summation model. Graphs on the right show Nobel laureate speeches. Original and translated texts are represented with different markers.

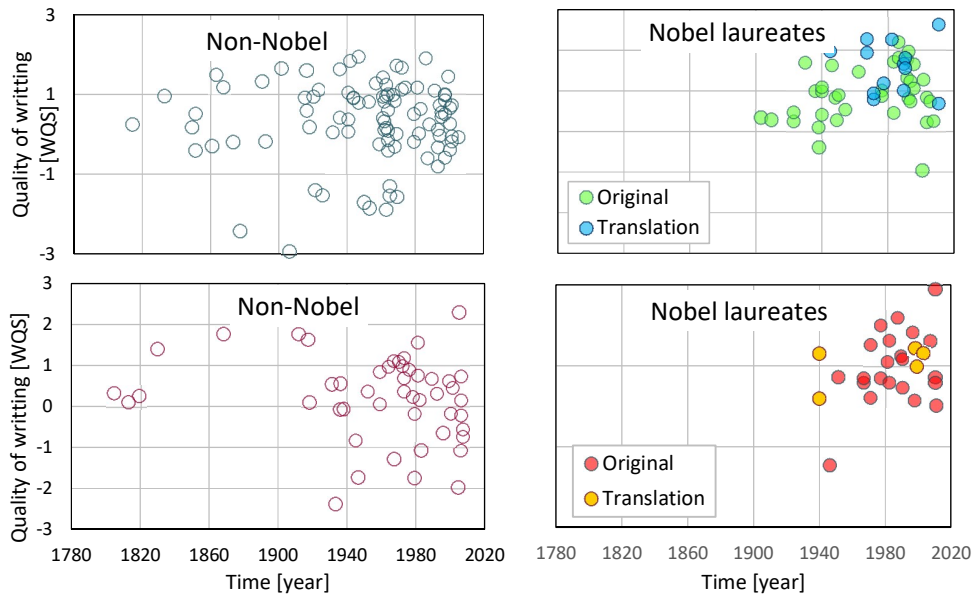


Figure 4.8: Writing Quality Scale WQS vs. year when the speech was written for English (top row), and Spanish (bottom row). Graphs on the left show texts written by non-Nobel writers. Graphs on the right show Nobel laureate speeches. Original and translated texts are represented with different markers.

None of the graphs of Figure 4.8 exhibit any important tendency of the WQS over time. But the average sentence length for non-Nobel writers graphed in Figure 4.7, does show a clear tendency to diminish over time for both, English and Spanish. Thus, we included a line to show the resulting regression from the minimization of the summation of the errors. Those line equations are $L = 187.1 - 0.0829 y$ for English, and $L = 189.8 - 0.0824 y$ for Spanish where L is the average length of sentences and y is the year.

4.3 Discussions

4.3.1 Diversity and entropy

In general, Literature Nobel laureates exhibit a richer vocabulary in their speeches when compared with other writers. Clearly, a necessary condition to win a Nobel Prize is the knowledge of an extended lexicon and the wisdom to use it appropriately and with a well-organized style. The higher diversity of words in exhibited by most texts from Nobel laureate shown in Figure 4.1, is thus, an expected result. Interestingly, Nobel laureates somehow handle this higher word diversity in such a way, that they produce texts with considerably lower entropy than the expected entropy value, at the corresponding specific diversity. Therefore, the lower entropy values exhibited by Nobel laureate's texts, must obey to the word's frequency distribution they use, which overcomes the natural effect of the larger diversity of words present in their texts.

4.3.2 Symbol frequency distribution profile

The difference between the Zipf's deviations $J_{1,D}$ for the two types of writers originating the texts, is small. However, the relatively small p-values indicate that Zipf's deviations $J_{1,D}$ express some of the differences between texts originated by Nobel and non-Nobel writers, and therefore, the inclusion of the Zipf's deviations $J_{1,D}$ as a writing quality sensitive factor, is justified.

4.3.3 Writing Quality Scale versus Readability Index

Readability indexes are intended to classify the ease with which a text can be read and understood. They are not directly associated to quality or style of writing. In fact, evaluating quality and style of writing is a highly subjective matter, difficult to submit to a quantifying procedure. It is a subtle and elusive task. However, good writing structure and style must include readability as an important characteristic of the resulting text. The measures of entropy explored here add information about more general aspects of writing quality.

Another factor influencing the readability indexes is the complexity of the idea being deployed with the text. A complex idea, probably, cannot be explained with the same high readability index of a simple idea. Thus the question is: What readability index can reach a writer when he or she writes a text to convey some complex idea? There is no obvious answer, among other reasons, because the complexity of the idea itself, is a subjective factor. But good writers should tend to produce more readable texts—with higher readability indexes: *RES* for English and *IPSZ* for Spanish— than those less talented for this activity. In fact, Figure 4.6 shows that for Spanish there is higher density of texts authored by Nobel laureates over the higher readability region, indicating that Spanish Nobel laureates tend to produce high readability texts. For English, we did not detect any important difference between the readability of the texts produced by Nobel and non-Nobel laureates.

Figure 4.5 illustrates how most of the texts in Spanish with high values of *WQS*, those which are reddish, lie in the lower right quadrant. This quadrant represents for texts with lower relative entropy and higher specific diversity; both tendencies formerly associated with the style of writing of Literature Nobel laureates. A similar orientation of the *WQS* is observed for English written texts. Even though it is not as notorious as it was for Spanish texts. This confirms that the *WQS* captures some of the properties associated quality and style of writing. Especially for Spanish writing.

4.3.4 Tendencies of the writing style

The change of the average sentence length estimated from the regression model shown in Figure 7, is a reduction of 8.29 and 8.24 words per century for sentences written in English and Spanish respectively; interestingly two values that are, in the practical sense, equal.

According to previous results by Sherman [8], the sentence length experienced a change of 22 words (from 45 to 23), in a time span of 293 years, from the times of Queen Elizabeth I (around the year 1600) to Sherman's times (around the year 1893). These numbers and dates result in a calculated decrease of 7.5 words per century for English; a figure consistent with our estimates, which validates the comprehensiveness of the data we used.

Independently of the results from Sherman's works, the reduction of the length of sentences observed in Figure 7, seems to be a sustained tendency for common writers. Perhaps the increasing need to produce more effective texts, leaving less space for words dedicated to embellish the texts, is partially responsible for the reduction of the number of words. The evolution of the natural languages may also contribute, by the acceptance of new words, to the

representation of concepts and ideas in a more compact fashion. Nevertheless, the decrease of the number of words in a typical sentence, is probably approaching a lower bound, since a certain number of words is needed to express precise and elaborated ideas.

The Nobel Prizes are awarded since 1901. The history records we have to evaluate the evolution of styles on Literature Nobel laureates, are shorter than the sample of speeches we have available for non-laureate writers. Yet, the average sentence length for Nobel laureate writers does not show any important tendency to change over time. This suggests that good style of writing is not necessarily aligned with the concept of readability. There is no obvious increase or decrease of the values of WQS in the graphs included in Figure 8. This suggests there is no direct incidence of the sentence length over the value of the Writing Quality Scale WQS .

4.4 Conclusions

Our analysis showed that some properties of texts written in English and Spanish, such as entropy, symbol diversity, and frequency distribution profiles, relate to aspects of what is considered by professionals as “good writing” in natural languages. In general, our method showed to work better for Spanish than for English language. Texts written in Spanish by Nobel laureates and non-Nobel, are easier to segregate than their counterpart in English. The visual assessment of graphs as well as the statistical evaluations, confirm this statement. However, even for the English language, the method is capable to classify a text according to its writing quality as compared with a text representative of those written by a Literature Nobel laureate. This is encouraging because it suggests the feasibility of using quantitative measures to characterize certain aspects related to the quality of writings.

This opens the door to eventually develop tools for automatic text evaluations. The fact that quality was related to higher specific diversity and less entropy, suggests that skillful writing involves incorporation of order into the text. The precise nature of this additional order is still unknown, but our method serves to detect its presence.

The results found so far are to be taken as insights of a preliminary exploration of the complexity of texts. Certainly, further studies applying these methods to a larger set of texts and extending the methods to other writing genres may lead to further refinements that may make WQS a useful tool for evaluation of writing capabilities. We believe, however, that feasibility of automated quantitative evaluation of writing quality is getting closer.

*"The words. Why did they have to exist?
Without them, there wouldn't be any of this."
Markus Zusak, The Book Thief*

Chapter V

The fundamental scale of descriptions

The understanding of systems and their complexity requires accounting for their entropy. The emergence of information upon the scale of observation has become a topic of discussion since it reveals much of the systems' nature and structure. Y. Bar Yam [12] and Y. Bar-Yam, D. Harmon, and Y. Bar-Yam [70] have proposed the concept of complexity profile as a useful tool to study systems at different scales. Among others, R. Lopez-Ruiz , H. L. Mancini and X. Calbet [9], and M. Prokopenko, Boschetti and Ryan [8] focus on the change of the balance between the system disorder and self-organization for different scales of observation. In a different approach, Murray Gell-Mann [11] considers complexity as a property associated to the irregularities of the physical system. But Gell-Mann sees both randomness and order as manifestations of regularity, and therefore quantities that offer the possibility for reducing the length of a description and hence the computed complexity of a system.

These complexity concepts are all evaluated using arbitrarily selected symbol scales. The selected observation scale depends on the communication system used in the description; for example, systems described with human natural languages are prone to be analyzed with the characters and words scales because they hold the most meaning for humans. When the analysis of information is in the context of its transmission, it is common to find binary codes as the base of study. A possible consequence of this preselected scale of observation is the possible inclusion of our assumptions about the system's structure, which skews our interpretation about system properties.

Many studies have evaluated the entropy of descriptions based on a preconceived scale; in 1997 I. Kontoyiannis [64] evaluates the description entropies at the scale of characters; in 2002 M. A. Montemurro and D. H. Zanette [41] study the entropy as a function of the word-role; more recently J. Savoy [65,66], G. Febres, K. Jaffe and C. Gershenson [63], and G. Febres and K. Jaffe [71], have studied the impact of the style of writing over entropy speeches using the word as the unit of the scale. In 2009 R. Piasecki and A. Plastino [72] study entropy as a function of a 2-dimensional domain. They explored the effects of multivariate distributions and calculate the entropy associated to several 2D patterns. All these studies share the same direction; assume a space for a domain and a scale and compute the entropy. The strategy of the present study is to set the same problem in a reversed fashion: given an entropy descriptor of a multivariate distribution defined for some domain space, what would be the best way to segment that domain space in order to reproduce the known entropy descriptor? The answer to this question would have a twofold value: (a) an indication to the scale that best represents the system expression as the distribution of sizes of the space segments, and (b) an approximation to the algorithmic complexity of the description.

Algorithmic complexity as a concept does not consider the observation scale [11,13]. Algorithmic complexity —also called Kolmogorov's complexity— is the length of the shortest string that completely describes a system. Since the shortest string is a characteristic impossible to guarantee, algorithmic complexity has been regarded as an unreachable figure. Nevertheless, estimating complexity by searching for a nearly uncompressible description of a system, would have the advantage of being independent of the observation scale. In fact, a method to search for a nearly uncompressible description could be achieved by adjusting the observation scale until the process discovers the scale that best comprises the original description. The result would lead to an approximation to the algorithmic complexity of the system.

While these previous studies assume symbols as characters or words, in our present study we leave freedom to group adjacent characters, to form symbols in order to comply with a higher hierarchy criterion, as is the minimization of the entropy. This study develops a series of algorithms to recognize the set of symbols that, according to their frequency, leads to a minimum entropy description.

The method developed in this study mimics a simplified communication system's evolution process. The proposed algorithm is tested with short example of English text, and two descriptions, the first is an English text and the second, a sound MIDI (Musical Instrument Digital Interface) file. This representation of the components may convey a description of a system and its structural essence.

5.1 A quantitative description of a communication system

A version of Shannon's entropy formula, generalized for communication systems comprised of D symbols, is used to compute quantity of information in a descriptive text. To determine the symbols that make up the sequential text, a group of algorithms were developed. These algorithms are capable of recognizing the set of symbols which form the language used in the textual description. The number of symbols D represents not only the diversity of the language but also the fundamental scale used for the system description.

5.1.1 Quantity of information for a D 'nary communication system

We refer to language as the set of symbols used to construct a written message. The number of different symbols in a language will be referred as the diversity D .

To compute the entropy h of a language, that is, the entropy of the set of D different symbols, used with a probability p_i to form a written message, we use the Shannon's entropy expression, normalized to produce values between zero and one:

$$h = - \sum_{i=1}^D p_i \cdot \log_D p_i , \quad (5.1)$$

Note that the base of the logarithm is equal to the language's diversity D , whereas classical Shannon's expression uses 2 as the base of the logarithm; also equal to the diversity of the binary language that he studied. Researchers as Zipf [2], Kirby [21], Kontoyiannis [64], Gelbukh and Sidorov [25], Montemurro and Zanette [41], Savoy [65,66], Febres, Jaffe and Gershenson [63], Febres and Jaffe [71], among others, have studied the relationship between the structure of some human and artificial languages, and the symbol probability distribution corresponding to written expressions of each type of language.

All these studies assume symbols as characters or words, in our present study we leave freedom to group adjacent characters, to form symbols in order to comply with the minimization of the entropy h as expressed in Equation (1). In the following sections we explain this optimization problem, and our approach to find a solution reasonably close to the set of symbols that produce an absolute minimum entropy.

5.1.2 Scale and resolution

We propose a quantitative concept of scale: the scale of a system equals the diversity of the language used for its description. Thus, for example, if a picture is made with all available colors in an 8-bit-color map of pixels, then the diversity

of the color language of the picture would equal 2^8 , and the scale of the picture description, considering each color as a symbol, would be also 2^8 . Another example would be a binary language, a scale 2 communication system made up of only two symbols. Notice we have used the term 'communication system' to refer to the media used to code information.

Interestingly, the system's description scale is determined, in first place, by the observer, and in a much smaller degree by the system itself. The presumably high complexity of a system, functioning with the actions and reactions of a large number of tiny pieces, simply dissipates if (a) the observer, or the describer, fails to see the details, (b) the observer or describer is not interested the details, and prefers to focus on the macroscopic interactions that regulate the whole system's behavior, or (c) the system does not have sufficient different components, which play the role of symbols here, to refer to each type of piece. It is clear that any observed system scale implies the use of a certain number of symbols. It is also clear that the number of different symbols used in a description is linked with our intuitive idea of *scale*. There being no other known quantitative meaning of the word *scale*, we suggest its use as a descriptor of languages by specifying the number of symbols forming them.

Resolution specifies the maximum accuracy of observation and defines the smallest observable piece of information. In the computer coded files we used to interpret descriptions, we consider the character as the smallest observable and non-divisible piece of information.

Let E denote the physical space that a symbol or a character occupies, and let the sub-index signal the object being referred to. Thus, considering a written message \mathbf{M} , constructed using D_M different symbols Y as $\mathbf{M} = \{Y_1, Y_2, \dots, Y_{D_M}\}$, we would say the message \mathbf{M} occupies the space E_M and each symbol Y_i occupies the space E_{Y_i} . We define the length of all characters equal one. Therefore $E_{C_i} \equiv 1$ for any i . Finally, if the number of characters in a message is N , each symbol Y_i appears F_{Y_i} times within the message, and the symbol diversity is D_M , we can write the following constraints over the number of characters, symbols and the space they occupy:

$$E_M = \sum_{i=1}^{D_M} F_{Y_i} \cdot E_{Y_i} = \sum_{i=1}^N E_{C_i} = N . \quad (5.2)$$

5.1.3 The minimum length description scale

We see the scale of a language as the set of finite symbols that 'best' serves to represent a written message. The qualification 'best' refers to the capacity of

the set of symbols to convey the message with precision in the most effective way.

Take for example the western natural languages. Among their alphabets, there are only minor differences; too few differences to explain how far from each other those languages are. As M. Newman [23] observes, some letters may be the basic units of a language, but there are other units formed by groups of letters.

Chomsky's syntactic structures [28], later called context-free grammar (CFG) [73] offers another representation of natural language structure. The CFG describes rules for the proper connections among words according to their specific function within the text. Thus, CFG is a grammar generator useful to study the structure of sentences. Chomsky himself treats a language as an infinite or finite set of sentences. CFG works at a much larger scale than the one we are looking for in this study.

Regarding natural languages it is common to think that a word is the group of characters within a leading and a trailing blank-space. At some time a meaning was assigned to that word, and thereafter the word's meaning, as well as its writing, evolves and adopts a shape that works fine for us, the users of that language. Zipf's principle of least effort [2] and Flesch's reading ease score [50] certainly give indications about the mechanisms guiding words, as written symbols, to reduce the number of characters needed to be represented.

From a quantitative linguistics perspective, this widely accepted method for recognizing words offers limited applicability. Punctuation signs, for example, have a very precise meaning and use. The frequency of their appearance in any western natural language compete with the most common words in English and Spanish [11] However, punctuation signs are very seldom preceded by a blank-space and are normally written with just a single character, which promotes the false idea that they function like letters from the alphabet; they do not. They have meaning as well as common words have.

Another situation revealing the inconvenience of this natural but too rigid conception of words, is the English contraction when using the apostrophe. It is difficult to count the number of words in the expression "they're". How many words are there, one or two? See G. Febres, K. Jaffe and C. Gershenson [63] for a detailed explanation on English and Spanish word recognition and treatment for quantification purposes.

Intuitively the symbols forming a description written using some language, should be those driving the whole message to low entropy when computed as the function of the symbols frequency. In this situation the message and the text are

fixed as is the quantity of information it conveys. Then, there appears to be a conflict: while the information is constant because the message is invariant, any change to the set of symbols considered as basic units, alters the computed message entropy, as if the information had changed; it has not. To solve this paradox, we return to the question asked at the beginning of this section about the meaning of 'best' in the context of this discussion. From the point of view of the message emitter, the term 'best' considers the efficiency to transmit an idea. This is what Shannon's work was intended for: to determine the amount of information, estimated as entropy, needed to transmit an idea. From the reader's point of view the economy of the problem works different. The reader's problem is to interpret the message received to maximize the information extracted. In other words, the reader focuses on the symbols which turn the script as an organized, and therefore easier to interpret message. If the reader is a human and there are words in the message, the focused symbols are most likely words because those are the symbols that add meaning for this kind of reader. But if there existed the possibility to select another set of symbols which makes the message look even more organized, the reader would rather use this set of symbols because it would require less effort to read.

In conclusion, what the reader considers 'best' is the set of symbols that maximizes the organization of the message while for the sender the 'best' means the set of symbols needed to minimize the disorder of the message and thus the quantity of information processed. These statements are expressed as objective functions in Equations (5.3a) a (5.3b) where the best set of symbols is named \mathbf{B} , the message is \mathbf{M} , the message entropy is $h_{\mathbf{M}}$ and the message organization is $(1 - h_{\mathbf{M}})$.

$$\text{Sender's objective:} \qquad \min_{\mathbf{B}} h_{\mathbf{M}} \qquad (5.3a)$$

$$\text{Receiver's objective:} \qquad \max_{\mathbf{B}} (1 - h_{\mathbf{M}}) = \min_{\mathbf{B}} h_{\mathbf{M}} \qquad (5.3b)$$

Following this reasoning, 'best' means the same for both sides of the communication process. This may have important implications when considering languages as living organism or colonies of organisms. Both parts of the communication process push the language to evolve in the same direction: augmenting self-organization and the reducing of entropy of the messages. Both come together. Self-organization can be seen as one of the evolving directions of languages. Thus, self-organization is an indirect way to measured how deeply evolved is a language and what its capacity is to convey complex ideas or sensations.

Finally, an objective function to search the most effective set of symbols —the set with minimal entropy— to describe a language has been found. It will be used

to recognize the set of symbols that best describes a language used to write a description.

5.1.4 Language recognition

Consider a description consisting of a message \mathbf{M} built up with a sequence of N characters or elementary symbols. The message \mathbf{M} can be treated as an ordered set of characters C_i as:

$$\mathbf{M} = \{ C_1, C_2, \dots, C_N \} . \quad (5.4)$$

No restriction is imposed over the possibility of repeating characters. Consider also the language \mathbf{B} , consisting of a set of D_B different symbols Y_i , each formed with a sequence of E_{Y_i} consecutive characters found with probability $P(Y_i) > 0$ in message \mathbf{M} . Thus

$$\mathbf{B} = \{ Y_1, Y_2, \dots, Y_{D_B}, P(Y_i) \} . \quad (5.5)$$

$$Y_i = \{ C_j, C_{j+1}, \dots, C_{j+E_{Y_i}-1} \} , \quad 1 \leq i \leq D_B , \quad 1 \leq j \leq N - E_{Y_i} + 1 \quad (5.6)$$

The symbol probability distribution $P(Y_i)$ can be obtained dividing the frequency distribution f_i by the total number of symbols N in the message:

$$P(Y_i) = \frac{f_i}{N} . \quad (5.7)$$

Language \mathbf{B} , used to convey the message \mathbf{M} , can now be specified as the set of D_B different symbols and the probability density function $P(Y_i)$ which establishes the relative frequencies of appearance of the symbols Y_i . Each symbol Y_i is constructed with a sequence of contiguous characters as indicated in Equation (5.6). The set of symbols that describes the message \mathbf{M} with the least entropy comes after the solution of the following optimization problem:

$$\min_{\mathbf{B}} - \sum_{i=1}^{D_B} \frac{F_{Y_i} \cdot E_{Y_i}}{N} \cdot \log_{D_B} \frac{F_{Y_i} \cdot E_{Y_i}}{N} , \quad (5.8a)$$

Subject to

$$\mathbf{B} = \{ Y_1, Y_2, \dots, Y_i, \dots, Y_{D_B}, P(Y_i) \} , \quad \text{for } i = 1, 2, \dots, D_B \quad (5.8b)$$

$$Y_i = \{ C_j, C_{j+1}, \dots, C_{j+E_{Y_i}-1} \} , \quad \text{for } i = 1, 2, \dots, D_B \quad \text{and} \quad (5.8c)$$

$$j = 1, 2, \dots, N - E_{Y_i} + 1 ,$$

$$\sum_{i=1}^{D_B} F_{Y_i} \cdot E_{Y_i} = N, \quad (5.8d)$$

$$F_{Y_i} \geq 1, \quad E_{Y_i} \geq 1, \quad \text{for } i = 1, 2, 3, \dots, D_B. \quad (5.8e)$$

The resulting language will be the best in the sense that it is the set of symbols that offers a maximum organization of the message. The symbol lengths will range from a minimum to a maximum defining a distribution of symbol lengths characteristic of this scale of observation which is referred to as the *Fundamental Scale*.

5.2 The Fundamental Scale Algorithm

The optimization problem (5.8a-e) is highly nonlinear and restrictions are coupled. A strategy for finding a solution has been devised. It is a computerized process compound of text-strings processing, entropy calculations, text-symbol ordering and genetic algorithms. Given a description consisting of a text of N characters, the purpose of the algorithm is to build a set of symbols \mathbf{B} whose entropy is close to a minimum. The process forms symbols by joining as many as V adjacent characters in the text.

A loop where V is kept constant, controls the size of the symbols being incorporated to language B. The process ends when the maximum symbol length of V_{mx} characters is considered to form symbols. We add a sub-index to language \mathbf{B}_V to indicate the symbol size V considered at each stage of its construction.

We have defined several sections of the algorithm and we named them according to their similarity with a system where each symbol appears and ends up being part of a language, only if it survives the competence it must stand against other symbols. A pseudo-code of the Fundamental Scale Algorithm is included in Appendix D.

5.2.1 Base language construction

In the first stage, the message \mathbf{M} is separated into single characters. The resulting set of characters along with their frequency distribution constitute the first attempt to obtain a good language and it will be denoted as \mathbf{B}_1 . The sub-index indicates the maximum length that any symbol can achieve.

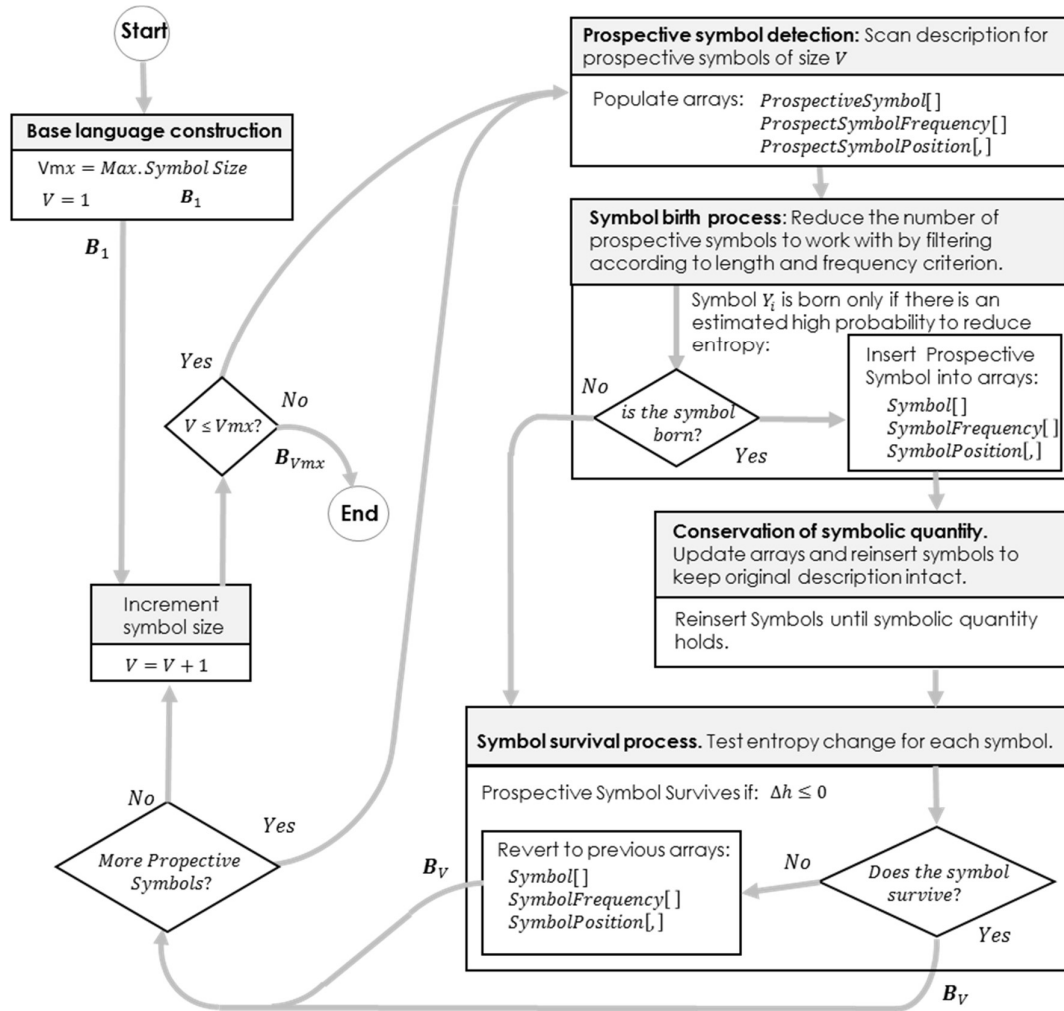
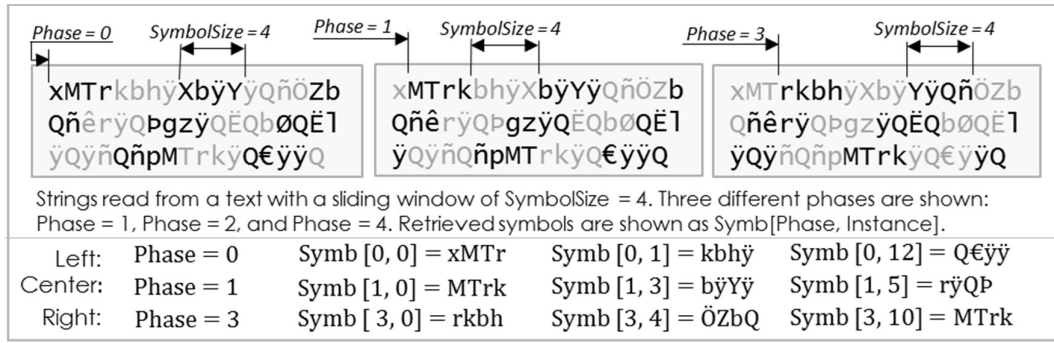


Figure 5.1: Major components of the Fundamental Scale Algorithm.

5.2.2 Prospective symbol detection

The Prospective Symbol Detection consists of scanning the text looking for strings of exactly V characters. All V -long strings are considered as prospective symbols to join the previously constructed language B_{V-1} made of strings of up to $V - 1$ characters. The idea is to find all possible different V -long strings present in the message M , which after complying with some entropy reduction criteria, would complement language B_{V-1} to form language B_V .



The Text: "xMTrkbhÿXbÿYÿQñÖZbQñêrÿQp̄gzÿQĒQbØQĒlÿQÿñQñpMTrkÿQ€ÿÿQ".

Figure 5.2: Examples of reading a text to recognize prospective symbols with a sliding window of SymbolSize = 4 and reading Phase = 0, 1, and 3. Phase = 2 not shown.

To cover all possibilities of character sequences forming symbols of length equal to V , several passes are done over the text. The difference from one pass to another is the character where the initial symbol starts, which will be called the *phase* of the pass. Figure 5.2 illustrates how the strategy covers all possibilities of symbol instances for any symbol size specification V .

5.2.3 Symbol birth process

Prospective Symbols detected in the previous stage whose likelihood to be an entropy reducer symbol is presumed too low, are discarded and never inserted as part of the language. Interpreting entropy Equation (5.1) as the summation of contributions of the uncertainty due to each symbol, we can intuit that minimum total uncertainty —minimum entropy— occurs when each symbol uncertainty contribution is about the same. Thus, any Prospective Symbol must be close to the average uncertainty per symbol in order to have some opportunity to actually reduce the entropy after its insertion. The average contribution of the uncertainty u_i for symbol i can be estimated as:

$$u_i = -p_i \log_{D_{B_V}} p_i = \frac{h}{D_{B_V}}, \quad (5.9)$$

This heads us to look for symbols complying with condition shown in Expression (5.10), and save processing time whenever a prospective symbol is not within a 2λ -width band of around the average uncertainty value,

$$\frac{h}{D_{B_V}} - \lambda < u_i < \frac{h}{D_{B_V}} + \lambda. \quad (5.10)$$

Parameter λ can be adjusted to avoid improperly rejecting entropy reducer symbols or to operate in the safe side at the expense on processing time.

5.2.4 Conservation of symbolic quantity

The inclusion of prospective symbols into the arrays of symbols representing the language \mathbf{B} , is performed to avoid the overlap of the newly inserted symbols and the previous language existing symbols. Therefore, every time a prospective symbol is inserted into the stack of symbols, the instances of former symbols occupying the space the new symbols, must be released. Sometimes this freed string is only a fraction of a previously existing symbol. Thus, the insertion of a symbol may produce a break up of other symbols, generating empty spaces for which recovered symbols must be reinserted in order to keep the original text intact.

5.2.5 Symbol survival process

A final calculation is performed to confirm the entropy reduction achieved after the insertion of a symbol into the language being formed. Those symbols not producing an entropy reduction, are rejected and the Language \mathbf{B} is reverted to its condition prior to the last insertion of a symbol.

5.2.6 Controlling computational complexity

The computational complexity of this algorithm is far beyond polynomial. A rough estimation sets the number of steps required above the factorial of the diversity of the language treated. Thus, segmenting the message into shorter pieces, allows the algorithm to find a feasible solution and to keep affordable processing times for large texts. This strategy is in fact a sort of parallel processing which significantly reduces the algorithm's computational complexity down to becoming an applicable tool. A complex system software platform has been developed along with this study to deal with the complexities of this algorithm, and the structure needed to maintain record of every symbol of each description within a core of very many texts. This experimental software, is named *Monet* and a brief description of it can be found in [6].

The noise introduced when chunking the original description in pieces, is limited. At most two symbols may be fractured for each segment. Very low compared to the number of symbols making each segment. The algorithm calculates the entropy of each description chunk. But, as Michael Grabchak, Zhiyi Zhang and D. T. Zhang explain [12], the estimation of the description's entropy must consider the bias introduced when short text samples are evaluated. Taking advantage of the extensive list of symbols and frequencies available and organized by means of the software *Monet*, we used the alternative of calculating the

description entropy using the joint sets of symbols for each description partition, and then forming the whole description. As a result, no bias has to be corrected.

5.3 Tests and results

In order to compare the differences obtained when observing a written message at the scales of characters, words and the fundamental scale, we designed the following Example Text. Table 5.1 shows the symbols obtained after the analysis of the Example Text at the three observation scales used in this study. The entropies calculated at the scales of characters and words were 0.81 and 0.90 respectively, the entropy at the fundamental scale was 0.76; an important reduction of the information required to describe the same message.

These results also get along with our intuition. Clearly, the selection of a certain character-string as a fundamental symbol, is favored by the frequency of appearance of the string of characters. As a result, the 'space character' (represented as \emptyset in the table) is recognized as the most frequent fundamental symbol. It indeed is an important structural piece in any English text, since it defines the beginning and the end of natural words.

The length of the string of characters also favors the survival of the symbol in its competence with other prospective symbols. The string '*describ*', for example, appears twice in the Example Text and the algorithm recognized it as a symbol. On the other hand, the 11-char long string '*. An adverb*' also appears 2 times, but the algorithm found it more effective in reducing the overall entropy, to break that phrase apart and increase the appearances of other symbols.

A similar case is that of the word '*adverb*', which appears in 9 instances (not including those written with the first capital letter) on the Example Text. But the entropy minimization problem found a more important entropy reduction by splitting the word '*adverb*' in shorter and more frequent symbols as '*dv*' (10 times), or the characters as '*e*' (70 times), '*a*' (40 times), '*,*', '*r*' (33 times), and '*b*' (12 times).

In another experiment, we contrasted two different types of communication systems by performing tests over full real messages. The first test is based on a text description written in English and the second in test based on the text file associated to music coded using the MIDI format. The English text is a speech by Bertrand Russell given in 1950 during the Nobel Prize ceremony. The MIDI music is a version of the 4th movement of Beethoven's ninth symphony. The sizes of these descriptions are near the limit of applicability of the algorithm.

Table 5.1: Results of the analysis of the Example Text at the three scales studied

Example Text: symbol sets at different scales.																
-What is an adverb? An adverb is a word or set of words that modifies verbs, adjectives, or other adverbs. An adverb answers how, when, where, or to what extent, how often or how much (e.g., daily, completely). Rule 1. Many adverbs end with the letters "ly", but many do not. An adverb is a word that changes or simplifies the meaning of a verb, adjective, other adverb, clause, or sentence expressing manner, place, time, or degree. Adverbs typically answer questions such as how?, in what why?, when?, where?, and to what extent?. Adverbs should never be confused with verbs. While verbs are used to describe actions, adverbs are used describe the way verbs are executed. Some adverbs can also modify adjectives as well as other adverbs.																
<i>FY</i> Frequency		<i>EY</i> Space occupier		<i>N</i> Message length [symbols]		\emptyset = space: max. symbol length = 13										
Chars Scale				Word Scale				Fundamental Scale								
<i>D</i> = 38 <i>h</i> = 0.8080 Diversity <i>D</i> = 82				Entropy <i>h</i> = 0.9033 Diversity <i>D</i> = 80				Entropy <i>h</i> = 0.7628								
<i>d</i> = 0.04 <i>N</i> 782 Spec. diversity <i>d</i> = 0.7033				Length <i>N</i> 171 Spec. diversity <i>d</i> = 0.1384				Length <i>N</i> 578								
Index	Symbol	<i>FY</i>	Index	Symbol	<i>FY</i>	Index	Symbol	<i>FY</i>	Index	Symbol	<i>FY</i>	<i>EY</i>	Index	Symbol	<i>FY</i>	<i>EY</i>
1	∅	169	1	,	21	41	complete	1	1	∅	##	1	41	ul	2	2
2	e	86	2	.	11	42	j	1	2	e	70	1	42	wi	2	2
3	a	45	3	or	7	43	Rule	1	3	a	40	1	43	io	2	2
4	s	44	4	adverbs	7	44	1	1	4	s	36	1	44	ie	2	2
5	r	44	5	?	6	45	end	1	5	t	36	1	45	im	2	2
6	t	39	6	adverb	5	46	letters	1	6	r	33	1	46	whē	2	3
7	o	34	7	verbs	4	47	ly	1	7	o	22	1	47	∅an	2	3
8	d	32	8	how	4	48	but	1	8	n	21	1	48	dif	2	3
9	n	30	9	an	4	49	do	1	9	,	18	1	49	uch	2	3
10	h	28	10	what	4	50	not	1	10	h	17	1	50	,∅c	2	3
11	i	25	11	is	3	51	changes	1	11	b	12	1	51	any∅	2	4
12	v	21	12	a	3	52	simplifies	1	12	dv	10	2	52	word∅	2	5
13	b	21	13	other	3	53	meaning	1	13	d	9	1	53	describ	2	7
14	w	21	14	to	3	54	verb	1	14	c	8	1	54	,∅Adverb	2	8
15	,	21	15	the	3	55	adjective	1	15	u	7	1	55	∅d	1	2
16	c	17	16	as	3	56	clause	1	16	l	6	1	56	∅v	1	2
17	l	16	17	are	3	57	sentence	1	17	?	6	1	57	word	1	4
18	.	11	18	word	2	58	expressing	1	18	wh	6	2	58	y∅	1	2
19	u	11	19	ot	2	59	manner	1	19	w	5	1	59	ma	1	2
20	m	10	20	that	2	60	place	1	20	i	5	1	60	t	1	1
21	y	10	21	adjective	2	61	time	1	21	.	4	2	61	ns	1	2
22	f	7	22	when	2	62	degree	1	22	g	4	1	62	An	1	2
23	?	6	23	where	2	63	typically	1	23	x	4	1	63	w	1	2
24	A	5	24	extent	2	64	answer	1	24	ly	4	2	64	b,	1	2
25	g	5	25	with	2	65	questions	1	25	m	4	1	65	v	1	1
26	p	5	26	"	2	66	such	1	26	verbs	4	5	66	-	1	1
27	x	4	27	used	2	67	in	1	27	y	3	1	67	(1	1
28	j	3	28	describe	2	68	why	1	28	p	3	1	68)	1	1
29	W	2	29	many	2	69	and	1	29	dj	3	2	69	R	1	1
30	"	2	30	-	1	70	should	1	30	∅of	3	3	70	l	1	1
31	-	1	31	set	1	71	never	1	31	ctiv	3	4	71	M	1	1
32	(1	32	words	1	72	be	1	32	,∅A	2	3	72	q	1	1
33)	1	33	modifies	1	73	confused	1	33	.	2	1	73	S	1	1
34	R	1	34	answers	1	74	While	1	34	W	2	1	74	ho	1	2
35	l	1	35	often	1	75	actions	1	35	"	2	1	75	∅m	1	2
36	M	1	36	much	1	76	way	1	36	ow	2	2	76	ng	1	2
37	q	1	37	€	1	77	executed	1	37	me	2	2	77	if	1	2
38	S	1	38	e	1	78	Some	1	38	le	2	2	78	in	1	2
			39	g	1	79	can	1	39	∅i	2	2	79	on	1	2
			40	daily	1	80	also	1	40	pl	2	2	80	si	1	2
						81	modify	1								
						82	well	1								

English descriptions of 1300 words or less can be processed in short times of less than a minute. Larger English texts have to be segmented using the Control Computational Complexity criteria mentioned in Section 5.3.6 to reach reasonable working times. Bertrand Russell's speech was fractioned in 7 pieces.

For MIDI music files, the processing times show an attitude of sharp increase starting for music pieces lasting about 3 minutes. The version of 4th movement of Beethoven's ninth symphony used, is a 25 minute long piece. It was necessary to process it by fractioning in 20 segments. To reveal the differences of descriptions when observed at different scales, symbol frequency distributions were produced. For the English text, characters, words and the fundamental scale were applied. For the MIDI music text distributions at character and fundamental scale were constructed. Words do not exist as scale for music. The corresponding detailed set of fundamental symbols can be seen in Appendix E. The frequency distributions were ordered upon the frequency rank of the symbols, thus the obtained were Zipf's profiles.

Table 5.2 shows the length N , the diversity D and the entropy h obtained for these two descriptions analyzed at several scales and Figure 5.3 shows the corresponding Zipf's profiles for Bertrand Russell's speech English speech and Beethoven's 9th Symphony's 4th movement. Both descriptions' profiles are presented at the scales they were analyzed: character-scale and the fundamental scale for both, English and music, and the word-scale only for English.

Table 5.2: Properties of two descriptions used to test the fundamental scale method

Text tag	Commtn. System	Name of Scale								
		Characters			Fundamental			Words		
		Length N	Diversity D	Entropy h	Length N	Diversit yD	Entropy h	Length N	Diversity D	Entropy h
1950.NL.Eng. BertrandRussell	English	32621	68	0.7051	26080	1227	0.5178	6476	1590	0.8215
Beethoven. Symphony9.Mov4	MIDI	103564	160	0.6464	84645	2824	0.4658	not defined		

In Figures 5.3a and 5.3b, the character scale exhibit the smallest diversity range. Taking only the characters as allowable symbols, leaves out any possibility of combination to form more elaborated symbols and excluding any possibility of representing how the describing information of a system arranges to create what could be loosely called the "language genotype". Allowing the composition of symbols as the conjunction of several successive characters, dramatically increases the diversity of symbols.

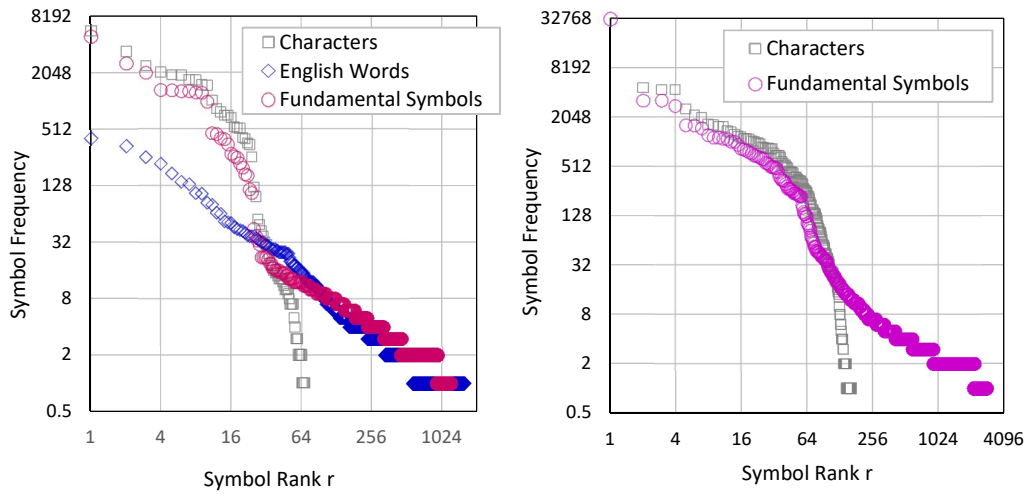


Figure 5.3: Symbol profiles for an English text (left) and a MIDI music text (right) at different scales of observation.

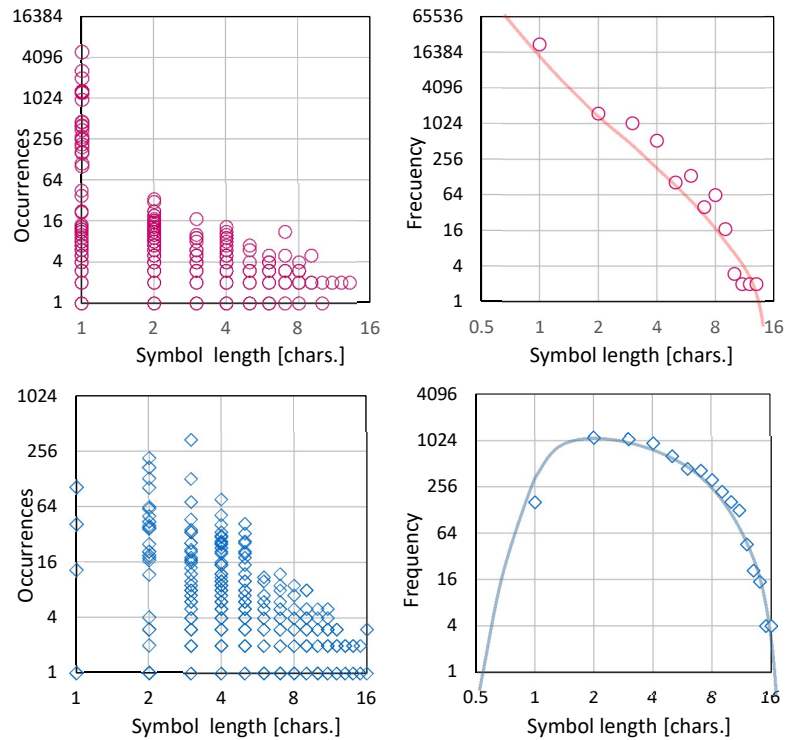


Figure 5.4: Bertrand Russell's 1950 Nobel ceremony speech behavior according to symbol length. Top row shows behavior at fundamental scale. Bottom row shows behavior at the scale of words. On the left each symbol is represented at its length on the horizontal axis and its occurrences at the vertical axis. On the right the graph shows the symbol length frequency distribution; the occurrences of all symbols sharing the same length are added and represented in the vertical axis. The horizontal axis represent the length of the symbols.

The selection of the symbols to constitute an observation scale holding the criteria of minimizing the resulting frequency distribution entropy, bounds the final symbolic diversity in a scale while capturing a variety of symbols that represents the way characters are organized to represent the language structure. The fundamental scale appears as the most effective scale, since with it, the original message can be represented with the most compressed information, expressed as the lowest entropy measured for all scales in both communication systems evaluated.

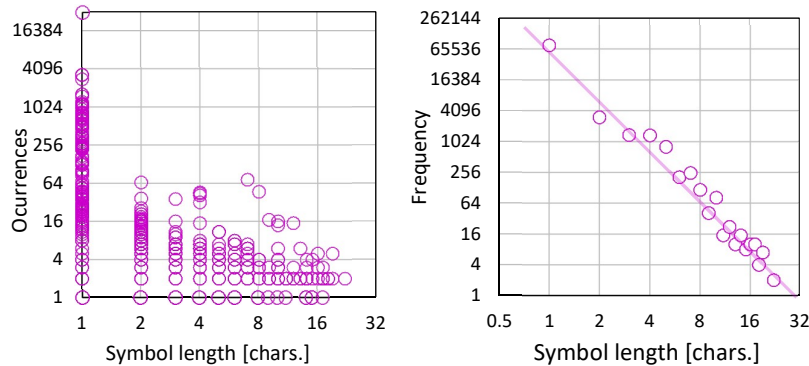


Figure 5.5: Beethoven's 9th symphony 4th movement MIDI music language behavior according symbol length. On the left each symbol is represented at its length on the horizontal axis and its occurrences at the vertical axis. On the right the graph shows the symbol length frequency distribution; the occurrences of all symbols sharing the same length are added and represented in the vertical axis. The horizontal axis represent the length of the symbols.

Any scale of observation has a correspondence with the size of the symbols focused at that scale. When that size is the same for all symbols, the scale can be regarded as a regular scale and specified indicating its size. If on the contrary, the scale does not correspond to a constant symbol size, then a symbol frequency distribution based on the sizes is a valid depiction of the scale. That is the case of the scales of words for English texts and the fundamental scale for our two examples. Figures 5.4 and 5.5 show those distributions and are useful to interpret the fundamental scales of both examples.

5.4 Discussions

The results clearly showed the calculus of the entropy content of a communication system varies in important ways, depending on the scale of analysis. Looking at a language at the scale of characters provides a different picture than examining it at the level of words, or at the here described fundamental scale. Thus, in order to compare different communication systems, we need to use a similar scale applicable to each communication system. We showed that the fundamental scale presented here is applicable to very

different communication systems, such as music, computer programs, and natural languages. This allows us to perform comparative studies regarding the systems entropy and thus to infer about the relative complexity of different communication systems.

In both examples analyzed, the profiles at the scale of characters and the fundamental scale run close to each other, within the range of the most frequent symbols to the symbols with a rank placed near the mid logarithmic scale. For points with lower ranking, the fundamental-scale profile extends its tail toward the region of low symbol frequencies. The closeness of fundamental and character scaled profiles in the high frequency region, indicates that the character-scaled language B_1 is a subset of the fundamental scale language. The language at fundamental scale, having a greater symbolic diversity and therefore more degrees of freedom, finds a way to generate a symbol frequency distribution with a lower entropy as compared to the minimal entropy distribution when the description is viewed at the scale of words. Focusing in the fundamental scale profiles, the symbols located in the lower rank region—the tail of the profile—tend to be longer symbols formed by more than one character. These multi-character symbols, which cannot exist at the character scale, are formed at the expense of instances of single character symbols typically located in the profile's head. This explains the nearly constant gap between the two profiles in the profiles' heads.

The English description, observed at the scale of words, produces a symbol profile incapable of showing short symbols—fragments of a word—which would represent important aspects of a spoken language as syllabus and other typical fundamental language sounds. On the opposite extreme, by observing at the character scale, the profile forbids considering strings of characters as symbols, thus meaningful words or structures cannot appear at this scale, missing important information about the structure of the described system.

The fundamental scale, on the other hand, appears as an intermediate scale capable of capturing the essence of the most elementary structure of a language, as its alphabet, as well as larger structures which represent the result of language evolution in its way to form more specialized and complex symbols. The same applies for music MIDI representation. There is no word scale for music, but clearly the character scale does not capture the richness that undoubtedly is present in this type of language.

Another difference between the fundamental scale, and other scales is the sensitivity to the order of the symbols as they appear in the text. At the scale of words or the scale of characters, the symbol frequency profile does not vary as the symbol order. The profiles depend only on the number of appearances of

each symbol, word or character, depending on the subject scale. The profile built at the fundamental scale does change as the symbol order is altered, not because of the symbol order itself, but because the symbol set recognized as fundamental, changes when the order or words or characters are modified. As a consequence, the character and word scales do not have any sense of grammar. The fundamental scale and its corresponding profile, on the other hand, is affected by the order in which words are organized—or disorganized—and is therefore sensitive to the rules of grammar. Other communication systems may not have words, but they must have some rules or the equivalence of a grammar. Assuming rigid rules as symbol size or symbol delimiters seems to be a barrier when studying the structure of system descriptions.

In the search for symbols, the fundamental scale method accounts for frequent sequences of strings which result from grammar rules. The string 'ing', for example appears at the end of words representing verbs or actions. Moreover, it normally comes followed by a space character (' '). As the sequence appears with noticeable frequency, the fundamental scale method recognizes the char sequence 'ing ' (ending with a space) as an entropy reducer token and therefore an important descriptive piece of English as a language. The observation of a description at its fundamental scale is therefore, sensitive to the order in which char-strings appear within the description. The fundamental scale method detects the internal grammar which has been ignored when analyzing Zipf's profiles at the scale of words in many previous studies.

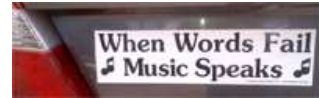
Despite the concept of fundamental scale being applicable to descriptions built over multidimensional spaces, the fundamental scale method and the algorithm developed is devised for 1-dimensional descriptions. The symbol search process implemented scans the description along the writing dimension of the text file being analyzed. This means that the fundamental symbols constituting 2D descriptions like pictures, photographs or plain data tables cannot be discovered with the algorithm as developed. To extend the fundamental scale algorithm to descriptions of more than one dimension, the restriction (8c) must be modified or complemented, to incorporate the sense of indivisible information unit—as has been the character in the development of this study—and the allowed symbol boundary shape in the description-space considered. This adjustment is a difficult task to accomplish because establishing criteria for the shapes of the boundaries becomes a hard to solve topology problem, especially in higher dimensional spaces.

There are other limitations for the analysis of descriptions of one dimension. Some punctuation signs which belong more to the writing system than to the language itself, work in pairs. Parenthesis, quotes, admiration and question marks are some of the written punctuation signs which work in couples. Intuition indicates that

each one of them is a half-symbol belonging to one symbol. In these cases, not considering each half as part of the same symbol most likely increases the entropy associated to the set of symbols discovered, thus becoming a deviation of the ideal application of the method. Nevertheless, for English, Spanish and human natural languages, in general, the characters which work in couples, appear unfrequently as compared to the rest of characters. Thus the minimal entropy distortion introduced by this effect is small.

Practical use of the algorithm is feasible up to some description lengths. The actual limit depends on the nature of the language used in the description. For syllabic human natural languages the algorithm can be directly applied to texts of 40000 characters or less. Longer texts, however, can be analyzed by partitioning. Thus the application limit for texts expressed in human natural languages, covers most needs. For the analysis of music, the use of the algorithm is limited to the MIDI format, result in large processing times even for powerful computers available today. The problem of scanning all possible sets of symbols in a sequence of characters grows as a combinatorial number. The Problem rapidly gets too complex in the computational sense, and its practical application is only feasible for representations of music in reduced sets of digitized symbols like the MIDI coding. Using more comprehensive formats like .MP3, a compressing technology capable of reducing the size of a music pack while keeping reasonably good sound quality, would be enough to locate the solution of the problem beyond our possibilities of performing experiments with large sets of musical pieces. Yet, the fundamental scale method provides new possibilities for discovering the most representative dimension of small sized textual descriptions, allowing us to advance in our understanding of languages. The Fundamental Scale method as developed, seems to be a great help for building evolution trees of languages and living organisms by allowing to quantify the degree of genetic description shared between two species belonging to different stages of an evolution process.

The Fundamental Scale, as a concept and as a method to find a quantitative approximation to the description of languages, promises interesting results for further research. Tackling the barriers of the algorithm by finding ways to reduce the number of loops and augmenting the assertiveness of the criteria used, may extend the space of practical use of the notion of a description's fundamental scale. Here we showed that the method reveals structural properties of languages and other communication systems, offering a path for comparative studies of the complexity of communication.



A sticker on a car in Caracas' traffic

"Music is a language that doesn't speak in particular words. It speaks in emotions, and if it's in the bones, it's in the bones."
Keith Richards, According to the Rolling Stones

Chapter VI

Several communication systems viewed at different scales

Ever since his influential paper was published in 1948, Shannon's entropy [4] has been the basis for quantifying information of symbolic systems descriptions. In summary, he postulated that the quantity of information of each description is proportional to the description text entropy. Shannon's work was developed over the basis of a binary communication system consisting of two symbols: zeros and ones. But the principle that relates entropy and information applies also to communication systems of more than two symbols.

Quantitative human natural language comparison has been a matter of study for decades. Some studies Schurmann and Grassberger [74] and Kontoyiannis [64], treat their text objects as large sets of characters. Other studies, as those by Savoy [65,66], Febres, Jaffe and Gershenson [63] and Febres and Jaffe [71], see texts as words. While the study of texts at the scale of characters misses the features of the natural languages that arise due to the structures composed of symbols to form words, the treatment of texts as exclusively built by words, ignores the presence of important elements that are part of the structure of the communication system. Neither the characters nor the words by themselves represent the actual symbol composition of communication systems.

The recently introduced concept of *Fundamental Scale* [75], offers the basis for a proper comparison of different types of communication systems. The representation of communication systems at the fundamental scale also allows to include in the comparison communication systems that do not use words.

In this Chapter we compare music, computer programming code, and two natural languages at several observation scales. For each type of communication system we consider a large number of texts. We measure the impact of changing the observation scale over the entropy measured for each text. Quantitative estimates of entropy were obtained at the scale of characters and at the fundamental scale calculated for each communication system [76]. When words are meaningful to the communication system, the scale of words was also included in the comparison.

6.1 Methods

We measured entropy h and specific diversity d for text descriptions expressed in four different communication systems: English, Spanish, computer programming code and MIDI music. Entropy calculations are used to estimate the quantity of information required for each text description at three different observation scales: characters, words and the fundamental scale.

6.1.1 Diversity and entropy

A version of Shannon's entropy [3], adapted for communication systems with more than two symbols [63,71,75] was used for these calculations. If communication system \mathbf{B} consists of as many as D symbols, then \mathbf{B} can be depicted as

$$\mathbf{B} = \{Y_1, \dots, Y_i, \dots, Y_D, \mathbf{P}(Y_i)\} , \quad (6.1)$$

where Y_i represents each symbol, from Y_1 to Y_D , used in the message, and $\mathbf{P}(Y)$ the probability density function which establishes the relative frequencies of appearance of symbols Y_i . It is worthwhile to mention that symbols Y_i do not have any syntactical meaning, thus they do not carry any information by themselves.

The quantity of information needed to convey a system description using the set of symbols included in language \mathbf{B} can be estimated as its entropy h . Thus,

$$h = - \sum_{i=1}^D p_i \log_D p_i , \quad (6.2)$$

where p_i refers to the probability of encountering symbol Y_i within a message described using communication system \mathbf{B} . Observe that the base of the logarithm is the symbol diversity D and therefore values of entropy h are normalized between zero and one, which is consistent with expressions for normalized entropy values established in previous works by Gershenson and Fernandez [10] and Febres, Jaffe and Gershenson [63].

The specific diversity d is the relation of the number of symbols D and the message length N as number of symbols.

$$d = \frac{D}{N} . \quad (6.3)$$

A value $d = 1$ means the description uses each symbol exactly once. In this case, and recalling these symbols are strictly symbolic —i.e. they have null syntactic meaning—, no pattern can be formed and thus, to reproduce the description, it would be needed the transcription of the whole set of symbols, employing the maximum quantity of information that possibly fits into a set of D different symbols. Equation (6.2) produces consistent results, since for this case $p_i = 1/D$ and therefore, all logarithms within the summation end up being 1 and the entropy reaches its maximum $h = 1$.

The lowest value the diversity can get is $D = 1$. This occurs when the description uses only one symbol and the message consists of a sequence of N identical symbols. When the diversity is $D = 1$, the description can be replaced by indicating the number of symbols N , therefore the information needed to express the description is just the number N . In this case, $p_1 = 1$, and the summation of Equation (6.2) will contain only one summand which leads to undetermined entropy value.

The cases where $D = N$ and $D = 1$ are extremes. In general the diversity D is an integer number between one and N . Typically D , is larger than 2 and thus, there are $D - 1$ different ways to modify the distribution of symbol probabilities p_i , as a result, the entropy h can be thought as a function defined over a dominion of $D - 1$ dimensions.

6.1.2 Language scale

As signaled in Expression (6.1), the specific language used in a message can be described as a symbol set along with the associated symbol frequency distribution. But the specific set of symbols considered as part of the language descriptor, depends on the way the whole message is divided in smaller pieces. The criteria used to segment the message into symbols is commonly called the *observation scale*; or simply the *scale*. Therefore an English text, for example, can be interpreted as a set of characters, a set of words, a set of sentences or any other way to rationally organize the information written in pieces. In this study we quantify the term scale as the number of symbols which the whole message is divided, thus the scale is, according to definition, equal to the language diversity D

We interpret descriptions written in different communication systems at several scales. When the communication system admits meaningful words as the natural languages, we split the messages into characters and words.

The nature of music differs radically from natural languages and thus has to be treated differently. Music is the superposition of simultaneously performed signals. In contrast to natural languages which can be described as a set of symbols formed by meaningful words, music digital records end up being a sequence of abstract characters. Without doubt, music is capable of transmitting senses about the frame of mind. Under some circumstances, and for some people, music can even be more effective to produce reactions in, and to reflex of, the frame of mind, than the natural languages can be. After the publication of Leonard Meyer's book "Emotion and meaning in music" [77], this field has been matter of study from the music technical view stand. For example, it is common to hear from conversations among musicians that melodies constructed over a minor scale⁴, transmit sadness, while those built over a major scale are joyful. Whether or not this associations are innate, some studies suggest that cultural exposure can overcome any innate initial bias [78]. Dana Wilson [79] emphasized the importance of patterning over the resulting effects of music pieces. But the patterns to which he refers, are not patterns of written recognizable symbols. Instead, he refers to components of music as rhythms, pitches and even instrument timbers. Yet he concludes: "paradoxically, the stronger the musical message, the less likely it ever will be explainable verbally or perhaps even understood rationally".

Definitely music transmits information. And music has structure and rules that must be respected in order to be regarded as music. But in spite of the use by musicians of the terms as 'word' and 'phrase', they are not equivalent to the sense of word and phrase in natural languages. In natural languages a word is easily recognized because it is preceded by a space and trailed by another space or a punctuation sign. It does not work the same way in music because music symbols are not strictly organized by spaces or silences. It is not either

⁴ In music a scale is a set of sounds, typically ordered according to their dominant sound frequency. For western music there are valid 12 basic sounds, each one characterized by its frequency and commonly called 'note'. The multiples and submultiples of each sound frequency are considered as instances of the corresponding note, thus all those instances are identified with the same note name. These basic notes are separated from the initial sound by a frequency factor of $(1 + \log_{12}(i/f_1))$, where f_1 is the lowest sound frequency at the base of the set or sounds. The complete set of 12 sounds is called the chromatic scale. In western music a melody typically uses less than the 12 notes of the chromatic scale. Actually most western melodies use 7 or less notes. There are several of these sets of seven notes in which it is possible to include most western melodies. These particular sets of seven notes are regarded as musical scales, and each one of them has an identifier, and there are categories of them as diatonic, major, minor and others.

organized by the size or duration of the symbols. Then music descriptions were interpreted at the scale at which the set of symbols lead to a minimal entropy [75].

6.1.2.1 The character's scale

To observe a description at the character's scale, the text is segmented as a sequence of N single characters. In Expression (6.1) there will be D different symbols Y_i , each of which will be an indivisible character. The random variable $P(Y_i)$ represents the probabilities of occurrence of each character Y_i . Since this scale is formed by symbols being of the same size, we classify this scale within the category of *regular size scale*.

6.1.2.2 The word's scale

At the scale of words, symbols are made by words or symbols having a comparable function like words within the written text. Words are sequences of characters (different from a space char) preceded and followed by a space or a punctuation sign. There are several considerations to make a precise interpretation of the elements of a text when the scale of words is adopted. Punctuation signs, as considered above, serve as word delimiters. But they also modify the context of the message. Therefore, the punctuation signs have some meaning and should be considered as words themselves. In general, words as symbols are written with lowercase. In English and Spanish the use of capital letters at the beginning of a word indicates it is a proper name or the beginning of an idea just after a period. Still, a word initiating a sentence and written with its first uppercase letter could be a proper name, and thus should be considered a different symbol from that written with the same sequence of letters with all its characters in lower case. An infallible criterion to recognize words, written with subtle variations, as different symbols is nearly impossible. Nonetheless, we built algorithms to recognize most cases where symbol disambiguation is possible. The criteria used for those algorithms is presented in greater detail in a previous study [63]. In our present study we use the same criteria to recognize and classify words as different symbols.

After recognizing all different words existing in a description, language **B** can be formed by assigning each word to the corresponding instance of variable Y_i . The random variable $P(Y_i)$ representing the probabilities of occurrence of each symbol, is determined according to the number of times the symbol Y_i appears in the text and the total number of symbols N . Symbols in the scale of words are basically determined by the presence of the space character which works as a symbol delimiter. The symbol lengths is not constant and thus we classify this scale within the category of *symbol delimited-irregular size scale*.

6.1.2.3 The fundamental scale

As explained in Chapters 5 and 6, the Fundamental Scale of a description is a set of symbols that minimizes the description's entropy as expressed in Equation (6.2). The set of symbols must not have any overlapping as they appear in the description's text. Additionally, when symbols are set one after another at their corresponding places within the text, the exact original description must be reproduced. An algorithm for the determination of the Fundamental Scale in one-dimensional descriptions have been presented by Febres and Jaffe [75]; we rely on it to evaluate the entropy and the complexity of descriptions at this scale.

6.1.3 Scale downgrading

The frequency profile associated to a language is a representation of the language. In a language made of D different symbols, this representation uses D values of symbol frequencies to describe the language. Plotting these values is useful because it permits to graphically observe an abstract description. Depending on the level of detail with which the observer appreciates the description, every value of the frequency of each symbol, may or may not be needed. If for some purpose a rough idea of the profile's shape is sufficient, a smaller number of values can be used. If on the contrary, the observer needs to detail tiny changes in the profile, a higher density of dots will be required to draw these changes of direction. Changing the number of symbols used to describe a system, constitutes a change of the scale of observation of the system; thus we refer to the process of reducing the number of values used to draw the frequency profile as *downgrading the language scale*.

If language \mathbf{B} introduced in Equation (6.1) is employed to build a N symbol long system description, then language \mathbf{B} can be specified as the set of D symbols Y_i and the probability density function $\mathbf{P}(Y_i)$ which establishes the relative frequencies of appearance of the symbols f_i . Thus, using p_i to represent the probability of finding symbol Y_i within the description, we have

$$p_i = p(Y_i) = \frac{f_i}{N}, \quad 1 \leq i \leq D. \quad (6.4)$$

At this point language \mathbf{B} is presented at scale D . To include the observation scale of a language as part of the nomenclature, we add a sub-index to the letter representing the language. Thus, language \mathbf{B} at some scale S , would be denoted as \mathbf{B}_S . To change the observation scale of language \mathbf{B}_D from its original scale D to another scale S ($S < D$), we use the transformation matrix $\mathbf{G}_{D,S}$. The sub-index indicate the original and the final observation scales. Whenever the index does not appear in the name of a language, it can be assumed that it is expressed at its original and maximum scale. That is $\mathbf{B} = \mathbf{B}_D$.

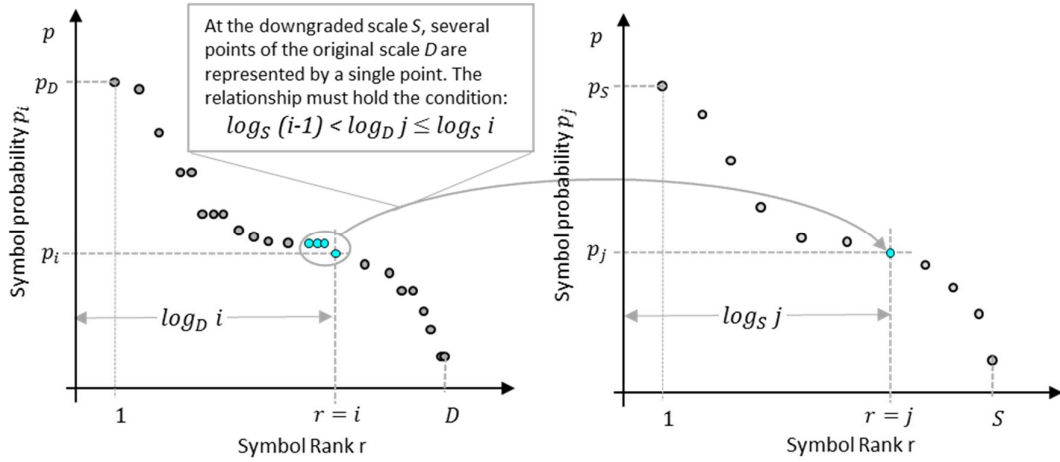


Figure 6.1: Graphic representation of a language scale downgrading from scale D to scale S ($S < D$). The total number of points at scale D , representing D symbols on the left graph, are transformed in S points when the language is represented at the scale S , as in the right graph.

Notice that in this context, a lower scale signals a smaller number of symbols to depict a communication system. That corresponds to a point of view from which less details—and therefore less symbols—are observed. Figure 6.1 illustrates how the symbols existing at the original scale D contribute to form groups of symbols which appear at such a probability, that the general shape of the frequency profile is reproduced at the smaller scale S .

Downgrading a language from scale D to scale S can be performed by multiplying the transpose of vector \mathbf{P}_D by the transformation matrix $\mathbf{G}_{D,S}$, as indicated below:

$$\mathbf{P}_S = \mathbf{P}_D^T \cdot \mathbf{G}_{D,S}, \quad (6.5)$$

$$\mathbf{G}_{D,S} = \begin{bmatrix} G_{1,1} & G_{1,2} & \cdots & G_{1,S} \\ G_{2,1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & G_{i,j} & G_{i,S} \\ \vdots & \vdots & \vdots & \vdots \\ G_{D,1} & \cdots & G_{D,j} & \cdots & G_{D,S} \end{bmatrix}, \quad (6.6a)$$

$$G_{i,j} = \begin{cases} 1 & \text{if } \log_D(i-1) \leq \log_S j < \log_D i \\ 0 & \text{otherwise.} \end{cases}, \quad 1 \leq i \leq D, 1 \leq j \leq S, \quad (6.6b)$$

$$j = \text{int}(S^{\log_D i}). \quad (6.6c)$$

This procedure for downgrading the language scale is useful given the frequent requirement of expressing text descriptions at the same scale.

6.1.4 Message selection

We applied our methods to messages expressed in three different types of communication system: natural languages, computer programming code and MIDI music. Table 6.1 shows the number of texts used for each type of communication system.

6.1.4.1 Natural languages

As messages written in natural languages we include 128 English and 72 Spanish speeches pronounced by politicians, military, writers, scientists, human right defenders and other public personages. For both, English and Spanish, the length of speeches range from about 200 words to more than 17000 words.

6.1.4.2 Computer programming code

Several computer programming codes, obtained from diverse programming languages are included as artificial language descriptions. Comments within the code are usually written in a natural language. Since recognizing these comments is easy, we could clean up most codes and leave them free of natural language comments. Nevertheless many programming language symbols are created after English and Spanish words. Therefore, avoiding the presence of some natural language words may not be possible. Table 6.1 shows the different programming languages represented in our experiment.

6.1.4.3 *MIDI* music

Polyphonic music is the result of the superposition of a vast variety of sounds. The information the sheet music may contain a relatively small number of sounds and effects. But the way music sounds, responds not only to the information written on the sheet music. It also brings information about the particular sound of the instrument, the ambient, minor deviations in the pitch and the rhythm, the addition of differences introduced by the interpreter, and innumerable effects, which despite not represented in the music score, are audible and part of the essence of music. Music as written in the sheet music is a discrete information phenomenon, but as it sounds is an analogous process which requires huge information packages to be faithfully recorded.

A discussion about the musical nature of *MIDI* is frequently found. At this point it is worthwhile to explain our dealing with *MIDI* music files and why we apply our

analysis to *MIDI* music instead of applying it to other types of music recorded files.

Table 6.1: Number of messages processed for English, Spanish, computer programming code, and *MIDI* music.

Meaningful Words				Abstract meaning				
		Genre/Class	Pieces	Authors				
English	Speech		128	108	MIDI Music	Total	438	> 93
	Sample: I have nothing to offer but blood, toil, tears, and sweat. We have before us an ordeal of the most grievous kind.					Medieval	38	12
Spanish	Speech		72	56		Renaissance	31	10
	Sample: Ni en el más delirante de mis sueños, en los días en que escribía Cien Años de Soledad, llegué a imaginar que podría asistir a este acto					Baroque	42	8
Programming Code	C		7			Classic	45	7
	C Sharp		20			Romantic	89	13
	HTML		2			Impressionistic	34	4
	Java		3			Twenty Century	35	8
	Mathlab		9			Movie Themes	18	> 4
	php		1			Rock	24	5
	Phyton		1		Hindu Raga	14	> 1	
Visual Basic		4		Chinese Traditional	12	> 1		
Sample: { class Program { void prime_num; long num; { bool isPrime = true; for jint i = 0; i = num; i++; { for jint j = 2; j = num; j++; { if ji != j && i < j == 0 { isPrime = false;				Venezuelan	56	> 20		
				Sample: #dnN Q E / # Nd Qd Ed -d !dn- !/_ #_nN Q E / # LX OX CX 1Z %Zn1 % 2U &UnL O C 2 & JL NL BL 4O (On4 (6J *JnJ N B 6 * E? I? L? @?7E +En7 + 4? (?nE I L @ 4 (E? J? >? 6? *?x6?n				

The digital musical interface *MIDI* is a way of digitizing music as is interpreted. The *MIDI* process converts music into synthetic music. The resulting sequences of discretized sounds are recorded in files with a large, though limited, number of symbols. These files can be regarded as *synthetic music* which provides a very compact code with obvious advantages and other not as obvious, as the possibility of adjusting, up to some degree, some of the components of the musical piece as for example the pitch or some instrument volume, and making the result more pleasant to the ear. Of course *MIDI* also has disadvantages. Typically the sound quality is far from the resulting from conventionally recorded music. However, even considering its quality as poor, *MIDI* music produce recognizable patterns of polyphonic sounds capable of preserving the pitch, the rhythm, the dynamics and even the timbre of the instruments coded in the recording. When listening to *MIDI* music files one can immediately recognize the piece, the instruments, the highs and lows and even detect any minor mistake

VI. Several Communication Systems viewed at different scales

or mistime introduced by the performer, or perhaps by the coding process due to the limits of digital resources. From the point of view of the sound produced, *MIDI* music has to be considered as a form of music; once we acknowledge the number of persons who today enjoy listening to it, what else would it be *MIDI* music?



Figure 6.2: A window showing a segment of the Maurice Ravel's Bolero. Three different arrangements of character strings indicate different passages within the musical piece. The entire file is about 180 times larger than the segment shown here.

Another terrain for discussion is about the *MIDI* code as a valid language for this study. Certainly inspecting the writing of any *MIDI* code in a file, it is possible to find some tokens referring to the instrument, rhythm, place where the sound appears or any other music parameter. *MIDI* files also include metadata at their beginnings and their ends, usually written in English or Spanish. The texts included in these tokens, headers and footers can be regarded as a sort of coding language with a very different nature from music, which may introduce fuzziness to the results. Fortunately, the length of the texts of the tokens, headers and footers are small compared to the total symbolic description length; since cleaning all files would represent a large non-automated task, we decided not to prune this small amount of noise and leave the files as they show when opened with a *.txt* extension. On the other hand, *MIDI* text files exhibit an organized sequence of characters associated with sounds. Just seeing any *MIDI* file the patterns of different passages of any music piece emerge.

Figure 6.2 illustrates a fraction of the file corresponding to a *MIDI* version of Ravel's Bolero. Here, as in any other piece observed, it is clear the symbols are

organized as time flows along with the music. Therefore these *MIDI* texts are codes of something that sounds when properly interpreted, therefore they are not encryptions. There is no need for preprocessing, as an encrypted script would have. Thus, these *MIDI* codes represent some sort of language susceptible to be directly read as the computer does. Whether or not *MIDI* music is regarded as real music, the coding of music in a symbolic *MIDI* structure represents a valid language to be studied and a splendid opportunity to test the concept of the Fundamental Scale as a method to recognize symbols in a language which operates with rules we are not unaware of.

Why not conventional music? One of the most popular encodings commonly regarded as conventional music. Is MP3. The same piece of *MIDI* music can be hundreds of times larger than its *MIDI* counterpart. Leaving no feasibility for applying the Fundamental Scale Algorithm as it is now. Then, taking advantage of the compactness of *MIDI* encoding, we calculated symbol diversity and entropy to hundreds of the almost unlimited *MIDI* music pieces available in Internet.

6.2 Results

Results are presented in three sets. The first compares diversity ranges. In a second set the resulting entropy is compared for the languages observed at different scales. In the third section we use entropies to calculate the complexity at the fundamental scale for the four types of language considered. In this section we also show an estimation of the length required for messages expressed in each language, so that the calculated properties settle down in a stable characteristic value.

6.2.1 Diversity

Figure 6.2 presents the diversity vs. the message length in symbols for the languages studied. Independently of the language, at the character scale only different elementary characters may represent symbols. Not being any possibility for combining characters to form strings, as the description length increases, the number of symbols rapidly saturates and cannot grow above certain number on some hundreds.

At the scale of words the graphs show similar results to those exposed in previous studies [63,71]; diversity behaves accordingly to the Heap's law. Since *MIDI* music descriptions do not contemplate meaning for words, therefore is no representation of diversity at this scale for music. At the fundamental scale diversity also follows the Heap's law; as the text length increases the diversity grows, but it grows at a lower speed for longer texts. Something to highlight is the

VI. Several Communication Systems viewed at different scales

dramatic reduction of diversity dispersion observable at the fundamental scale. Especially for English and Spanish, there seems to be a narrow band of diversity where the symbol diversity should fit in order to achieve a low entropy. For few texts the diversity falls outside this narrow band, but they should be considered as exceptional cases.

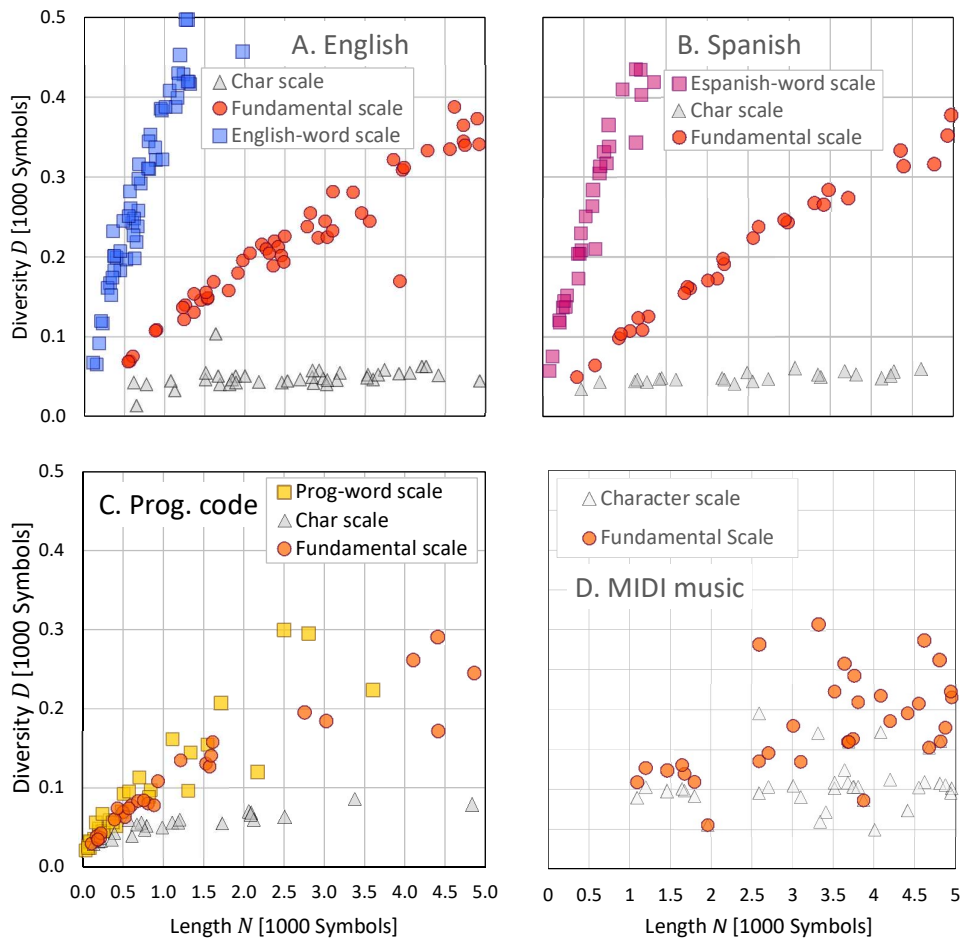


Figure 6.2: Diversity of as a function of description length measured in symbols. Descriptions expressed in several types of languages. A: English, B: Spanish. C: Programming code and D: MIDI music.

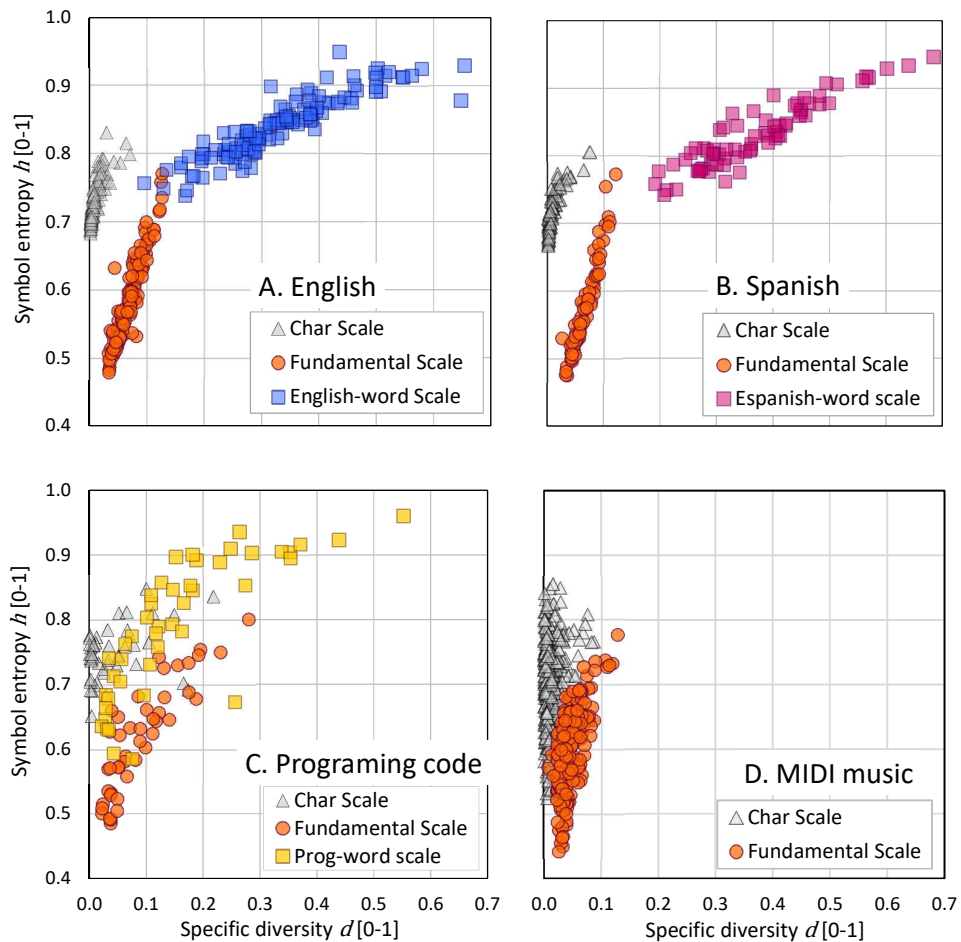


Figure 6.3: Entropy h as a function of specific diversity d . Entropies are shown for different observation scales for several types of languages: A: English, B: Spanish. C: Programing code and D: *MIDI* music.

6.2.2 Entropy

For each message represented in Table 6.1, we computed the entropy measured at the scales considered. Figure 6.3 shows four graphs where entropy is graphed against specific diversity for each language. The markers are shaped differently to facilitate the observation of the corresponding scales at which the calculus belong.

As with the diversity, at the fundamental scale the entropy occupies a narrow band in the space entropy-specific diversity. Not surprising here since the symbols were selected precisely to minimize the resulting entropy. For programing code and music the minimal entropy space found is not as reduced as for English and Spanish. Perhaps a consequence of the less restrictive subtypes of languages considered as computer programing code and music.

6.2.3 Symbol frequency profiles

We compared the profiles of the four communication systems considered. Each communication system is represented as the set of symbols resulting from the union of all the independent pieces belonging to each system. Thus, four large descriptions were processed to produce a profile corresponding to English, Spanish, programming code and *MIDI* music.

Table 6.2 shows some properties of these communication systems considered as the union of all descriptions within each class. In order to reduce the number of points from the original number of symbols to 129, the Scale Downgrading calculation explained in Section 6.1.3 was applied. The Scale Downgrading is useful to normalize the scale of observation bringing the symbol diversity to a specified number while keeping the general shape of the frequency profile shown with log-log axes. Figure 6.4 shows the resulting symbol frequency profiles representing each communication system studied.

Graphs in Figure 6.4 allow to compare different communication systems at their fundamental scale. English and Spanish's profiles are very similar. The most frequent symbol is the space ' ', revealing that this particular character is for these languages, more than an actual symbol, part of the protocol used to indicate the start and the end of words. Both profiles exhibit two clearly differentiable ranges of behavior: a first rank range where the profile's slope increases its negative value, and a second rank range where the log-log profile's slope keeps nearly constant until no additional symbol exist and the frequency profile drops suddenly. Even though programming code and *MIDI* music exhibit a softer transition between these phases of behavior, they do show changes in their profile shapes according to the range of symbol ranking where it is observed.

For natural languages, English and Spanish, the transition between the two profile ranges appears as a nearly straight segment connecting them. Since single-character symbols fit in every place the character appears, they are useful to fill the interstice left in between longer symbols formed by several characters. Thus, it should be expected the communication system's alphabet and the punctuation signs to occupy, most of the head of the frequency distribution range, leaving the range of the tail for the longer and less frequent symbols. To ease the visualization of this effect, Figure 6.4 shows tags with the lowest ranked single-character symbols as well as the highest ranked multi-char symbols. Notice how these tags indicate the location of the profile's transition for English, Spanish and programming code, suggesting that the change of profile behavior is related to the number of characters forming each symbol.

Table 6.2: Properties of different communication systems considered as the union of all messages expressed in English, Spanish, computer programming code, and MIDI music.

Communication systems' properties								
	Length N	Diversity D	specific diversity d	Entropy h	Approx. transition from single-char. symbol to multiple-char. symbol			
					Highest ranked multi- char symbol		Lowest ranked single- char symbol	
					Symbol	Rank	Symbol	Rank
English	1626927	7597	0.00467	0.430	of	34	Y	145
Spanish	984044	4688	0.00476	0.440	que	32	G	152
Prog. code	958547	2964	0.00309	0.541	maxChild	52	Y	247
MIDI music	14592192	22266	0.00153	0.564	Ã,Â _i	1	=	432

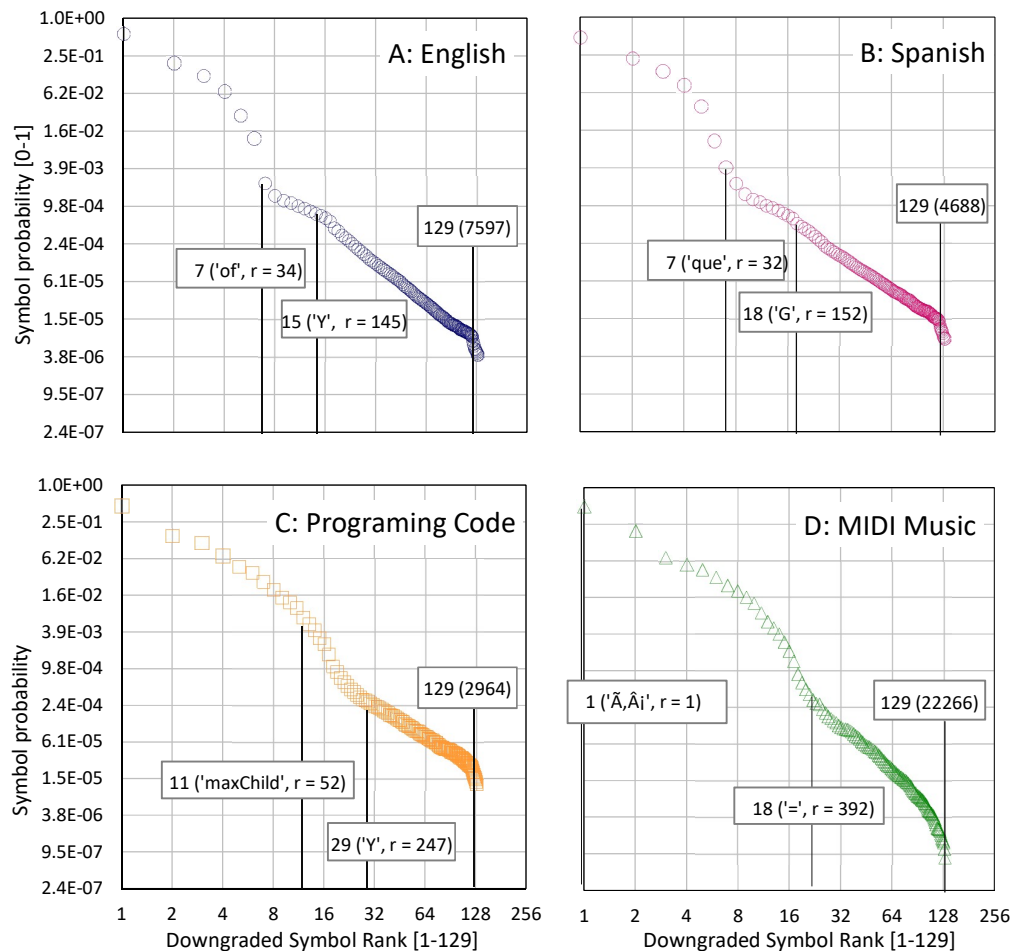


Figure 6.4: Probability profiles for several communication systems. A: English, B: Spanish, C: Programming code, D: MIDI music. All symbol rank axes are downgraded to the scale of 129. Numbers in tags show the symbol rank of the first multi-char. symbol and the last single-char. symbol as well as their corresponding symbol rank at the original scale (before downgrading). At the right end of each profile, the tags show the downgraded scale and the non-downgraded (original) scale.

The profile for MIDI music, represented in Figure 6.4D, shows a different behavior from those formerly viewed. For music the tendency of single-character symbols to occupy the heist ranked positions is not as dominant as it is for the natural languages and programming code. In fact the most frequent symbol—the symbol ranked as $r = 1$ —is the 4-character symbol ‘ \tilde{A}, \hat{A}_1 ’. Thus, the *MIDI* music profile starts right away with the transition from a phase with only single-character symbols, which does not manifest, to a phase dominated by longer and less frequent symbols at the profile’s tail. For music as represented in computer files, there is no alphabet. Single characters symbols are not limited to the 26 or 28 letters of any alphabet. *MIDI* files, on the contrary, employ about 400 characters available in the Unicode character set. This explains why the transition range for music, ending with the least frequent single-char. Symbol, is around the 400th ranked symbol.

6.2.4 Stabilization length

Figure 6.5 shows the values of entropy for the communication systems studied. English, Spanish and programming code are observed at the word, character and fundamental scales. *MIDI* music is observed at the character and the fundamental scale. Graphs included in Figure 6.5 show how entropy at all scales tend to decrease with the description length. For character and word scales, entropy seems to diminish indefinitely. At the fundamental scale all communication systems require some text length in order to ‘develop’ the value of entropy. There appears to an asymptotic value at which the entropy to settles. We will call the entropy stabilization value h_s .

In order to estimate the stabilization value, we built models of entropy as a function of the message length N measured in symbols

$$h \approx h_{st} + \frac{1}{\mu \cdot N^\nu} . \quad (6.7)$$

The parameters μ and ν are adjusted to minimize the error respect the real values presented in Figure 6.5. The values determined for the best minimal squared error fit at the fundamental scales are the following:

English:	$h_s = 0.421$	$\mu = 0.301$	$\nu = 0.348$
Spanish:	$h_s = 0.419$	$\mu = 0.315$	$\nu = 0.348$
Programing code:	$h_s = 0.439$	$\mu = 0.997$	$\nu = 0.225$
MIDI music:	$h_s = 0.479$	$\mu = 0.213$	$\nu = 0.407$

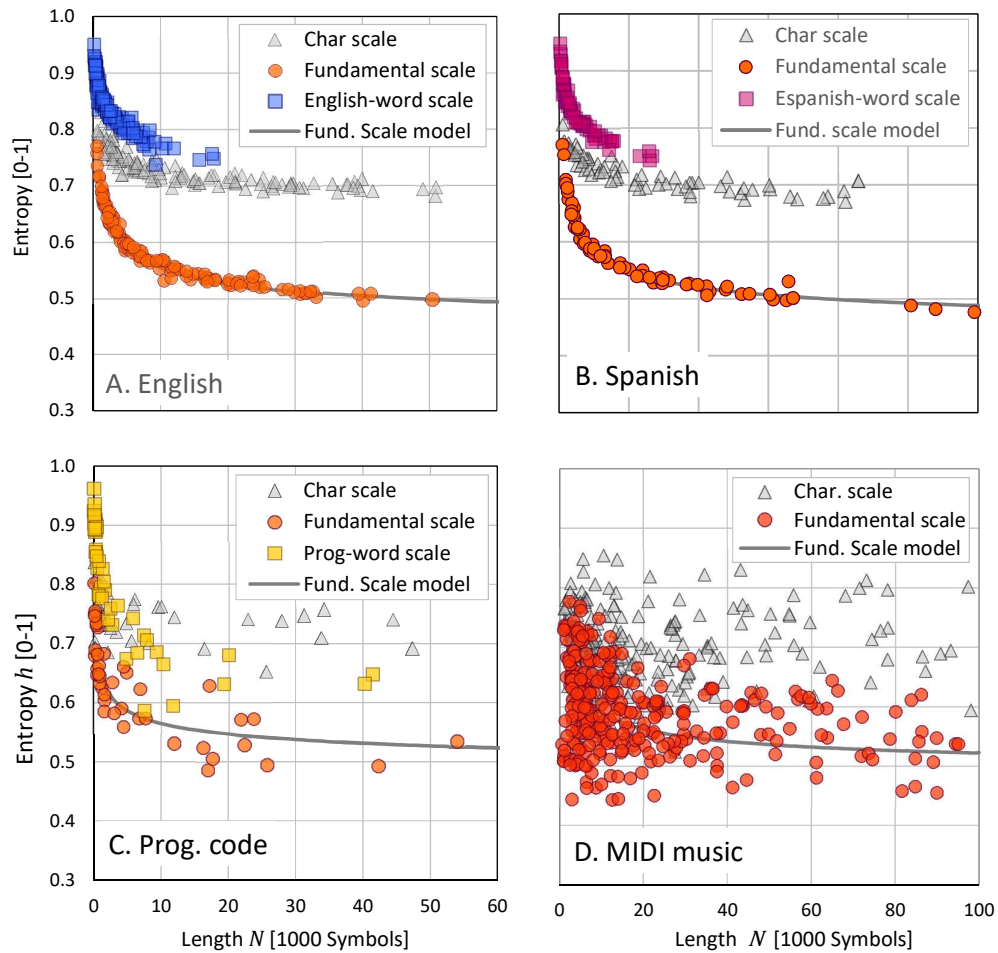


Figure 6.5: Entropy h vs. description length N in symbols. Graphs show the relationship between entropy and length for descriptions expressed in several types of communication systems: A. English, B. Spanish, C. Programming code and D. MIDI music.

Figure 6.6 shows the expected entropy values from very short messages to the range of long messages, where expected entropy values become almost static. The rate at which the expected entropy approximates the established value h_s , is an indication of the length needed for a communication system to organize itself and reduce the entropy to convey the message. We arbitrarily set the lower limit of this range as the length at which the communication system's entropy reaches 80% of its settlement value. We refer to that value as the Stabilization Length N_s and we measure it in characters. Once the considered stable range of entropy is numerically defined, the communication systems can be characterized by the specific diversity and the entropy found within their respective ranges.

VI. Several Communication Systems viewed at different scales

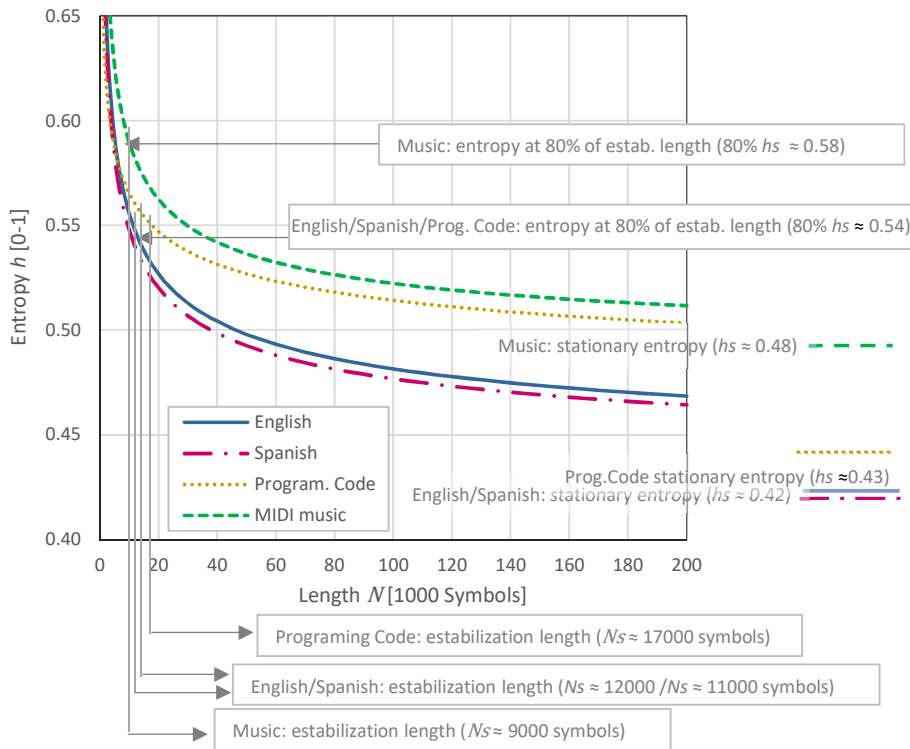


Figure 6.6: Model of entropy h vs. description length N in symbols. Graphs show the relationship between entropy and length for descriptions expressed in several types of communication systems.

Table 6.3: Average and standard deviation of the specific diversity and entropy for different types of communication systems, measured at the fundamental scale.

Specific diversity and entropy for several types of communication systems at their fundamental scales						
		Stabilization				
		length N_s	specific diversity d		entropy h	
Language Type	Specific Language	[symbols]	Average	Std.Dev.	Average	Std.Dev.*
Human natural	English	12000	0.0460	0.00720	0.5239	0.0066
	Spanish	11000	0.0459	0.00755	0.5167	0.0060
Artificial	Comp. prog. Code	17000	0.0321	0.00824	0.5173	0.0221
Music	MIDI music	9000	0.0398	0.01266	0.5732	0.0517

Student t-tests					
		specific diversity d		entropy h	
Distributions 1-2		n_{s1}	n_{s2}	p-value	p-value
English	Spanish	47	32	0.9598	0.0946
English	Comp.prog.	47	11	0.0011	0.4666
English	Music	47	190	< 0.0001	< 0.0001
Spanish	Comp.prog.	32	11	0.0011	0.9534
Spanish	Music	32	190	0.0005	< 0.0001
Comp. prog.	Music	11	190	0.0324	0.0001

* Standard deviation for entropy h is calculated with respect to the value returned by the Fundamental Scale entropy model (Equation 6.7) at each description length L measured in symbols.

The results of this characterization are included in Table 6.3. The Student t-tests p-values indicate that only for English and Spanish, their specific diversity and entropy of could come from similar distributions. All other combination of communication systems studied, are definitively different since the null hypothesis is discarded by the Student t-tests. The upper section of Table 6.3 includes averages and standard deviations for the specific diversity d and the entropy h of the communication systems studied. We shall highlight how smaller the standard deviation for the entropy of natural languages is as compared to the entropy's standard deviation of music and artificial languages.

6.3 Discussions

Human natural languages like English, Spanish or Chinese are syntactical. In their written form they are constructed with symbols with an assigned meaning. The meaning of a symbol may vary slightly from an interpreter to another. And it may slowly change with time. In fact, the symbol itself may even lay in disuse and totally disappear from the active version of the language.

Music language, in contrast, is made out of the effects obtained with combinations of sounds produced at different pitches, durations and time phases [13]. In spite of the audible essence of music, the sounds it is made of, can be coded as texts. In the context of this study we consider those music text-codes as musical language. Whether this conception of music language actually represents the musical phenomena, is one the questions this study intends to answer.

Artificial languages, on the other hand, are represented in this study as algorithms coded in different programming languages. Computer programming languages are designed to give instructions to machines, but they are constructed using human natural languages symbols to produce the structures which can be designed by humans. Computer programming languages are therefore a sort of instructional language based on fundamental language symbols. Artificial languages are specialized but very precise.

6.3.1 Diversity and entropy

Figure 6.1 shows diversity as a function of message length. For natural languages at the fundamental scale this function exhibits very little deviations, suggesting that at fundamental scale the diversity is a function almost exclusively dependent of the description's length. Similarly, Figure 6.2 shows fundamental symbol entropy as a function specific diversity. Again, for natural languages the entropy is dominated by the specific diversity.

At the scale of words, Spanish shows a slightly lower entropy than English; confirming previous results presented in [63]. But at the fundamental scale and at character scale, English and Spanish do not show any important difference. For programming code the diversity, as well as the entropy, both measured at the scale of words, show high dispersion. An indication that words have little or no meaning for this kind of language.

At its fundamental scale the dispersion of entropy reduces considerably but still is high compared to its counterpart for natural languages. The mix of many different programming languages in the same category may be an explanation. There is no word's scale for music. At its fundamental scale *MIDI* music shows the lowest specific diversity of all languages studied. This may be due to the nature of *MIDI* coding which in fact simplifies information during the digitizing process which basically consists of limiting the diversity of symbols associated to sound spectrum.

The values of diversity and the entropy computed at the fundamental scale, do not carry any distortion that may have been introduced by assuming a size of the scale, as is the case of the character's scale, or by assuming a symbol delimiter, as is the case of the scale of words. Graphs of diversity and entropy at the fundamental scale have been included In Figure 6.4 and Figure 6.5 to highlight the differences of those language properties at that scale.

For English and Spanish at their fundamental scales, the symbol entropy is proportional to the specific diversity. For music, on the other hand, entropy shows more dispersion. Perhaps a consequence of the diversity of music types included in the study which may behave like having several subtypes of languages into the same language group.

The diversity of natural languages grows with the message length, behaving with small dispersion around an average value which follows the Heaps law; as expected. For computer programming code, the diversity is definitively lower than the diversity for natural languages. But *MIDI* music, as a language, while exhibits a large dispersion of symbolic diversity values, shows its capability to incorporate much richer variety of symbols than any other of the languages studied here. Entropy variance is conspicuously low in natural languages compared to artificial language and music. This hints to special structure or order in natural languages that is absent in the other two communication systems.

As signaled in Table 6.1, we included different genres and styles of music; from all over the world and covering more than 700 years of music transformation. Thus the *MIDI* music set studied is itself, a very diverse data set. As with the computer programming codes, here we could regard our set of musical pieces as

expressions of many musical sublanguages, and therefore, a considerable deviation should be expected for most language properties studied. However, there must be other important sources of deviations for the values of language properties, since for English and Spanish, in spite of being different expressions of natural languages, the fundamental scale showed overlapping curves of diversity and entropy, as if they were languages structurally equivalent.

6.3.2 Symbol frequency profiles

The number of characters of the symbols, combined with their relative frequency, are definitively related to the shape of the symbol frequency profile. While in the most frequent symbol region—the head of the ordered distribution—the increasingly negative slope corresponds to a Gaussian distribution of the symbols frequency, the tail of the distribution—where syllabus, word segments, complete words and other multi-character symbols appear—shows a power-law distribution resembling the qualitative profile shape announced by Zipf [2] in his early work.

6.3.3 Description length

Languages evolve to respond to '*stimuli*' of different kind exerted by the environment. For a human natural language for example, it is intuitively clear how people are prone to use more frequently those words which are written and pronounced using less space and time; this effect is the main argument behind the Zipf's principle of least effort [2] and the readability formulas for English [50] and Spanish [60]. The connection between word-entropy and readability for English and Spanish was explored by Febres and Jaffe [71]. Their findings signal a relationship between average word-length and sentence-length and symbolic entropy. Moreover, Febres and Jaffe show this numeric relationship goes beyond a mere quantitative effect and actually represents the possibility for evaluating styles of writing.

Of all languages studied here, English and Spanish show the lowest symbolic entropy. Reinforcing the idea that natural languages have evolved to be effective using resources in the transmission process, and being effective in the mutual understanding and coherence of the information shared by emitter and receiver. On the other side, *MIDI* music shows a more entropic distribution of symbols. This implies less compact messages. Music patterns are interconnected in such a way that anyone can detect a sound that does not correspond to a melody or polyphonic set of sounds. Even when listening a musical work for first time, a reasonably trained ear person may have a precise idea of what the short-time horizon sounds are expected. This effect of music language may explain why the complexity establishment length is so little for music; most music

pieces expose their main theme rapidly; in very few compasses. Thus, the language of music requires shorter string lengths to develop its message.

For music the compactness of the message is not as important as it is for human natural languages. Music pursues other objectives not necessarily constrained in time and text length as English and Spanish do. Thereof musical expressions has evolved not to be short but to produce certain feelings and sensations.

Entropy stabilization value h_s is approximately 0.45 (range 0.419 to 0.479) suggesting an optimal complexity for all communication systems studied.

6.3.4 About the forces shaping languages

Our results show a connection among communication system properties as specific diversity and entropy. Observing the description at their fundamental scale, and obtaining the set of fundamental symbols, we were able to calculate the characteristic properties of communication systems presented in Table 6.2.

When focusing in the entropy of messages written in English and Spanish as represented in Figures 6.5 and 6.6, it can be seen that Spanish has a slightly lower entropy when compared with English. A result that suggests that Spanish, in spite of being formed by a number of words representing only a fraction of English words, is a more structured language. This is a consistent result with those obtained when language complexity was compared at the scale of words in [11], and with results shown in Figures 6.3 and 6.5 at the scale of words, where entropy values indicate a little more order for Spanish than for English. Explaining this slight but consistent difference between the complexities of English and Spanish, requires the inclusion of rigorous linguistics analysis and it lies beyond the purpose of this study. Nevertheless, we are tempted to mention that Spanish is the result of diversification process from Greek and Latin from which it inherited parts of its complex grammar from Latin. On the other hand, Modern English is the result of a conjunction of many dialects and old languages. Its evolution is then, characterize by its capacity to borrow words and to simplify, or to lose, grammar structures.

Natural languages and artificial languages have evolved to transfer and to record, precise complex ideas, the former, and precise instructions the latter. The effectiveness of both types of language rely on the presence of symbols with preconceived and shared meaning articulated by complex grammar rules to ensure the description in the contexts of place, time, actions, conditions, and all other elements that contribute to specify an idea. This functionality, together with the time they have had to evolve, explains the consistency of the entropy values obtained for natural languages.

In contrast, musical language uses its capacity to trigger emotions and sensations rather than to convey concepts with preconceived meaning. Music is perceived as sequences of sounds patterns. Even though an almost unlimited sort of sounds can be incorporated to music, and explicit rules govern the essence of music and must be present in any pattern of sounds, for it to be considered as music. But in music the meaning of sounds or patterns of sounds do not have to be predefined. Certainly, there is a connection between sounds, harmonies and music scales with emotions. But that is not the result of a conscious and rational decision; any listener is free to feel and interpret music in a particular and personal fashion. Having a different function from that of the natural languages, music is not anchored to keep its consistency as time passes; there is no meaning nor structure that music as a language has to maintain for long periods of time or large geographical areas. We think this *freedom* of music, specially manifested during the last two centuries, is the major factor that explains the vast variety of musical classes, genres, styles and even music definitions. Yet, within any branch of the music 'tree' at any time and region, music, as an audible phenomena, must obey a rather rigid network of relationships among its symbols which perhaps bounds the possibility from music being even more complex than it already is. In any case, Figure 6.5 shows how music exhibit a wide range of entropy values, at any range of the description length. Music initially results from the composer's feelings and inspiration; the composer *designs* his or her music to produce the desired emotions from the pattern of sounds. After being written in the music record, music tend to stick to the established sound structure defined as the style or genre. In music this structure seems to be governed by precise mathematical relations of sound duration and sound frequencies within the rhythms and superimposed accords which make polyphony music. When an instrument is played at an improper time or at an improper pitch, or plays an improper accord, the sound is considered to lose its beauty and in fact it may cause uncomfortable sensations for most listeners [77,80]. Yet some space remains free for the interpreter to alter the sound strictly described in the original music sheet, thereof any different interpretation of a musical piece adds —or subtracts— information to the musical description. Thus, the resulting entropy of a musical piece results from a personal way of using the language, initially imposed by the composer and then adjusted by the musicians who play the piece.

6.4 Conclusions

The character and word scales are the way we understand human natural languages; they allow us to learn and teach about them. The character scale and the word scale let us to organize complex languages into manageable components, but those scales do not seem to be the way languages, as

adaptive entities, organize themselves. On the contrary, the symbols forming a fundamental scale, while being a difficult to set to determine, reveal much of the essence of each language and is a good basis to establish comparisons among languages of different types.

The fundamental scale is formed by those symbols having a dominant role within a description. Being the result of a computation with no assumptions about the size or delimiters of symbols, and being capable of representing the original description at a minimal length, the Fundamental Scale represents the best single-scale to study one-dimensional languages. Other observation scales may alter the evaluation of languages with the assumptions about their scale and structure, and thus results could be biased or misleading.

Human natural languages have evolved to transmit complex description in a precise and organized way. Being breve without diminishing content and precision, have been always an important aspect of the symbol generation and survival in the for natural language evolution process. This principles, captured in Zipf's law and Flesch's readability formulas, have molded natural languages up to their current status.

Natural languages are more symbolic diverse than music. But natural languages are dominated by grammar in a degree so high, that they show very thin dispersion around average property values. Music language, in its written form has a very limited number of symbols. Yet, due to the variations introduced when music is played, the assembly of sounds which constitute polyphonic music, the different instruments timbers, the rhythm syncopation, and many other effects of music as it sounds, the resulting symbolic diversity of music is, by a wide difference, the highest of all the languages studied.

The objective of music is not evolve to be effective in the sense of transmitting information. It probably evolves with another underlying sense of beauty, very difficult to describe in a quantitative manner. However, there is a fundamental scale for the music language which drives the sound patterns to constitute music. The possibility of knowing about the fundamental scale for specific music types, allows for a deeper studies of music as a language and detailed comparisons of the different types and styles with which music can be written and played.

Finally, being complexity dependent on entropy, an optimal complexity for all communication systems might exist.

"La música pone orden al silencio."
Gabriel García Márquez

"After silence, that which comes nearest to expressing the inexpressible is music"
Unknown

Chapter VII

Music entropy models

We all share the intuitive idea of music as a flow of ordered sound waves. Formally the presence of order in music was studied by Leonard Meyer [77], who pioneered the analysis of music as a phenomenon capable of creating emotions. Meyer analyzed in depth the expectancy experienced by the listener. In his explanations Meyer used musical concepts and technical notations which are difficult to represent in quantitative mathematical terms. But the idea of music as a means to create specific sensations as tension, sadness, euphoria, happiness, rest and completeness, is always present along his study. Meyer described the emotions caused by music as the result of the interaction between the sound patterns perceived and the brain. In his words [77]:

"The mind, for example, expects structural gaps to be filled; but what constitutes such a gap depends upon what constitutes completeness within a particular musical style system. Musical language, like verbal language, is heuristic in the sense "that its forms predetermine for us certain modes of observation and interpretation."⁵ Thus the expectations which result from the nature of human mental processes are always conditioned by the possibilities and probabilities inherent in the materials and their organization as presented in a particular musical style."

Meyer's referral to conditional probabilities implies, at least from his point of view, the possibility of capturing some the essence of musical style by observing the

⁵ Edward Sapir, "Language," Encyclopedia of the Social Sciences, IX (New York: Macmillan Co., 1934), 157.

values of entropy associated with each music style. But music style has proved to be a difficult concept to handle; as for other languages, style is a way of classifying specific pieces of music according to many characteristics describing them and their source. Some researchers have set a style framework for music starting from values of those characteristics. In 1997 R. Dannenberg, B. Thom and D. Watson [80] produced readable *MIDI* files by recording trumpet 10-second-long performances. Dannenberg et al used neural networks to classify the style of each recorded performance according to several features of music. In 2004 P. J. Ponce de León and J. M. Iñesta [81] measured music components as pitch, note duration, silence duration, pitch interval, non-diatonic notes, syncopation, and other to build statistical characterizations of jazz and classical melody pieces. Perez-Sancho, J. M. Iñesta and J. Calera-Ruiz [82] approached the same problem by categorizing the texts of music *MIDI* files. They extracted the melodies from the *MIDI* files and segmented the resulting texts into sequences of characters representing different lengths of music beats. In 2004, P. van Kranenburg and E. Backer [83] study music styles starting from some music properties. But they include the entropy of some parameters as properties. All these studies indicate that it is possible to recognize properties related to the musical style in an automated fashion, but, none fulfills the required generality as to be considered a true style recognizer. Music style is just a too fuzzy concept to serve as a quantitative reference framework to classify with a single value something as complex as music.

From a more theoretical perspective, some researchers have provided useful schemas about the structures underlying music. In 2006 Mavromatis [84] presented models of Greek Chants depicting the melodic component of music as a process dominated by Markov chains. Later, in 2011 Rohrmeier [85] argues that that Markovian processes are too limited to properly model the complexity that arises when harmonies are added to melody. Rohrmeier proposes a Generative Theory of Tonal Harmony (GTTH) [85] as a set of recursive rules based on the Chomskian grammar and on the Generative Theory of Tonal Music (GTTM) by Lerdahl and Jackendorf [86]. Both branches of study, music as a phenomenon governed by Markovian processes, and the recursive context-free rules to model harmonies, are developed for music as written on the music-sheet. That is, music as an abstract entity represented by a set of meaningful symbols written on the music-sheet which are supposed to produce the sonic effects pretended by the composer. Even for this conception of music —its description on the music-sheet, which is simpler than recorded actual sounds— Rohrmeier points out in one of his notes in 2012 [87], that GTTH does not suffice to properly model the polyphonic music. On the other hand, in 2009 Mavromatis [88] suggested the application of the Minimal Description Length Principle (MDL) as an alternative to the Markovian models of melodies, and explained why MDL should be a powerful tool to describe music. Yet he announces these

advantages are subject to the huge computational complexity foreseen of the algorithms associated to this type of analysis.

Even for the most intricate pieces of music, the music-sheet is rather simple when compared with the actual music and the recorded file that can be produced when it is interpreted—the physical sounds we hear. The quantitative analysis of music is even more demanding if polyphonic music is the subject of study. Polyphony adds more dimensions to an already almost unmanageable problem. To deal with polyphonic music Cox [89] measured the entropy of the sound for each time beat. Cox represents his results in two time-dependent entropy profiles: one for pitch and another for rhythm. The polyphonic music can be described as the superposition of many monophonic sound streams. The result is an overwhelmingly large number of combinations of sound frequencies. Luckily, all these sound streams are synchronized in time and therefore its record in a file leads to a one-dimensional text where some character sequences may appear forming patterns that represent the musical elements contained in the text-file.

Working independently, Febres and Jaffe [75] developed the Fundamental Scale Algorithm (FSA). A method based on the MDL Principle applicable not only to music, but to most problems in which the recognition of patterns in a large string of written symbols, is an issue. The FSA is capable of unveiling the 'dominant' symbols of a description. In the present work we apply the FSA to 453 MIDI files containing academic, traditional, and popular music. For each piece, the Fundamental Symbols—the set of symbols leading to the description minimal symbolic entropy—was determined, and the symbol frequency profiles built. In order to compare the shape of profiles based on different number of symbols, a method is devised and presented.

Additionally, a measure of Higher Order Entropy and a method for its calculation, is proposed. We used these methods to represent different types of music in an entropy-diversity space. The dependence between the type of music and the selected representation-space is analyzed.

7.1 Methods

Understanding the structures underlying music is an old restlessness, always present among researchers. Starting with Meyer [77] and more recently Huron [90] link music structure with our emotions and expectations. Their description of musical structure and its influence in our emotions is based on the explicit musical language considerations. Using other analytical resources, a group of researchers, Mavromatis[91] among them, offer models for the construction of melodies assuming that a Markovian process is behind the specific melody's

style. These models, based on Finite State Machines (FSM) generalized in stochastic terms by a Hidden Markov Model (HMM) [92], which after being properly trained, are able to produce melodies that fit into a certain music style. Extending the HMM to harmonies requires the identification of an inconveniently large number of states. As an alternative method Rohrmeier [85] proposes a system of grammar rules to model harmonic progressions, an important extension of Lerdahl and Jackendoff [86] previous work and their Generative Theory of Tonal Music (GTTM).

Music can be seen as a recursively nested group of structures (Rohrmeier [93]). Even considering just melody, music consists of kinds of fractal structures leading any attempt for its analysis, to a very complex task. Attempting to model polyphonic music 'amplifies' these difficulties so much, that Rohrmeier [93] considers it impossible.

In this study we propose a radically different method to study the structure of music. Instead of analyzing the symbols written on the music sheet which represent how the composer wanted it to sound —instruments, rhythms and tempo, scales, note pitches, keys, chords, temperament, volume, etc. — we look at the sound recorded from an actual performance, by reading the text associated to the computerized file containing the recording. To do this we inspect the sequence of characters of the computerized files viewed as texts. No matter how long the file is, this is not a simple task.

Music files contain character strings to represent sounds according to the coding system used and the selected discretization level. But, as opposed to natural language text files, the music files do not show words or symbols that we humans can recognize without the help of some decoding device. Therefore, to find some order within these symbols —sequences of characters— that are camouflaged with the surrounding text, we consider the entropy of each possible set of symbols. We claim that the set of symbols whose frequency distribution corresponds to the lowest possible entropy value (or is near to), is a good representation of the structure of the language used for the description. We call this set The Fundamental Symbols, and the method used to its determination the Fundamental Scale Algorithm [75].

7.1.1 Language recognition

We applied the concept of Language Fundamental Scale. The fundamental scale concept let us obtain the set of symbols Y_i which can reproduce the description with such a frequency distribution $\mathbf{P}(Y_i)$ that the entropy associated results minimal. We refer to those symbols Y_i as fundamental symbols. The set grouping the fundamental symbols is regarded as the fundamental

language \mathbf{B}_* . The asterisk as sub-index is used to recall that \mathbf{B} is the result of an entropy minimization process.

$$\mathbf{B}_* = \{Y_1, \dots, Y_i, \dots, Y_D, \mathbf{P}(Y_i)\}, \quad (7.1)$$

In Expression (7.1) the diversity—the number of different symbols—is represented as D .

7.1.2 Specific diversity and entropy

The specific diversity is calculated as:

$$d = \frac{D}{N}, \quad (7.2)$$

where D is the diversity of language \mathbf{B} —the number of different symbols in the description—and N is the total number of symbols, repeated or not. A version of Shannon's entropy, generalized for languages compound of D symbols, is used to compute quantity of information for each music piece. The probabilities of occurrence of symbols Y_i are the components of the 1-dimensional array \mathbf{P} :

$$h = -\mathbf{P} \log_D \mathbf{P}, \quad (7.3)$$

7.1.3 The fundamental scale of a description

In those studies where the focus is on the music sheet, the analysis is limited to the music as the composer intended it to sound, but leaving out of the assessment of many other effects of real music which are present when it is performed with musical instruments. This study, on the contrary, is done with the recording of sounds as expressed in computerized music files. Subtleties as the effects of relative position of the instruments, their timber, syncopation, little mistuning, the performer's style and even errors, are represented in these files, up to some degree depending on the recording quality and resolution.

A music file read as a text, is a long sequence of characters which does not exhibit recognizable patterns, resulting in a code extremely difficult to interpret. Not knowing the rules of a grammar system it is not possible to decide a priori the scale to interpret the description. There are no words in the sense we are used to, and the characters we see do not indicate any meaning for us. We cannot even be sure about the meaning of the space character " ". To overcome this barrier we used the concept of Fundamental Scale [75]. The Fundamental Scale of a description is a set of symbols that minimizes the description's entropy as expressed in Equation (3). The set of symbols must not have any overlapping as they appear in the description's text. Additionally, when symbols are set one after another at their corresponding places within the text, the exact original description, must be reproduced. Once the fundamental

scale is obtained, the result is the set of symbols—strings of characters— close to the most efficient representation of the original description, and thus, useful to analyze the music there contained. The determination of the Fundamental Scale of a description is a combinatorial-order problem. An algorithm for its determination in one-dimensional descriptions have been presented by Febres and Jaffe [75]. We rely on this algorithm to evaluate the entropy and symbolic diversity of the music pieces included in this study.

7.1.4 Scale downgrading

The scale downgrading method as explained in Section 6.1.3 is extensively applied to compare the symbol probability profiles of music pieces described at different scales.

7.1.5 Higher order entropy

For an ordered symbol frequency distribution, entropy can be used as a general concavity –or convexity– profile index. To obtaining an indication about the oscillations of the profile around the middle line represented by the Zipf's distribution reference line, a new index must generated. We propose the entropy of the distance between the distribution profile and the Zipf's reference as the new index. Figure 7.1 illustrates the basis for the definition of this new entropy level.

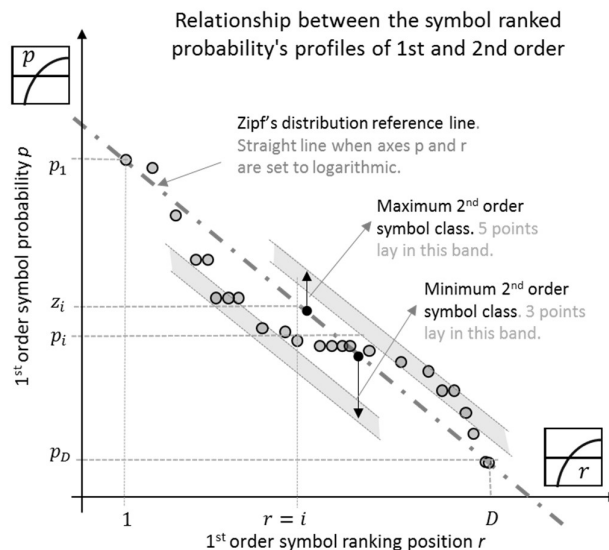


Figure 7.1: Typical symbol ranked probability profile with examples of 2nd order symbol bands. Each dot represent the probability of finding a symbol within all the symbols forming a system description. 1st order symbols are ranked according to their probability of appearance. The most common symbol appears in first place ($r = 1$) and the least frequent symbols appear at the end or tail of the ordered probability distribution representation ($r = D$).

To differentiate these two entropy calculations, we will call this *first order entropy*, or simply, entropy. We refer to the newly created concept as the *second order entropy*. For an ordered probability distribution profile, its first order entropy of is sensitive to its overall shape. Since any change of the profile slope needs to run along a wide range of the horizontal axis in order to impact the weighted area calculation, local changes in the profile slope are not effectively captured with the entropy. Second order entropy, on the contrary, is an index that focuses in the gap between the ordered symbol frequency distribution and the reference Zipf's distribution, it senses therefore the shape of the oscillations of the symbol probability profile.

To obtain a measure sensitive to small oscillations – or slope changes- we focus the distance E between the symbols probability and the imaginary perfect Zipf's distribution z_i that best fits the profile subject to study. The distribution z_i is calculated as follows:

$$z_i = \frac{k}{i^g}, \quad g = \frac{p_1 - p_D}{D}, \quad (7.4a) \quad (7.4b)$$

where g is the Zipf's distribution slope and k is a real number to establish the starting point on the Zipf line for the first ranked symbol. Notice that k is not necessarily equal to p_1 , as is usually presented. Here the value of k have to be adjusted to lead to a unitary area under the Zipf's distribution. The distance E_i between a symbol probability p_i and the imaginary Zipf's distribution z_i is presented as a one-dimensional array.

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_D \end{bmatrix} = \begin{bmatrix} p_1 - z_1 \\ p_2 - z_2 \\ \vdots \\ p_D - z_D \end{bmatrix} \quad (7.5)$$

As depicted in Figure 1, the size of these deviations around the Zipf's profile can define a new language: the *second order language*. To obtain the 2nd order language we need to define the smallest E_{min} and the largest E_{Max} and a resolution q to establish the size of the bands to classify the symbols between the values of E_{min} and E_{max} . After some arithmetic, these band boundaries can synthetized as the one-dimensional array \mathbf{B} as:

$$\Delta q = \frac{E_{max} - E_{min}}{q}, \quad B_i = B_{i-1} + \Delta q, \quad B_1 = E_{min} - \frac{\Delta q}{2} \quad (7.6a) \quad (7.6b) \quad (7.6c)$$

$$\mathbf{B} = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_q \end{bmatrix} \quad (7.6d)$$

Vectors and distributions associated to an order u are represented using a supra-index enclosed by squared brackets. The transition matrix \mathbf{U} to relate the distribution at order u with the distribution at order $u - 1$, is represented using the supra-index formed by the supra-index $[u, u - 1]$. The symbol probability distribution associated to the 2nd order language is represented by array $\mathbf{P}^{[2]}$ and obtained as indicated by Expression (7.8).

$$\mathbf{P}^{[2]} = \mathbf{U}^{[1,2]} \cdot \mathbf{P}^{(1)} . \quad (7.7)$$

$$\mathbf{U}^{[1,2]} = \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & & U_{1,D} \\ U_{2,1} & \ddots & \cdots & \cdots & \vdots \\ \vdots & \vdots & U_{i,j} & & U_{i,D} \\ \vdots & \vdots & & \ddots & \vdots \\ U_{q,1} & \cdots & U_{q,j} & \cdots & U_{q,D} \end{bmatrix} , \quad (7.8a)$$

$$U_{i,j} = \begin{cases} 1 & \text{if } B_i \leq E_j < B_{i+1} \\ 0 & \text{else} \end{cases} . \quad (7.8b)$$

In general, specifying the desired resolution at some distribution order q_u the distribution of any order u can be obtained starting from the preceding order $u - 1$ as:

$$\mathbf{P}^{[u]} = \mathbf{U}^{[u-1,u]} \cdot \mathbf{P}^{[u-1]} . \quad (7.9)$$

$$\mathbf{U}^{[u-1,u]} = \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & & U_{1,q_{u-1}} \\ U_{2,1} & \ddots & \cdots & \cdots & \vdots \\ \vdots & \vdots & U_{i,j} & & U_{i,q_{u-1}} \\ \vdots & \vdots & & \ddots & \vdots \\ U_{q_u,1} & \cdots & U_{q_u,j} & \cdots & U_{q_u,q_{u-1}} \end{bmatrix} \quad (7.10a)$$

$$U_{i,j} = \begin{cases} 1 & \text{if } B_i \leq E_j < B_{i+1} \\ 0 & \text{else} \end{cases} \quad (7.10b)$$

$$\mathbf{B}_u = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_{q_u} \end{bmatrix} \quad (7.10c)$$

$$\Delta q_u = \frac{E_{max} - E_{min}}{q_u} , \quad B_{u_i} = B_{u_{i-1}} + \Delta q_u , \quad B_{u_1} = E_{u_{min}} - \frac{\Delta q_u}{2} \quad (7.10d) \quad (7.10e) \quad (7.10f)$$

7.1.6 Music Selection

Music is the result of the superposition of a vast variety of sounds. But music sounds respond not only to the information written on the music sheet, but also the addition of small differences introduced by the interpreter. Music is then the

result of a vast number of different symbols to form sounds sequences. We relied on file *MIDI* coding to discretize the number of symbols. Most *MIDI* files include metadata at their beginnings and their ends, usually written in English or Spanish. The length of these headers and footers can be considered small compared to the total symbolic description length; since cleaning all files would represent a large non-automated task, we decided not to prune this small amount of noise and leave the files as they show when opened with a .txt extension.

Table 7.1 shows a synthesis of the music selection we used as subject to apply the entropy measurement method. The selection includes pieces from classical and popular music of different genres. Our music library is organized in a tree. To have some reference of the place where a music piece, or group of pieces, is located within the tree, we assigned a name to each tree level Table 7.1 shows this classification structure fed with more than 430 pieces from 71 composers and 15 different periods or types of music. As mentioned formerly, the Language Recognition algorithm is not tractable. Thus, most pieces were segmented in fragments of about 5 Kbytes long. The Fundamental-Scale Algorithm (FSA) was applied over about 3800 music fragments.

Table 7.1: Music classification tree and the data associated to the musical pieces considered.

MusicNet.												
Class	Type	Period/Style	Region	Genre			Spec. diversity		Entropy		2nd Ord. Ent.	
					Composers	Pieces	Ave.	Std.Dev.	Ave.	Std.Dev.	Ave.	Std.Dev.
Total					71	453						
Western	Academic	Medieval			12	40	0.062	0.026	0.649	0.048	0.949	0.037
		Reinainssance			10	31	0.048	0.016	0.622	0.037	0.935	0.041
		Baroque			8	55	0.039	0.013	0.581	0.057	0.911	0.050
		Classical			7	45	0.040	0.019	0.566	0.059	0.896	0.049
		Romantic			13	89	0.049	0.021	0.602	0.068	0.914	0.061
		Impressionistic			4	34	0.050	0.015	0.582	0.052	0.921	0.044
		20th Century			8	35	0.052	0.017	0.559	0.057	0.888	0.062
	Traditional	Venezuelan	Tradition	>20	56	0.049	0.014	0.540	0.056	0.929	0.036	
	Popular / Contemp.			Movie Themes		18	0.048	0.010	0.615	0.051	0.934	0.033
				Rock	5	24	0.041	0.010	0.585	0.043	0.919	0.045
			Jazz									
			Regie Tecno									
Asian	Traditional		Hindu-Raġ Raga	Several	14	0.083	0.019	0.697	0.061	0.974	0.026	
			Chinese	Several	12	0.048	0.015	0.582	0.038	0.915	0.046	

7.2 Results

All pieces and fragments of music were organized in a classification tree to which we refer to as *MusicNet*. By computing the Fundamental Scale to all leaves of *MusicNet*, we were able to obtain the fundamental symbols of each music piece included in our dataset, as well as for each music subset defined by composer, type, genre, period, or any other characteristic property of the included music.

MusicNet is too lush to be extensively presented here. But we include the upper levels of the tree in Table 1 and a link that allows access to the whole tree in Appendix G. Table 7.1 displays the datasets of *MIDI* music used for our tests and values of specific diversity, entropy and 2nd order entropy accompanied with their respective standard deviations.

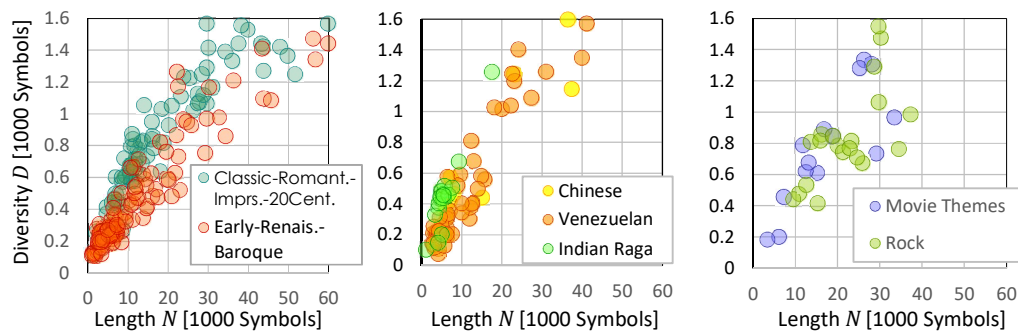


Figure 7.2: Diversity as a function of piece length measured in symbols for different classes of music.

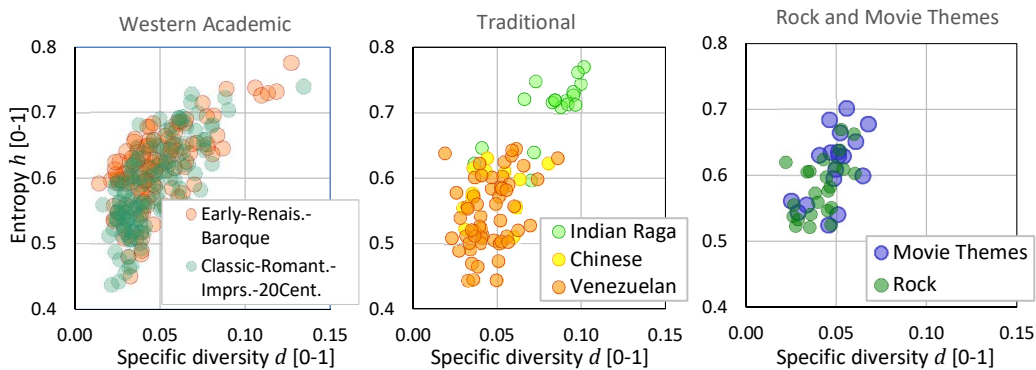


Figure 7.3: Entropy as a function of specific diversity for different classes of music.

7.2.1 Diversity and entropy

Diversity and entropy are quantitative characterizations of languages. Within the scope of a language, the diversity and the entropy may reveal differences regarding style or even period of its evolution. All pieces of our music library are

organized in three groups: occidental academic, traditional and Rock/Movie Themes. Diversity vs. length and entropy vs. length graphs are shown in Figures 7.2 and 7.3.

7.2.2 Information profiles

We are interested to know the effect of degrading the scale of observation of musical descriptions. Prior to this computation, we know that degrading the scale, equivalent to viewing the system from a remoter perspective, means observing less details, therefore, as the number of symbols use in the description decreases, we expect to get less information. Nevertheless, there are at least two reasons to inspect these information profiles: (a) to evaluate if they capture information about the music's type or class. (b) To obtain a sense of the minimal degraded diversity that maintains some of the essence of the system, by showing a shape that resembles the description at its original symbol diversity. Using this *minimal degraded diversity* allowed us to compare the shapes of many music frequency profiles at the same diversity; a condition needed for a fair comparison.

We included three examples of these information profiles. To obtain them we started from the description at their original symbol diversity D , and degraded the observation scale s by applying the methods explained in section 6.1.3. The results are presented in Figure 7.4. Downgraded values of the diversity were selected, so that at any scale the number of degrees of freedom⁶ of the information profile is a power of 2.

When comparing the information profiles at different scales for the example *Hindu-Raga.Miyan ki Malhar* with the other two music pieces, it is visually clear that, the Hindu-Raga piece differentiates showing a promontory in the profile at a diversity $S = 17$, that none of the other present at that scale. But the diversity $S = 17$ is not detailed enough to recognize the slight differences between the profiles of *Beethoven.Symph9.Mov_4* and *LAURO.Antonio.ValsVenezolanoNro3-Natalia*. In order to keep visually different profile shapes, among the three samples analyzed, we had to inspect the profiles with a diversity $S = 129$. With that level of refinement in the profile drawing, we were able to distinguish each music pieces' profile from another; we thus selected this diversity value ($S = 129$) as the diversity we should downgrade all pieces in order to obtain characteristic property values for each piece.

⁶ The number of degrees of freedom of any probability distribution is $k - 1$, being k the number of different categories in the distribution. Thus, the number of different symbols considered for each degraded symbol diversity is $S = 2^i + 1$, where i is a positive integer.

VII. Music entropy models

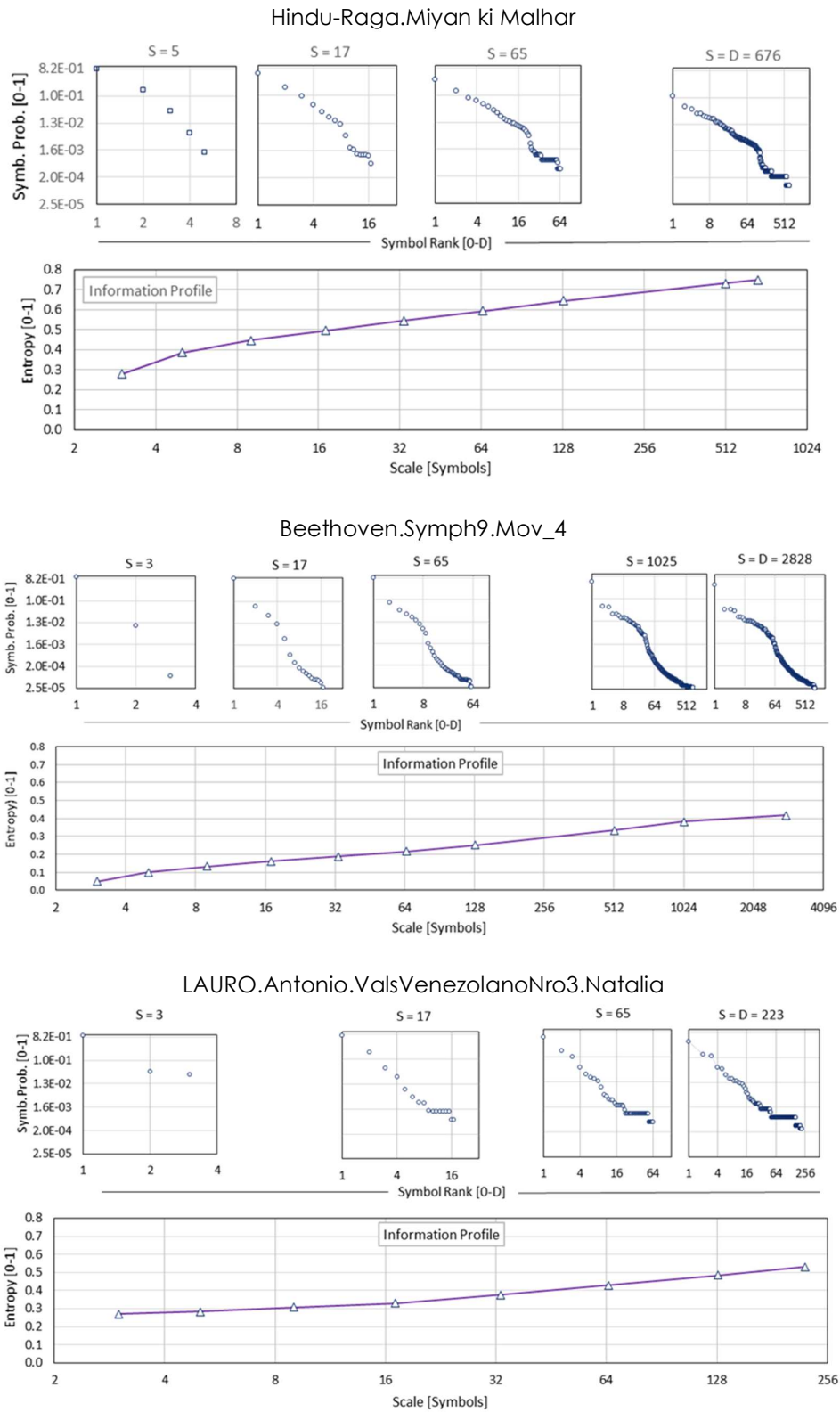


Figure 7.4: Variation of frequency profiles for several degraded scales and Information profiles calculated for three musical pieces.

7.2.3 Symbol frequency profiles

diversity we should downgrade all pieces in order to obtain characteristic property values for each piece.

A way to visualize the differences between two classes of music is to draw the ranked symbol frequency profile. Each profile has $D - 1$ degrees of freedom. That means the profile's shape can be altered in $D - 1$ different ways by modifying the frequency of the D different symbols which make the musical piece description.

Figure 7.5 shows the frequency profiles computed for each group of music included in our data set. All profiles were computed at a scale or downgraded diversity $D = 129$. The first seven graphs show academic music profiles ordered according to the chronological periods to which they belong. The last five profiles represent different genres and styles of music which represent popular music. A reasonable question with regard these downgraded symbol frequency profiles, would be: Are these profiles capable of depicting the organized change that might be produced by an evolution process of music? The seven graphs from Medieval to 20th Century music, suggest that the answer is yes. For most periods and music styles, the frequency profiles exhibit two easily recognizable regions: a higher ranked frequency region located toward the head of the ranked distribution, and a second region at the right of the ranked distribution, which extends until the symbol rank's cut-off value where sometimes an elbow shaped profile appears near the last ranked symbol at rank = $D = 129$. For Medieval music the distribution head's region occupies most of the profile range, showing a bow shaped profile. While the academic type of music covers the time until the classical period, this bowed section progressively shortens until the transition of the two regions reaches the middle of the logarithmic horizontal axis. The last tail elbow also softens till it disappears at the classical music profile. The slope at the transition zone also shows a gradual increase from the Medieval music, where transition zone is very soft, up to the 20th Century music, which shows a rather stiff transition zone. The vertical range of the profiles also grows as the time period progresses; with the only exception of Impressionistic music, all other considered styles of academic music, require a larger range of different frequency values in the vertical axis when compared with its previous music period.

When looking at traditional and popular music, we observe a shorter vertical range of values if compared against the academic music profiles. From all non-academic music considered, Hindu-Raga music exhibits the flattest profile while Chinese music has the steepest one.

VII. Music entropy models

Differences in the profiles suggest that it is possible to capture structural music differences by observing these shapes. On the other hand, profile similitude exists between some pairs of classes of music. Baroque music and Rock music have similar shaped profiles. Also, music from Impressionistic period and Chinese displays similar overall profiles. However, reducing the profile shapes down to a quantifiable index proves to be difficult.

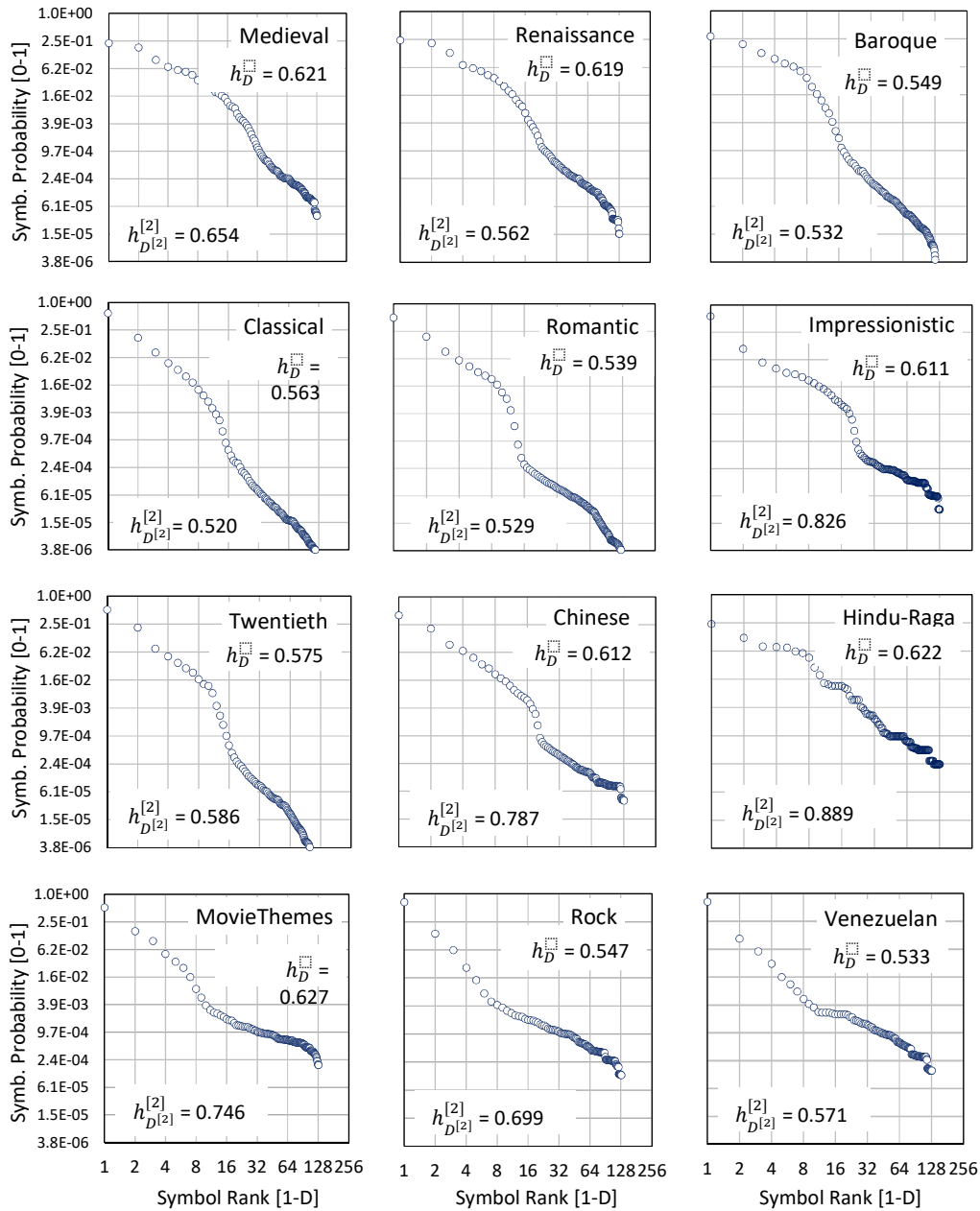


Figure 7.5: Symbol ranked frequency profiles for 12 different types of music.

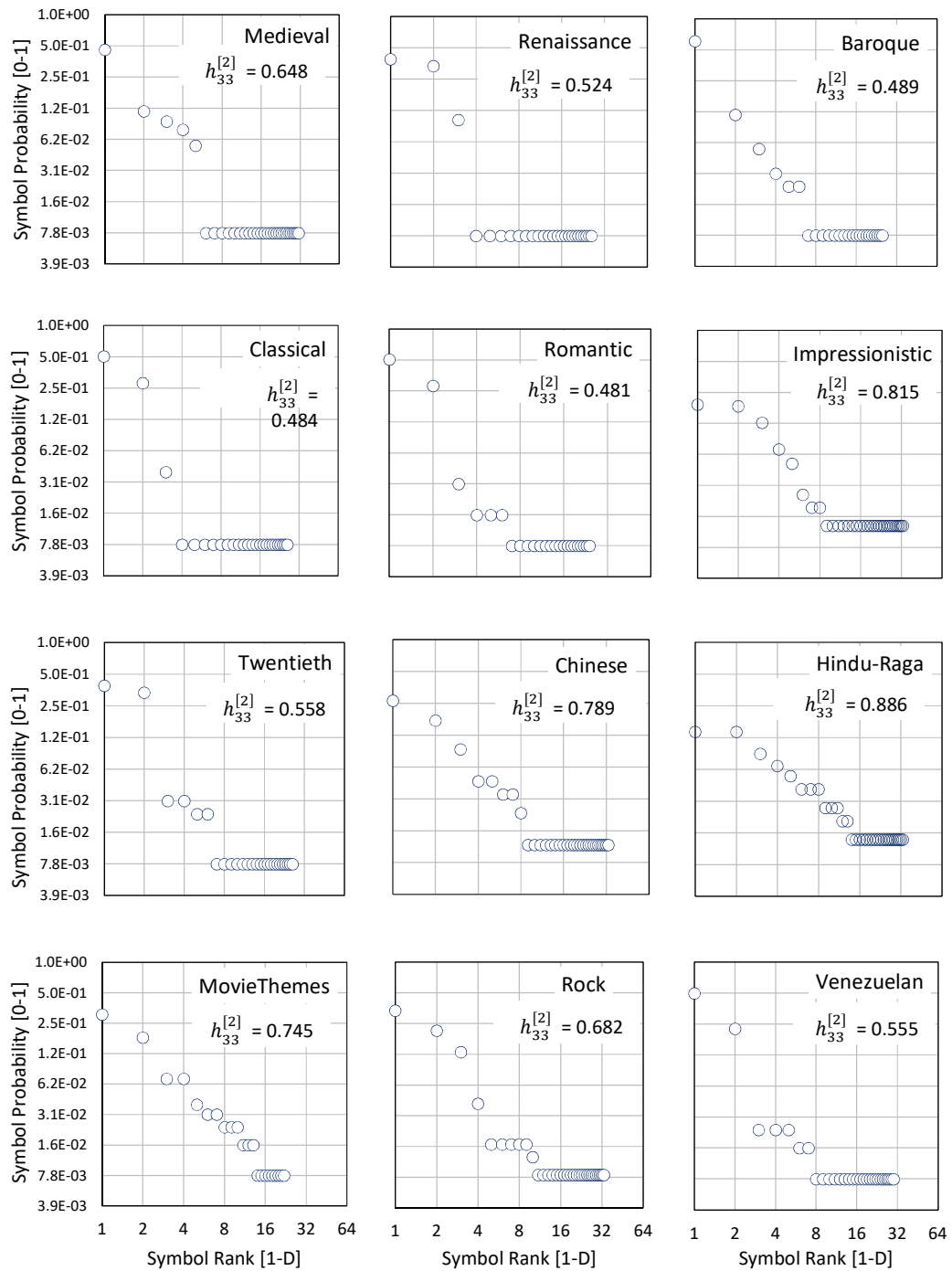


Figure 7.6: 2nd order symbol ranked frequency profiles for 12 different types of music.

7.2.4 Clusters and tendencies

Music can be regarded as a flow of information patterns perceivable as sounds. Different from languages built over meaningful semantic symbols as natural and artificial languages, music is made of totally abstract symbols. Perhaps a set of sounds forming a pattern can be assigned of certain meaning, but each symbolic sound has not meaning a priori. Yet, when we hear a song, we experience emotions and we can even describe the sounds with an endless list of adjectives as for example: soft, hard, violent or lovely. Thus, we should expect that any sound patterns can be characterized by its representation in the symbol profile and way the profile oscillates around its general downward slope, setting a sort of profile '*temperament*'. To capture this profile '*temperament*' we suggest evaluating 2nd order entropy $h^{[2]}$ as explained in section 7.1.5 and to extend the observation space from 2-dimensional, as was used in previous chapters, to a 3-dimensional space. Averages and standard deviation of the properties that characterize each type of music in our sample, were calculated. Tables 7.2 and 7.3 show the results.

Table 7.2: Properties of western academic music.

Properties of western academic music								
		Medieval	Renaissance	Baroque	Classical	Romantic	Impress.	20th Century
Num.Elem.		40	31	55	89	45	34	35
Specific diversity d	Average	0.0618	0.0479	0.0388	0.0403	0.0485	0.0500	0.0518
	Std.Dev.	0.0258	0.0159	0.0127	0.0190	0.0210	0.0150	0.0168
Entropy h	Average	0.6489	0.6219	0.5806	0.5661	0.6023	0.5819	0.5592
	Std.Dev.	0.0475	0.0373	0.0566	0.0592	0.0676	0.0521	0.0570
2nd order entropy $h^{[2]}$	Average	0.9446	0.9014	0.9085	0.8664	0.8521	0.8829	0.8917
	Std.Dev.	0.0320	0.0629	0.0499	0.0700	0.0945	0.1153	0.0679

Table 7.3: Properties of some traditional and popular academic music.

Properties of popular and traditional music						
		Hindu Raga	Chinese	Venezuelan	Movie Thms.	Rock
Num.Elem.		14	12	56	18	24
Specific diversity d	Average	0.0828	0.0476	0.0493	0.0485	0.0415
	Std.Dev.	0.0189	0.0153	0.0143	0.0104	0.0103
Entropy h	Average	0.6971	0.5818	0.5398	0.6150	0.5853
	Std.Dev.	0.0607	0.0380	0.0558	0.0511	0.0431
2nd order entropy $h^{[2]}$	Average	0.9539	0.8608	0.9259	0.8915	0.8577
	Std.Dev.	0.0288	0.0777	0.0614	0.0104	0.0706

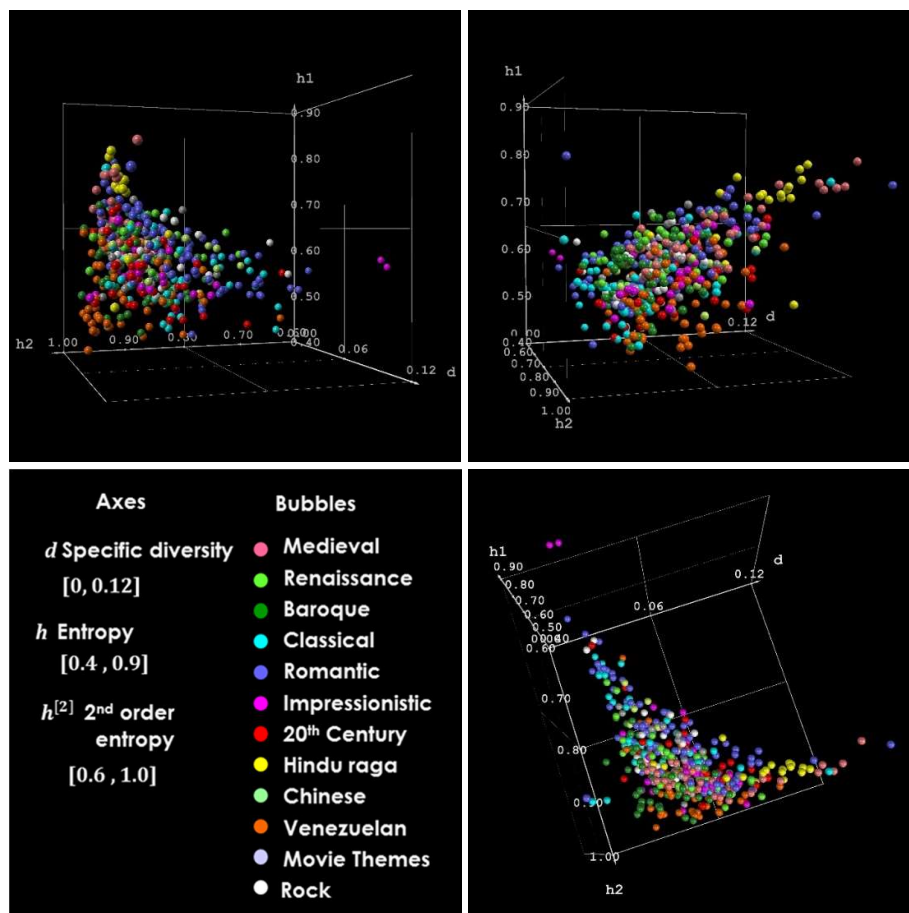


Figure 7.7: Three views of the representation of music pieces in the space specific diversity, entropy, 2nd order entropy ($d, h_D^{[1]}, h_D^{[2]}$). Each bubble represents a music piece. Each color represent a music style/period.

Plotting pieces of all periods from academic music in 3D graphs for diversity d , entropy h and 2nd order entropy $h^{[2]}$, reveal that different periods/styles tend to localize in different sector of this space.

To appreciate any tendency of specific diversity d and entropies h and $h^{[2]}$ over time, we plotted these variables as functions of time. The resulting graphs are included in Figure 7.9. For Chinese and Hindu-Raga music pieces we do not have information about the time when they were composed. We, therefore, did not include those types of music in these graphs.

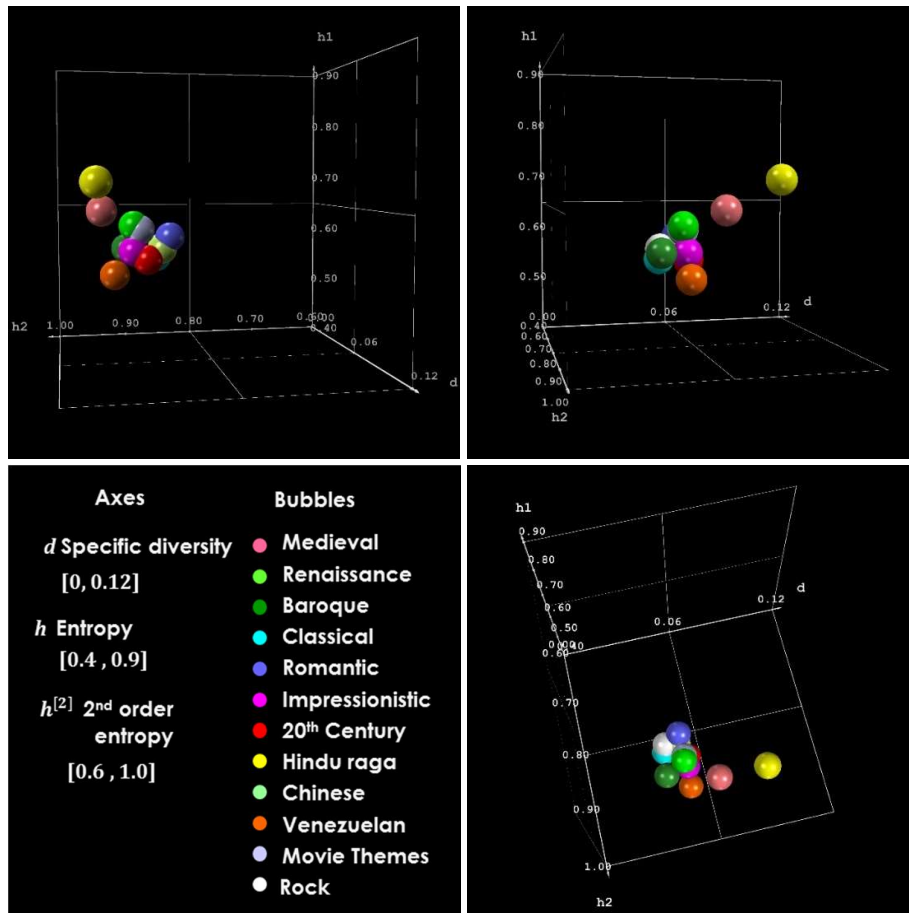


Figure 7.8: Three views of the representation of music period/style groups in the space specific diversity, entropy, 2nd order entropy ($d, h_D^{[1]}, h_{D[2]}^{[2]}$). Each bubble represents a group of music pieces sharing the same style/period.

7.3 Discussions

Music can be transmitted by sounds and by writing. But, the communication of music by writing lacks of its essence and does not produce, at least not for most people, the emotions and sensations associated to a pattern of sounds. Music writing shall be considered as a useful tool for composing, making arrangements, recording, and teaching music. Transferring musical information is also possible by means of music sheets or other kinds of music written representation. However, written forms of music, convey information instead of music. Yet, digital computers must create some sort of written version of sounds, in order to capture, reproduce, and register them or apply any other imaginable process to music. Depending on the variety of elementary symbols and the possibilities

for creating new ones by combining them, this coded representation of music may reach the number of degrees of freedom needed to actually represent music with the complexity that characterizes it.

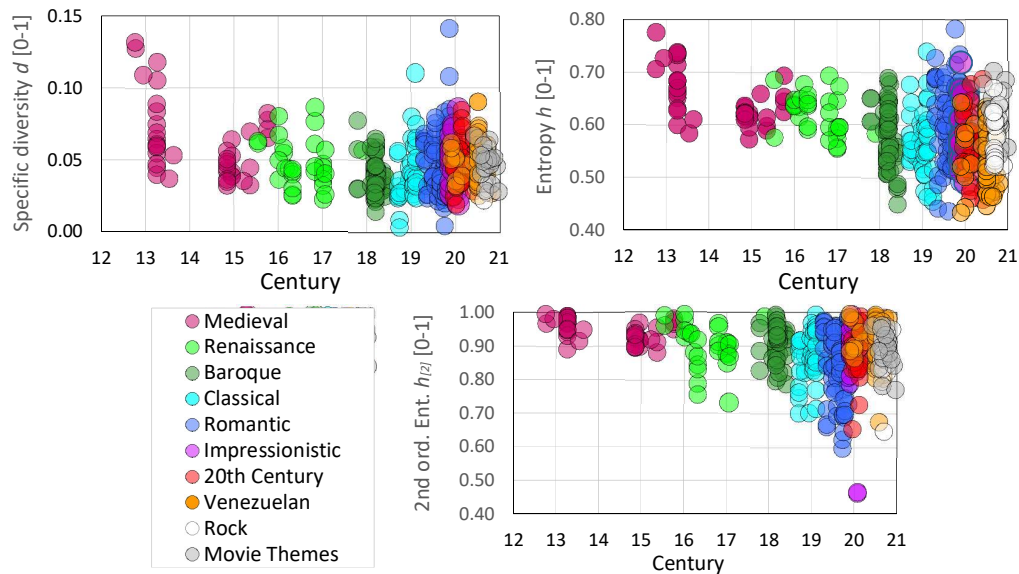


Figure 7.9: Variation of 2nd order entropy over time for several types of music

7.3.1 Diversity and entropy

The dependence of Diversity D vs. Length N is nearly linear. Only for short music pieces, the Diversity-Length curve shows slight concavity. For all other ranges, the Diversity D of music can be modelled as linear relationship with the length N of the music description. The slope change observed near the origin may be due to the English and Spanish overhead texts included as headers and footers in most *MIDI* files. These natural language segments are considered as noise and its presence should not have an important affect over the overall music description when the music piece is reasonably large in terms of symbols. Nevertheless, the specific diversity represented by slope D/N keeps close to a constant value for every type of music, becoming a characteristic value that may distinguish one type or style of music from another. Figure 7.2 illustrates how the point clusters for different types of music, tend to group around different lines, leading to different averages of specific diversity as shown in Tables 7.2 and 7.3. The value specific diversity measured for individual pieces ranges from 0.0183 (*Academic: Impressionistic: RAVEL.Maurice, Bolero2*) to 0.1341 (*Academic: Romantic: SAINTSAENS.Camille: CarnavalDesAnimaux: 08.Personnages LonguesOreilles*). Complete set of values can be found in Appendix G.

As can be seen in Figure 7.2, the graphs show that entropy is aligned to a very stiff slope in the plane entropy-specific diversity, and even though music entropy represented fills a wide range of values from 0.45 to 0.8, it seems to closely follow an average curve of the form $h = d^\alpha$, similar to those found for human natural languages in a previous work [11]. The large dispersion of entropy is then a consequence of the small range of symbol specific diversities where music establishes. Nevertheless, the values of the entropy standard deviation observed in Tables 7.2 and 7.3 are, in general, comparable with the range of entropy averages, thus entropy values capture some of the essence of the type or period of music and therefore justify its inclusion in a music entropy model. Values of the 2nd order entropy average go from 0.89 (*Academic: Classical*) to 0.97 (*Asian: Traditional: Indian Raga*). The standard deviation is about 0.05 and, in general, smaller than the range of variation of the average 2nd order entropy from one group to another.

7.3.2 Frequency profiles

Figure 7.4 shows three examples of musical pieces viewed from several scales of observation. For each example a series of graphs, each one with an observation scale, is presented. These sequence of observation scales graphs allows us to build the information profile included as the larger graph. These graphs have received two different, non-coherent, names. Researchers who consider Shannon's information [4] as a direct measure of complexity [14, 15], call it *Complexity Profile*. Those who consider complexity as the pseudo-equilibrium [2, 1, 16, 3] that the system reaches when it bounds its disorder by self-organizing its symbols, —we ourselves included—, call these graphs *Information Profile*. These names referring to the same graph, arise from the different interpretation of complexity. The first see complexity as proportional to the length of the symbolic description while the latter pays more attention to the system's activity to keep itself organized. Despite these names refer to different concepts, both seem to be valid. They just pay attention to different manifestations of complexity, whatever it is. The relatively low diversity of medieval, renaissance and baroque music, as well as its limited standard deviation, could be explained as the result of a music consisting of short pieces, played with few instruments and ruled by rather rigid musical rules. The growth of the number of musicians in the orchestra, the increasing appeal of more sophisticated sounds and arrangements, and the more flexible conception of music are the sources, perhaps, of the more complex expression of academic music in the later times. However, these explanations are difficult to prove since there is no evident connection between the fundamental scale and the audible and recognizable properties of music as it is perceived. The fundamental scale provides a way to determine the dominant structural symbols within a language, but those symbols generally do not describe the macroscopic, visible or audible, properties of the description.

Yet, there are reasons for us to consider our methods as consistent procedures to classify music styles and to quantify differences among them.

We traversed two paths for our calculus. In a path we inspected the shapes of the ordered frequency profiles for all types of music included in this study. By visually comparing them, we found similitudes between the profiles of different types of music; Baroque and Rock showed very similar shaped profiles, as well as the chronologically successive periods, Romantic and impressionistic, did. We also found that Hindu-Raga and Venezuelan music have the flattest and the steepest profile shapes respectively, locating their shapes at opposite extremes of a scale somehow built to evaluate these shapes. In another route for our method, we joined the types of music for which we could not prove a different statistical behavior. By applying this more rigorous method, we consolidated the groups shown in the header of Table 5. The two ways of looking at the original music groups, produces consistent results.

7.3.3 About the evolution of music

Figures 7.7 and 7.8 shows how each type of music occupy different sector of the space diversity-entropy. Focusing in the academic music it can be seen a progressive move from medieval, located in the sector of high diversity and entropy, to classical and impressionistic, located at relatively lower specific diversity and entropy. The ordered locations of each type of academic music upon the time parameter, suggests that some types of music evolve in a way that can be detected in the mentioned space; $(d, h^{[1]}, h^{[2]})$.

Hindu traditional raga and Venezuelan traditional music are easily recognizable. There must be some properties that make them well defined and characteristic. The fact Hindu-Raga and Venezuelan music appear far from any other style of music in Figure 7.7, does not surprise. On the contrary, it should be taken as sign of goodness of the space $d, h^{[1]}, h^{[2]}$ to represent music subtle differences, and confirming the prominent distinctions between the profile shapes seen for these types of music in Figures 7.5 and 7.6.

Graphs included in Figure 7.9 clearly indicate there is a general tendency of 1st and 2nd order entropies h and $h^{[2]}$. Both entropies show a tendency to lower with time. Despite the evident increase of dispersion of these indexes, which may hide the overall change over time, academic music's entropy has lowered from about $h = 0.7$ at medieval period to $h = 0.52$ at present. Figure 8, showing how the 2nd order entropy has changed, evidences that at all times the shapes of the profiles may oscillate around an imaginary Zipfian profile, in such a way to produce 2nd order entropies near the maximum possible. But beginning with the Baroque, music has developed to produce profiles associated with a lower

value of the 2nd order entropy; for academic music this tendency seems sustained from the medieval music up to the impressionistic period. Traditional and popular music exhibit a 2nd order entropy comparable to the academic 20th Century's music. The specific diversity d , on the other hand, reveals a slight reduction with time but an increase of dispersion of this variable, shown starting from the classical music and the romantic period, does not allow us to make a clear statement about the sustained tendency of a reduction of the specific diversity over time. On the side of traditional and popular music, specific diversity and entropy show less dispersion than their counterpart from academic music at comparable times.

Music is a reflex of social and cultural likes. We have strived to compare music styles over a quantitative basis. Our results reveal that for all the indexes used to characterize musical genres and styles, there is an increasing dispersion over time; perhaps the image of a society constantly committed to overcome any cultural barrier, thus making music an expanding phenomenon which grows in any direction of the space we use to observe it.

7.4 Conclusions

Music coded with the MIDI synthesizer produce texts susceptible to be analyzed using specific symbol diversity and entropy as variables which specify music type and even more subtle properties as style. The inclusion of higher order entropies accentuates the detectable differences between music styles.

We did not use any knowledge of the mechanisms of the MIDI coding process. We started looking at file texts that seemed to be totally meaningless and not decipherable. Discovering the set of fundamental symbols for each music text description we proved several important facts: 1: There is a fundamental symbol set that describes each piece of music. 2: The fundamental scale concept, presented in former works, is capable of determining the fundamental scale of machine coded texts as MIDI music text descriptions. 3: The scale downgrading method proposed allows for comparison of properties of systems of different nature and at different scale.

By applying the Fundamental Scale Algorithm, we have gone beyond the theoretical considerations about the Minimal Description Length Principle. We built frequency symbol profiles which work as quantitative descriptions for several hundreds of musical pieces. As the shape of these profiles is almost unique, they represent a 'signature' of the complete polyphonic sound, with all its subtleties and complexity, of each musical piece. After comparing our results for musical pieces according to their music style and period of time, we can affirm that our method works as a consistent procedure to classify music styles

and to quantify differences among them. Representing text descriptions in the space specific diversity, entropy and 2nd order entropy, presents as a promising tool for classifying system descriptions, with applications in many research fields as quantitative linguistics, pattern recognition, machine learning, and automated experimental design.

This novel quantitative way of analyzing music might eventually allow us to gain a deeper insight into the musical structures that elicit emotions, illuminating the working of our brains and getting a better handle on music.

*"Things are either black or white,
You see them grey when you are not focusing at the proper scale."*
Gerardo Febres Añez

Chapter VIII

Where is the information?

The 'anatomy' of a description has been the subject of intense discussion. Three abstract entities have been recognized as essential [94] [95] for the construction of descriptions in any language or communication system: resolution, scale and scope.

Surprisingly, there is not a unified definition of scale of a description. If scale is treated as a quantifiable concept; one that can be managed by the computer, the available definitions are even fuzzier.

Evolution is commonly understood as an adaptive process where the parts of a system change to reach a condition more suited to their immediate environment. Since the environment is made of other parts, competence for limited resources commonly arise as the center of the adaptation process. In some systems, some parts are unable to keep control of their required resources, and eventually fall in an inert condition or even disappear.

Complex systems offer difficulty to the recognition of all their components. Depending on the focus of the observer, some parts may shadow other and the boundaries of each part of the system overlap.

This study links the sense and meaning of *scale* with the set of symbols participating in a descriptive process. The concept of scale, along with its relationship to emergence and complexity, have been subject of research and discussion. Heylighen [94] presented emergence as a measure of the change of dynamics after a system transition. This measure cannot be directly made over the system itself but over models or observations of it. Bar-Yam [12,95] associated

complexity with information profiles⁷. In this sense, Bar-Yam identifies the relevance of the information that emerges when the system is observed from different detail level, as the essential cause of complexity. Ryan [96] depicts the relationship of scope, resolution and self-organization. Ryan considers emergence as the apparition of novel properties that a system exhibits when changes from a condition to another. The discussion focuses on scope and resolution, but *scale* is left as a slave property of resolution.

Prokopenko, Bochetti and Ryan [8] consider scale as parameter defining the emergence phenomena. However in their treatment state is almost the same as degree of detail or levels of resolution, thus diminishing the degree of independence that scale, as a concept, should have in relation to scope and resolution.

Fernandez, Maldonado and Gershenson [14] indicate that any change of the system's structure is reflected on the quantity of information needed to describe the system before and after the change. The change of the amount of information is a measure of the emergence between any two states. Fernandez et al. show how four numbers initially expressed in a sequence of binary digits, can be presented in a sequence of numbers expressed at different basis. The resulting entropy, computed for each string, clearly suggests there is an important impact of the language used in the effort the reader must apply to interpret the message.

The treatment of this problem —the observation of a system description— has been typically restricted to the idea of considering the scale, as a representation of the ways of grouping information elements into groups of regular shapes and equal size; in other words, topologically equivalent. This vision has proved to be of limited utility since it is a linear simplification of the resolution and therefore does not add freedom to model the consequences of varying parameters within the process of interpreting descriptions.

Here we offer a quantitative conception of scale, establishing clear differences with the concepts of resolution and scope. These concepts are intrinsically involved in the description of systems. However not making the appropriate distinction between them may restrict the possibility of studying systems at several scales.

⁷ The information profile of a system description is the graph presenting how entropy —Shannon's information— varies as the degree of detail seen from the point of view, changes. Thus, the information profile is the function of entropy vs the scale of observation.

8.1 Properties of descriptions: Resolution, Scale and Scope

8.1.1 Resolution R

Resolution is a human created artifact to split system descriptions in regular, equally sized, pieces. It results from the process of discretizing the description of a system. The original description can be discrete or can be an analogous depiction directly taken from physical reality. Resolution ends up being the number of equally sized pieces in which we divide the original description and thus it refers to the smallest piece of information of a description. It is commonly specified as the number of smallest information pieces that make each dimension of the description. Resolution can be regarded as the total number of elementary information pieces. In that case we use the letter R to refer to it. Resolution can also be specified as the density of information contained in a physical dimension. When this is the case, resolution is specified as the number information pieces r that fit into the dimension j considered, thus $r_j = R_j/Dim_{uj}$.

As an example we can consider a 16" x 9" computer screen with 1920 pixels in the horizontal longitudinal dimension and 1080 pixels in the vertical longitudinal dimension. The resolution R is regarded as a screen with a resolution 1920 x 1080 [pixels x pixels] and r would be 120 x 120 [pixels/in x pixels/in]. If the description refers to a 60-seconds long sequence of 36000 very short sounds, then $R = 36000$ [sounds] and $r=600$ [sounds/sec].

The concept of resolution loses meaning when the mesh of information elements is not regular—an information structure formed by a set of symbols with diverse sizes— can hardly be describe using the resolution as a characteristics parameter because the density of the resolution would not be constant and thus the resolution density becomes variable.

8.1.2 Scale D

All of us have an intuitive notion of scale. Commonly scale is associated with the 'distance' from which the system is observed. Thus the term scale is typically used to mean the system is being interpreted at a closer scale (higher scale with finer detail) or at a farther scale (lower scale with less detail) as if scale were exclusively defined by the distance between the observer and the object. As a consequence, the word 'level' found as a synonym of scale. Surprisingly, there is not a unified definition of scale of a description. Definitions of scale, treated as a quantifiable concept—one that can be managed by the computer—are rather fuzzy.

The concept of scale has been typically related with the number of adjacent pieces of information required in order to get a discernible meaning about the image or the description observed. Thus, there has been a tendency to treat scale as a measure of the number of information pieces needed to build meaningful information tokens. This conception of scale leads to its treatment as something linked to spaces depicted by topological congruent components, that is, regular lattices. This is a retracting way of interpreting the symbols we perceive.

During the last decade, several studies reflected the relevance the concept of scale in our interpretation of descriptions. In 2004 Bar-Yam [12,95] presents complexity as a property intimately related to the scale. However, his treatment of scale as a variable capable of varying continuously rather than discretely, perhaps leads him to present scale profiles, where entropy is a strictly monotonically decreasing function of scale. Piasecki and Plastino [72] showed entropy as a function of scale length—the size of the group of pixels making each object component—in a pattern of regular distributed greyscale pixels. The graphs show local minimal values of entropy when the scale length is a multiple of the characteristic size of the pattern, measured in pixels.

The scale is not absolutely dominated by the system properties. Even more, the scale is a product of our 'understanding processes'. When we consider a description, our brain probably scans several interpretations of the observed description. At each interpretation, we combine raw information by joining adjacent information elements and forming with them hypothetical larger symbols. Simultaneously we look for patterns which we can associate with previous experiences and learned notions, or even with our personal conception of beauty, thus giving certain meaning to a message that was initially abstract. Therefore, the scale is a property of the way the observer looks at the system. Once the observed system conception is organized in our brain, a clear account of the symbols resulting from our interpretation, along with their frequency of appearance and their relative position, constitute our model of the system.

This arguments let me introduce a definition of scale that does not contradict our previous intuitive notion: The scale of the system, as it is observed, is the set of different symbols used to create the system's model. Thus, when symbols fit into a regular lattice of pixels, for example, the number of pixels forming each one of these regular symbols specify the shape and the size of them. If this were the case, saying the system exhibits rectangles formed by $n \times m$ pixels could be appropriate to specify how we are looking at the system description, in short, to specify our scale of observation. But if we are looking at the countries of a map, there is no constant number of symbols that can be assigned to indicate we are looking at countries. The symbols here should be the countries shown in the map,

disregarding any number of pixels contained in any country; the scale would be the number of different countries seen in the map. If we see the same map at the scale of continents, the symbols become the continents and the scale the number of them.

Finally, I emphasize the quantitative notion of scale is identical to the symbolic diversity and therefore the designation of D is interchangeably used to refer to both Diversity and Scale.

8.1.3 Scope L

The scope refers to the total number of information units contained in the description. When the description is done over arrays on elements regularly distributed, the scope equals de product of the resolution of each dimension of the description. Up to this point scope this seems to be a redundant concept with resolution. However, when the information vessel is not a regular sized mesh, and resolution losses its meaning and utility, we still can use scope to characterize the description just counting the number of information elements contained.

Once the size and shape of the symbols have been established by the selected scale, the entire description conveys an amount of information determined by the total number of symbols repeated or not, included in the description. In this sense the scope equal the length of the description measured as number of symbols and thus, length and scope are both represented by the letter L .

8.2 Balance of information content

Encoding, transmission, decoding and interpretation are all different phases of the communication process. Information emerges from an idea and needs to be transformed and transported to achieve its function. In the transmission of a message representing an idea, a balance of the total ways information manifest must hold:

$$M = Y + S + E + N , \quad (8.1)$$

where M is the total information comprising the idea, Y the symbolic information, S the spatial information, E the semantic information and N any noise that could exist. Symbolic information Y is due to the relative frequency of the symbols used in the message. Semantic information E due to previously assigned meaning to symbols. Spatial information S is due to the relative position of the symbols within the message; the relative position of a sequence or in an array of symbols, may add or destroy some or all the interpretability of the set of symbols, spatial

information is therefore the vessel for grammatical content. Noise N is meaningless, arbitrary and random portions of information.

Some of the components of the overall message information are within our possibilities for estimation. To address these estimations, we can assume a communication system based on b different elementary types of pulses. Representing symbols as sequences of these pulses, each symbol within the message can be specified and transmitted using a string of $\log_b D$ *pulses_b* of information. The units of pulses have the sub index b to highlight each of these pulses occur in a transmitting system based on b different type of pulses.

A description made of a set of L_D *symbols* being represented with D different symbols, requires L_D signals. Therefore, a description consisting of a sequence of L_D symbols requires $L_D \cdot \log_b D$ *pulses*, where each pulse may adopt any one of b different signals. But, this sequence of pulses may not be organized in the best way, and the effort e_h to transmit the message might be greater than the actual information conveyed. According to Shannon [3] the compression factor that relates the effort of transmission and the actual symbolic information is the entropy h_Y based on the probability distribution of the appearance of the symbols. Therefore, the symbolic information Y_D using a serial transmission device is

$$Y_D = -L_D \cdot \log_b D \cdot \sum_{i=1}^D p_i \cdot \log_D p_i \text{ [pulses}_b\text{]}, \quad (8.2)$$

where p_i represents the probability of encountering symbol i within the description. The transmission effort e_h has the sub index h to indicate this it has been already reduced by a factor signaled by the symbolic entropy h of the original message. Thus, the information amount e_h is the result from compressing the message, and therefore the redundant—and not effective—amount of information that was present at the beginning, has already been removed from the account. Equation (8.2) may seem to defy from Shannon's information equation. But when the base of the transmission system is $b = 2$ and the diversity of symbols is $D = 2$ —as in a binary system—the *pulse_b* and *bits* are equivalent and expression (8.2) adopts the more familiar Shannon's expression

$$Y_D = -L_D \cdot \sum_{i=1}^D p_i \cdot \log_D p_i \text{ [bits]}, \quad (8.3)$$

Proper interpretation of the message is feasible if the symbols appear in certain order or position relative one another. After someone has organized the symbols, the communication process has to transfer symbols in a sequential manner,

while keeping the previously established relative position of the symbols. This is not always possible for descriptions of more than one dimension and certainly is not possible when, on purpose, the message has been disordered—encrypted—before its transmission. As a result, the description of the relative position of symbols is crucial to maintain the proper interpretation of the message at the end of its transmission. Given the fact the description of the relative position of the symbols is actually a description of the topology of the symbol-structure, I call this the spatial information S .

Equation (8.2) Indicates that symbolic information Y_D is dependent on the number of combinations of symbols. The number of possible states of a set of L symbols, each having as much as D different values, can be computed as L^D . This can be an enormous number. Fortunately the specification of the exact symbol-value frequency distribution, represented by p_i in Expression (8.2), enables us to reduce from that huge number of possible states, to a smaller, but still large, number of states which share the same symbol frequency distribution. Then, when the spatial information S is added to the symbolic information Y , the number of possible combinations of symbol sequences is reduced to the exact original description.

The number of coordinates needed to specify the position of a symbol in a space, depends on the number of degrees of freedom G characteristic of that space. A way to compute the spatial information S is to consider the number of degrees of freedom G of each symbol within the description space. Each degree of freedom for each symbol's position has to be specified with a number of $\log_b R_j$ pulses _{b} , where R_j is the resolution of the j 'th degree of freedom of the space. Thus, the positional information of a symbol can be transmitted with the product of the number of pulses required for each degree of freedom. For the whole set of symbols of the description, the spatial information requires

$$S_D = (L_D - 1) \cdot \prod_{j=1}^G \log_b R_j \quad [\text{pulses}_b]. \quad (8.4)$$

It is obvious that S_D can quickly rise up to big numbers, especially for descriptions expressed in multidimensional spaces. However, spatial information is just as compressible as the symbolic information was. In fact, since typically the descriptions refer to symbols located at the nodes of multidimensional nets, the regularity and symmetries of these spaces often represent a highly organized—with low entropy—structure, offering therefore, the possibility of high compressibility with which it is feasible to shrink this fraction of information into small code segments, usually called protocols of communication. Referring to the compressed spatial information as S_D , we can write

$$S_D = h_S \cdot (L_D - 1) \cdot \prod_{j=1}^G \log_b R_j \quad [\text{pulses}_b] , \quad (8.5)$$

where h_S is the entropy of the spatial information.

Multiplying each one of the pulses required to specify S_D by $\log_b D$, and setting the base of the transmitting system to 2, the spatial information can be expressed in bits as:

$$S_D = h_S \cdot \log_2 D \cdot (L_D - 1) \cdot \prod_{j=1}^G \log_b R_j \quad [\text{bits}] . \quad (8.6)$$

The semantic information E is related to the pre-established meaning assigned to the symbols. When we read an English written text, our interpretation relies in the meanings of the symbols—represented by words—included in the text. Since semantic information has an important degree of subjectivity, it results difficult to quantify.

Noise N is generally the result of losses or interference of information occurring during the transmission process. In experiments with controlled and fixed descriptions like those used in this study, noise can be neglected.

Information as a concept has been the center of dense and long discussions and polemics. The argument between Shannon [3] and Wiener [97] is perhaps the most outstanding example of it. While Shannon focused on the engineering problem of transmitting information, Wiener [98] argued that information is the knowledge the receiver could use on his or her favor. Obviously this is not the only discussion. During the mid-fifties Chomsky [28,99] started a field where information, or at least a portion of it, is regarded as the result of a syntactic phenomenon. Later Jackendoff [29,100], and Lerdahl and Jackendoff [86] viewed the phenomenon of grammar in a broader sense and extended their application to other types of languages as music. Whatever the truth is, a received message carries some information, but reproduction of the original idea depends on the receiver's ability to select the proper scale for its interpretation. This ability is the sum of the languages needed to understand symbolic, spatial, and semantic information.

Given the same ability to interpret the message presented at various scales, the total information M must add up to the same value. Thus the total information must be the same across several observation scales.

$$Y_{B1} + S_{B1} + E_{B1} = Y_{B2} + S_{B2} + E_{B2} . \quad (8.7)$$

We do not have yet an expression to directly determine the semantic information E . But we know that in binary language, based on only two symbols, there is 'space' to assign meaning only to those two symbols, therefore at the base of two, semantic information can be neglected, and then the semantic information at any base B can be estimated using information at $B_2 = 2$ as a reference.

$$E_B = Y_2 + S_2 - Y_B - S_B \quad (8.8)$$

Thus, Equation (8.8) says that semantic information is a function of the symbolic information Y and spatial information S . Having expressions to determine the symbolic and spatial information at several observation scales, allows us to quantify the elusive notion of semantic information E .

8.3 An information flow model

The information balance presented implies that information is integrated by components of different types which do not travel synchronously from the sender to the receiver. Nor the balance implies a unique sender as the source of information to be processed in order to interpret the idea being communicated. In fact all knowledge needed to interpret the symbolic information transferred, that is the language, including its components semantic and grammatical, must have been acquired by the receiver from a different source of information during a prior process. Figure 8.1 illustrates this information model. When the sender intends to convey an idea to the receiver, the first step is to convert the idea into an organized sequence of symbols. To attempt this, the sender must select a coding protocol—or language—so that the idea is represented synthetically by the mentioned sequence of symbols. Paradoxically, when we say or write something, we are actually filtering from the information that makes our idea, all the information needed to feel and understand the idea. After this filtering, we are just conveying a string of certain symbols—Shannon's information. At the other end of the process, the string of symbols will be properly interpreted if the receiver knows about the language selected to code the idea. While interpreting the set of symbols, the receiver adds the information that was filtered out just before the transmission by using his knowledge of the language he assumes the message is coded in. In the case of this language being coherent with the sender's language conception, the receiver interpretation will reproduce an idea congruent with its original version at the sender's mind.

The knowledge of the language forms from long term processes of experience and explicit learning. Language learning begins with a process of associations of experiences with symbol patterns. This forms an initial core with which it is

possible to continue the language learning by using the information transfer mechanism that now works thanks to the language being warehoused as knowledge. Language perfection is acquired by explicitly information which explains its syntaxes and grammar, transferred by means of a previous core of language already handled. Therefore, language is an ever perfecting entity intimately linked to our way of thinking and our intelligence. In a more external feedback loop, language itself evolves every time a new tested pattern of symbols shows to be more effective than its precedent version. In such cases, the incremental effectivity attracts the sender and the receiver to agree on the use of the new structure, and after this knowledge spreads out, it becomes part of the language.

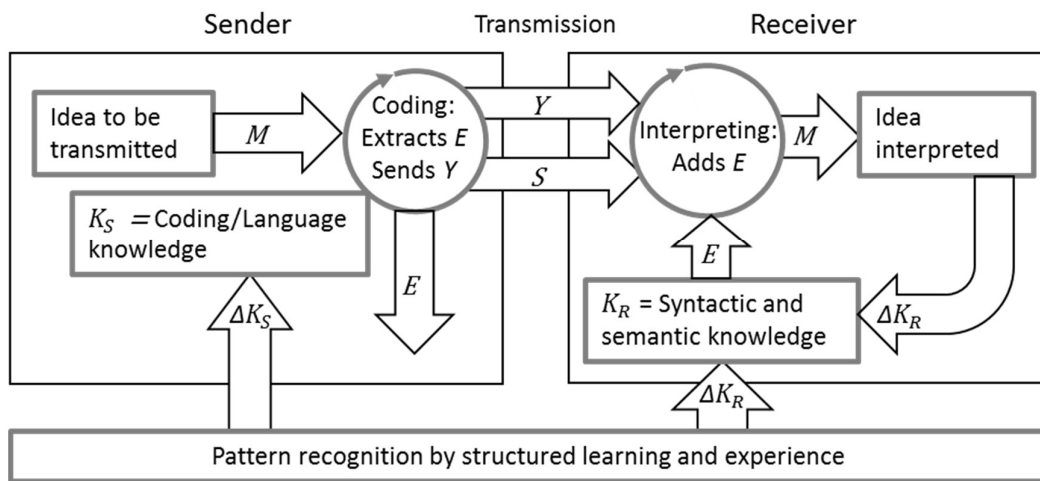


Figure 8.1: Information flow graphical model.

8.4 Finding the fundamental scale

The scale of a description has been commonly considered as groups of the most elemental information component used within the description. If, for example, the description consists of an image projected on a computer screen, pixels would be the most elemental component because each pixel shows the same color and therefore it does not make sense to divide it into smaller pieces. Now assume the screen is showing a bunch of letters, one after another forming an array of characters that fills up the screen. If we observe the letters, the scale would be represented by groups of adjacent pixels, each group forming images of larger size than a single pixel and containing a graphical description of a letter. If we decide that we are observing characters instead of pixels—which we have never stopped seeing—then we can say that our observation is at the Character Scale. It is also possible to think of the same computer screen description being represented by the binary codes of each characters instead

of the characters themselves. In fact, any text being represented in a digital computer, as we know them today, is only the reflex of some binary code physically stored as the orientation of micro magnets or the shine grade of micro mirrors contained in a magnetic or optical disk existing somewhere. If we could magnify the disk where this message is stored, we would actually see the same message at the scale of bits.

But the observation scale does not have to be made at a regular space. We actually think that when nature decides how to group its elements and form organisms and societies, it rarely pays attention to the regularity of the spaces that serve as frames for those compound entities. There are of course cases that can be seen as exceptions to this statement. The hexagonal pattern in a panel of bees or the fractal describing the appearance of a fern could be considered regular spaces, but in a description combining bees and ferns the resulting space would not be regular any more. Therefore, it would not be a good idea for a Fundamental Scale search strategy, to rely on the existence of regular symbol-space patterns.

When the communication system works on unknown rules it is not possible to decide a priori the scale to interpret the description. That is the case of the texts of any file containing recorded music. There are no words in the sense we are used to, and the characters we see do not indicate any meaning for us. Thus, we cannot even be sure about the meaning of the space character " ". In natural languages, a space is used as delimiter for words, but in music a space does not mean a silence. Fortunately, even having no idea about the 'grammar' of a communication system, we still can rely on the Minimal Description Length Principle (MDL) to reveal the symbols with which that communication system is built. The algorithm Fundamental Scale Algorithm (FSA), described in Chapter V, is essential to the determination of the set of symbols that minimizes a description's entropy.

8.5 Comparing languages at different scales

Comparing languages at different scales is like comparing apples and oranges; they 'live' in different dimensional spaces and therefore their nature can be radically different. Hence, if we represent a language with its n point symbol frequency profile, we can think of the shape of the profile as the *shape* of the language; there will be $n - 1$ independent ways to change the shape of the profile, thus say this is a space with $n - 1$ degrees of freedom.

It is possible to smooth a frequency profile while preserving its overall shape. This is done by removing some points from the profile. The selection of points to be removed must be done considering the density of points in each profile

segment, in order to keep the same level of detail in all sectors of the profile; this is the basis of the formulation of scale downgrading presented in Chapter VI. The standardization of scales down to a common language scale —or equivalent language symbol diversity— is essential to a proper comparison among them.

8.6 Some tests with different language expressions

The following sections present tests where the information of several descriptions are computed at different scales of observations. These tests allow the estimation of the information *flows* from a type to another in order to preserve the total message information implied in Equation (8.1). After recognizing the scope L , resolution R , and the scale D of the situation of each description or interpretation, the symbolic information Y , the spatial information S and the semantic information E are computed using Equations (8.3), (8.6) and (8.8).

8.6.1 Natural languages

We can read messages expressed in natural languages in two scales: the character and the word scales. Additionally we know the scale used to store a message in a computer file and to transmit it from a device to another, is the binary scale. The fundamental scale is an additional scale that has to be added to our possibilities for interpreting the message. Table 8.1 shows the results of evaluating information at these four scales for the small English text presented in Chapter V, Table 5.1 and for the Nobel lecture given by Bertrand Russell in 1950.

Table 8.1: Effects of different observation scales over the quantity of information of the little English text presented in Chapter V and Bertrand Russell 1950 Nobel lecture.

	Text: Little text Chapter V				Bertran Russell: 1950.Nobel Lecture			
Scale name	Binary	Chars.	Words	Fund.	Binary	Chars.	Words	Fund.
Data representation	0's & 1's	Letters and signs	English words	Symb. Min. entropy	0's & 1's	Letters and signs	English words	Symb. Min. entropy
Res. R [Symbols]	6256	782	varies	varies	260968	32621	varies	varies
Scope L_D [Symbols]	6256	782	171	578	260968	32621	6476	26080
Scope L₂ [bits]	6256	6256	1368	4624	260968	260968	51808	208640
Scale (Diversity) D	2 (0, 1)	38	82	80	2 (0, 1)	68	1590	1227
Symbolic Entropy h	near 1	0.808	0.903	0.763	near 1	0.705	0.822	0.518
Specific Diversity d	0.0003	0.049	0.703	0.138	0.00001	0.002	0.246	0.047
Symb. info. Y_h [bits]	6256	5055	1236	3527	260968	184009	42560	108034
Spatial info. S_h [bits]	78882	7506	1261	5294	4695713	489088	81979	382596
Semantic info. E [bits]	0	72577	82641	76317	0	4283584	4832142	4466051
Total info. M [bits]	85138	85138	85138	85138	4956681	4956681	4956681	4956681

8.6.2 Same symbolic structure. Different perceptions

Both mosaics shown in Figure 8.2 are built with identical number of pixels. Each is an array of 60x60 pixels some of which are dark or light colored. For both mosaics there are white or grey pixels inducing the interpretation of the figures in one or other manner. But the number of pixels and different colors used are the same, implying their symbolic information is the same.

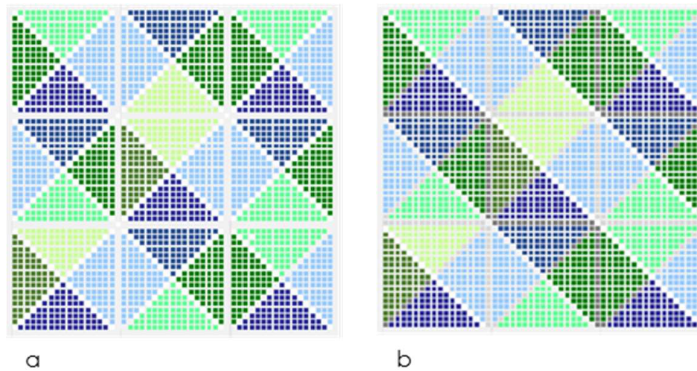


Figure 8.2: Two perceptions of a 2D mosaic with a resolution 60 x 60 pixels. Mosaic (a) shows pixels with 4 different colors. Same color pixels are grouped and separated by white pixels forming triangles. Mosaic (b) shows vertical and horizontal white lines dimmed to grey.

When estimating the account for information expressed in each type, however, there are different numbers which depend on the type of information considered. Interpreting the mosaics as set of 3136 pixels (56x56) each one represented with one out of four possible colors, we account for a symbolic information of 6272 bits, obtained applying Equation (8.2) and a spatial information of 18972800 bits, obtained applying Equation (8.6). The semantic information that can be stored in four different color shades is negligible. Thus the total information of any one of these mosaics is about 18979072. However, changing the focus from single pixels to the larger tiles suggested by the arrangements, Triangles in Figure 8.2a and bands in Figure 8.2b, the distribution of the types of information settles on the amounts shown in Table 8.2. Notice that resolution, scope and scale have different values for the three interpretations of these mosaics, but entropy equal one for all of them due to the uniform distribution of symbol frequencies.

Table 8.2: Properties of each interpretation of 2D patterns shown in Figure 8.2.

Figure	Fig. 8.2a	Fig. 8.2b
Scale name	Symbols	Symbols
Data representation	Triangles	bands
Resolution Rhorz	56	3
Rvert	56	3
Rangle	ⁱ	4 ⁱⁱ
Rcolor	4	4
Scope (Length) L	3136	36
Scale (Diversity) D	4	4 ⁱⁱⁱ
Entropy h	1.000	1.000
Specific diversity d	0.001	0.111
Symbolic info. Yh [bits]	6272	72
Spatial info. Sh [bits]	19662720	630
Semantic info. E [bits]	0	19668290
Total info. M [bits]	19668992	19668992

ⁱ This degree of freedom doesn't exist for single pixels
ⁱⁱ Only four angular positions are required.
ⁱⁱⁱ Square, triangle, trapezoids, large rectangle.
⁺ Square, triangle, trapezoids, large rectangle, small rectangle.

8.6.3 Partial changes of resolution and scope

Figure 8.3 uses a 2D example to illustrate the same picture observed for different combinations of resolution and scope. Fig. 8.3.a shows a set of 2D symbols over a 'surface' of 27 x 46 pixels. Here the squares have the role of elementary information and each of them may have one of two values: black or white, therefore the number of possible states for each square is $D = 2$. There are about 10 squares per inch, thus the resolution can be estimated as $r = 0.1 \text{ pxl/in}$. The number of possible states of this array of squares is determined by the length L and the diversity D , is $c = L^D$.

Thus, transmitting a message describing Fig. 8.3a requires L^D bits. However, according to Shannon the message could be compressed by a factor equal to the entropy h of the distribution of values of the squares.

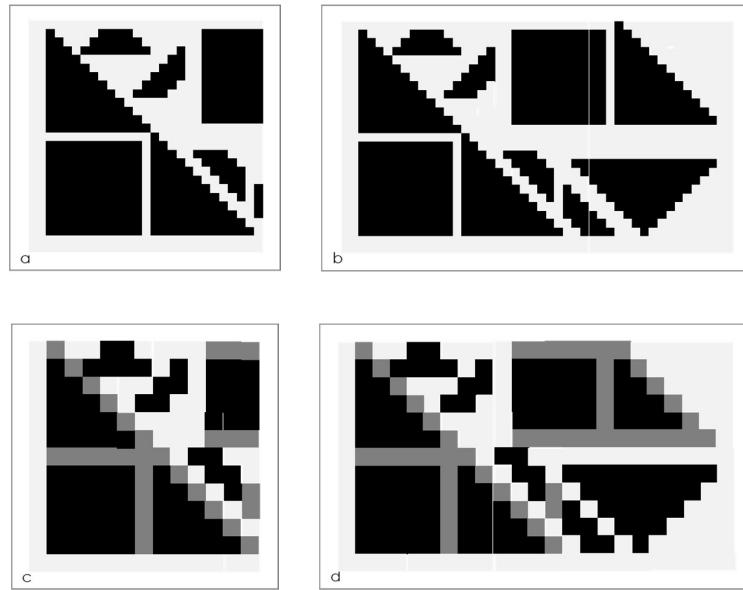


Figure 8.3: Effects of changes of resolution and scope over a 2D representation of polygons. Graphic representation of a language scale downgrading from scale D to scale S ($S < D$). The total number of points at scale D , representing D symbols on the left graph, are transformed in S points when the language is represented at the scale S , as in the right graph.

Table 8.3: Balance of information for the 2D example presented Figure 8.3.

Figure	Fig. 8.3a		Fig. 8.3b		Fig. 8.3c		Fig. 8.3d	
	Pixels	Symbols	Pixels	Symbols	Pixels	Symbols	Pixels	Symbols
Data representation	0's & 1's	Polygons	ones	Polygons	0's & 1's	Polygons	0's & 1's	Polygons
Resolution R_{horz}	27	27	46	46	14	14	23	23
R_{vert}	27	27	27	27	14	14	14	14
R_{angle}	- ⁱ	8 ⁱⁱ	- ⁱ	8 ⁱⁱ	- ⁱ	8 ⁱⁱ	- ⁱ	8 ⁱⁱ
Scope (Length) L	729	8	1242	10	196	8	322	8
Scale (Diversity) D	2	5ⁱⁱⁱ	2	3⁺	3	4⁺⁺	3	5⁺⁺⁺
Entropy h	0.985	0.928	1.000	0.646	0.929	0.813	0.931	0.861
Specific diversity d	0.003	0.625	0.002	0.300	0.015	0.500	0.009	0.625
Symbolic info. Y_h [bits]	718	17	1242	10	288	13	475	16
Spatial info. S_h [bits]	530712	94790	1541322	141734	60577	21952	163825	41869
Semantic info. E [bits]	0	436622	0	1400820	0	38901	0	122415
Total info. M [bits]	531430	531430	1542564	1542564	60866	60866	164300	164300

ⁱ This degree of freedom doesn't exist. Single pixels' angular position concept degenerates.

ⁱⁱ Only eight angular positions are required to describe symbols represented.

ⁱⁱⁱ Square, triangle, trapezoids, large rectangle, small rectangle.

⁺ Square, triangle, trapezoids.

⁺⁺ Large Triangle, rectangle, and noise: trapezoids, stairs-like polygon.

⁺⁺⁺ Large Triangle, small triangle, wedge, and noise: trapezoids, stairs-like polygon.

8.6.4 The impact of reorganizing

In this section I present a little experiment. The purpose is to assess the impact of organizing the symbols with which we interpret a description. Figure 5.4 shows arrays of 30 squares colored with five different intensities; the lightest named as 1 and the darkest named as 5.

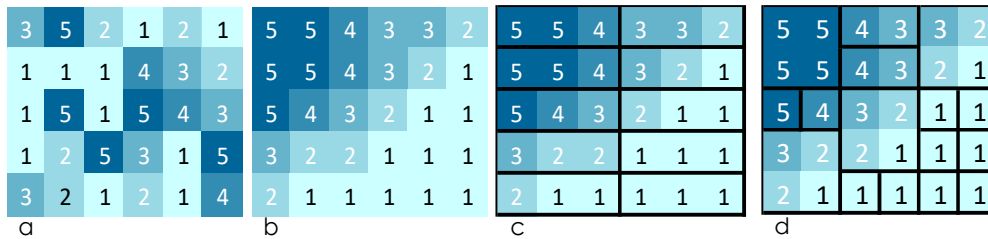


Figure 8.4: Four views of the same distribution of 30 squares colored with five different tones of blue. Number indicate each tone used. The lightest is represented by 1 and the darkest with 5. Each tone appears with the same frequencies in the three graphs. (a) Shows the 30 squared randomly ordered. (b) Orders the squares according to the rule indicating that no darker square can appear below or at the right of another square. (c) Shows groups of symbols formed by regular shaped lattice of 1 x 6 elementary bricks. (d) Shows with black borders the groups of squared forming symbols to reduce the entropy of this description.

In the leftmost array, Figure 5.4a, the squares are randomly organized while Figure 8.4b shows the colored squares ordered with the darkest at the top-left corner of the array and the lightest at the bottom-right corner. Figures 8.4c and 8.4d show the organized distribution of squares indicating different symbols formed by grouping several squares into each type of symbol.

This test shows the impact of the interpretation over the distribution of the different types of information. Interpreting Figure 8.4b as an array of 30 squares, requires just as much symbolic information as the disorganized squares presented in Figure 8.4a. Despite our unavoidable tendency to appreciate order in Figure 8.4b, if we consider all squares as independent single symbols, transmitting this information would require the same effort as for transmitting Figure 8.4a. But if we let our brain to group the squares in repeated patterns by degrees of color intensity (see Figure 8.4c), we reduce the number of symbols we have to consider and the spatial information associated to them. The possibility for associating semantic information also appears along with the variety of different symbols that now can be arranged. This transference of information from one type to another may be augmented (as illustrated in Figure 8.4d) or diminished with the grouping of the squares to form symbols of any shape.

Table 8.4: Properties of each interpretation of 2D patterns shown in Figure 8.4

Figure	Fig. 8.4a	Fig. 8.4b	Fig. 8.4c	Fig. 8.4d
Scale name	Symbols	Symbols	Symbols	Symbols
Data representation	Single squares	Single squares	Organized squares	Organized squares
Resolution R _{horz}	6	6	2	3
• R _{vert}	5	5	5	3
• R _{color}	5 ⁱ	5 ⁱ	varies ⁱⁱ	varies ⁱⁱⁱ
Scope (Length) L	30	30	10	16
Scale (Diversity) D	5 ⁱ	5 ⁱ	7 ⁺	6 ⁺⁺
Entropy h	0.943	0.943	0.970	0.812
Specific diversity d	0.167	0.167	0.700	0.375
Symbolic info. Y_h [bits]	66	66	27	34
Spatial info. S_h [bits]	4350	4350	630	810
Semantic info. E [bits]	0	0	3758	3572
Total info. M [bits]	4416	4416	4416	4416

ⁱ Different colors for 1x1 array of squares
ⁱⁱ Aproximation of different combinations of ordered 3x1 squares
ⁱⁱⁱ Different combinations of 2x2, 2x1, 1x2 and 1x1 arrays of ordered squares
ⁱⁱⁱⁱ $5^4/4 + 10 + 10 + 5$.
⁺ 5 5 4 | 5 4 3 | 3 3 2 | 3 2 2 | 3 2 1 | 2 1 1 | 1 1 1.
⁺⁺ 5 5 - 5 5 | 4 3 | 3 2 - 2 1 | 5 | 4 | 1.

8.6.5 Music

Table 8.5 shows a comparison of information balance of two segments of music recorded in .MP3 format. The two segments correspond to the same fraction of Beethoven's 5th symphony 1st mov. They are different in the instruments used to play them. One is the full orchestra this piece was written for. The other is played with a piano solo. For each version of the music segment analyzed, three observation scales are used: binary, characters and The Fundamental. The characters' scale consists of splitting the music-text in single characters. Each character exhibits a frequency with which entropy is computed. The binary observation can be obtained by substituting each character with its corresponding ASCII number expressed in the binary base. The Fundamental observation scale is obtained applying the Fundamental Scale Algorithm [75] which finds the sequences of characters that minimize the overall entropy of the text.

The results show that for music, semantic information accounts for almost all the information forming a musical message. Polyphonic music is the superposition of sounds and effects which makes it extremely complex. Yet, music can be regarded as a unidimensional phenomenon because all those sounds and

VIII. Where is the information?

sonorous effects must occur synchronically. Thus there is only one degree of freedom to alter the order of the symbols and the relative weight of the spatial information is expected to be rather limited.

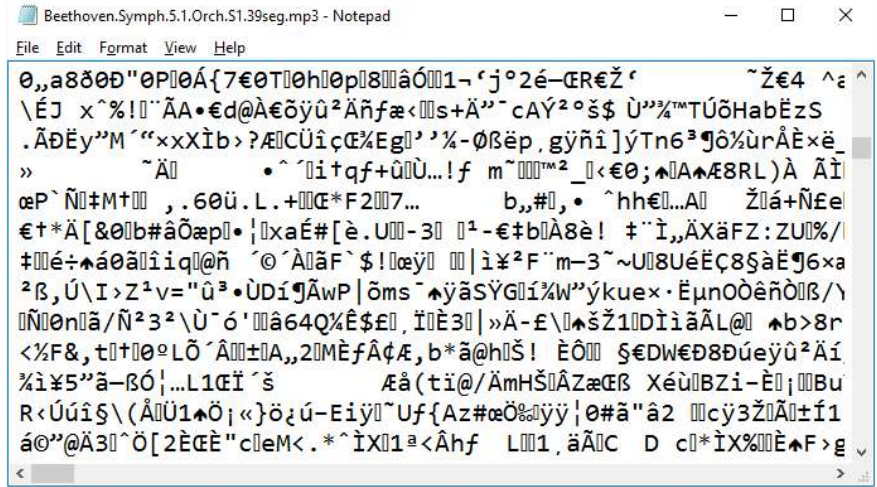


Figure 8.5: A tiny fraction of the text which constitutes the Beethoven's 5th symphony 1st movement segment interpreted with orchestra.

Table 8.5: Effects of different observation scales over the quantity of information of segment of Beethoven's 5th Symphony versioned with a full orchestra and piano solo.

Beethoven: 5th Symph.1st Mov.Segment.Orch.	Fundamental			5th Symph.1st Mov.Segment.Pianc		
	Binary zeroes and ones	Characters alphabet letters, punctuation & other signs	Minimal entropy Symbols	Binary zeroes and ones	Characters letters, punctuation & other signs	Min. entropy Symbols recognized
Resolution R [Symb./sec]	188948	23618	4517	192669	24084	4241
Scope L_D [Symbols]	5668432	708554	135519	7514080	939260	165387
Scope L₂ [bits]	5668432	5668432	1084152	7514080	7514080	1323096
Scale value (Diversity) D	2 (0 and 1)	252	4635	2 (0 and 1)	257	13808
Symbolic Entropy h	near 1	0.990	0.893	near 1	0.990	0.722
Specific Diversity d	0.049	0.00036	0.03420	0.049	0.00027	0.08349
Symbolic info. Y_h [bits]	5668432	701432	120972	7514080	929819	119461
Spatial info. S_h [bits]	99354166	82114998	20037890	#####	109450063	27409161
Semantic info. E [bits]	0	22206168	84863736	0	29049604	111900865
Total information M [bits]	105022598	105022598	105022598	#####	139429486	139429486

The emotions triggered by music is a phenomenon so complex that it is rarely possible to express them in words. We feel them but we do not have precise descriptions of them in terms of natural languages.

8.6.6 Mathematics as a language

Mathematics is commonly considered a language. The language of science. Beyond that actual but informal statement, here we apply our analysis to Mathematics as formally as we did with languages of different nature or category. Five famous mathematical models were selected. They are shown in Figure 8.6.

a	$F = m \cdot a$
b	$a^2 = b^2 + c^2$
c	$P(A B) = \frac{P(B A)P(A)}{P(B)}$
d	$f(w) = \int_{-\infty}^{\infty} f(x) \cdot e^{-2\pi i x w} \cdot dx$
e	$\frac{\partial u}{\partial t} + u \cdot \nabla u = \frac{1}{\rho} \cdot \nabla \bar{p} + \nu \cdot \nabla^2 u + \frac{1}{3} \cdot \nu \cdot \nabla(\nabla \cdot u) + g$

Figure 8.6: Five examples of mathematical descriptions: (a) The Newton's 3rd Law equation. (b) The Pythagoras' Theorem. (c) The Bayes' Theorem. (d) The Fourier's Transform equation. (e) The Navier-Stokes momentum equation.

8.7 Information component fractions

Using the results for the four examples presented, we can see where the information content is stressed for each type of information transmission media evaluated. The individual results for the description used in the previous section are plotted and shown in Figure 8.7. The graph uses different color for the messages according to their nature: 2D graphs, music, mathematics and natural languages, but each bubble represents one description.

Table 8.6: Properties of the mathematical descriptions shown in Figure 8.6

Expression name:		Newton's 3rdLaw		Pythagoras' Thm.		Bayes' Thm.		Fourier's Trnsf.		Navier-Stokes	
Scale name		Binary	Chars.	Binary	Chars.	Binary	Chars.	Binary	Chars.	Binary	Chars.
Data representation		0's, 1's	Math Signs	0's, 1's	Math Signs	0's, 1's	Math Signs	0's, 1's	Math Signs	0's, 1's	Math Signs
Res.	Rhorz [Symbols]	40	5	64	8	176	22	192	24	248	31
	Rvert [Symbols]	8	1	16	2	24	3	24	3	24	3
Scope	Ld [Symbols]	40	5	128	16	528	66	576	72	744	93
	L2 [bits]	40	40	128	128	528	528	576	576	744	744
Scale (Diversity) D		2	5	2	6	2	8	2	15	2	18
Symbolic Entropy h		near 1	1	near 1	0.562	near 1	0.492	near 1	0.460	near 1	0.501
Specific Diversity d		0.050	1	0.016	0.375	0.004	0.121	0.003	0.208	0.003	0.194
Symb.info. Yh [bits]		40	12	128	23	528	97	576	129	744	194
Spatial info. Sh [bits]		594	27	2989	124	17797	1399	19744	2044	26750	3045
Semantic info. E [bits]		0	595	0	2970	0	16828	0	18146	0	24255
Total info. M [bits]		634	634	3117	3117	18325	18325	20320	20320	27494	27494

$$\begin{aligned}
 & \blacksquare F = m \cdot a & \square f(w) = \int -\infty x \cdot e^{2\pi i d} \\
 & \bullet a^2 = b + c \\
 & \circ P(A|B) = \frac{P(A \cap B)}{P(B)} + \partial u_t + \nabla = 1 \rho p - v^2 3() g
 \end{aligned}$$

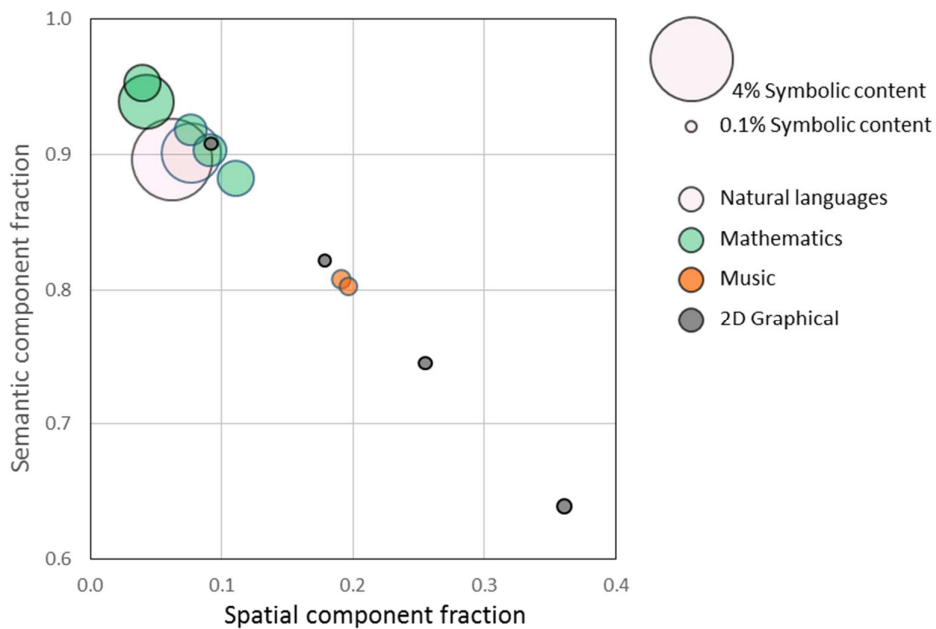


Figure 8.7: Fraction of semantic information vs. fraction of spatial information for 13 descriptions expressed in languages of different nature. The bubbles' area are proportional to the fraction of symbolic information.

Table 8.7: Relative weight of information type content for three information transmission media.

Language nature	2D Graphs	Music	Math Language	Natural Languages
Data representation	Polygons	Polyphonic Sounds	Math Signs	Words
Symbolic info. Y_h [fraction]	0.0011	0.0020	0.0098	0.0326
Spatial info. S_h [fraction]	0.2214	0.1937	0.0721	0.0697
Semantic info. E [fraction]	0.7775	0.8043	0.9181	0.8977

Table 8.7 presents an estimated of the symbolic, spatial and semantic information for respective fractions in 2D graphs, music, mathematics and natural languages. These representative numbers were obtained averaging the values plotted in Figure 8.7.

8.8 Discussion

8.8.1 Implications of scale, scope and resolution

The term *scale* is commonly used in a qualitative manner. Expressions like “individual scale”, “massive scale”, “microscopic scale”, “astronomical scale” and many other similar ones, are typically used to characterize the type of interpretation that should be given to certain descriptions. However, their utility relies on our subjective criteria to adequately apply those expressions. Subsequently, this rather diffused conception of *scale* is of little use for our purposes. We then propose a quantitative conception of *scale*. The scale of a system equals the scale of the language used for its description; the scale of the language equals the number of different symbols which constitute the language.

Interestingly, the system’s description scale is determined, in first place, by the observer, and in a much smaller degree by the system itself. The presumably high complexity of a system, functioning with the actions and reactions of a large number of tiny pieces, simply dissipates if (a) the observer, or the describer, fails to see the details, (b) the observer or describer is not interested the details, and prefers to focus on the macroscopic interactions that regulate the whole system’s behavior, or (c) the system does not have sufficient different components, which play the role of symbols here, to refer to each type of piece. It is clear that any observed system scale implies the use of a certain number of symbols. Thus the number of different symbols used in a description is linked with our intuitive idea of *scale*. Therefore, the term *Scale* can be used as a descriptor

of the combination language-observation by specifying the number of different symbols required to describe the language at any specific observation type. English, for example, is a 600 thousand word language if describe in terms of words, but a 26 letter language if described in terms of letters.

Another important concept with a close relationship to *scale* is *resolution*. Resolution is the *density* of symbols, repeated or not, that participate in a description. Therefore, resolution separates the description space in many smaller space segments, as many as indicated by the resolution itself. Each space segment must be occupied with a symbol, in other words, even an empty space is a valid symbol to be considered. In spite of the general use of the term *resolution*, the space-segments need not to be equally sized.

8.8.2 The balance of information

The symbolic, spatial and semantic components of information work in a distinctive way according to the nature of the language. This is reflexed in the results of Graph 8.7 and Table 8.7 where the values of the relative weight of information components tend to form clusters around the corresponding average values. Only semantic and spatial information for 2D graph languages show considerable dispersion in Graph 8.7. Nonetheless the dispersion for 2D-graphic language values, there is a noticeable alignment of the bubbles in the plane semantic-spatial information component fraction. A consequence of the almost negligible fractions of the symbolic components for music and 2d-graphical languages, leaving a direct linear relationship between semantic and spatial information for these types of language. The higher relevance of symbolic information for mathematics and natural languages, introduces the small deviations that can be observed the upper region of the graph. Music and 2D-graph languages are dominated by the spatial and semantic components. This does not mean their symbolic component is irrelevant. Actually pictures and sound files are so large that even the small percentage of symbolic information implies some effort to transfer it. But the semantic content defines the overall effect of these types of language which can cause direct emotions when the message is seen or listened.

As their counterpart, natural languages and mathematics evolved to be written languages. They rely on primitive information units, as for instance the alphabet, which make them able to create easily recognizable discrete symbols. The stability and controllability of the written version of these symbols, has been a crucial factor in the deep development of these families of languages which have become the fundamental basis for human progress and its technology.

Natural languages have developed the mechanism of forming words to adopt specific meanings. Words are a so powerful and controllable way to represent meanings that they have become the standard way to communicate complex ideas. The combinatorial nature of the way new words are generated, allow natural languages to cope with the need for creating a new symbol each time a specific meaning required its representation. Thus making the communication process more precise with time and the language more effective.

The exact reproduction that the writing system allows, has converted words into meaningful symbols. Yet, in order to combine these symbols to generate complex ideas, words must follow certain rules according to their relative positions. Thus grammar appears as a need for natural languages to reach an agreement between the sender and the receiver of the message in order to articulate words and form even more complex and precise ideas. Grammar is implemented by the language by defining the correct order in which symbols should appear. Thus grammar must be part or all the information stored as spatial. For natural languages there are many options to order the symbols and still comply with grammar rules, therefore the modest content of information in the spatial component for natural languages, should be expected. The congruence in the use of the language by both ends of the communication process is crucial for natural languages. The huge number of symbols in any natural language and their stability in terms of their meaning, and the importance of their relative position in every sentence, explains why none of the components of information in natural languages can be neglected.

Mathematics also has its set of grammar rules. As in natural languages, there is some flexibility that allow witting the same idea with a different order or positions of the symbols involved. Thus, regarding spatial information, the behavior of mathematics is similar to the natural languages. But, focusing in the semantic information, mathematics shows the highest values of all the languages tested. Perhaps due to the fact that mathematical expressions hold deeper and more extensive meaning and repercussion than expressions written with any other known type of language.

For music, the order of sounds is a rigorous condition. The spatial information content is therefore expected to be larger than for natural languages and mathematics. The high spatial information value that exhibits some of the 2D-graphic messages, is not the product of grammar since these expression do not have to hold any precedence rule. The high spatial information value is the consequence of the two dimensions involved which augments the degrees of freedom of the symbols formed.

The relative *weight* of the three components of information, indicate that semantic information accounts for most of the information of any message, regardless of the nature of the language. This result is consistent with the information flow model presented in Figure 8.1 where semantic information is presented as the cumulative of formal language learning plus the ability to recognize patterns obtained along our live experience. Following this model, symbolic and spatial information together form a pattern of symbols which triggers the receiver's interpretation up to the level allowable by his ability to understand the meaning of the pattern received; that is his knowledge about the language used to transmit the message. Clearly, the semantic information is then, the measurement of the potential *meaning* obtainable from the message received.

8.9 Summary

A quantitative conception of scale is introduced. This conception allows for generalizing Shannon's information expression to include transmission systems based on more than two symbols. An expression that may be useful when evaluating the convenience of transmitting information by means of non-binary systems.

The notion of spatial information is presented as the type of information which contains the relative position of the symbols and the language's grammar. Methods for its estimation, based on the degrees of freedom of the space where the language expresses, are also provided.

Having expressions to compute the amount of symbolic and spatial information from different scales of observation, a balance of total information is used to determine the maximum semantic information content in a message. Tests with languages of five different natures were performed. The results indicate that the model provided leads to coherent results.

After introducing the quantitative concept of scale and the information flow model, the tests, applied to languages of radically different nature, have offered coherent results. Now I feel confident enough to announce definitions for the components of information used along this work. Information: any set of recognizable patterns of symbols which may cause some effect on the viewer or the environment. Spatial information: the description of the relative position of symbols within a message. Symbolic information: the distribution of symbol frequencies within a message. Semantic information: the interpretable meaning associated with the symbolic and spatial information.

These concepts, along with the information flow model offers a solution for the old argument about what is information and how to compute each information component. The model is rather simple. It has been presented here at the risk of being considered naive or trivial. Yet, it is a powerful idea which provides possibilities not only to quantify distinct facets of information and how they behave according to the observer focus, but also as a novel approach to help in making compatible the various, so far contradicting, conceptions of information.

Extending our understanding about the nature of information and languages may help us to unveil the mechanisms underlying collective phenomena surrounding us. We may be closer to model the physical complex systems as languages working according to the symbols represented by molecular sequences, mass and electromagnetic attractions, substance concentrations, and other physical real entities.

Finally, considering scale as a model component reinforces the idea about our capacity of thinking as linked to our ability to build internal languages of useful symbols. This is a motivating thought since it paves the road to conceive intelligence as the capacity of finding observation scales –or interpretation scales– which lead to a more effective way of understanding nature and its phenomena.

"Every day we know more and understand less"
Albert Einstein

*"The more you say, the less people remember.
The fewer the words, the greater the profit."*
Francois de Salignac Fenelon

Chapter IX

Conclusion

Several experiments were conducted. Languages of different types and natures were focused as objects of experimentation. Natural languages, represented by English and Spanish can be perceived by means audible signals as well as writing systems. Artificial languages, included as several programming language codes, express only by writing. Music, can transmitted by sounds and by writing. But the communication of music by writing lacks of its essence and does not produce, at least not for most people, the emotions and sensations associated to a pattern of sounds. Human natural languages, computer programming code and music belong to absolutely different types of language.



Obtained from twitter.com: @ScienceAllDay. Sep 21 2014.

Figure 9.1: The message's interpretation obeys the aspect the observer is interested in.

Despite their radical different nature, the analysis and comparison of these languages was possible. The method developed consists of splitting description texts in symbols recognized at several scales —characters, words, and fundamental symbols— and computing characteristic language properties as

symbolic specific diversity, entropy and complexity. By applying the method to hundreds of text descriptions, expressed in several languages, it was possible to sense more subtle properties as the style of use, and time-period where the analyzed piece corresponds. Results proved the capacity of the method to analyze text messages and to compare the effectivity of languages.

Languages are vehicles used in accordance to the objectives of the message emitter and the receiver. There are an indefinite number of ways of separating symbols –observation scales– in any description. It is the observer's choice to adopt one or another criteria for symbol selection. The observer selects what pattern within the message is important, in order to insert it into a meaningful context. In certain way, it would be valid to say the observer (receiver) decides with which languages to interpret the complex integral message perceived, despite the restrictions imposed by the sender's (if any) language choice.

Finally, in the search for ways to compute the impact of different conceptions of information and complexity, the notion of compressibility has been always close. It is perhaps useful to contrast the classical information compression target, with the objective of our study. While the field of traditional compression techniques point towards the economy of transferring information, the study of languages by means of their diversity, entropy, structural shape, and other quantifiable properties, is directed to what we could call the synthesis of our understanding of languages as living organisms, capable of self-organize the symbols which constitute them. Languages evolve within our minds, up to a so highly sophisticated level, that can be regarded as the most important factor for the accomplishments of humankind and of our intelligence.

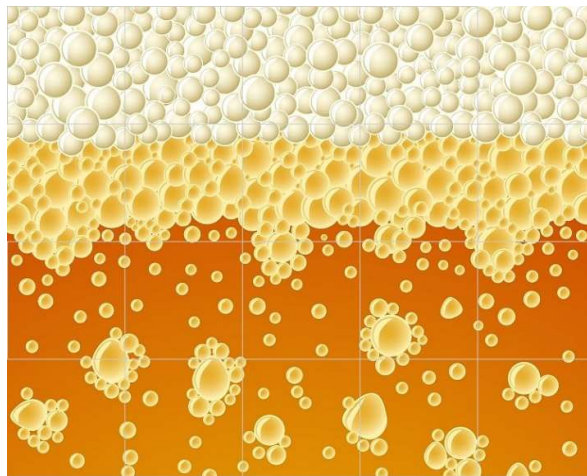


Figure 9.2: Languages are self-organizing sets of symbols.

9.1 Main results and contributions

By decomposing system's descriptions at a specified scale, it has been possible to compute revealing properties of a system studied. Our findings show these properties represent the actual system's behavior and therefore represent a valid framework for their study.

Proper calculation and handling of diversity and entropy lead to the possibility of automated classification of system descriptions and messages. The process is sensible to characteristics as style, lexicon and repetitiveness that have been traditionally considered too subtle to be treated in an automated fashion.

9.1.1 Language quantitative analysis

Quantitative linguistics have been an active field of study for some decades. Quantitative linguistics is devoted to study human natural languages. Despite the common view of languages as a highly interrelated combination of words, letters and other symbols, they are seldom presented as systems. An aspect we regard as an innovation from this study, is considering languages not only as the expression of systems but also as systems themselves.

The methods presented are the basis for comparing different kind of languages or particular expressions using the same language. The comparison rely on the quantitative measurement of general properties as symbolic diversity and entropy, which can be implemented in an automated fashion, thus opening the possibility for massive or distant evaluation of system descriptions and natural language texts.

9.1.2 The notion of scale as a numerical property

Resolution and scale are closely related concepts. Sometimes confused. Offering precise definitions, highlighting their differences and relationship was a relevant effort that made possible the construction of methods for representing descriptions at a reduced scale.

Resolution has been typically a quantitative description of the number of pieces in which some space is divided to discretely represent a continuous phenomenon, whereas the notion of scale, is traditionally used in a qualitatively to specify the kind of major components of a description. Consider the literal expressions "reading the speech at the scale of words" and "analyzing the speech at the scale of characters". In both expressions, our intuition recognizes the number of different words and the number of different characters, as the scale at which we refer to, in the corresponding expression. Thus, being no reason there to avoid the use of the number of different symbols as the

numerical value of the observation scale, we present the quantitative notion of Scale as the number of different symbols used to form a description that is the diversity of the language used to interpret the description.

Even though this subtle distinction between scale as a qualitative concept and scale as a precise number, may appear to be only an semantic accessory issue, we think it is a relevant contribution to the precision of the language used within the fields of quantitative linguistics, information theory and complexity, where the term 'scale' has lacked of the precise meaning needed to allow its use in formal way. Complex systems arguably have some specific way of working. Independently of our ability to understand them, they change, react and evolve accordingly to the shape and the DNA of their internal structure, whatever it is. But when we model them, we try to discover the internal structure that dominates their behavior and their identity. In this regard, the role of scale in describing a system, has to be acknowledged as of crucial importance.

9.1.3 The Fundamental Scale

The fundamental scale, as a set of symbols that best serve as the basis to produce a description, has proven to be a powerful concept. The fundamental scale was used to analyze music coded in the form of texts. These texts did not have recognizable character strings. They did not even show a recognizable alphabet. The idea that there exists a set containing symbols that works as the best to construct a system description, makes possible to achieve the quantification of some properties of those human uninterpretable, texts.

In a broader sense, the concept of the fundamental scale promises usefulness to deepen not only the understanding of the forces involved of language evolution, but even to depict the mechanisms by which information builds itself until it becomes an entity capable of transporting information.

9.1.4 Notions of spatial and semantic information

The *anatomy* of information introduced in Chapter VIII is a novel approach to an old subject of discussion. The three-section of information in its components, symbolic, spatial and semantic allows not only for a better understanding of the internal structure of languages. The methods provided for the quantification of each information component, promises being helpful in the forecoming language related studies

9.1.5 An information flow model

The information flow model introduced is a useful tool that let us a better understanding of the communication process, with potentially high impact in

the fields of social sciences and education, but also in the modeling of language and nature evolution.

9.1.6 A complex-experiment software platform

A software platform to perform simulation experiments has been devised and developed up to a useful and working version. The software, referenced as *MoNet*, incorporates parsing languages to ease the handling of complex structures, allowing the researcher to focus in the upper scale levels of the experiment, while keeping the modeling of the complex interaction that occur at smaller scales. Administering the instances of experiments replicated with changing conditions, is also a valuable characteristic of *MoNet* which allows the realization of extended experiments as the one presented in this work.

9.2 Possible future works

9.2.1 The concept of fundamental scale at multidimensional languages

The concept of fundamental scale was applied for languages expressed in a one-dimensional space. English and Spanish texts, as well as the computer codes subject to analysis in this thesis, are clearly descriptions presented in the dimension of the flow of the words. A completely different language as music is also ordered in the one dimension represented by time.

But the notion of fundamental scale is not be limited to one-dimensional languages as those studied here. All perceptions obtained by means of the sight, are bi-dimensional perceptions; they are in fact, the projections of segments of the 3-dimensional world surrounding us that creates a 2-dimensional image in our retinas. Then, our brains interpret the contrasts and shapes to convert every 2-dimensional image in a source of sensations, notions, orientations and any other cognitive process around the visual information, as to create the conception of the 3-dimensional world that is out there, and that we actually do not see but conceive.

Upgrading the Fundamental Scale algorithm from one-dimensional descriptions to description of two or three dimensions is a challenge. The number of possible ways to scan the space where the fundamental scale is sought, increases dramatically when the search is performed in a higher dimensional space. Thus, not only the solution, but also setting the optimization problem to find the fundamental scale, makes the computational problem, which was already extremely complex for the one-dimensional description case, several orders of magnitude more complex. Yet, even in the most likely case we do not find a practical solution for these problems, this patterns of reasoning may lead us to

improve our understanding of how the brain interprets related signal from our senses, as well as the languages we use to interpret sensations.

9.2.2 Applying the method in other fields

The unveil of the specific diversity and the entropy of a system at different observation scales, including its fundamental scale, allows for determining the dominant symbols of a complex systems description. Thus, these methods become a promising tool which may contribute to the understanding of the information content and structure of descriptions in fields as bioinformatics and cladistics.

Bibliography

- [1] D. Rybski, Auerbach's legacy, *Environ. Plan. A.* 45 (2013) 1266–1268. doi:10.1068/a4678.
- [2] G.K. Zipf, *Human Behavior and the principle of least effort: An introduction to human ecology*, Addison-Wesley, New York, 1949.
- [3] C.E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* 27 (1948) 379–423. doi:10.1145/584091.584093.
- [4] R.J. Solomonoff, *A Preliminary Report on a General Theory of Inductive Inference*, Cambridge, MA, 1960.
- [5] W.D. L., C.S. Dowe, Minimum Message Length and Kolmogorov Complexity, *Comput. J.* 42 (1999) 270–283.
- [6] G. Chaitin, On the Simplicity and Speed of Programs for Computing Infinite Sets of Natural Numbers, *J. ACM.* 16 (1969) 407–422.
- [7] E. Schrodinger, *What Is Life?*, Cambridge University Press, 1944.
- [8] M. Prokopenko, F. Boschetti, A.J. Ryan, An information-theoretic primer on complexity, self-organisation and emergence, *Complexity.* 15 (2008) 11–28. doi:10.1002/cplx.20249.
- [9] R. Lopez Ruiz, H. Mancini, X. Calbet, Statistical Measure Of Complexity, *Phys. Lett. A.* 209 (1995) 321–326.
- [10] C. Gershenson, N. Fernandez, Complexity and Information: Measuring Emergence, Self-organization, and Homeostasis at Multiple Scales, *Complexity.* 18 (2012) 29–44.
- [11] M. Gell-mann, *What is Complexity?*, Complexity. (1995).
- [12] Y. Bar-Yam, Multiscale Complexity/Entropy, *Adv. Complex Syst.* 07 (2004) 47–63. doi:10.1142/S0219525904000068.
- [13] P. Funes, Complexity measures for complex systems and complex objects, (2014) 1–11. www.cs.brandeis.edu/~pablo/complex-maker.html.
- [14] N. Fernandez, C. Maldonado, C. Gershenson, Information Measures of Complexity, Emergence, Self-organization, Homeostasis, and Autopoiesis,

- Springer, Berlin. Heidelberg, 2014. <http://arxiv.org/abs/1304.1842> (accessed December 21, 2014).
- [15] A.-L. Barabasi, R. Albert, Emergence of Scaling in Random Networks, *Science* (80-.). 286 (1999) 509–512.
- [16] R. Albert, A.L. Barabasi, Statistical mechanics of complex networks, *Mod. Phys.* 74 (2002) 47–97.
- [17] D. Watts, *Six degrees. The Science of a connected age*, 1st ed., W. W. Norton & Company, New York, NY, 2003.
- [18] J. Jastrow, A Study of Mental Statistics, *New Rev.* 5 (1891) 559–568.
- [19] J. Jastrow, Community of Ideas of Men and Women, *Psychol. Rev.* 3 (1896) 68–71.
- [20] H.S. Heaps, *Information Retrieval*, Comput. Theor. Orlando Acad. Press. (1978).
- [21] G. Kirby, Zipf 's Law, 10 (1985) 180–185.
- [22] P. Bak, *How nature works*, Oxford: Oxford university press, 1997.
- [23] M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemp. Phys.* 46 (2005) 323–351. doi:10.1016/j.cities.2012.03.001.
- [24] L. Lu, Z.-K. Zhang, T. Zhou, Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems, *PLoS ONE* 5(12). e14139 (2010).
- [25] A. Gelbukh, G. Sidorov, Zipf and Heaps Laws' Coefficients Depend on Language, *Comput. Linguist. Intell. Text Process.* 2004 (2001) 332–335.
- [26] Y. Sano, H. Takayasu, M. Takayasu, Zipf's Law and {Heaps}' Law Can Predict the Size of Potential Words, *Prog. Theor. Phys. Supp.* 194 (2012) 202–209. doi:10.1143/PTPS.194.202.
- [27] F. Font-Clos, G. Boleda, Á. Corral, A scaling law beyond Zipf's law and its relation to Heaps' law, *New J. Phys.* 15 (2013) 93033. doi:10.1088/1367-2630/15/9/093033.
- [28] N. Chomsky, *Syntactic Structures*, Mouton, The Hague, 1957.
- [29] R. Jackendoff, *X-Bar Syntax: A Study of Phrase Structure*, MIT Press, Cambridge, MA, 1977.
- [30] G. Markowsky, An Introduction to Algorithmic Information Theory, *Complexity*,. 2 (1997) 12–22.
- [31] G. Febres, MoNet: Complex experiment modeling platform, (2014). [www.gfebres.com\F0IndexFrame\F132Body\F132BodyPublications\MoNET\Multisc](http://www.gfebres.com/F0IndexFrame\F132Body\F132BodyPublications\MoNET\Multisc) (accessed February 26, 2015).
- [32] M. Newman, *Networks: An Introduction*, Oxford University Press, Inc., New York, 2010.
- [33] I. Forum, *Spanish Language @ Usage*, (2013). <http://spanish.stackexchange.com/questions/1508/comparing-number-of-words-in-spanish->.

- [34] J. Ling, Number of Words in the English Language, (2001). <http://hypertextbook.com/facts/2001/JohnnyLing.shtml> (accessed October 6, 2012).
- [35] G.R. Alberich, No Title, (2012). <http://dirae.es/about> (accessed October 9, 2012).
- [36] O. University, Oxford English Dictionary, (2012). <http://public.oed.com/about/> (accessed January 13, 2013).
- [37] F. Heylighen, J.-M. Dewaele, Variation in the contextuality of language: An empirical measure, *Found. Sci.* 7.3 (2002) 293–340.
- [38] C. Gershenson, S. a. Kauffman, I. Shmulevich, The Role of Redundancy in the Robustness of Random Boolean Networks, *Proc. Tenth Int. Conf. Simul. Synth. Living Syst.* (2006) 7. <http://arxiv.org/abs/nlin/0511018>.
- [39] A. Moore, The Structure of English Language, (n.d.). <http://www.universalteacher.org.uk/lang/engstruct.htm> (accessed August 27, 2013).
- [40] M. Gerlach, E.G. Altmann, Stochastic Model for the Vocabulary Growth in Natural Languages, *Phys. Rev. X.* 3 (2013) 021006. doi:10.1103/PhysRevX.3.021006.
- [41] M.A. Montemurro, D.H. Zanette, Entropic Analysis of the Role of Words in Literary Texts, *Adv. Complex Syst.* 5 (2002) 7–17.
- [42] L.A. Sherman, *Analytics of literature: A manual for the objective study of English prose and poetry*, Boston: Ginn & Co., 1893.
- [43] W. DuBay, *The principles of readability*. 2004, Costa Mesa Impact Inf. (2008) 77. doi:10.1.1.91.4042.
- [44] E.L. Thorndike, *A Teacher's Word Book*, Teachers College, Columbia University, New York, 1921.
- [45] E.L. Thorndike, *A Teacher's Word Book of 20,000 Words*, Teachers College, Columbia University, New York, 1932.
- [46] E.L. Thorndike, *The Teacher's Word Book of 30,000 Words*, Teachers College, Columbia University, New York, 1944.
- [47] R. Flesch, *Marks of a readable style*, Columbia University contributions to education, no. 897. New York: Bureau of Publications, New York, 1943.
- [48] R. Flesch, *The art of plain talk*, Harper & Brothers, New York, 1946.
- [49] R. Flesch, *The art of readable writing*, Harper & Brothers, New York, 1949.
- [50] R. Flesch, *How to test readability*, Harpe & Brothers, New York, 1951.
- [51] R. Flesch, *The art of clear thinking*, Harper & Brothers, Barnes and Nobel Books, n.d.
- [52] R. Flesch, *How to write, speak and think more effectively*, Harper & Brothers, New York, 1958.
- [53] P.J. Kincaid, R.P. Fishbourne, S.R. Rogers, B.S. Chissom, *Derivation of new*

- readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Millington, Tennessee. (1975).
- [54] K.M. Sheehan, I. Kostin, Y. Futagi, F. Michael, *Generating Automated Text Complexity Classifications That Are Aligned with Targeted Complexity Standards*, Princeton, New Jersey, 2010.
- [55] I.M. Barrio Cantalejo, *Legibilidad y salud*, (2007). [http://digitool-uam.greendata.es/R/SAM4\\$N4YMYHBJ9NK5RANEV9N1V4BH82BVV9G5LTP DTT21FE7SJ-01133?func=dbin-jump-full&object_id=3907](http://digitool-uam.greendata.es/R/SAM4$N4YMYHBJ9NK5RANEV9N1V4BH82BVV9G5LTP DTT21FE7SJ-01133?func=dbin-jump-full&object_id=3907).
- [56] S. Trauzettel-Klosinski, K. Dietz, Standardized assessment of reading performance: The new international reading speed texts IReST, *Investig. Ophthalmol. Vis. Sci.* 53 (2012) 5452–5461. doi:10.1167/iovs.11-8284.
- [57] O. Gröne, *Inventario de instrumentos para medir la legibilidad de un texto*, (2009) 1–23.
- [58] J.S. Chall, *Readability: An appraisal of research and application*, Epping, Essex, England: Bowker Publishing Company, Columbus, OH, 1958.
- [59] J. Fernández-Huerta, *Medidas sencillas de lecturabilidad*, Consigna, 1959.
- [60] F. Szigrisz Pazos, *Sistemas Predictivos de Legibilidad del Mensaje Escrito: Formula de Perspicuidad.*, Universidad Complutense de Madrid, 1993.
- [61] K. Tanaka-Ishii, S. Tezuka, H. Terada, *Sorting Texts by Readability*, *Comput. Linguist.* 36 (2010) 203–227. doi:10.1162/coli.2010.09-036-R2-08-050.
- [62] G. R. Klare, *The Role of Word Frequency in Readability*. *Elementary English.*, *Natl. Council. Teach. English.* 45 (1968) 12–22.
- [63] G. Febres, K. Jaffé, C. Gershenson, *Complexity measurement of natural and artificial languages*, *Complexity.* 20 (2015) 429–453. doi:10.1002/cplx.21529.
- [64] I. Kontoyiannis, *The Complexity and Entropy of Literary Styles*. Department of Statistics, Stanford University. NSF Technical Report No. 97. 1997., 1997.
- [65] J. Savoy, *Vocabulary Growth Study□: An Example with the State of the Union Addresses*, *Quant. Linguist.* (in press) (n.d.).
- [66] J. Savoy, *Text Clustering□: An Application with the State of the Union Addresses*, *J. Am. Soc. Inf. Sci. Technol.* (in press) (n.d.).
- [67] M. Grabchak, Z. Zhang, D.T. Zhang, *Authorship Attribution Using Entropy*, *J. Quant. Linguist.* 20 (2013) 301–313. doi:10.1080/09296174.2013.830551.
- [68] H.S. Eaton, *An English-French-German-Spanish Word Frequency Dictionary*, Dover Publications, 1940.
- [69] J. A. Gualda-Gil, *Densidad de información del español vs el inglés*, (n.d.). <http://www.elcastellano.org/noticia.php?id=2230> (accessed November 3, 2013).
- [70] Y. Bar-Yam, D. Harmon, Y. Bar-Yam, *Computationally tractable pairwise complexity profile*, *Complexity.* 18 (2013) 20–27. doi:10.1002/cplx.21437.

- [71] G. Febres, K. Jaffé, Quantifying literature quality using complexity criteria, *J. Quant. Linguist.* (in press) (n.d.).
- [72] R. Piasecki, A. Plastino, Entropic descriptor of a complex behaviour, *Physica A.* 389 (2010) 397–407.
- [73] M. Sipser, *Introduction to the Theory of Computation*, Thomson Course Technology, Boston, 2006.
- [74] T. Schurmann, P. Grassberger, Entropy estimation of symbol sequences., *Chaos.* 6 (1996) 414–427. doi:10.1063/1.166191.
- [75] G. Febres, K. Jaffé, A Fundamental Scale of Descriptions for Analyzing Information Content of Communication Systems, *Entropy.* 17 (2015) 1606–1633. doi:10.3390/e17041606.
- [76] G. Febres, K. Jaffe, Calculating entropy at different scales among diverse communication systems, *Complexity.* early view (2015). doi:10.1002/cplx.21746.
- [77] L. Meyer, *Emotion and meaning in music*, The University of Chicago Press, Chicago, 1956.
- [78] D.J. Levitin, *This Is Your Brain on Music*, Penguin Group (USA) Inc., 2006.
- [79] D. Wilson, The Role of Patterning in Music, *Leonardo.* 22 (1989) 101–106.
- [80] R.B. Dannenberg, B. Thom, D. Watson, A Machine Learning Approach to Musical Style Recognition A Machine Learning Approach to Musical Style Recognition, in: *Int. Comput. Music Conf.*, 1997.
- [81] P.J. Ponce de Leon, J.M. Inesta, Pattern Recognition Approach for Music Style Identification Using Shallow Statistical Descriptors, *IEEE Trans. Syst. Man. Cybern.* 37 (2007) 248–257. doi:10.1109/TSMCC.2006.876045.
- [82] Perez Sancho C., J.M. Inesta, J.C. Rubio, A text categorization approach for music style recognition, *Lect. Notes Comput. Sci.* 3523 (2005) 649–657. doi:10.1007/11492542_79.
- [83] P. van Kranenburg, E. Backer, Musical style recognition - a quantitative approach, in: *Conf. Interdiscip. Musicol.*, 2004: pp. 1–10.
- [84] P. Mavromatis, A Hidden Markov Model of Melody Production in Greek Church Chant, *Comput. Musicol.* 14 (2005).
- [85] M. Rohrmeier, Towards a generative syntax of tonal harmony, *J. Math. Music.* 5 (2011) 35–53. doi:10.1080/17459737.2011.573676.
- [86] F. Lerdahl, R. Jackendoff, *A Generative Theory of Tonal Music*, MIT Press, Cambridge, MA, 1983.
- [87] M. Rohrmeier, *Musical syntax and its cognitive implications*, (2012). http://www.musica.ed.ac.uk/wp-content/uploads/Manual_Upload/Martin_Rohrmeier_MusicalSyntax.pdf.
- [88] P. Mavromatis, Minimum Description Length Modeling of Musical Structure, *J. Math. Music.* 00 (2009) 1–21.

- [89] G. Cox, On the Relationship Between Entropy and Meaning in Music: An Exploration with Recurrent Neural Networks, *Proc. Annu. Meet. Cogn. Sci. Soc.* (2010).
- [90] D. Huron, Sweet Anticipation: Music and the Psychology of Expectation, *Music Percept.* 24 (2007) 511–514. doi:10.1525/mp.2007.24.5.511.
- [91] P. Mavromatis, A hidden markov model of melody production in Greek church chant, *Comput. Musicol.* 14 (2005) 93–112.
- [92] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *IEEE*. 77 (1989) 257–286. doi:10.1109/5.18626.
- [93] M. Rohrmeier, *Music Syntax Theory*, (2012). <http://www.musica.ed.ac.uk/archive/2012/martin-rohrmeier/> (accessed December 9, 2015).
- [94] F. Heylighen, Modelling Emergence, *World Futur. J. Gen. Evol.* 31 (1991) 89–104. doi:10.1080/02604027.1991.9972256.
- [95] Y. Bar-Yam, A mathematical theory of strong emergence using multiscale variety, *Complexity.* 9 (2004) 15–24. doi:10.1002/cplx.20029.
- [96] A. Ryan, Emergence is coupled to scope, not level, *Complexity.* 13(2) (2007) 67–77. <http://arxiv.org/abs/nlin/0609011>.
- [97] N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*, The MIT Press, 1961.
- [98] K.-E. Sveiby, What is Information?, *Inf. Serv. Use.* 18 (1998) 243–254. <http://www.sveiby.com/articles/Information.html>.
- [99] N. Chomsky, Three models for the descriptions of languages, (1956).
- [100] R. Jackendoff, What is the human language faculty?: Two views, *Language (Baltim).* 87 (2011) 586–624. doi:10.1353/lan.2011.0063.
- [101] A. van den Bosch, A. Content, W. Daelemans, B. de Gelder, Measuring the complexity of writing systems, *J. Quant. Linguist.* 1 (1994) 178–188.

"Why repeat the same old mistakes with so many new errors to commit?"
Bertrand Russel

Appendix A

MoNet: Complex experiment modeling platform

A.1 Overview

MoNet is a bundle of scripts, interpretations, programs and visual interfaces designed to analyze complex systems descriptions at different scales. *MoNet* describes a system as a collection of objects and object families connected by hierarchical and functional relationships. Any object exists within the system as a description constituted by an object-related file containing the object descriptive attribute names and their values as well as the file addresses of others intimately related objects. Thus, the *MoNet*'s description of an object holds information about the object constitution and the relationship with other objects.

The attributes describing an object can adopt many possible dimensionalities: single or multiple values. When multiple values is the case, the arrangement can be a regular array of up to 5 dimensions, a tree or an irregular array forming a net of single values. The scope of each object description can be adjusted adding attributes or modifying their representation and dimensionality. *MoNet* can treat every text included in a library as well as the library itself, offering results for text-objects as independent elements or as groups. For every component of the system modeled, descriptions at different scales can co-exist. Individual objects can be selected combining logical conditions based on properties or attribute values. The library holds descriptions of each existing object with its attribute values.

MoNet makes possible to represent amazingly complex systems by describing the behavior of its parts at an immediate level. The researcher may then describe some of these parts according to the behavior of their integrating elements. And successively it is possible to describe any system as a fractal-set of interconnected rules hierarchically organized.

A.2 Major Components. Architecture

As any modern software, *MoNet*'s architecture has several layers from the most disperse piece of data up to the most comprehensive visualization outputs. Figure 1 shows a diagram with the architecture of *MoNet*. Some layer are superimposed to represent that not all the coding of some layers are completely independent from others, yet they may share some of their contents. *MoNet*'s limits are defined more by the computer used and the researcher conditions than the software itself

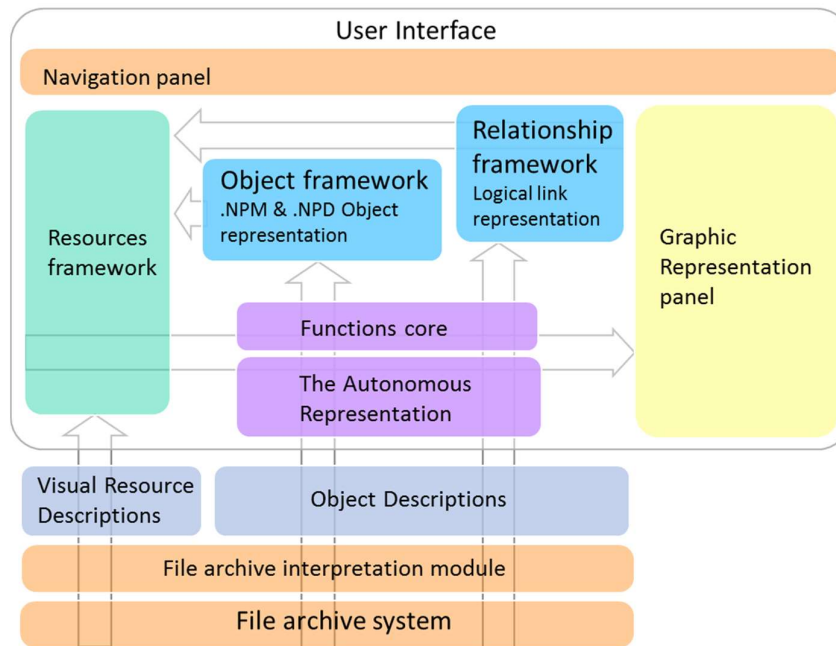


Figure A.1: *MoNet*'s general architecture.

A.2.1 Environment

Monet runs in a Windows environment. There are no special requirements. Due to the typical complexity of the systems being modelled, the use a computer with at least 4GB of memory RAM and a high resolution monitor, is recommended. Multiple extended monitors ease the visibility of all aspects of *Monet* and allows for running multiple instances of the program simulating several components of the complex model, simultaneously.

A.2.2 Data Storage. File-Object Types

MoNet does not rely in a Data Base file. MoNet drives the file-objects of a system by reading each object description and the connections to other objects. Traditional data bases are orthogonal too rigid data organizers to properly and efficiently represent complex and irregularly shaped systems. Therefore MoNet dispenses the direct use of data bases. Instead, MoNet uses internally configured files organized in a logical file structured called *NetPlex*. *NetPlex* is made of two types of files: the *.NPD* (*NetPlex Data*) and the *.NPM* (*NetPlex Model*). Figure A.2 resembles a hypothetical model file structure showing the relationships of files and the logical connections that build up the notion of data web which is able to reproduce, mimic and store the system behavior.

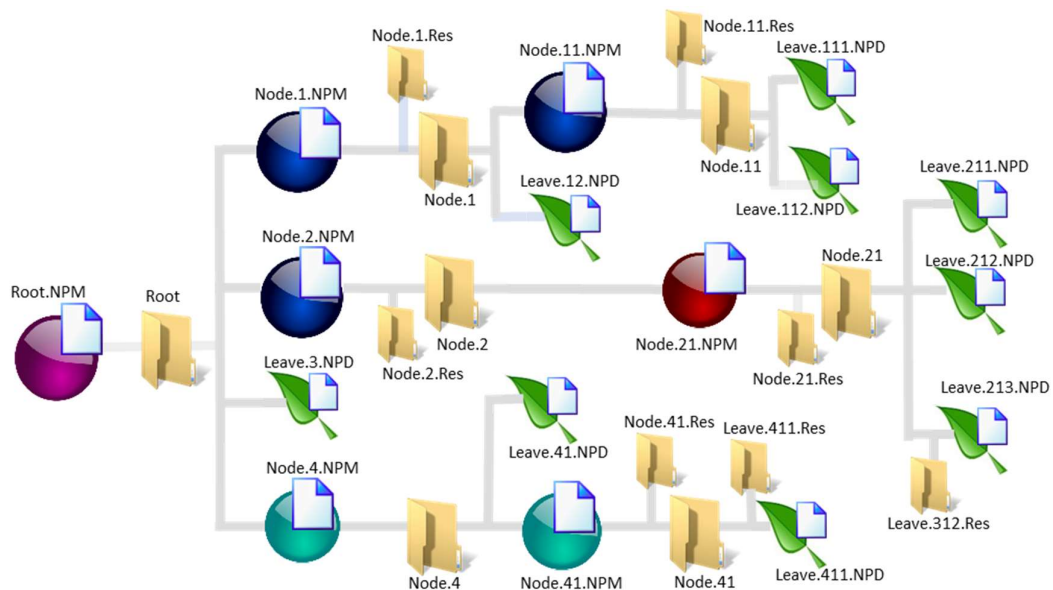


Figure A.2: MoNet's hypothetical model file structure showing the relationships of files and the hierarchical ownership connections.

A.2.2.1 The *.NPD* extension

Files with the *.NPD* extension serve to describe specific nodes within the network model; thus for each node in the network model there should be a corresponding file with the extension *.NPD* describing the node I as much detail a the model designer considered convenient for a stage of the modelling process.

A.2.2.2 The *.NPM* extension

Files with the *.NPM* extension identifies the center node of cluster and the relative file paths of the surrounding nodes. File with the *.NPM* extension are used

to crawl outbound from the center of a cluster until a specified radius is reached. All nodes included within that logical space can be marked and selected to build a sub-network of the whole model.

A.3 Object nature types

Several types of abstract objects are used to assembly complex models of a real physical or conceptual systems in *MoNet*. Each type of object serves to represent the nature of an abstract data function within the model. So far the types of objects that *MoNet* recognizes are the following:

- Nodes
- Arcs
- Filters
- Graph Resources

A Node can be seen as a set of property and their corresponding values which compound the description of an entity. In that sense, the Node is an instance of the type of entity and its description includes not only the static values its permanent properties but also the rules and relations to other entities that govern its dynamics and its evolution. A node also 'knows' which other nodes it is made of and which other node it is a part of, so a node contains information about the scale at which it exists as part of the whole model. The other types of objects, Arcs, Filters and Graph Resources, serve as auxiliary information to allow the operation and useful interpretation of the model behavior. It is worthwhile to mention that these auxiliary components are also nodes themselves; they belong to specially designed class of nodes to adequately perform the tasks they serve for.

A grid serves to present objects or agents with their identification information pieces and attribute values. Each element occupies a row of the grid. Columns are headed with the name of attributes, thus every cell on a row will present the value of the element attribute which corresponds to the cell column.

A.4 User interface

MoNet's graphic interface is designed around a single window. This window is capable of representing all those abstract objects that form a node. Each other object related with an open window object, can be accessed in an additional instance of the program. This gives *MoNet*'s operator the possibility of crawling thru the nodes of an extended complex system representation.

Figure A.3 shows the main form of *MoNet*'s graphic interface. The form is organized with panels devoted nodes, arcs, filters, graphic properties, graphic

There are panels for specifying nodes and arcs graphic visual properties. Using these devices it is possible to design powerful graph that can be projected on the Graph panel. When cell from any spread sheet is focused by clicking it, the content is shown in the value panel. If the cell contains a computable expression, it is shown in the Expression panel.

Table A.1: *MoNet*'s inherent attributes.

Inherent Attribute Name	Function	Format
ID.STRN	Node Unique Identification	YYYY.MM.DD.hh.mm.ss.mmm
Select.BOOL	Operational selection	true/false
Tag.STRN	Node Name	Unformatted string
Node Type.STRG	Type of Node: LEAF or BRANCH	Unformatted string
IsCenter.BOOL	true when the node acts as the center of the	true/false
IsVersion.BOOL	true when the node is a version of other node existing in the network	true/false
IsCopy.BOOL	true when the node is a copy of other node existing in the network	true/false
Path.LINK	Node file address	File Path. URL
OwnerNode Tag.STRN	Owner node file address	File Path. URL
Node Degree.INTG	Number of nodes directly linked	Integer
Node Nature.STRG	Real system node nature	Unformatted string
Last Change	Time and attribute of last change of node	1-Dimensional <i>MoNet</i> Script
Last Update	Time and attribute of last update of node	1-Dimensional <i>MoNet</i> Script
Node Roll.STRG		Unformatted string
Node PropertyList.LIST	Sequence of attribute tags and values sepatated by the 0-Dimensional splitting-string ']'0['	1-Dimensional <i>MoNet</i> Script Format

A.5 Object description

Every object has attributes to establish identification and localization of the node-file. We give these attributes the category of '*Inherent*' thus, any time a new node is created by a system operator or is somehow '*born*' by the model dynamics, it is generated with these *inherent attributes* as the node core information.

A.6 Model description and data input

Information input in *MoNet* is actually inputting system elements description. Different from conventional software, in *MoNet* it is a matter of building up a network of files that resembling the real system. Each file looks and behaves as an element of the real system. Also, it may be formed or integrated by many other smaller-scaled elements living as files registered in a proper file folder. Therefore, to introduce a new object into the model, the user must first create

a property formatted element file descriptor, and record it in the final location within the file system. This is usually a design task which success often relies on the experience of the researcher.

A.7 Internal languages and syntaxes

Specific interpreters and multidimensional object representations had to be devised in order to build a platform capable of modeling complex systems at different scales. In addition to the conception of an architecture that allows for distributing the processing of data among several computers, the handling of complex abstract objects required the development of specialized representation of multidimensional structures and some non-conventional arithmetic to perform operations with them in a way that otherwise would be rather cumbersome.

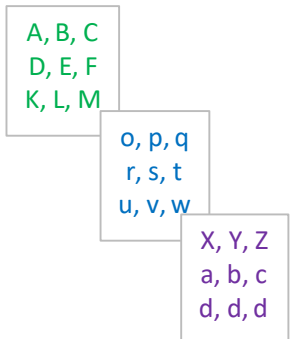
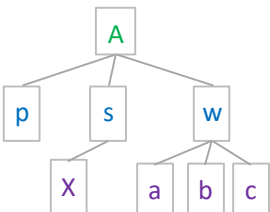
A.7.1 The Autonomous Multidimensional Object Representation

MoNet's treatment of dimensionality of complex data object is one of its important differences with conventional modeling programs. MoNet treats data objects as dimensional spaces. A number or a string is classified as a zero dimensional value. If the value is a number, then it worries about determining whether it is an INTeger or a FLOaTing type of number. When the object consists of a string of values, MoNet handles it as a 1-Dimensional list of values. Matrices are 2-dimensional structures. More complex data objects can also be represented. In general the string LIST or STRCT is used to indicate the most likely type of object an attribute name or function refers to. But this should not be interpreted as a rigorous rule. MoNet handles objects operations in spite of the names given to attributes and objects within the system; a direct consequence of MoNet's capability of operating over data objects, with no need for prior definition of the object spatial shape or dimensionality. MoNet's object representation carries itself the information about the object's topology. This characteristic is the main reason to call this syntax *the Autonomous Multidimensional Object Representation*, and it is crucial factor that determines the flexibility and power this software platform needs to effectively model complex systems where some objects can evolve and therefore change their own topology.

The Autonomous Representation integrates information about the dimensional shape of the object, and its size measured along each one of the object's dimensions. It is in fact a compact, self-contained, topological description of any object. The Autonomous Representation uses appropriately designed *Splitter-Strings*. These splitter-strings indicate where the boundaries among the smaller parts forming an object. The splitter-strings also say the dimensionality of

the smaller parts that end and begin just where the splitter-string is. A splitter-string id compound of three pieces: A splitter-string start signal (the closing squared bracket ']'), a sub-object dimensionality indicator (an integer number) and a splitter-string end signal (the opening squared bracket '['). Thus, a splitter-string that indicating the boundary between two 3-dimensional structures would be '3['.

Table A.2: Description of some structured objects coded in the Autonomous Representation.

Multidimensional structure representation			
Struct. Name	Struct. Dims.	Structure Depiction	Autonomous Representation
Scalar	0	A	A
Tuple	1	A,B	A]0[B
Vector	1	G, F, D, S, A	G]0[F]0[D]0[S]0[A
Matrix	2	G, F, D, S, A 1, 2, 3, 4, 5 v, w, x, y, z	G]1[F]1[D]1[S]1[A]0[1]1[2]1[3]1[4]1[5]0[v]1[w]1[x]1[y]1[z
Matrix	3		A]2[B]2[C]1[D]2[E]2[F]1[K]2[L]2[M]0[o]2[p]2[q]1[r]2[s]2[t]1[u]2[v]2[w]0[X]2[Y]2[Z]1[a]2[b]2[c]1[d]2[d]2[d
Tree	>1 <2		A]0[p]1[s]2[X]1[w]2[a]3[b]3[c

A.7.2 The Localizer pseudo-language syntax

MoNet uses a specially adapted syntax to indicate a function where to look for a value. The Localizer language uses tags to identify each dimension where the sought parameter can be found. A list of the delimiter tags is shown following:

Delimiter tag based on Attribute's name:	start: <@>	End: </@>
Delimiter tag based on Node's Path:	start: <~>	End: </~>
Indication that ANY value is valid:	<*Any*>	
Indication of self-element:	<*ThisVeryElement*>	
Look for the minimum value:	<*min*>	
Look for the maximum value:	<*max*>	
No action is needed:	<*Free*>	
Anything is valid:	<*Any*>	
Empty variable, property or descriptor:	<*Empty*>	

An example: the localizer string

```
<~><Path.LINK></~><LocalFrag Length.INTG><@><Tag.STRN> = <*Any*></@>
```

Would be interpreted as: Open the *Node.File* located where the tag <Path.LINK> indicates, and look for the *SubNode* where the attribute <Tag.STRN> is named as indicated in the right side of the conditional phrase (Any name in this case). Then extract the value of the attribute *LocalFrag Length.INTG* and return that value.

A.7.3 Functions and complex operations

Table A.3: List of transcendental functions routines

Transcendental functions	
Function Name	Returns
Abs	(Arg0). Returns the Absolute value of Arg0.
ArcTan	(Arg0). Returns the Arc tangent of Arg0.
Exp	(Arg0, Arg1). Returns the result of the expression $\text{Arg0}^{\text{Arg1}}$. Arg0 and Arg1 can be structures of zero to 4 dimensions.
Int	(Arg0). Returns the integer part of Arg0.
Lg2	(Arg0). Returns the Log of Arg0 with base 2 .
Log(Arg1,Arg2)	(Arg0, Arg1). Returns the Log of Arg0 with base Arg1.
PI	Returns the number π .
Pow	(Arg0, Arg1). Returns the result of $\text{Arg0}^{\text{Arg1}}$. Arg0 and Arg1 may be scalars or compatible structure.
Rand	Return a random number between 0 and 10000.
Sin	(Arg0). Returns the result of the expression $\text{Sin}(\text{Arg0})$. Arg0 can be structures of zero to 4 dimensions.
Sqrt	(Arg0). Returns the result of the expression $\text{SQRT}(\text{Arg0})$. Arg0 can be a structure of zero to 4 dimensions.
Sum	(Arg0). Returns the summation of all elements of structure Arg0.
Tan	(Arg0). Returns the result of the expression $\text{Tan}(\text{Arg0})$. Arg0 can be a structure of zero to 4 dimensions.

Table A.4: List of matrix operation functions

Matrix Operations	
Function Name	Returns
VectorPrune	(Arg0, Arg1, Arg2). Deletes the exeding elements from vector represented in Arg0. The result preserves the Arg1 elements from the side indicated in Arg2 (start or end).
MatrixFromVectorsSTRC	(Arg0, Arg1). Returns the matrix resulting from joining column vectors Arg0 and Arg1.
MultiplyMatrixSTRC	(Arg0, Arg1). Returns the product of matrixes Arg0 and Arg1.
VectorUnityLIST	(Arg0). Returns a vector of Arg0 elements withn value 1.
VctorNegOneLIST	(Arg0). Returns a vector of Arg0 elements withn value -1.
VectorOfFunctionLIST	Builds a vector with the values of the Arg1 coordinate of matrix Arg0.

Table A.5: List of probability distribution routines

Probability Distributions	
Function Name	Returns
Average	Returns the average of the scalars contained in argument's structure
CDF	(Arg0). Returns the addition of the probability density functions contained in Arg0 and Arg1. The returned value is presented as a 1D structure.
CDFInverseFunctionPointValue	(Arg0). Returns the values of the inverse function contained in the autonomus 2D string Arg0.
Count	Returns the count of single elements contained in the argument's structure
Normalize	(Arg0). Forms a Probability Density Function with the autonomus 2D string Arg0.
PDFAdd	(Arg0, Arg1). Returns the Cumulative Density Function of the probability density function contained in Arg0. The returned value is presented as a 1D structure.

Table A.6: List of file relative position functions

File Relative Position functions	
Function Name	Returns
SUBSETAverage	(Arg0). Returns the Average of the values whose location is specified in Arg0 according to the Localizer psuedo-language.
SUBSETStdDev	(Arg0). Returns the Standard Deviation of the values whose location is specified in Arg0 according to the Localizer psuedo-language.
SUBSETSumPDFs	(Arg0). Returns the Sum of the Probability Density Functions whose location is specified in Arg0 according to the Localizer psuedo-language.
SUBSETSumValues	(Arg0). Returns the Sum of the values whose location is specified in Arg0 according to the Localizer psuedo-language.
SUBSETSumProbHistograms1	(Arg0). Returns the Sum of the Histograms whose location is specified in Arg0 according to the Localizer psuedo-language.

Table A.7: List of discrete functions and structures

Discrete functions and Structures	
Function Name	Returns
AssembleStructDim	(Arg0, Arg1). Assembles a string with the values of the components of Arg0 indicated in Dimension Arg1.
Discrete1DFunctionMakeSTRC	(Arg0). Returns a string with the values of the components of 1D array Arg0 as an autonomuos structure.
DiscretelInnerFunctMakeSTRC	(Arg0). Returns a string with the values of the inner envelope of the components of the 2D autonomuos structure Arg0.
DiscreteFunctDerivativeSTRC	Returns de derivative of a discrete function presented in Arg0.
DiscreteFunctLargestRoot	(Arg0). Returns the largest root of the discrete function presented in structure Arg0.
DiscreteFunctSubstractionSTRC	(Arg0, Arg1). Returns the result of subtracting Arg1 from structure Arg0.
DiscreteOuterFunctMakeSTRC	(Arg0). Returns a string with the values of the outer envelope of the components of the 2D autonomuos structure Arg0.
GetDimValueSTRC	(Arg0). Returns the LIST of elements contained in a dimension of structure Arg0.
GetFractalDimValueSTRC	Returns a structure of 1 dimension contained in a tree. Arg1 specifies the criteria of selection of the elements within the tree.
GETLISTDIM	Builds a LIST from the dimension indicated of Arg0
GetPointValAtDimForValue	(Arg0). Returns the value of an element of structure in Arg0 located at the coordinates indicated.
GetScalarValueAtDimValueSTRC	(Arg0, Arg1, Arg2). Returns the LIST of Symbols of the structure expressed in Arg0. The resulting vector is built with the dimension expressed in Arg2.
GetStructDimSize	(Arg0). Returns the size of a dimension of structure in Arg0.
Ladder	Orders the elements of an 1-dimensional structure, in an increasing order
MakeRankSTRC	Builds an ordered 1-dimensionall structure with elements starting from 1 to the last element indicated in Arg0.
JoinStruct	Returns the struct resulting from joining Arg1 and Arg2. The resulting struct will increase its dimension
PropertyBasedHistogram	Returns the probability distribution of the elements represented in Arg0.
SegmentStruct	Segments the struct contained in Agr0. The segmentation is done based on the parameter Arg1.
SumStructElements	(Arg0). Returns the sumation of the values of the components of 1D array Arg0 as an autonomuos structure.
SymbolStructSplitByLength	(Arg1, Arg2, Arg3). Splits the string contained in Arg1 in strings of length Arg2-characters. The result is returned in the 1-dimesional array Arg3.
SymbolStructSplitBySymbol	(Arg1, Arg2, Arg3). Splits the string contained in Arg1 using Arg2 as a splitter string. The result is returned in the 1-dimesional array Arg3

Table A.8: List of language description functions

Language Description	
Function Name	Returns
Entropy	(Arg0). Returns the entropy of language Arg0
LANGFromNumericSet	Builds the set of numerical symbols according to their frequencies in the values of a function.
ZIPFRefProfile	(Arg0). Returns the list of v values corresponding to a Zipf reference distribution.
LANGScaleDowngrade	Returns the Downgraded list of symbol probabilities after changing the observation scale to a lower value.
LANGSymbolDowngrade	Returns the Downgraded list of symbol probabilities after changing the observation scale to a lower value.
SpecialCharAdjust	(Arg0). Replaces troublesome chars found in Arg0
LanguageComplexSymbolLIST	(Arg0, Arg1, Arg2). Returns the LIST of Fundamental Symbols of the Description expressed in Arg0. The maximum number of characters a symbol can be built with is Arg1. The routine starts assuming the set of symbols expressed in Arg2.
SymbolStruct1stCapitalEquiv	(Arg0). Returns a modified version of text Arg0. To modify the text Arg0, all capital letters after a period, are replaced with lower case letters, except when the word appears in other context of the same text with upper case as its first character.
EngApostrophesReplc	(Arg0). Returns a modified version of text Arg0. To modify the text Arg0, all contraction in English is replaced with the corresponding two words.
IsolateNonLetterChars	(Arg0). Surrounds with spaces any nonletter character contained in the string Arg0.

Table A.9: List of discrete functions and structures

Discrete functions and Structures	
Function Name	Returns
AssembleStructDim	(Arg0, Arg1). Assembles a string with the values of the components of Arg0 indicated in Dimension Arg1.
Discrete1DFunctionMakeSTRC	(Arg0). Returns a string with the values of the components of 1D array Arg0 as an autonomuos structure.
DiscreteInnerFunctMakeSTRC	(Arg0). Returns a string with the values of the inner envelope of the components of the 2D autonomuos structure Arg0.
DiscreteFunctDerivativeSTRC	Returns de derivative of a discrete function presented in Arg0.
DiscreteFunctLargestRoot	(Arg0). Returns the largest root of the discrete function presented in structure Arg0.
DiscreteFunctSubstractionSTRC	(Arg0, Arg1). Returns the result of subtracting Arg1 from structure Arg0.
DiscreteOuterFunctMakeSTRC	(Arg0). Returns a string with the values of the outer envelope of the components of the 2D autonomuos structure Arg0.
GetDimValueSTRC	(Arg0). Returns the LIST of elements contained in a dimension of structure Arg0.
GetFractalDimValueSTRC	Returns a structure of 1 dimension contained in a tree. Arg1 specifies the criteria of selection of the elements within the tree.
GETLISTDIM	Builds a LIST from the dimension indicated of Arg0
GetPointValAtDimForValue	(Arg0). Returns the value of an element of structure in Arg0 located at the coordinates indicated.
GetScalarValueAtDimValueSTRC	(Arg0, Arg1, Arg2). Returns the LIST of Symbols of the structure expressed in Arg0. The resulting vector is built with the dimension expressed in Arg2.
GetStructDimSize	(Arg0). Returns the size of a dimension of structure in Arg0.
Ladder	Orders the elements of an 1-dimensional structure, in an increasing order
MakeRankSTRC	Builds an ordered 1-dimensionall structure with elements starting from 1 to the last element indicated in Arg0.
JoinStruct	Returns the struct resulting from joining Arg1 and Arg2. The resulting struct will increase its dimension
PropertyBasedHistogram	Returns the probability distribution of the elements represented in Arg0.
SegmentStruct	Segments the struct contained in Agr0. The segmentation is done based on the parameter Arg1.
SumStructElements	(Arg0). Returns the sumation of the values of the components of 1D array Arg0 as an autonomuos structure.
SymbolStructSplitByLength	(Arg1, Arg2, Arg3). Splits the string contained in Arg1 in strings of length Arg2-characters. The result is returned in the 1-dimesional array Arg3.
SymbolStructSplitBySymbol	(Arg1, Arg2, Arg3). Splits the string contained in Arg1 using Arg2 as a splitter string. The result is returned in the 1-dimesional array Arg3

Table A.10: List of file system functions and other functions

File System	
Function Name	Returns
OnlyThePath	(Arg1). Returns the File Path where the file of Arg1 is located
CreateNodeCopy	(Arg1). Copies the File of Node represented in Arg1
Other	
Function Name	Returns
OptimBySteepestSlope	(Arg0, Arg1, Arg2, Arg3, Arg4, Arg5).
	Arg0: <min* 0 <max*> Type of optimization
	Arg1: <Objective Function> Objective Function
	Arg2: <Control Variable> Control Variable
	Arg4: <Lower Bound> Lower Bound Restriction
	Arg5: <Upper Bound> Upper Bound Restriction
LeadingTrim	(Arg0)

Appendix B

Properties of natural languages and programming languages texts

Properties of artificial and natural language texts. Details of actual texts can be seen in the link indicated below.

[Artificial texts](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Artificial.Properties.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Artificial.Properties.htm>

[English texts](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/English.Properties.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/English.Properties.htm>

[Spanish texts](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Spanish.Properties.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Spanish.Properties.htm>

Table B.1: Artificial texts.

Artificial language texts properties							
<i>L</i> Text Length		<i>d</i> Specific diversity [0-1]		<i>J_{1,D}</i> Zipf's diviation			
<i>D</i> Diversity		<i>h</i> Entropy [0-1]		<i>J_{θ,D}</i> Tail Zipf's diviation			
		<i>g</i> Zipf's exponent					
Text Name	<i>L</i>	<i>D</i>	<i>d</i>	<i>h</i>	<i>g</i>	<i>J_{1,D}</i>	<i>J_{θ,D}</i>
FibonacciNumbers.CSharp	62	27	0.435	0.921	0.788	0.100	-0.045
Math.Mime2d.MathLab	11376	120	0.011	0.681	1.599	2.334	0.249
Levenberg.MathLab	567	99	0.175	0.823	0.993	0.376	0.502
IsPrime.C	158	56	0.354	0.902	0.823	0.146	0.108
InsertAfterBefore.CSharp	141	37	0.262	0.935	0.738	0.615	0.632
MathLab.Fr.MathLab	1707	207	0.121	0.788	1.041	0.561	0.583
MathLab.pplane8.MathLab	68788	2157	0.031	0.586	1.252	0.435	1.751
MatrixLUDecomp.CSharp	416	52	0.125	0.856	1.007	1.011	0.413
MatrixFuncts.CSharp	8069	194	0.024	0.663	1.487	0.704	0.565
MathLab.Taller.MathLab	2162	122	0.056	0.740	1.190	1.383	0.472
MathLab.programa2.MathLab	9324	254	0.027	0.692	1.345	1.555	0.476
HeapSort.Java	314	59	0.188	0.857	0.956	0.534	0.010
HeapSort.CSharp	247	46	0.186	0.900	0.894	0.721	0.388
HanoiTowers.Java	484	92	0.190	0.847	0.892	0.624	0.266
CopyFolderNContent.CSharp	195	49	0.251	0.908	0.800	0.488	0.473
ChainedScatterTable.CSharp	201	46	0.229	0.890	0.842	0.570	0.046
BoolFunctWithMultiplexerLogic.C	1030	163	0.158	0.796	1.000	0.479	0.238
BlowfishEncryption.C	3574	669	0.187	0.706	1.254	-0.181	0.058
ExtendedEuclidean.C	86	24	0.279	0.902	0.846	0.443	0.149
GameOfLife.C	247	46	0.186	0.893	0.864	0.882	0.799
FTPFunctions.CSharp	11505	312	0.027	0.593	1.467	0.286	1.184
FiniteElements.MathLab	2748	295	0.107	0.731	1.079	0.551	0.467
Factorial.CSharp	36	21	0.583	0.965	0.578	0.063	0.078
MatrixLUDecomp.Phyton	298	40	0.134	0.882	0.956	1.250	-0.023
MetaWords.FormsAnsClasses.CSharp	1279	144	0.113	0.828	0.963	1.065	0.544
Sociodynamica.Module1	9617	290	0.030	0.666	1.271	1.705	0.496
Sociodynamica.Forms	2428	297	0.122	0.759	1.075	0.461	0.532
SnakeGame.C	1515	157	0.104	0.803	1.061	0.816	0.512
QuickSort.CSharp	364	56	0.154	0.896	0.882	0.934	0.216
Sociodynamica.Module2	7672	428	0.056	0.706	1.176	0.873	0.824
Sociodynamica.Module3	3363	223	0.066	0.770	1.086	1.288	0.822
Sumation.CSharp	71	25	0.352	0.895	0.850	0.208	0.051
WebSite.TiempoReal.Html	7495	565	0.075	0.586	1.264	0.250	0.434
WebSite.RistEuropa.Html	11713	503	0.043	0.595	1.321	0.668	0.852
WebSite.Inmogal.php	19299	647	0.034	0.632	1.261	0.966	1.185
ViscomSoft.Sca nnerActivex.CSharp	11275	623	0.055	0.534	1.405	-0.084	0.471
QuadraticPrograming.CSharp	433	72	0.166	0.848	0.927	0.713	0.427
Polinom.CSharp	86	33	0.384	0.916	0.760	0.228	0.049
PermutationAlgorithm.Java	1227	96	0.078	0.775	1.149	1.032	0.610
NetPlexMainForm.CSharp	59496	1218	0.020	0.531	1.432	0.232	1.449
NetPlex.Forms.CSharp	66281	1482	0.022	0.644	1.226	1.492	1.449
NetPlex.Classes.CSharp	18662	652	0.035	0.689	1.139	1.880	1.416
ModularInverse.C	91	31	0.341	0.902	0.865	0.248	0.148
MetaWordsMainForm.CSharp	65492	1081	0.017	0.531	1.462	0.279	1.478
PartDifEqtnsHeatEq.MathLab	677	104	0.154	0.774	1.053	0.613	-0.041
PartDifEqtnsLaplaceEq.MathLab	825	96	0.116	0.780	1.131	0.741	0.030
PermutationAlgorithm.Csharp	777	88	0.113	0.841	1.015	1.041	0.334
PartDifEqtnsWaveEqtn.MathLab	249	67	0.269	0.855	0.897	0.336	-0.017
mail.log.2	147418	3195	0.022	0.560	1.297	1.297	0.460
Apache.Access.log	168355	2870	0.017	0.534	1.464	0.445	0.542

Table B.2: English texts (1/3):

English texts properties							
<i>L</i> Text Length	<i>d</i> Specific diversity [0-1]			<i>J_{1,D}</i> Zipf's diviation			
<i>D</i> Diversity	<i>h</i> Entropy [0-1]			<i>J_{θ,D}</i> Tail Zipf's diviation			
	<i>g</i> Zipf's exponent						
Text Name	<i>L</i>	<i>D</i>	<i>d</i>	<i>h</i>	<i>g</i>	<i>J_{1,D}</i>	<i>J_{θ,D}</i>
1381.JohnBall	227	117	0.5154	0.9143	0.7680	-0.1156	-0.0788
1588.QueenElizabethI	359	156	0.4345	0.8786	0.9376	-0.1528	-0.0414
1601.Hamlet	150	97	0.6467	0.9501	0.7449	-0.1987	-0.0760
1601.QueenElizabethI	1140	388	0.3404	0.8647	0.8467	0.0056	0.1264
1606.LancelotAndrewes	9285	1540	0.1659	0.7374	1.0936	-0.0690	0.1691
1814.NapoleonBonaparte	182	95	0.5220	0.9202	0.7380	-0.0406	-0.1394
1833.ThomasBabington	15668	2647	0.1689	0.7460	0.9988	0.0602	0.4564
1849.LucretiaMott	7575	1720	0.2271	0.7705	0.9759	-0.0457	0.2653
1851.ErnestineLRose	8301	1630	0.1964	0.7643	0.9851	0.0630	0.3239
1851.SojournerTruth	436	180	0.4128	0.9185	0.7197	0.0651	0.1175
1861.AbrahamLincoln	4007	1018	0.2541	0.8077	0.9268	0.0116	0.1815
1863.AbrahamLincoln	292	143	0.4897	0.9270	0.6524	0.0557	-0.0018
1867.ElizabethCadyStanton	5847	1481	0.2533	0.7846	0.9869	-0.1277	0.2027
1873.SusanBAnthony	626	255	0.4073	0.8617	0.9043	-0.1505	-0.0215
1877.ChiefJoseph	183	92	0.5027	0.9257	0.7401	-0.0375	0.0102
1890.RusselConwell	17766	2207	0.1242	0.7483	0.9861	0.4742	0.6646
1892.FrancesEWHarper	4395	1244	0.2830	0.8053	0.9417	-0.0930	0.1430
1901.MarkTwain	660	255	0.3864	0.8889	0.7560	0.0641	0.0714
1903.BS.Eng.BjornstjerneBjornson	1651	573	0.3471	0.8496	0.8244	0.0487	0.0336
1906.MaryChurch	1558	585	0.3755	0.8524	0.8963	-0.1499	0.0423
1909.BS.SelmaLagerlof	2296	626	0.2726	0.8247	0.9300	0.0186	0.1479
1915.AnnaHoward	10633	1425	0.1340	0.7747	0.9506	0.5420	0.7245
1916.CarrieChapman	6120	1542	0.2520	0.7943	0.9360	-0.0303	0.2165
1916.HellenKeller	2557	854	0.3340	0.8294	0.9202	-0.1370	0.1029
1918.WoodrowWilson	2753	769	0.2793	0.8177	0.9053	-0.0112	0.1972
1920.CrystalEastman	2131	669	0.3139	0.8482	0.8341	0.0675	0.1753
1921.MarieCurie	921	307	0.3333	0.8642	0.8356	0.0582	0.1353
1923.BS.Eng.WilliamButlerYeats	320	167	0.5219	0.9197	0.6626	0.0119	-0.0906
1923.JamesMonroe	1178	417	0.3540	0.8493	0.8830	-0.0816	0.0324
1923.NL.Eng.WilliamButlerYeats	4257	1127	0.2647	0.8194	0.9079	0.0306	0.3058
1925.MaryReynolds	4300	852	0.1981	0.8022	0.9134	0.2870	0.4510
1930.NL.Eng.SinclairLewis	5707	1609	0.2819	0.7986	0.9534	-0.1360	0.1034
1932.MargaretSanger	1162	399	0.3434	0.8468	0.9113	-0.1212	0.1130
1936.EleanorRoosevelt	1966	457	0.2325	0.8301	0.8796	0.2744	0.3562
1936.KingEdwardVIII	596	243	0.4077	0.8747	0.7740	0.0342	0.0039
1936.NL.Eng.EugeneO'Neill	1177	407	0.3458	0.8381	0.8972	-0.0955	0.0109
1938.BS.PearlBuck	520	197	0.3788	0.8935	0.8255	-0.0226	0.0538
1938.NL.PearlBuck	10270	1825	0.1777	0.7666	0.9751	0.1511	0.4402
1940.05.WinstonChurchill	703	292	0.4154	0.8730	0.8546	-0.0931	-0.0164
1940.06.A.WinstonChurchill	3762	1067	0.2836	0.8228	0.9142	-0.0642	0.1929
1940.06.B.WinstonChurchill	4899	1189	0.2427	0.8022	0.9318	0.0320	0.2297
1941.AdolfHitler	10901	2228	0.2044	0.7691	0.9748	-0.0076	0.4059
1941.FranklinDRoosevelt	574	261	0.4547	0.8806	0.9119	-0.1906	-0.0467
1941.HaroldIckes	2448	720	0.2941	0.8221	0.8930	0.0248	0.1185
1942.MahatmaGandhi	1234	428	0.3468	0.8554	0.8496	0.0138	0.0437
1944.DwightEisenhower	208	120	0.5769	0.9248	0.7615	-0.1486	-0.1436
1944.GeorgePatton	873	313	0.3585	0.8861	0.8021	0.0323	0.1694
1945.BS.Eng.GabrielaMistral	370	196	0.5297	0.8833	0.9373	-0.2446	-0.1705

C. Literature Nobel laureates and non-laureates text properties

Table B.2: English texts (cont. 2/3):

1946.WinstonChurchill	1285	498	0.3875	0.8496	0.9141	-0.1538	0.0101
1947.GeorgeCMarshall	1606	582	0.3624	0.8422	0.9275	-0.1892	0.0749
1947.HarryTruman	2445	716	0.2928	0.8218	0.9296	-0.0612	0.1564
1948.BS.Eng.ThomasEliot	1467	504	0.3436	0.8448	0.8791	-0.0470	0.1376
1949.BS.Eng.WilliamFaulkner	622	248	0.3987	0.8838	0.7906	-0.0064	0.0733
1950.MargaretChase	1717	561	0.3267	0.8450	0.8948	-0.0602	0.0507
1950.NL.Eng.BertrandRussell	6476	1590	0.2455	0.7893	0.9457	-0.0280	0.1745
1953.DwightEisenhower	2906	830	0.2856	0.8108	0.9130	-0.0310	0.1212
1953.NelsonMandela	4967	1433	0.2885	0.8010	0.9572	-0.1801	0.1607
1954.BS.Eng.ErnestHemingway	367	183	0.4986	0.9195	0.7180	-0.0366	-0.1006
1957.MartinLutherKing	7952	1261	0.1586	0.7797	0.9508	0.3431	0.6792
1959.RichardFeynman	8135	1300	0.1598	0.7855	0.9460	0.3665	0.5540
1961.01.JohnFKennedy	1519	529	0.3483	0.8523	0.8439	-0.0371	0.1181
1961.04.JohnFKennedy	1715	605	0.3528	0.8454	0.8870	-0.0980	0.0729
1961.05.JohnFKennedy	6584	1535	0.2331	0.7991	0.8770	0.1528	0.3033
1961.11.JohnFKennedy	680	316	0.4647	0.8922	0.8045	-0.1246	-0.0369
1962.09.JohnFKennedy	2428	751	0.3093	0.8272	0.8962	-0.0497	0.1076
1962.10.JohnFKennedy	2772	811	0.2926	0.8287	0.8727	0.0344	0.1864
1962.12.MalcomX	17199	1640	0.0954	0.7573	1.0099	0.7080	0.8545
1962.BS.Eng.JohnSteinbeck	952	385	0.4044	0.8589	0.8830	-0.1454	-0.0257
1963.06.10.JohnFKennedy	3680	1019	0.2769	0.8149	0.8815	0.0568	0.1859
1963.06.26.JohnFKennedy	662	237	0.3580	0.8752	0.8416	-0.0101	0.1583
1963.09.20.JohnFKennedy	3986	1089	0.2732	0.8043	0.9247	-0.0233	0.1468
1963.MartinLutherKing	1731	527	0.3044	0.8366	0.8858	0.0428	0.1703
1964.04.MalcomX	3288	660	0.2007	0.8198	0.8880	0.3468	0.5379
1964.05.LyndonBJohnson	1168	430	0.3682	0.8484	0.8902	-0.0695	-0.0292
1964.LadybirdJohnson	818	353	0.4315	0.8762	0.8278	-0.1008	-0.0384
1964.MartinLutherKing	1266	498	0.3934	0.8616	0.8238	-0.0458	0.0172
1964.NelsonMandela	11929	2152	0.1804	0.7667	0.9602	0.1212	0.4337
1965.03.LyndonBJohnson	4166	975	0.2340	0.8052	0.9006	0.1485	0.2551
1965.04.LyndonBJohnson	1286	419	0.3258	0.8486	0.8581	0.0454	0.0569
1967.BS.Eng.MiguelAngelAsturias	1039	435	0.4187	0.8489	0.8953	-0.2004	0.0091
1967.MartinLutherKing	7360	1743	0.2368	0.7939	0.9337	0.0198	0.2773
1967.NL.Eng.MiguelAngelAsturias	5026	1479	0.2943	0.7899	0.9974	-0.2295	0.1170
1968.MartinLutherKing	5022	986	0.1963	0.7928	0.9192	0.2791	0.4393
1968.RobertFKennedy	627	197	0.3142	0.8978	0.7639	0.2170	0.2682
1969.IndiraGhandi	1058	408	0.3856	0.8674	0.8591	-0.0709	0.0118
1969.RichardNixon	5056	1105	0.2186	0.8048	0.9198	0.1308	0.3209
1969.ShirleyChisholm	966	382	0.3954	0.8671	0.8312	-0.0465	0.0192
1971.BS.Eng.PabloNeruda	503	209	0.4155	0.8701	0.8653	-0.1067	-0.1051
1971.NL.Eng.PabloNeruda	4114	1150	0.2795	0.8115	0.9101	-0.0349	0.1604
1972.JaneFonda	792	340	0.4293	0.8741	0.8661	-0.1875	-0.0853
1972.RichardNixon	5362	920	0.1716	0.7938	0.9442	0.3032	0.5676
1974.RichardNixon	1959	536	0.2736	0.8331	0.8944	0.0518	0.1618
1976.BS.Eng.SaulBellow	395	199	0.5038	0.9118	0.7613	-0.0981	-0.0452
1976.NL.Eng.SaulBellow	5625	1499	0.2665	0.7990	0.9128	-0.0142	0.1551
1977.NL.Eng.VicenteAleixandre	2618	845	0.3228	0.8267	0.9052	-0.0723	0.0825
1979.MargaretThatcher	3217	1002	0.3115	0.8204	0.9330	-0.1461	0.0988
1979.MotherTeresa	4349	652	0.1499	0.8055	0.9244	0.5909	0.4403
1981.RonaldReagan	1175	448	0.3813	0.8553	0.8701	-0.0838	0.0317
1982.NL.Eng.GabrielGarciaMarquez	2132	833	0.3907	0.8401	0.9209	-0.2080	0.0088
1982.RonaldReagan	5037	1392	0.2764	0.8062	0.9200	-0.0617	0.2275
1983.BS.Eng.WilliamGolding	369	201	0.5447	0.9131	0.8124	-0.1896	-0.1445
1983.NL.Eng.WilliamGolding	5140	1375	0.2675	0.8124	0.8904	0.0502	0.2202
1983.RonaldReagan	5160	1257	0.2436	0.8137	0.8787	0.1348	0.2692
1986.BS.Eng.WoleSoyinka	482	245	0.5083	0.8925	0.8611	-0.1972	-0.1240
1986.NL.Eng.WoleSoyinka	9033	2530	0.2801	0.7783	0.9740	-0.1930	0.1335

Table B.2: English texts (cont. 3/3)

1986.RonaldReagan	784	305	0.3890	0.8603	0.8460	-0.0231	0.0221
1987.RonaldReagan	3160	935	0.2959	0.8218	0.9151	-0.0794	0.1962
1988.AnnRichards	3109	863	0.2776	0.8265	0.8720	0.0951	0.2281
1989.BS.Eng.CamioJoseCela	464	227	0.4892	0.8917	0.8719	-0.1957	-0.0973
1989.NL.Eng.CamioJoseCela	5826	1541	0.2645	0.8058	0.9200	-0.0364	0.2301
1990.BS.Eng.OctavioPaz	636	300	0.4717	0.8686	0.8820	-0.1903	-0.0978
1990.NL.Eng.OctavioPaz	5704	1549	0.2716	0.7870	0.9619	-0.1224	0.1249
1991.BS.Eng.NadineGordimer	562	280	0.4982	0.8924	0.8368	-0.1779	-0.1242
1991.GeorgeBush	1771	582	0.3286	0.8438	0.8640	-0.0115	0.0453
1991.NL.Eng.NadineGordimer	4384	1254	0.2860	0.8021	0.9389	-0.0878	0.1359
1992.BS.Eng.DerekWalcott	104	68	0.6538	0.9300	0.8842	-0.1905	-0.2184
1992.NL.Eng.DerekWalcott	7403	1965	0.2654	0.7743	0.9897	-0.1805	0.1731
1993.BS.Eng.ToniMorrison	368	201	0.5462	0.9122	0.7638	-0.1078	-0.1185
1993.MayaAngelou	794	311	0.3917	0.8349	0.9974	-0.2499	-0.0170
1993.NL.Eng.ToniMorrison	3486	1023	0.2935	0.8116	0.9188	0.0002	0.0533
1993.SarahBrady	949	316	0.3330	0.8697	0.7814	0.1311	0.1014
1993.UrvashiVaid	1315	414	0.3148	0.8399	0.8832	0.0114	0.0994
1994.MotherTeresa	3953	638	0.1614	0.8150	0.9124	0.5166	0.7301
1994.NelsonMandela	1010	388	0.3842	0.8479	0.8798	-0.0936	-0.0319
1995.BS.Eng.SeamusHeaney	287	161	0.5610	0.9150	0.7679	-0.1065	-0.1995
1995.ErikaJong	2356	601	0.2551	0.8298	0.8624	0.1873	0.2864
1995.HillaryClinton	2483	715	0.2880	0.8225	0.8776	0.0696	0.1309
1995.NL.Eng.SeamusHeaney	7050	1897	0.2691	0.7871	0.9577	-0.1276	0.1895
1997.BillClinton	1302	417	0.3203	0.8451	0.8357	0.1107	0.0794
1997.EarlOfSpencer	1327	508	0.3828	0.8574	0.8216	-0.0041	-0.0359
1997.NancyBirdsall	2312	644	0.2785	0.8328	0.8783	0.0783	0.2596
1997.PrincessDiana	1753	602	0.3434	0.8498	0.8286	0.0174	0.1282
1997.QueenElizabethII	449	207	0.4610	0.8998	0.7511	-0.0231	-0.0017
1999.AnitaRoddick	2012	634	0.3151	0.8420	0.8630	0.0101	0.1252
2000.CondoleezzaRice	1510	517	0.3424	0.8540	0.8794	-0.0503	0.0159
2000.CourtneyLove	8166	1627	0.1992	0.8000	0.9275	0.1890	0.3911
2000.PopeJohnPaulII	823	327	0.3973	0.8747	0.8438	-0.0937	0.0713
2001.09.11.GeorgeWBush	670	296	0.4418	0.8807	0.7944	-0.0727	-0.0107
2001.09.13.GeorgeWBush	550	251	0.4564	0.8744	0.8626	-0.1634	0.0185
2001.BS.Eng.VSNaipaul	347	172	0.4957	0.8982	0.8301	-0.1708	-0.0516
2001.HalleBerry	636	219	0.3443	0.8496	0.8505	0.0575	-0.0875
2001.NL.Eng.VSNaipaul	6303	1220	0.1936	0.7873	0.9287	0.2345	0.4531
2002.OprahWinfrey	603	226	0.3748	0.8653	0.8559	-0.0308	0.0252
2003.BethChapman	877	334	0.3808	0.8764	0.8497	-0.0435	0.1807
2003.BS.Eng.JMCoetzee	329	151	0.4590	0.9138	0.7498	-0.0398	0.0728
2003.NL.Eng.JMCoetzee	4593	1105	0.2406	0.7934	0.9693	-0.0467	0.1642
2005.NL.Eng.HaroldPinter	5787	1478	0.2554	0.8023	0.9161	-0.0002	0.2048
2005.SteveJobs	2584	706	0.2732	0.8321	0.8750	0.0857	0.2561
2007.NL.Eng.DorisLessing	5881	1244	0.2115	0.7924	0.9689	0.0405	0.3426
2010.BS.Eng.MarioVargasLlosa	452	204	0.4513	0.8911	0.7897	-0.0722	-0.0271
2010.NL.Eng.MarioVargasLlosa	7073	2034	0.2876	0.7740	1.0218	-0.2736	0.0711
ErnestHemingway.TheOldManAndTheSea	14915	1804	0.1210	0.7438	1.0362	0.3077	0.6836
ErnestHemingway.TheOldManAndTheSea	16130	1775	0.1100	0.7385	1.0420	0.3865	0.8258
ErnestHemingway.TheSunAlsoRises	22951	2099	0.0915	0.7049	1.0839	0.4104	0.7308
IsaacAsimov.IRobot.Cap2	8772	1636	0.1865	0.7678	0.9552	0.1894	0.3930
IsaacAsimov.IRobot.Cap6	12812	1977	0.1543	0.7597	0.9808	0.2709	0.5145

C. Literature Nobel laureates and non-laureates text properties

Table B.3: Spanish texts (1/3)

Spanish texts properties							
<i>L</i> Text Length	<i>d</i> Specific diversity [0-1]		<i>J_{1,D}</i> Zipf's diviation		<i>J_{θ,D}</i> Tail Zipf's diviation		
<i>D</i> Diversity	<i>h</i> Entropy [0-1]		<i>g</i> Zipf's exponent				
Text Name	<i>L</i>	<i>D</i>	<i>d</i>	<i>h</i>	<i>g</i>	<i>J_{1,D}</i>	<i>J_{θ,D}</i>
1755.PatrickHenry	313	151	0.482	0.910	0.817	-0.131	-0.057
1805.Simón Bolívar	462	230	0.498	0.878	0.942	-0.190	-0.196
1813.Simón Bolívar	739	332	0.449	0.864	0.835	-0.086	-0.122
1819.Simón Bolívar	11502	2629	0.229	0.751	0.962	-0.030	0.226
1830.Simón Bolívar	201	121	0.602	0.930	0.745	-0.112	-0.120
1863.AbrahamLincoln	305	152	0.498	0.923	0.760	-0.115	0.010
1868.CarlosMCespedes	1457	591	0.406	0.836	0.945	-0.214	-0.082
1873.SusanBAnthony	594	237	0.399	0.870	0.829	-0.005	-0.042
1899.Vladimir Lenin	1920	644	0.335	0.817	0.948	-0.140	-0.002
1912.Emiliano Zapata	2590	935	0.361	0.811	0.974	-0.251	-0.051
1917.Emiliano Zapata	1619	653	0.403	0.826	0.985	-0.260	-0.094
1918.Emiliano Zapata	1438	593	0.412	0.830	0.987	-0.243	-0.131
1918.WoodrowWilson	303	171	0.564	0.909	0.813	-0.158	-0.160
1919.Georges Clemenceau	209	126	0.603	0.928	0.792	-0.179	-0.174
1919.Lloyd George	135	92	0.681	0.950	0.746	-0.191	-0.209
1921.Marie Curie.Esp	563	239	0.425	0.895	0.785	-0.010	0.022
1931.Manuel Azaña	297	152	0.512	0.906	0.832	-0.153	-0.118
1933.JAntonioPrimoDeRivera	3190	972	0.305	0.803	0.963	-0.158	0.099
1934.AdolfHitler	347	163	0.470	0.893	0.870	-0.146	-0.100
1936.Dolores Ibarruri	537	193	0.359	0.864	0.807	0.138	-0.061
1936.José Buena Ventura Durruti	690	305	0.442	0.877	0.796	-0.049	-0.065
1938.Dolores Ibarruri	774	318	0.411	0.846	0.962	-0.218	-0.042
1938.Leon Trotsky	1023	416	0.407	0.860	0.835	-0.063	-0.044
1938.Neville Chamberlain	638	302	0.473	0.883	0.827	-0.131	-0.077
1940.B.Winston Churchill	68	36	0.529	0.892	1.048	-0.136	-0.191
1940.Benito Mussolini	736	338	0.459	0.868	0.924	-0.221	-0.115
1940.Charles de Gaulle	122	69	0.566	0.928	0.756	-0.098	-0.153
1940.Winston Churchill	395	195	0.494	0.900	0.819	-0.123	-0.046
1941.Franklin Roosevelt	280	158	0.564	0.922	0.803	-0.151	-0.131
1941.Joseph Stalin	880	341	0.388	0.878	0.832	-0.021	0.036
1942.08.Mahatma Gandhi	2588	864	0.334	0.825	0.864	-0.025	0.049
1943.Heinrich Himmler	350	184	0.526	0.919	0.806	-0.197	-0.048
1943.Joseph Goebbels	1173	414	0.353	0.864	0.798	0.101	0.031
1945.Harry Truman	768	315	0.410	0.869	0.890	-0.110	-0.125
1945.Hirohito	766	352	0.460	0.868	0.871	-0.163	-0.144
1945.Juan Domingo Perón	1059	414	0.391	0.865	0.851	-0.069	0.053
1946.Jorge Eliécer Gaitán	3544	986	0.278	0.811	0.893	0.039	0.150
1947.George Marshall	754	328	0.435	0.867	0.871	-0.128	-0.051
1948.David Ben Gurion	1178	417	0.354	0.826	0.977	-0.149	-0.064
1950.Robert Schuman	998	385	0.386	0.840	0.956	-0.211	-0.076
1950.William Faulkner	533	233	0.437	0.895	0.761	-0.006	-0.002
1952.Eva Perón	1124	344	0.306	0.839	0.899	0.030	0.071
1953.Dwight D Eisenhower	1732	622	0.359	0.847	0.886	-0.115	0.058
1956.Gamar Abdel Nasser	839	337	0.402	0.868	0.821	-0.050	-0.059
1959.Fidel Castro	2892	853	0.295	0.810	0.908	-0.012	0.090
1959.Fulgencio Batista	85	58	0.682	0.947	0.783	-0.168	-0.172
1959.Nikita Krushchev	404	198	0.490	0.889	0.870	-0.161	-0.123
1961.J F Kennedy	1613	602	0.373	0.838	0.894	-0.136	0.014
1961.Nelson Mandela	5350	1373	0.257	0.778	0.945	-0.032	0.198
1962.J F Kennedy	319	160	0.502	0.905	0.836	-0.168	-0.061
1963.J F Kennedy	651	256	0.393	0.891	0.738	0.070	0.077

Table B.3: Spanish texts (cont. 2/3)

1963.Martin Luther King Jr	1746	578	0.331	0.828	0.945	-0.114	0.077
1964.Ernesto Che Guevara	7172	1911	0.266	0.779	0.961	-0.135	0.168
1964.Malcom X	824	321	0.390	0.877	0.776	0.028	0.034
1964.Nelson Mandela	5347	1372	0.257	0.778	0.944	-0.032	0.198
1964.Ronald Reagan	1062	450	0.424	0.875	0.789	-0.049	-0.026
1967.BS.Esp.MiguelAngelAsturias	804	339	0.422	0.845	0.959	-0.203	-0.127
1967.Ernesto Che Guevara	5868	1696	0.289	0.788	0.937	-0.120	0.135
1967.Fidel Castro	5519	1232	0.223	0.788	0.953	0.019	0.304
1967.Martin Luther King	7418	1924	0.259	0.786	0.944	-0.067	0.197
1967.NL.Esp.MiguelAngelAsturias	4901	1533	0.313	0.787	0.967	-0.184	0.038
1969.Richard Nixon	4501	1200	0.267	0.800	0.925	-0.026	0.210
1970.Salvador Allende	1865	718	0.385	0.834	0.898	-0.147	-0.044
1971.BS.Esp.PabloNeruda	468	209	0.447	0.859	0.946	-0.193	-0.120
1971.Pablo Neruda	3683	1290	0.350	0.806	0.948	-0.223	-0.019
1972.Salvador Allende	10046	2540	0.253	0.766	0.971	-0.141	0.228
1973.Augusto Pinochet	4191	1318	0.314	0.797	0.935	-0.121	0.040
1973.Bando Nro 5	801	366	0.457	0.860	0.925	-0.209	-0.143
1973.Salvador Allende	700	314	0.449	0.868	0.893	-0.174	-0.093
1974.Richard Nixon	741	302	0.408	0.879	0.775	0.009	0.002
1976.Jorge Videla	604	264	0.437	0.875	0.916	-0.183	-0.034
1977.BS.Esp.VicenteAleixandre	241	137	0.568	0.917	0.850	-0.209	-0.134
1977.NL.Esp.VicenteAleixandre	2379	859	0.361	0.818	0.988	-0.253	-0.010
1978.Juan Carlos I	973	411	0.422	0.848	0.925	-0.188	-0.092
1979.Adolfo Suárez	13201	2799	0.212	0.751	0.990	-0.102	0.407
1979.Ayatolá Jomeini	254	126	0.496	0.918	0.762	-0.049	-0.106
1979.Fidel Castro	12832	2668	0.208	0.743	0.989	-0.049	0.345
1981.Adolfo Suárez	1348	420	0.312	0.842	0.818	0.088	0.099
1981.Roberto Eduardo Viola	3823	1288	0.337	0.799	0.929	-0.174	0.043
1982.BS.Esp.GabrielGarciaMarquez	522	251	0.481	0.876	0.892	-0.157	-0.118
1982.Felipe González	6592	1818	0.276	0.782	0.940	-0.099	0.192
1982.Gabriel García Márquez	2095	856	0.409	0.831	0.949	-0.242	-0.073
1982.Leopoldo Galtieri	119	76	0.639	0.934	0.896	-0.243	-0.130
1982.Margaret Thatcher	586	242	0.413	0.895	0.776	-0.003	0.034
1983.Raúl Alfonsín	3309	976	0.295	0.805	0.896	0.010	0.094
1984.Ronald Reagan	790	339	0.429	0.864	0.825	-0.073	-0.059
1986.Ronald Reagan	729	323	0.443	0.879	0.862	-0.153	-0.091
1987.Camilo José Cela	1591	621	0.390	0.830	0.944	-0.205	-0.092
1987.Ronald Reagan	3150	1016	0.323	0.816	0.924	-0.144	0.083
1988.Gorbachov	1017	416	0.409	0.859	0.847	-0.093	-0.085
1989.Carlos Saúl Menem	1199	404	0.337	0.845	0.864	0.008	0.067
1989.NL.Esp.CamiloJoseCela	6291	1803	0.287	0.777	0.965	-0.148	0.085
1990.BS.Esp.OctavioPaz	613	284	0.463	0.878	0.850	-0.131	-0.109
1990.George H. W. Bush	654	269	0.411	0.881	0.854	-0.114	0.049
1990.NL.Esp.OctavioPaz	4804	1452	0.302	0.788	0.933	-0.076	0.050
1991.Boris Yeltsin	466	219	0.470	0.889	0.865	-0.160	-0.070
1991.Gorbachov	197	126	0.640	0.936	0.715	-0.133	-0.161
1992.Rafael Caldera	2504	832	0.332	0.810	0.932	-0.150	0.048
1992.Severn Suzuki	1001	403	0.403	0.869	0.876	-0.157	0.040
1993.Bill Clinton	2010	703	0.350	0.827	0.914	-0.098	-0.002
1996.Jose María Aznar	5069	1383	0.273	0.782	0.951	-0.104	0.197
1998.José Saramago	6235	1775	0.285	0.781	0.978	-0.182	0.106
1999.Elle Wiesel	806	328	0.407	0.854	0.971	-0.218	-0.068
1999.Hugo Chavez	12766	2441	0.191	0.760	1.002	-0.051	0.442
2000.Vicente Fox	7417	1998	0.269	0.778	0.929	-0.066	0.173
2001.Fernando de la Rúa	1129	436	0.386	0.853	0.825	-0.042	0.004
2001.George W. Bush	340	173	0.509	0.905	0.808	-0.155	-0.049
2001.Osama Bin Laden	455	215	0.473	0.891	0.801	-0.105	-0.062
2002.A.George W. Bush	590	271	0.459	0.887	0.820	-0.095	-0.032
2002.Barack Hussein Obama	983	379	0.386	0.840	0.900	-0.068	-0.073
2003.B.George W. Bush	564	237	0.420	0.886	0.823	-0.056	0.013
2003.George W. Bush	741	352	0.475	0.879	0.854	-0.164	-0.140
2003.José Saramago	1110	441	0.397	0.849	0.870	-0.083	-0.062

C. Literature Nobel laureates and non-laureates text properties

Table B.3: Spanish texts (cont. 3/3)

2004.Pilar Manjón	209	118	0.565	0.917	0.867	-0.209	-0.065
2005.Daniel Ortega	7593	1516	0.200	0.779	0.943	0.125	0.347
2005.Gerhard Schroeder	1547	559	0.361	0.843	0.875	-0.076	-0.011
2005.Steve Jobs	2524	832	0.330	0.831	0.880	-0.061	0.077
2006.Alvaro Uribe	4555	1552	0.341	0.776	0.969	-0.244	-0.030
2006.Dianne Feinstein	1503	525	0.349	0.841	0.888	-0.088	0.018
2006.Evo Morales	3391	890	0.262	0.812	0.981	-0.154	0.287
2006.Gastón Acurio	4348	1276	0.293	0.803	0.953	-0.148	0.122
2006.Hugo Chavez	3353	948	0.283	0.808	0.969	-0.150	0.164
2007.Al Gore	1319	580	0.440	0.859	0.903	-0.228	-0.097
2007.Cristina Kirchner	5004	1228	0.245	0.795	0.918	0.047	0.262
2007.Daniel Ortega	3373	857	0.254	0.805	0.969	-0.082	0.282
2008.Barack Hussein Obama	309	159	0.515	0.897	0.843	-0.120	-0.078
2008.J. L. Rodriguez Zapatero	449	204	0.454	0.886	0.803	-0.040	-0.120
2008.Julio Cobos	280	138	0.493	0.907	0.768	-0.049	-0.062
2008.Randy Paush	1817	624	0.343	0.847	0.875	-0.080	0.062
2009.Barack Hussein Obama	2834	978	0.345	0.817	0.894	-0.089	-0.002
2010.BS.Esp.MarioVargasLlosa	424	204	0.481	0.888	0.882	-0.217	-0.091
2010.Hillary Clinton	2426	832	0.343	0.831	0.874	-0.107	0.088
2010.NL.Esp.MarioVargasLlosa	7034	2215	0.315	0.763	1.035	-0.318	0.007
2010.Raúl Castro	260	145	0.558	0.912	0.877	-0.229	-0.141
2010.Sebastian Piñera Echenique	432	173	0.400	0.890	0.819	-0.037	0.025
CamiloJoseCela.LaColmena.Cap1	17409	3089	0.177	0.736	1.021	0.003	0.332
CamiloJoseCela.LaColmena.Cap2	15370	2943	0.191	0.741	1.000	-0.006	0.339
CamiloJoseCela.LaColmena.Cap6	3629	1117	0.308	0.798	0.990	-0.223	0.056
CamiloJoseCela.LaColmena.Notas4E	1623	596	0.367	0.829	0.954	-0.171	-0.031
ErnestHemingway.ElViejoYElMar.Par	13979	2498	0.179	0.751	0.975	0.116	0.452
ErnestHemingway.ElViejoYElMar.Par	15446	2424	0.157	0.743	0.993	0.186	0.542
ErnestHemingway.Fiesta.Libro1	17642	3064	0.174	0.733	1.016	0.018	0.422
GabrielGMarquez.CronMuerteAnunc	12454	2621	0.210	0.754	0.948	0.080	0.248
GabrielGMarquez.CronMuerteAnunc	12680	2760	0.218	0.754	0.944	0.058	0.246
GabrielGMarquez.CronMuerteAnunc	6751	1586	0.235	0.774	0.933	0.088	0.193
GabrielGMarquez.DicursoCartagena	1443	579	0.401	0.844	0.910	-0.175	-0.081
GabrielGMarquez.MejorOficioDelMu	2949	1059	0.359	0.808	0.948	-0.186	-0.051
IsaacAsimov.YoRobot.Cap2	8080	1856	0.230	0.767	0.967	-0.020	0.220
IsaacAsimov.YoRobot.Cap6	12235	2391	0.195	0.754	0.968	0.075	0.380
JorgeLuisBorges.ElCongreso	6656	1926	0.289	0.774	0.963	-0.140	0.014
JorgeLuisBorges.ElMuerto	2109	753	0.357	0.814	0.950	-0.174	-0.067
JorgeLuisBorges.ElSur	2746	948	0.345	0.800	0.984	-0.193	-0.044
JorgeLuisBorges.LasRuinasCirculare:	2238	824	0.368	0.826	0.920	-0.138	-0.046
JoseSaramago.Valencia	3711	1126	0.303	0.786	1.045	-0.290	0.033
MarioVargasLlosa.DicursoBuenosAi	1984	776	0.391	0.819	0.967	-0.246	-0.081
MiguelAAsturias.SrPresidente.Parte:	4352	1269	0.292	0.786	0.975	-0.143	0.057
OctavioPaz.DicursoZacatecas	2238	711	0.318	0.810	0.949	-0.101	0.013
OctavioPaz.LaberintoSoledad.Part3	7054	1843	0.261	0.757	0.991	-0.143	0.065

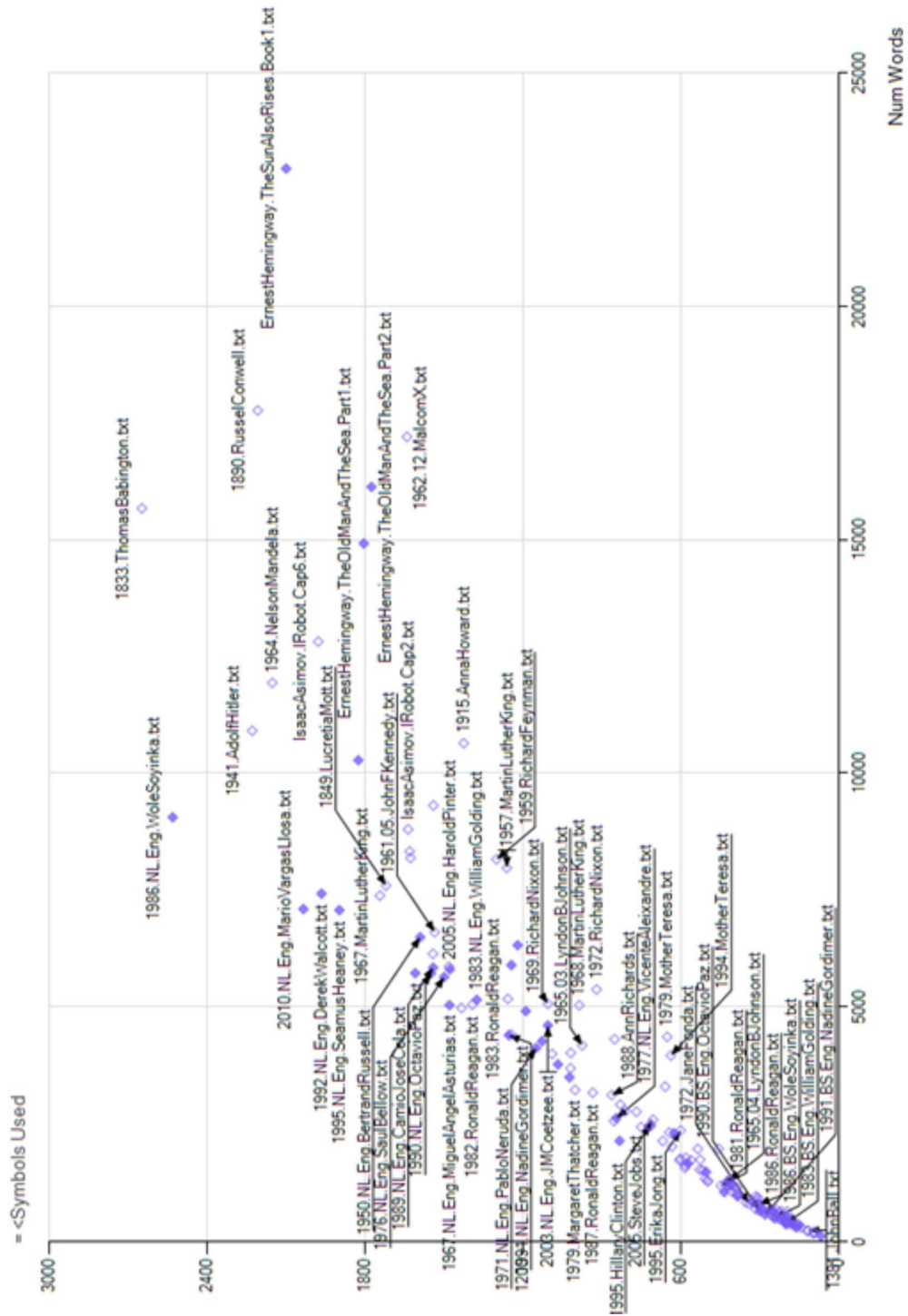


Figure B.1: Diversity of words used in Spanish speeches and novel segments vs. text length in words.

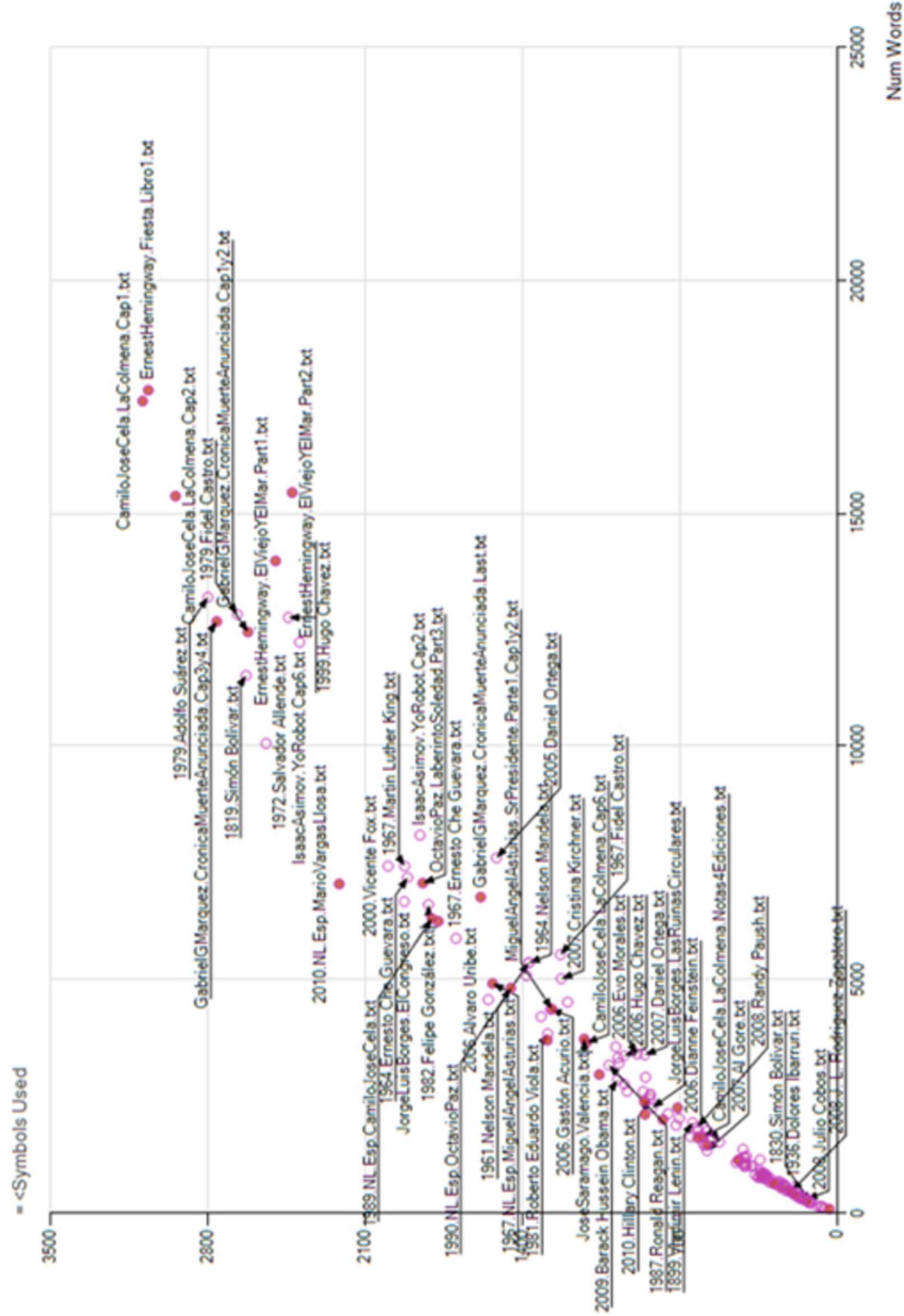


Figure B.2: Diversity of words used in Spanish speeches and novel segments vs. text length in words.

Appendix C

Literature Nobel laureates and non-laureates text properties

Properties of English and Spanish texts from literature Nobel laureates and non-Nobel writers. Details of actual texts can be seen in the link indicated below.

[English texts. Non-Nobel texts.](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/English.WQS-RES.NonNobels.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/English.WQS-RES.NonNobels.htm>

[English texts. Literature Nobel laureate texts.](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/English.WQS-RES.LiterNobels.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/English.WQS-RES.LiterNobels.htm>

[Spanish texts. Non-Nobel texts.](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Spanish.WQS-IPSZ.NonNobels.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Spanish.WQS-IPSZ.NonNobels.htm>

[Spanish texts. Literature Nobel laureate texts.](http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Spanish.WQS-IPSZ.LiterNobels.htm)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/NatArtifLangs/Whole/Spanish.WQS-IPSZ.LiterNobels.htm>

C. Literature Nobel laureates and non-laureates text properties

Table C.1:

English texts: non-Nobel laureates. Readability and Writing Quality Scale comparisson												
Index	Text Name	Genre	Lang	h Entropy [0-1]		drel relative specific diversity [0-1]		hrel relative Entropy [0-1]		J1,D Zipf's deviation	RES Flesch Reading Easy Score	WQS Writing Quality Scale
				d	h	drel	hrel	J1,D	RES	WQS		
ET1	1814.NapoleonBonaparte	S	T	0.522	0.920	-0.2170	0.0087	-0.0406	55.3122	0.4525		
ET2	1921.MarieCurie	S	T	0.333	0.864	-0.1462	0.0233	0.0582	71.7192	-0.7511		
ET3	1941.AdolfHitler	S	T	0.204	0.769	0.1833	-0.0174	-0.0076	76.4677	0.7855		
ET4	1979.MotherTeresa	S	T	0.150	0.805	-0.3591	0.0423	0.5909	52.0252	-9.2807		
ET5	1994.MotherTeresa	S	T	0.161	0.815	-0.3314	0.0469	0.5166	68.3648	-8.1092		
ET6	2000.PopeJohnPaulII	S	T	0.397	0.875	-0.0194	0.0078	-0.0937	85.2700	1.8933		
E1	1381.JohnBall	S	O	0.515	0.914	-0.1684	0.0049	-0.1156	59.9047	1.8183		
E2	1588.QueenElizabethI	S	O	0.435	0.879	-0.1844	-0.0027	-0.1528	74.4296	2.4497		
E3	1601.Hamlet	S	O	0.647	0.950	-0.0899	0.0139	-0.1987	73.3751	3.2760		
E4	1601.QueenElizabethI	S	O	0.340	0.865	-0.0646	0.0209	0.0056	53.0996	0.2296		
E5	1606.LancelotAndrewes	S	O	0.166	0.737	-0.0893	-0.0326	-0.0690	55.0698	1.5793		
E6	1833.ThomasBabington	S	O	0.169	0.746	0.1025	-0.0253	0.0602	65.1686	-0.3280		
E7	1849.LucretiaMott	S	O	0.227	0.771	0.1658	-0.0256	-0.0457	83.3508	1.4091		
E8	1851.ErnestineRose	S	O	0.196	0.764	0.0391	-0.0187	0.0630	48.8446	-0.4255		
E9	1851.SojournerTruth	S	O	0.413	0.919	-0.1738	0.0455	0.0651	54.4254	-1.0505		
E10	1861.AbrahamLincoln	S	O	0.254	0.808	0.0572	0.0001	0.0116	53.1160	0.3574		
E11	1863.AbrahamLincoln	S	O	0.490	0.927	-0.1414	0.0259	0.0557	57.3512	-0.8227		
E12	1867.ElizabethCadyStanton	S	O	0.253	0.785	0.1940	-0.0227	-0.1277	88.6036	2.7269		
E13	1873.SusanBAnthony	S	O	0.407	0.862	-0.0814	-0.0091	-0.1505	71.8227	2.7051		
E14	1877.ChiefJoseph	S	O	0.503	0.926	-0.2445	0.0203	-0.0375	54.3828	0.2927		
E15	1890.RusselConwell	S	O	0.124	0.748	-0.1550	-0.0039	0.4742	75.7833	-7.0522		
E16	1892.FrancesEWHarper	S	O	0.283	0.805	0.2143	-0.0153	-0.0915	39.9448	2.1982		
E17	1901.MarkTwain	S	O	0.386	0.889	-0.1134	0.0263	0.0641	72.7968	-0.8170		
E18	1906.MaryChurch	S	O	0.375	0.852	0.1440	-0.0058	-0.1499	61.9777	3.0566		
E19	1915.AnnaHoward	S	O	0.134	0.775	-0.2305	0.0183	0.5420	61.0696	-8.2541		
E20	1916.CarrieChapman	S	O	0.252	0.794	0.2057	-0.0124	-0.0303	60.8585	1.1782		
E21	1916.HellenKeller	S	O	0.334	0.829	0.1983	-0.0118	-0.1370	45.3861	2.9046		
E22	1918.WoodrowWilson	S	O	0.279	0.818	0.0269	-0.0006	-0.0112	56.7073	0.6883		
E23	1920.CrystalEastman	S	O	0.314	0.848	0.0606	0.0153	0.0675	56.1624	-0.5422		
E24	1923.JamesMonroe	S	O	0.354	0.849	-0.0166	-0.0002	-0.0816	36.3421	1.7338		
E25	1925.MaryReynolds	S	O	0.198	0.802	-0.1561	0.0184	0.2870	84.2138	-4.2377		
E26	1932.MargaretSanger	S	O	0.343	0.847	-0.0503	0.0017	-0.1212	53.0309	2.2962		
E27	1936.EleanorRoosevelt	S	O	0.232	0.830	-0.2353	0.0317	0.2744	56.2773	-4.2295		
E28	1936.KingEdwardVIII	S	O	0.408	0.875	-0.0954	0.0037	0.0342	65.3628	-0.2582		
E29	1941.FranklinDRoosevelt	S	O	0.455	0.881	-0.0036	-0.0082	-0.1906	48.2128	3.4850		
E30	1941.Haroldickes	S	O	0.294	0.822	0.0402	-0.0024	0.0248	62.8989	0.1461		
E31	1942.MahatmaGandhi	S	O	0.347	0.855	-0.0215	0.0089	0.0138	63.5228	0.2040		
E32	1944.DwightEisenhower	S	O	0.577	0.925	-0.0956	-0.0011	-0.1486	61.7413	2.5448		
E33	1944.GeorgePatton	S	O	0.359	0.886	-0.0977	0.0348	0.0323	80.0396	-0.2950		
E34	1947.GeorgeCMarshall	S	O	0.362	0.842	0.1152	-0.0106	-0.1892	47.2257	3.6418		
E35	1947.HarryTruman	S	O	0.293	0.822	0.0353	-0.0022	-0.0612	46.4032	1.4898		
E36	1950.MargaretChase	S	O	0.327	0.845	0.0279	0.0068	-0.0602	48.1757	1.4439		
E37	1953.DwightEisenhower	S	O	0.286	0.811	0.0689	-0.0101	-0.0310	59.9117	1.0738		
E38	1953.NelsonMandela	S	O	0.289	0.801	0.2887	-0.0211	-0.1801	43.0789	3.6483		
E39	1957.MartinLutherKing	S	O	0.159	0.780	-0.1727	0.0128	0.3431	65.5787	-5.0861		
E40	1959.RichardFeynman	S	O	0.160	0.785	-0.1600	0.0180	0.3665	66.4433	-5.4560		
E41	1961.01.JohnFKennedy	S	O	0.348	0.852	0.0522	0.0052	-0.0371	56.0718	1.1232		
E42	1961.04.JohnFKennedy	S	O	0.353	0.845	0.1094	-0.0036	-0.0980	54.3368	2.1805		
E43	1961.05.JohnFKennedy	S	O	0.233	0.799	0.1429	0.0004	0.1528	47.2569	-1.7914		
E44	1961.11.JohnFKennedy	S	O	0.465	0.892	0.0769	-0.0003	-0.1246	52.3144	2.5821		
E45	1962.09.JohnFKennedy	S	O	0.309	0.827	0.0909	-0.0038	-0.0497	56.0597	1.3826		
E46	1962.10.JohnFKennedy	S	O	0.293	0.829	0.0780	0.0048	0.0344	47.9640	0.0194		
E47	1962.12.MalcomX	S	O	0.095	0.757	-0.3583	0.0174	0.7080	72.7674	-10.9145		

C. Literature Nobel laureates and non-laureates text properties

Table C.1: (cont.)

English texts: non-Nobel laureates. Readability and Writing Quality Scale comparison										
Index	Text Name	Genre	Lang	d		hrel		J _{1,D}	RES	WQS
				d	h	drel	hrel			
E48	1963.06.10.JohnFKennedy	S	O	0.277	0.815	0.1203	-0.0023	0.0568	56.3615	-0.2739
E49	1963.06.26.JohnFKennedy	S	O	0.358	0.875	-0.1777	0.0241	-0.0101	60.7442	0.2292
E50	1963.09.20.JohnFKennedy	S	O	0.273	0.804	0.1349	-0.0114	-0.0233	52.3290	1.0192
E51	1963.MartinLutherKing	S	O	0.304	0.837	-0.0396	0.0077	0.0428	61.7331	-0.2700
E52	1964.04.MalcomX	S	O	0.201	0.820	-0.2175	0.0349	0.3468	72.6743	-5.3061
E53	1964.05.LyndonBJohnson	S	O	0.368	0.848	0.0199	-0.0068	-0.0695	60.2291	1.6193
E54	1964.LadybirdJohnson	S	O	0.432	0.876	0.0629	-0.0040	-0.1008	61.7123	2.1837
E55	1964.MartinLutherKing	S	O	0.393	0.862	0.1191	-0.0037	-0.0458	50.7437	1.3926
E56	1964.NelsonMandela	S	O	0.180	0.767	0.0760	-0.0095	0.1212	53.1394	-1.3414
E57	1965.03.LyndonBJohnson	S	O	0.234	0.805	-0.0136	0.0061	0.1485	64.1089	-1.8824
E58	1965.04.LyndonBJohnson	S	O	0.326	0.849	-0.0682	0.0108	0.0454	68.0541	-0.3693
E59	1967.MartinLutherKing	S	O	0.237	0.794	0.2044	-0.0064	0.0198	55.1386	0.3628
E60	1968.MartinLutherKing	S	O	0.196	0.793	-0.1198	0.0098	0.2791	75.3576	-4.0453
E61	1968.RobertFKennedy	S	O	0.314	0.898	-0.2911	0.0649	0.2170	63.6677	-3.6741
E62	1969.IndiraGhandi	S	O	0.386	0.867	0.0340	0.0051	-0.0709	65.4832	1.6338
E63	1969.RichardNixon	S	O	0.219	0.805	-0.0180	0.0123	0.1308	57.8786	-1.6256
E64	1969.ShirleyChisholm	S	O	0.395	0.867	0.0290	0.0009	-0.0465	53.8575	1.2548
E65	1972.JaneFonda	S	O	0.429	0.874	0.0462	-0.0053	-0.1875	44.3886	-4.5441
E66	1972.RichardNixon	S	O	0.172	0.794	-0.2140	0.0213	0.3032	69.9506	3.5186
E67	1974.RichardNixon	S	O	0.274	0.833	-0.1010	0.0172	0.0518	54.6775	-0.5165
E68	1979.MargaretThatcher	S	O	0.311	0.820	0.2055	-0.0114	-0.1461	54.8761	3.0375
E69	1981.RonaldReagan	S	O	0.381	0.855	0.0584	-0.0052	-0.0838	60.9046	1.9008
E70	1982.RonaldReagan	S	O	0.276	0.806	0.2401	-0.0108	-0.0617	52.5699	1.7139
E71	1983.RonaldReagan	S	O	0.244	0.814	0.1019	0.0105	0.1348	55.2849	-1.5653
E72	1986.RonaldReagan	S	O	0.389	0.860	-0.0551	-0.0034	-0.0231	73.2132	0.7504
E73	1987.RonaldReagan	S	O	0.296	0.822	0.1384	-0.0035	-0.0794	58.8383	1.8934
E74	1988.AnnRichards	S	O	0.278	0.826	0.0623	0.0089	0.0951	68.2167	-0.9664
E75	1991.GeorgeBush	S	O	0.329	0.844	0.0445	0.0049	-0.0115	61.5888	0.7076
E76	1993.MayaAngelou	S	O	0.392	0.835	-0.0447	-0.0298	-0.2499	69.3620	4.4012
E77	1993.SarahBrady	S	O	0.333	0.870	-0.1386	0.0289	0.1311	70.7339	0.1381
E78	1993.UrvashiVaid	S	O	0.315	0.840	-0.0930	0.0067	0.0114	63.9636	-1.8950
E79	1994.NelsonMandela	S	O	0.384	0.848	0.0144	-0.0138	-0.0936	55.1333	2.0087
E80	1995.ErikaJong	S	O	0.255	0.830	-0.1092	0.0217	0.1873	56.5715	-2.6580
E81	1995.HillaryClinton	S	O	0.288	0.822	0.0232	0.0005	0.0696	54.1626	-0.5856
E82	1997.BillClinton	S	O	0.320	0.845	-0.0803	0.0096	0.1107	71.2576	-1.4077
E83	1997.EarlOfSpencer	S	O	0.383	0.857	0.1062	-0.0037	-0.0041	57.9058	-0.8324
E84	1997.NancyBirdsall	S	O	0.279	0.833	-0.0333	0.0148	0.0783	45.2409	0.7152
E85	1997.PrincessDiana	S	O	0.343	0.850	0.0878	0.0047	0.0174	61.3143	0.3149
E86	1997.QueenElizabethII	S	O	0.461	0.900	-0.0684	0.0087	-0.0231	64.9646	0.6688
E87	1999.AnitaRoddick	S	O	0.315	0.842	0.0446	0.0087	0.0101	60.3981	0.3569
E88	2000.CondoleezzaRice	S	O	0.342	0.854	0.0324	0.0093	-0.0503	63.3798	1.2903
E89	2000.CourtneyLove	S	O	0.199	0.800	0.0486	0.0157	0.1890	71.2193	-2.4893
E90	2001.09.11.GeorgeWBush	S	O	0.442	0.881	0.0188	-0.0034	-0.0727	65.8042	1.6648
E91	2001.09.13.GeorgeWBush	S	O	0.456	0.874	-0.0139	-0.0150	-0.1634	57.2094	3.0551
E92	2001.HalleBerry	S	O	0.344	0.850	-0.2194	0.0042	0.0575	52.0924	-0.8636
E93	2002.OprahWinfrey	S	O	0.375	0.865	-0.1652	0.0074	-0.0308	83.1871	0.6137
E94	2003.BethChapman	S	O	0.381	0.876	-0.0401	0.0161	-0.0435	79.1260	1.0474
E95	2005.SteveJobs	S	O	0.273	0.832	-0.0163	0.0163	0.0857	62.0435	-0.9303
E96	IsaacAsimov.IRobot.Cap2	N	O	0.187	0.768	0.0050	-0.0110	0.1894	73.2634	-2.4601
E97	IsaacAsimov.IRobot.Cap6	N	O	0.154	0.760	-0.0577	-0.0054	0.2709	82.1102	-3.7988

C. Literature Nobel laureates and non-laureates text properties

Table C.2:

Spanish texts: literature Nobel laureates. Readability and Writing Quality Scale comparisson										
Index	Text Name	Genre [S = Speech : N = Novel segment/Story] h Entropy [0-1]				Lang [S = Spanish : T = Translation to Spanish] drel relative specific diversity [0-1]		J1,D Zipf's deviation		
		Genre	Lang	d	h	drel	hrel	J1,D	IPSZ	WQS
EN1	1903.BS.Eng.BjornstjerneBjornson	S	O	0.347	0.850	0.0778	0.0030	0.0487	37.7653	-0.1832
EN2	1909.BS.SelmaLagerlof	S	O	0.273	0.825	-0.0559	0.0092	0.0186	63.3497	0.0911
EN3	1923.BS.Eng.WilliamButlerYeats	S	O	0.522	0.920	-0.0569	0.0083	0.0119	51.9322	0.1080
EN4	1923.NL.Eng.WilliamButlerYeats	S	O	0.265	0.819	0.1238	0.0073	0.0306	53.5183	0.1300
EN5	1930.NL.Eng.SinclairLewis	S	O	0.282	0.799	0.3184	-0.0208	-0.1360	45.8358	2.9699
EN6	1936.NL.Eng.EugeneO'Neill	S	O	0.346	0.838	-0.0396	-0.0080	-0.0955	49.2763	1.9359
EN7	1938.BS.PearlBuck	S	O	0.379	0.893	-0.1964	0.0339	-0.0226	61.4490	0.3313
EN8	1938.NL.PearlBuck	S	O	0.178	0.767	0.0088	-0.0085	0.1511	56.8045	-1.8622
EN9	1940.05.WinstonChurchill	S	O	0.415	0.873	-0.0268	-0.0010	-0.0931	62.9192	1.8917
EN10	1940.06.A.WinstonChurchill	S	O	0.284	0.823	0.1559	0.0027	-0.0642	56.0341	1.6485
EN11	1940.06.B.WinstonChurchill	S	O	0.243	0.802	0.0792	-0.0006	0.0320	52.6761	0.0567
EN12	1946.WinstonChurchill	S	O	0.388	0.850	0.1080	-0.0135	-0.1538	51.2463	3.0963
EN13	1948.BS.Eng.ThomasEliot	S	O	0.344	0.845	0.0261	-0.0003	-0.0470	52.6855	1.2559
EN14	1949.BS.Eng.WilliamFaulkner	S	O	0.399	0.884	-0.1028	0.0163	-0.0064	64.5715	0.3323
EN15	1950.NL.Eng.BertrandRussell	S	O	0.246	0.789	0.1970	-0.0147	-0.0280	55.3765	1.1375
EN16	1954.BS.Eng.ErnestHemingway	S	O	0.499	0.919	-0.0572	0.0154	-0.0366	66.3650	0.8770
EN17	1962.BS.Eng.JohnSteinbeck	S	O	0.404	0.859	0.0473	-0.0108	-0.1454	55.3406	2.8707
EN18	1976.BS.Eng.SaulBellow	S	O	0.504	0.912	-0.0241	0.0061	-0.0981	64.1628	1.9462
EN19	1976.NL.Eng.SaulBellow	S	O	0.266	0.799	0.2402	-0.0139	-0.0142	59.1397	0.9673
EN20	1983.BS.Eng.WilliamGolding	S	O	0.545	0.913	0.0317	-0.0049	-0.1896	62.6378	3.5411
EN21	1983.NL.Eng.WilliamGolding	S	O	0.268	0.812	0.2085	-0.0009	0.0502	64.8120	-0.1032
EN22	1986.BS.Eng.WoleSoyinka	S	O	0.508	0.892	0.0515	-0.0148	-0.1972	43.6372	3.7238
EN23	1986.NL.Eng.WoleSoyinka	S	O	0.280	0.778	0.5240	-0.0404	-0.1930	43.2831	4.0432
EN24	1991.BS.Eng.NadineGordimer	S	O	0.498	0.892	0.0842	-0.0116	-0.1779	56.6878	3.4734
EN25	1991.NL.Eng.NadineGordimer	S	O	0.286	0.802	0.2261	-0.0191	-0.0878	53.4544	2.1420
EN26	1992.BS.Eng.DerekWalcott	S	O	0.654	0.930	-0.1846	-0.0066	-0.1905	24.2505	2.8317
EN27	1992.NL.Eng.DerekWalcott	S	O	0.265	0.774	0.3525	-0.0381	-0.1805	52.4729	3.7205
EN28	1993.BS.Eng.ToniMorrison	S	O	0.546	0.912	0.0336	-0.0062	-0.1078	56.9804	2.2667
EN29	1993.NL.Eng.ToniMorrison	S	O	0.293	0.812	0.1663	-0.0127	0.0002	65.4967	0.6959
EN30	1995.BS.Eng.SeamusHeaney	S	O	0.561	0.915	-0.0220	-0.0073	-0.1065	53.5405	2.1145
EN31	1995.NL.Eng.SeamusHeaney	S	O	0.269	0.787	0.3492	-0.0268	-0.1276	49.3848	2.8628
EN32	2001.BS.Eng.VSNaipaul	S	O	0.496	0.898	-0.0800	-0.0049	-0.1708	70.4715	2.9811
EN33	2001.NL.Eng.VSNaipaul	S	O	0.194	0.787	-0.0647	0.0054	0.2345	69.6769	-3.2749
EN34	2003.BS.Eng.JMCoetzee	S	O	0.459	0.914	-0.1630	0.0234	-0.0398	77.9339	0.6442
EN35	2003.NL.Eng.JMCoetzee	S	O	0.241	0.793	0.0472	-0.0085	-0.0467	65.3898	1.2826
EN36	2005.NL.Eng.HaroldPinter	S	O	0.255	0.802	0.1998	-0.0058	-0.0002	65.9238	0.6858
EN37	2007.NL.Eng.DorisLessing	S	O	0.212	0.792	-0.0010	0.0029	0.0405	69.1879	-0.1665
EN38	ErnestHemingway.TheOldManAndTheSea.Part1	N	O	0.121	0.744	-0.2234	-0.0070	0.3077	82.9544	-4.4847
EN39	ErnestHemingway.TheOldManAndTheSea.Part2	N	O	0.110	0.738	-0.2750	-0.0076	0.3865	82.6875	-5.7503
EN40	ErnestHemingway.TheSunAlsoRises.Book1	N	O	0.091	0.705	-0.3231	-0.0333	0.4104	94.1593	-6.0639
ENT1	1945.BS.Eng.GabrielA.Mistral	S	T	0.530	0.883	0.0042	-0.0305	-0.2446	80.0260	4.4109
ENT2	1967.BS.Eng.MiguelAngelAsturias	S	T	0.419	0.849	0.1159	-0.0264	-0.2004	40.2646	3.8880
ENT3	1967.NL.Eng.MiguelAngelAsturias	S	T	0.294	0.790	0.3196	-0.0347	-0.2295	57.9760	4.4907
ENT4	1971.BS.Eng.PabloNeruda	S	T	0.416	0.870	-0.1283	-0.0040	-0.1067	47.9302	1.8938
ENT5	1971.NL.Eng.PabloNeruda	S	T	0.280	0.811	0.1733	-0.0069	-0.0349	52.8923	1.2278
ENT6	1977.NL.Eng.VicenteAleixandre	S	T	0.323	0.827	0.1671	-0.0098	-0.0723	49.7503	1.2278
ENT7	1982.NL.Eng.GabrielGarciaMarquez	S	T	0.391	0.840	0.3202	-0.0242	-0.2080	56.0816	4.2274
ENT8	1989.BS.Eng.CamiloJoseCela	S	T	0.489	0.892	-0.0006	-0.0093	-0.1957	44.9517	3.5743
ENT9	1989.NL.Eng.CamiloJoseCela	S	T	0.265	0.806	0.2453	-0.0063	-0.0364	52.4467	1.2975
ENT10	1990.BS.Eng.OctavioPaz	S	T	0.472	0.869	0.0693	-0.0264	-0.1903	45.4568	3.6718
ENT11	1990.NL.Eng.OctavioPaz	S	T	0.272	0.787	0.2697	-0.0280	-0.1224	58.2610	2.7308
ENT12	2010.BS.Eng.MarioVargasLlosa	S	T	0.451	0.891	-0.0860	0.0035	-0.0722	57.6237	1.4151
ENT13	2010.NL.Eng.MarioVargasLlosa	S	T	0.288	0.774	0.4434	-0.0478	-0.2736	52.6690	5.2930

Table C.3:

Spanish texts: non-Nobel laureates. Readability and Writing Quality Scale comparisson										
Index	Text Name	Genre	Lang	Entropy [0-1]				Zipf's deviation		
				d	h	drel	hrel	J1,D	IPSZ	WQS
				relative specific diversity [0-1]				Szigritsz perspicuity inde		
				relative Entropy [0-1]				Writing Quality Scale		
ST1	1755.PatrickHenry	S	T	0.482	0.910	-0.0895	0.0204	-0.1309	77.192	-0.331
ST2	1863.AbrahamLincoln	S	T	0.498	0.923	-0.0655	0.0268	-0.1147	63.068	-0.204
ST3	1873.SusanBAnthony	S	T	0.399	0.870	-0.1162	0.0178	-0.0055	66.189	-1.091
ST4	1899.Vladimir Lenin	S	T	0.335	0.817	-0.0038	-0.0047	-0.1405	50.814	0.481
ST5	1918.WoodrowWilson	S	T	0.564	0.909	0.0565	-0.0111	-0.1577	61.409	1.081
ST6	1919.Georges Clemenceau	S	T	0.603	0.928	0.0285	-0.0040	-0.1792	73.798	0.922
ST7	1919.Lloyd George	S	T	0.681	0.950	0.0423	0.0049	-0.1906	46.434	1.077
ST8	1921.MarieCurie.Esp	S	T	0.425	0.895	-0.0722	0.0306	-0.0097	65.611	-0.714
ST9	1934.Adolf Hitler	S	T	0.470	0.893	-0.0903	0.0084	-0.1460	70.292	-0.257
ST10	1938.Leon Trotsky	S	T	0.407	0.860	0.0319	0.0040	-0.0634	59.841	0.446
ST11	1938.Neville Chamberlain	S	T	0.473	0.883	0.0675	-0.0032	-0.1313	59.572	1.051
ST12	1940.Benito Mussolini	S	T	0.459	0.868	0.0733	-0.0121	-0.2212	62.692	1.497
ST13	1940.Charles de Gaulle	S	T	0.566	0.928	-0.1566	0.0075	-0.0981	64.763	-1.029
ST14	1941.Franklin Roosevelt	S	T	0.564	0.922	0.0357	0.0010	-0.1514	55.837	0.858
ST15	1941.Joseph Stalin	S	T	0.388	0.878	-0.0530	0.0307	-0.0206	56.737	-0.503
ST16	1942.08.Mahatma Gandhi	S	T	0.334	0.825	0.0684	0.0045	-0.0251	78.587	0.596
ST17	1943.Heinrich Himmler	S	T	0.526	0.919	0.0203	0.0114	-0.1970	70.493	0.905
ST18	1943.Joseph Goebbels	S	T	0.353	0.864	-0.0732	0.0335	0.1007	70.605	-1.198
ST19	1945.Harry Truman	S	T	0.410	0.869	-0.0312	0.0110	-0.1099	65.903	0.093
ST20	1945.Hirohito	S	T	0.460	0.868	0.0847	-0.0119	-0.1625	52.164	1.345
ST21	1947.George Marshall	S	T	0.435	0.867	0.0228	-0.0020	-0.1276	54.028	0.651
ST22	1948.David Ben Gurion	S	T	0.354	0.826	-0.0695	-0.0050	-0.1490	40.487	-0.045
ST23	1950.Robert Schuman	S	T	0.386	0.840	-0.0271	-0.0058	-0.2110	46.696	0.583
ST24	1950.William Faulkner	S	T	0.437	0.895	-0.0575	0.0243	-0.0062	74.087	-0.594
ST25	1953.Dwight D Eisenhower	S	T	0.359	0.847	0.0395	0.0140	-0.1146	53.995	0.716
ST26	1956.Gamar Abdel Nasser	S	T	0.402	0.868	-0.0300	0.0141	-0.0501	58.889	-0.156
ST27	1959.Nikita Khrushchev	S	T	0.490	0.889	-0.0141	-0.0041	-0.1613	60.915	0.481
ST28	1961.J F Kennedy	S	T	0.373	0.838	0.0613	-0.0017	-0.1359	60.538	1.016
ST29	1961.Nelson Mandela	S	T	0.257	0.778	-0.0152	-0.0029	-0.0317	64.026	-0.083
ST30	1962.J F Kennedy	S	T	0.502	0.905	-0.0489	0.0069	-0.1678	75.548	0.194
ST31	1963.J F Kennedy	S	T	0.393	0.891	-0.1087	0.0408	0.0699	69.876	-1.382
ST32	1963.Martin Luther King Jr	S	T	0.331	0.828	-0.0398	0.0090	-0.1139	68.380	0.039
ST33	1964.Walcom X	S	T	0.390	0.877	-0.0635	0.0288	0.0284	75.042	-0.799
ST34	1964.Nelson Mandela	S	T	0.257	0.778	-0.0155	-0.0028	-0.0318	63.910	-0.085
ST35	1964.Ronald Reagan	S	T	0.424	0.875	0.0854	0.0112	-0.0485	59.968	0.831
ST36	1967.Martin Luther King	S	T	0.259	0.786	0.0801	0.0035	-0.0666	62.289	0.874
ST37	1969.Richard Nixon	S	T	0.267	0.800	-0.0202	0.0138	-0.0257	61.460	-0.175
ST38	1974.Richard Nixon	S	T	0.408	0.879	-0.0459	0.0220	0.0089	62.354	-0.556
ST39	1979.Ayatolá Jomeini	S	T	0.496	0.918	-0.1114	0.0225	-0.0487	73.056	-0.873
ST40	1982.Margaret Thatcher	S	T	0.413	0.895	-0.0883	0.0358	-0.0031	66.736	-0.888
ST41	1984.Ronald Reagan	S	T	0.429	0.864	0.0208	-0.0030	-0.0730	71.437	0.402
ST42	1986.Ronald Reagan	S	T	0.443	0.879	0.0330	0.0055	-0.1528	68.187	0.835
ST43	1987.Ronald Reagan	S	T	0.323	0.816	0.0842	0.0015	-0.1437	64.967	1.241
ST44	1988.Gorbachov	S	T	0.409	0.859	0.0365	0.0015	-0.0930	54.310	0.615
ST45	1990.George H. W. Bush	S	T	0.411	0.881	-0.0667	0.0222	-0.1144	62.149	-0.209
ST46	1991.Boris Yeltsin	S	T	0.470	0.889	-0.0203	0.0045	-0.1596	55.557	0.408
ST47	1991.Gorbachov	S	T	0.640	0.936	0.0752	-0.0035	-0.1333	68.773	1.126
ST48	1992.Severn Suzuki	S	T	0.403	0.869	0.0161	0.0144	-0.1567	75.017	0.694
ST49	1993.Bill Clinton	S	T	0.350	0.827	0.0508	-0.0020	-0.0983	66.216	0.766
ST50	1999.Elise Wiesel	S	T	0.407	0.854	-0.0271	-0.0022	-0.2175	69.381	0.606
ST51	2001.George W. Bush	S	T	0.509	0.905	-0.0196	0.0045	-0.1551	59.143	0.394
ST52	2001.Osama Bin Laden	S	T	0.473	0.891	-0.0208	0.0052	-0.1053	74.907	0.171
ST53	2002.A.George W. Bush	S	T	0.459	0.887	0.0158	0.0067	-0.0949	65.758	0.438
ST54	2002.Barack Hussein Obama	S	T	0.386	0.840	-0.0313	-0.0060	-0.0680	67.871	-0.062
ST55	2003.B.George W. Bush	S	T	0.420	0.886	-0.0811	0.0235	-0.0557	65.029	-0.585
ST56	2003.George W. Bush	S	T	0.475	0.879	0.1121	-0.0077	-0.1636	57.980	1.578
ST57	2005.Gerhard Schroeder	S	T	0.361	0.843	0.0168	0.0084	-0.0765	64.530	0.366

C. Literature Nobel laureates and non-laureates text properties

Table C.3: (cont.)

ST58	2005.Steve Jobs	S	T	0.330	0.831	0.0484	0.0123	-0.0615	75.390	0.567
ST59	2006.Dianne Feinstein	S	T	0.349	0.841	-0.0241	0.0125	-0.0883	70.932	0.059
ST60	2007.Al Gore	S	T	0.440	0.859	0.1891	-0.0125	-0.2285	64.044	2.522
ST61	2008.Barack Hussein Obama	S	T	0.515	0.897	-0.0320	-0.0056	-0.1204	76.954	0.155
ST62	2008.Randy Paush	S	T	0.343	0.847	0.0061	0.0216	-0.0798	80.083	0.269
ST63	2009.Barack Hussein Obama	S	T	0.345	0.817	0.1298	-0.0095	-0.0886	64.936	1.413
ST64	2010.Hillary Clinton	S	T	0.343	0.831	0.0800	0.0063	-0.1075	49.720	1.043
ST65	IsaacAsimov.YoRobot.Cap2	N	T	0.230	0.767	-0.0228	0.0003	-0.0198	84.229	-0.204
ST66	IsaacAsimov.YoRobot.Cap6	N	T	0.195	0.754	-0.0778	0.0049	0.0754	81.050	-1.088
S1	1805.Simón Bolívar	S	O	0.498	0.878	0.0356	-0.0184	-0.1901	68.058	1.051
S2	1813.Simón Bolívar	S	O	0.449	0.864	0.0510	-0.0116	-0.0863	56.710	0.730
S3	1819.Simón Bolívar	S	O	0.229	0.751	0.0621	-0.0153	-0.0302	59.935	0.592
S4	1830.Simón Bolívar	S	O	0.602	0.930	0.0171	-0.0020	-0.1123	68.505	0.536
S5	1868.CarlosMCéspedes	S	O	0.406	0.836	0.1245	-0.0194	-0.2141	46.264	1.917
S6	1912.Emiliano Zapata	S	O	0.361	0.811	0.1556	-0.0228	-0.2512	45.508	2.347
S7	1917.Emiliano Zapata	S	O	0.403	0.826	0.1480	-0.0289	-0.2595	44.732	2.326
S8	1918.Emiliano Zapata	S	O	0.412	0.830	0.1394	-0.0292	-0.2428	42.040	2.182
S9	1931.Manuel Azaña	S	O	0.512	0.906	-0.0467	0.0037	-0.1526	61.079	0.152
S10	1933.JAntonioPrimoDeRivera	S	O	0.305	0.803	0.0275	-0.0025	-0.1575	64.229	0.819
S11	1936.Dolores Ibarruri	S	O	0.359	0.864	-0.2237	0.0311	0.1382	66.835	0.253
S12	1936.José Buenaventura Durruti	S	O	0.442	0.877	0.0165	0.0049	-0.0492	69.053	-2.646
S13	1938.Dolores Ibarruri	S	O	0.411	0.846	-0.0276	-0.0124	-0.2177	58.630	0.617
S14	1945.Juan Domingo Perón	S	O	0.391	0.865	0.0007	0.0165	-0.0695	67.165	0.186
S15	1946.Jorge Eliécer Gaitán	S	O	0.278	0.811	-0.0368	0.0192	0.0390	52.904	-0.602
S16	1952.Eva Perón	S	O	0.306	0.839	-0.2049	0.0325	0.0302	66.053	-2.026
S17	1959.Fidel Castro	S	O	0.295	0.810	-0.0295	0.0095	-0.0116	66.794	1.052
S18	1959.Fulgencio Batista	S	O	0.682	0.947	-0.0703	0.0014	-0.1683	46.517	0.020
S19	1964.Ernesto Che Guevara	S	O	0.266	0.779	0.1002	-0.0072	-0.1346	64.026	1.353
S20	1967.Ernesto Che Guevara	S	O	0.289	0.788	0.1351	-0.0095	-0.1203	49.783	1.593
S21	1967.Fidel Castro	S	O	0.223	0.788	-0.1367	0.0240	0.0186	56.669	-1.379
S22	1970.Salvador Allende	S	O	0.385	0.834	0.1352	-0.0119	-0.1467	56.961	1.711
S23	1972.Salvador Allende	S	O	0.253	0.766	0.1358	-0.0129	-0.1413	45.267	1.694
S24	1973.Augusto Pinochet	S	O	0.314	0.797	0.1354	-0.0142	-0.1215	51.864	1.608
S25	1973.Bando Nro 5	S	O	0.457	0.859	0.0000	0.0000	-0.2092	48.134	1.606
S26	1973.Salvador Allende	S	O	0.449	0.868	0.0353	-0.0075	-0.1743	65.375	0.965
S27	1976.Jorge Videla	S	O	0.437	0.875	-0.0277	0.0043	-0.1833	56.853	0.445
S28	1978.Juan Carlos I	S	O	0.422	0.848	0.0586	-0.0153	-0.1884	45.654	1.236
S29	1979.Adolfo Suárez	S	O	0.212	0.751	0.0198	-0.0071	-0.1023	42.830	0.524
S30	1979.Fidel Castro	S	O	0.208	0.743	-0.0071	-0.0122	-0.0490	50.503	0.074
S31	1981.Adolfo Suárez	S	O	0.312	0.842	-0.1529	0.0327	0.0881	61.705	-1.827
S32	1981.Roberto Eduardo Viola	S	O	0.337	0.799	0.1887	-0.0232	-0.1744	48.864	2.304
S33	1982.Felipe González	S	O	0.276	0.782	0.1150	-0.0086	-0.0994	48.228	1.331
S34	1982.Leopoldo Galtieri	S	O	0.639	0.934	-0.0535	-0.0057	-0.2430	45.919	-7.158
S35	1983.Raúl Alfonsín	S	O	0.295	0.805	0.0038	0.0043	0.0097	58.433	0.808
S36	1989.Carlos Saúl Menem	S	O	0.337	0.845	-0.1103	0.0231	0.0081	41.623	-1.107
S37	1992.Rafael Caldera	S	O	0.332	0.810	0.0546	-0.0098	-0.1498	46.915	1.030
S38	1996.Jose María Aznar	S	O	0.273	0.782	0.0330	-0.0079	-0.1035	62.146	0.643
S39	1999.Hugo Chavez	S	O	0.191	0.760	-0.0880	0.0129	-0.0507	58.903	-0.650
S40	2000.Vicente Fox	S	O	0.269	0.778	0.1217	-0.0098	-0.0660	63.589	1.248
S41	2001.Fernando de la Rúa	S	O	0.386	0.853	0.0044	0.0066	-0.0422	48.038	0.116
S42	2004.Pilar Manjón	S	O	0.565	0.917	-0.0368	-0.0032	-0.2086	66.429	0.486
S43	2005.Daniel Ortega	S	O	0.200	0.779	-0.1637	0.0275	0.1246	54.243	-2.068
S44	2006.Alvaro Uribe	S	O	0.341	0.776	0.2560	-0.0477	-0.2444	57.430	3.215
S45	2006.Evo Morales	S	O	0.262	0.812	-0.1013	0.0277	-0.1543	53.388	-0.343
S46	2006.Gastón Acurio	S	O	0.293	0.803	0.0693	0.0034	-0.1481	57.135	1.129
S47	2006.Hugo Chavez	S	O	0.283	0.808	-0.0346	0.0136	-0.1495	54.065	0.229
S48	2007.Cristina Kirchner	S	O	0.245	0.795	-0.0739	0.0199	0.0469	46.475	-0.955
S49	2007.Daniel Ortega	S	O	0.254	0.805	-0.1312	0.0252	-0.0824	59.970	-0.903
S50	2008.J. L. Rodríguez Zapatero	S	O	0.454	0.886	-0.0616	0.0078	-0.0399	74.018	-0.462
S51	2008.Julio Cobos	S	O	0.493	0.907	-0.0954	0.0130	-0.0492	44.405	-0.720
S52	2010.Raúl Castro	S	O	0.558	0.912	0.0048	-0.0069	-0.2289	65.089	0.935
S53	2010.Sebastian Piñera Echenique	S	O	0.400	0.890	-0.1808	0.0363	-0.0374	75.856	-1.537
S54	JorgeLuisBorges.ElCongreso	N	O	0.289	0.774	0.1727	-0.0242	-0.1398	75.849	2.021
S55	JorgeLuisBorges.ElMuerto	N	O	0.357	0.814	0.0857	-0.0185	-0.1740	73.621	1.412
S56	JorgeLuisBorges.ElSur	N	O	0.345	0.800	0.1213	-0.0263	-0.1925	70.208	1.808
S57	JorgeLuisBorges.LasRuinasCircularé	N	O	0.368	0.826	0.1363	-0.0121	-0.1380	72.796	1.683

Table C.4:

Spanish texts: literature Nobel laureates. Readability and Writing Quality Scale comparisson											
Index	Text Name	Genre	Lang.	h Entropy [0-1]		drel relative specific diversity [0-1]		hrel relative Entropy [0-1]		J1,D Zipf's deviation	
				d	h	drel	hrel	J1,D	ISZ	WQS	
SN1	1967.BS.Esp.MiguelAngelAsturias	S	O	0.422	0.845	0.0074	-0.018	-0.2033	72.818	0.865	
SN2	1967.NL.Esp.MiguelAngelAsturias	S	O	0.313	0.787	0.1743	-0.023	-0.1844	58.872	2.223	
SN3	1971.BS.Esp.PabloNeruda	S	O	0.447	0.859	-0.068	-0.016	-0.1928	63.482	0.169	
SN4	1971.Pablo Neruda	S	O	0.35	0.806	0.2243	-0.023	-0.2226	59.511	2.815	
SN5	1977.BS.Esp.VicenteAleixandre	S	O	0.568	0.917	0.005	-0.005	-0.2092	64.889	0.851	
SN6	1977.NL.Esp.VicenteAleixandre	S	O	0.361	0.818	0.1315	-0.016	-0.2534	67.929	2.140	
SN7	1982.BS.Esp.GabrielGarciaMarquez	S	O	0.481	0.876	0.0313	-0.013	-0.1567	58.857	0.864	
SN8	1982.Gabriel Garcia Márquez	S	O	0.409	0.831	0.2403	-0.026	-0.2419	54.692	3.039	
SN9	1987.Camilo José Cela	S	O	0.390	0.830	0.1061	-0.0181	-0.2049	65.607	1.718	
SN10	1989.NL.Esp.CamiloJoseCela	S	O	0.287	0.777	0.1453	-0.02	-0.1478	61.068	1.813	
SN11	1990.BS.Esp.OctavioPaz	S	O	0.463	0.878	0.0344	-0.004	-0.1311	54.83	0.767	
SN12	1990.NL.Esp.OctavioPaz	S	O	0.302	0.788	0.1291	-0.016	-0.076	64.924	1.363	
SN13	2010.BS.Esp.MarioVargasLlosa	S	O	0.481	0.888	-0.02	-0.001	-0.2166	69.764	0.658	
SN14	2010.NL.Esp.MarioVargasLlosa	S	O	0.315	0.763	0.294	-0.048	-0.3184	63.788	3.857	
SN15	CamiloJoseCela.La Colmena.Cap1	S	O	0.177	0.736	-0.085	-0.003	0.0027	49.293	-0.832	
SN16	CamiloJoseCela.La Colmena.Cap2	S	O	0.191	0.741	-0.043	-0.005	-0.0057	88.945	-0.433	
SN17	CamiloJoseCela.La Colmena.Cap6	S	O	0.308	0.798	0.0719	-0.009	-0.2226	88.846	1.488	
SN18	CamiloJoseCela.La Colmena.Notas4Ediciones	S	O	0.367	0.829	0.0458	-0.008	-0.1715	82.21	1.043	
SN19	GabrielGMarquez.CronMuerteAnunciada.Cap1y2	N	O	0.21	0.754	-0.002	-0.003	0.0802	71.553	-0.451	
SN20	GabrielGMarquez.CronMuerteAnunciada.Cap3y4	N	O	0.218	0.754	0.0364	-0.006	0.0578	70.095	-0.017	
SN21	GabrielGMarquez.CronMuerteAnunciada.Last	N	O	0.235	0.774	-0.044	0.0048	0.0882	75.775	-0.857	
SN22	GabrielGMarquez.DicursoCartagena	S	O	0.401	0.844	0.1097	-0.009	-0.1753	67.05	1.61	
SN23	GabrielGMarquez.MejorOficioDelMundo	S	O	0.359	0.808	0.1874	-0.025	-0.1859	55.767	2.345	
SN24	MarioVargasLlosa.DicursoBuenosAires	S	O	0.391	0.819	0.1713	-0.029	-0.2463	50.209	2.47	
SN25	MiguelAAsturias.SrPresidente.Parte1.Cap1y2	N	O	0.292	0.786	0.0627	-0.013	-0.1428	76.649	1.074	
SN26	OctavioPaz.DicursoZacatecas	S	O	0.318	0.81	-0.02	-0.002	-0.1014	68.929	0.175	
SN27	OctavioPaz.LaBerintoSoledad.Part3	N	O	0.261	0.757	0.0744	-0.027	-0.1427	70.767	1.1927	
SNT1	1940.B.Winston Churchill	S	T	0.529	0.892	-0.3179	-0.0166	-0.1364	72.469	-2.214	
SNT2	1940.Winston Churchill	S	T	0.494	0.900	-0.0125	0.0058	-0.1234	57.863	0.319	
SNT3	1998.José Saramago	S	T	0.285	0.781	0.1351	-0.0141	-0.1817	67.259	1.862	
SNT4	2003.José Saramago	S	T	0.397	0.849	0.0290	-0.0028	-0.0835	83.911	0.516	
SNT5	ErnestHemingway.ElViejoYElMar.Part1	N	T	0.179	0.751	-0.1281	0.0106	0.1155	85.557	-1.700	
SNT6	ErnestHemingway.ElViejoYElMar.Part2	N	T	0.157	0.743	-0.2150	0.0147	0.1858	90.348	-2.750	
SNT7	ErnestHemingway.Fiesta.Libro1	N	T	0.174	0.733	-0.1019	-0.0046	0.0184	77.504	-1.038	
SNT8	JoseSaramago.Valencia	S	T	0.303	0.786	0.0626	-0.0193	-0.2904	49.823	1.711	

C. Literature Nobel laureates and non-laureates text properties

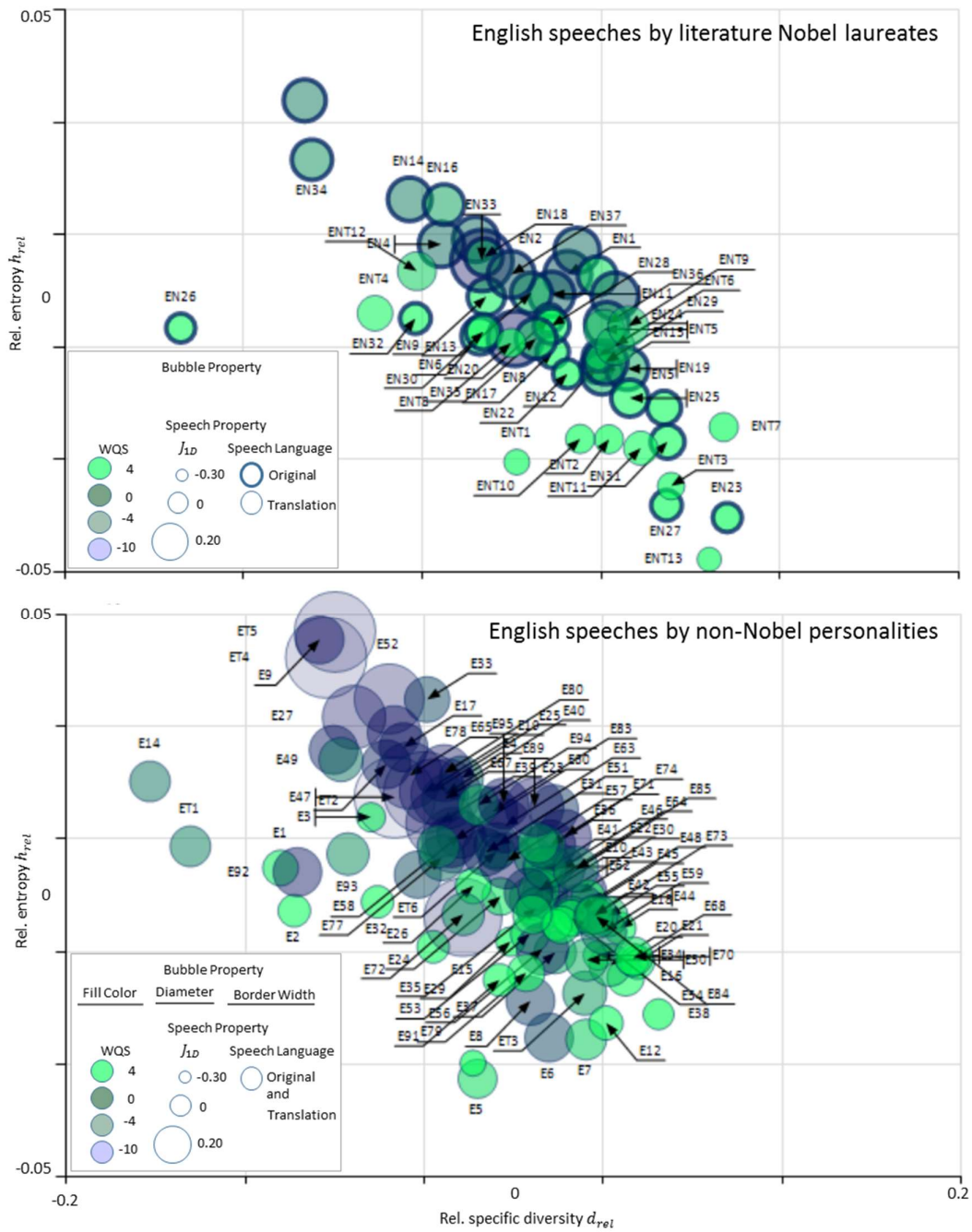


Figure C.1: Writing style for English speeches

C. Literature Nobel laureates and non-laureates text properties

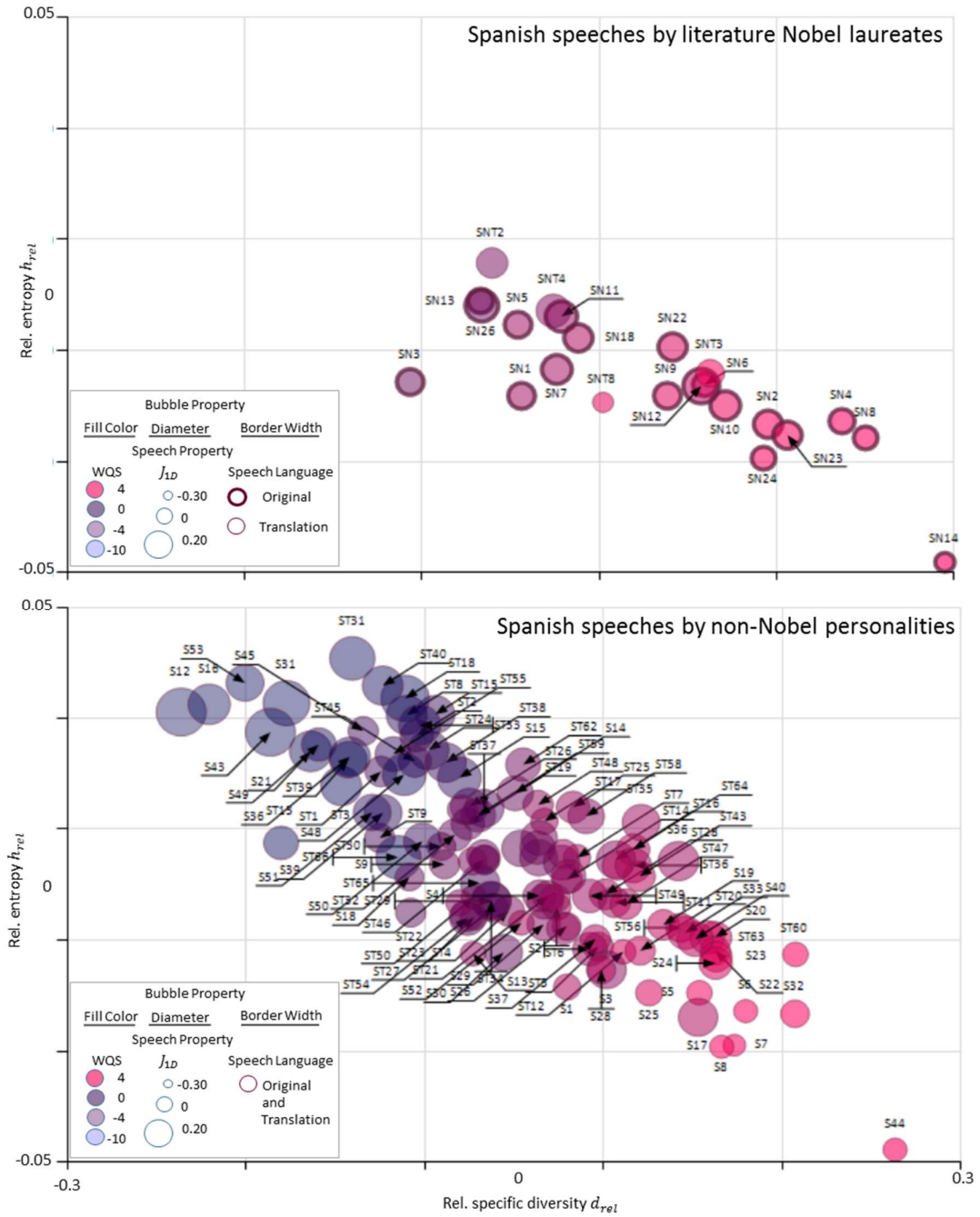


Figure C.2: Writing style for Spanish speeches

Appendix D

The Fundamental Scale Algorithm

The following are a series of pseudo-codes of routines to determine the Fundamental Scale of any sequence of characters.

BlueLetterPhrases: refers to computer instructions: Routine names, control loops and conditional statements.

BlackItalicLetterPhrases: refers to variables. 1D Arrays are followed by [], and 2D arrays are followed by [.,].

GreyLetterPhrases: Comments and NAME OF PROCESS STAGES.

FundamentalScale(*TheText*, *MaximumSymbolSize*, *Symbol*[], *SymbolFrequency*[], *SymbolPosition*[.,])

Scans *TheText* looking for symbols formed by adjacent characters. Returns the array of different symbols *Symbol*[], their frequency of appearance *SymbolFrequency*[] and their position of appearance *SymbolPosition*[.,] within *TheText*.

SymbolSize = 1

UncertLowerLimit = 1, *UncertUpperLimit* = 0

while *SymbolSize* ≤ *MaximumSymbolSize*

" BASE LANGUAGE CONSTRUCTION (When *SymbolSize* = 1)

ProcessTextForASymbolSize(*TheText*, *SymbolSize*, *Symbol*(*i*), *SymbolFrequency*[], *SymbolPosition*[.,])

" CONSTRUCTION OF LANGUAGES WITH LONGER SYMBOLS (When *SymbolSize* > 1)

BirthAndSurvival(*Symbol*[], *SymbolFrequency*[], *SymbolPosition*[], *PropectiveSymbol*[],
PropectSymbolFreq[], *PropectSymbolPosition*[], *N*)

SymbolSize = *SymbolSize* + 1

end while

D. The Fundamental Scale Algorithm

ProcessTextForASymbolSize(*TheText*, *SymbolSize*, *Symbol[]*, *SymbolFrequency[]*, *SymbolPosition[,]*)

Scans *TheText* looking for a characters sequences of *SymbolSize* characters, varying the position of the cursor at the beginning of the reading process. Returns the array of different symbols with size = *SymbolSize*, their frequency of appearance *SymbolFrequency[]*, and the *SymbolPosition[,]* at any scan

Phase = 0

while *Phase* ≤ *SymbolSize*

ScanTextStartingAtAPhase(*TheText*, *SymbolSize*, *Phase*, *Symbol[]*, *SymbolFrequency[]*, *SymbolPosition[,]*)

Phase = *Phase* + 1

end while

for each *i*

" for each *ProspectiveSymbol[]*

N = *N* + *SymbolFrequency[]*

end for

ConsolidateSymbolsFromDifferentPhases(*ProspectiveSymbol[]*, *ProspectSymbolFreq[]*, *ProspectSymbolPosition[]*)

PROSPECTIVE SYMBOL DETECTION

ScanTextStartingAtAPhase(*TheText*, *SymbolSize*, *Phase*, *Symbol[]*, *SymbolFrequency[]*, *SymbolPosition[,]*)

Scans *TheText* looking for a characters sequences of *SymbolSize* characters, starting the reading at the character position *Phase*. Returns the array of different *Symbol[]* with size = *SymbolSize*, their frequency of appearance at this scan *Phase*, *SymbolFrequency[]*, and the array of *SymbolPosition[,]*.

CursorPosition = *Phase* " *CursorPosition* = the position of the cursor in the process of reading a text

StillSomeCharsToRead = true

if *CursorPosition* > *TextLength* - *Phase* **then**

StillSomeCharsToRead = false

i = 0

end if

while *StillSomeCharsToRead*

SymbolJustRead = **TheSequenceOfSymbolSize CharsReadAt** *CursorPosition*

ThisIsANewSymbol = true

if *SymbolJustRead* **IsNotAnElementOfArray** *Symbol[]* **then** *ThisIsANewSymbol* = false

if *ThisIsANewSymbol* **then**

i = *i* + 1

Symbol(*i*) = *SymbolJustRead*

SymbolFrequency[*i*] = 1

else

IndexOfExistingSymbol = **IdentifyIndexOfSymbolJustRead**

SymbolFrequency[*IndexOfExistingSymbol*] = *SymbolFrequency*[*IndexOfExistingSymbol*] + 1

end if

CursorPosition = *CursorPosition* + *Phase*

MoveCursorToPosition *CursorPosition*

if *CursorPosition* > *TextLength* - *Phase* **then** *StillSomeCharsToRead* = false

end while

PROSPECTIVE SYMBOL OVERLAP REDUCTION

ConsolidateSymbolsFromDifferentPhases(*ProspectiveSymbol*[], *ProspectSymbolFreq*[], *ProspectSymbolPosition*[,])

Filters the instances of the *ProspectiveSymbols*[] by deleting the instances being partially overlapped by other, with higher priority *ProspectiveSymbols*[] .

```

OrderArrays ProspectiveSymbol[], ProspectSymbolFreq[], ProspectSymbolPosition[,] by the value of
    ProspectSymbolFreq[], " Higher Frequency is more priority.
i = 0 " for each ProspectiveSymbol[i]
for each i " for each ProspectiveSymbol[i]
    j = 0 " for each ProspectSymbolPosition[i,j]
    for each j " for each ProspectSymbolPosition[i,j]
        Locate the ProspectiveSymbol[k] and the instances m affected by (conflicting with) the insertion
            of the ProspectiveSymbol[i] " There may be more than one ProspectiveSymbol[] affected.
        for each k " for each ProspectiveSymbol[k] affected by insertion of ProspectiveSymbol[i] .
            for each m " for each instance m of the ProspectiveSymbol[k] affected by some insertion.
                Delete the instance m of ProspectiveSymbol[k] located at ProspectSymbolPosition[k,m]
                Update arrays ProspectSymbolFreq[] and ProspectSymbolPosition[,]
            end for
        end for
    end for
end for
for each i " for each ProspectiveSymbol[i]
    if ProspectSymbolFreq[] < 2
        Delete elements i of arrays ProspectiveSymbol[i], ProspectSymbolFreq[i] and ProspectSymbolPosition[i, ]
        and ProspectSymbolPosition[i, ]
    end if
end for

```

EntropyOfASymbolSet(*SymbolFrequency*[], *Entropy*)

Computes the entropy of the symbols present in set. Each symbol is present in the set with the quantity indicated in the array *SymbolFrequency*[] .

```

Entropy = 0
N = UpperBound of array SymbolFrequency[] (+1 depending on the coding language)
For i = 1 to N
    Entropy = Entropy - SymbolFrequency[i] / N · log (SymbolFrequency[i] / N )
endfor
Entropy = Entropy / log (N)

```

D. The Fundamental Scale Algorithm

BIRTH AND SURVIVAL PROCESSES

BirthAndSurvival(*Symbol*[], *SymbolFrequency*[], *SymbolPosition*[], *ProspectiveSymbol*[], *PropectSymbolFreq*[], *PropectSymbolPosition*[], *N*)

Inserts *ProspectiveSymbol*[*i*] into the arrays *Symbol*[] if favorable condirions for an entropy reduction are observed. Every time a *ProspectiveSymbol*[*i*] is inserted into the *Symbol*[] array, an entropy test is performed. If no entropy decrease is observed, the lastly inserted symbol is deleted and arrays are reverted to their condition prior to the insertion. Returns the updated arrays *Symbol*[], *SymbolFrequency*[], *SymbolPosition*[] .

```

i = 0
Entropy = 0
EntropyOfASymbolSet(SymbolFrequency[], Entropy )
D = UpperBound of array Symbol[] (+1 depending on the coding language)
UncertaintyPerSymbol = Entropy / D
Lambda = 0.01
for each i                                     " for each ProspectiveSymbol[]
    P[i] = SymbolFrequency[] / N
    PropectSymbolUncertainty[i] = - P[i] * log(P[i]) / log(D)
    if UncertaintyPerSymbol - Lambda < PropectSymbolUncertainty[i] < UncertaintyPerSymbol + Lambda
        BIRTH PROCESS
        j = 0
        for each PropectSymbolPosition[i,j]       " for each PropectSymbolPosition[i,j]
            Copy established arrays
            EstablishedSymbol[] = Symbol[]
            EstablishedSymbolFrequency[] = SymbolFrequency[]
            EstablishedSymbolPosition[,] = SymbolPosition[,]

            CONSERVATION OF SYMBOLIC QUANTITY
            Locate the Symbol[k] and the instaces m affected by (conflicting with) the insertion
                of the Symbol[i]                               " There may be more than one Symbol[] affected
            for each k                                     " for each Symbol[k] affected by insertion of ProspectiveSymbol[i]
                for each m                               " for each instance m of Symbol[k] affected by some insertion.
                    Delete the instace m of Symbol[k] located at ProspectSymbolPosition[k,m]
                    Update arrays SymbolFrequency[] and SymbolPosition[,]
                end for
            end for
            Insert ProspectiveSymbol[i] into array Symbol[]
            Update arrays SymbolFrequency[] and SymbolPosition[,]

            SURVIVAL PROCESS
            EntropyOfASymbolSet(SymbolFrequency[], Entropy )
            if Entropy < EstablishedEntropy then           Entropy decreased
                EstablishedEntropy = Entropy
            else                                           Entropy increased
                Reject ProspectiveSymbol just inserted and Revert to Previous arrays
                Symbol[] = EstablishedSymbol[]
                SymbolFrequency[] = EstablishedSymbolFrequency[]
                SymbolPosition[,] = EstablishedSymbolPosition[,]
            end if
        end for
        EntropyOfASymbolSet(SymbolFrequency[], Entropy )
        EstablishedEntropy = Entropy
    else                                                 " PropectSymbolUncertainty[i] out of band
        " ProspectSymbol has no oportunity to survive
    end if
end for

```


Appendix E

Symbols of two descriptions at the Fundamental Scale

- E.1 Bertrand Russell's speech given at the 1950 Nobel Award Ceremony:
Word-scale profile [Complete List](#), [Speech text](#).
Total number of symbols [words]: 5716. Diversity: 1868
- E.2 Bertrand Russell's speech given at the 1950 Nobel Award Ceremony:
Fundamental-scale profile: [Complete profile](#), [Speech text](#).
Total number of symbols [Fundamental Symbols]: 25362. Diversity: 1247
- E.3 Beethoven 9th Symphony, 4th movement:
Fundamental -scale profile: [Complete profile](#), [Complete text](#), [Listen MIDI Version](#).
Total number of symbols [Fundamental Symbols]: 84645. Diversity: 2824

E. Symbols of two descriptions at the Fundamental Scale

Table E.1: Word-scale profile of Bertrand Russell's speech at the 1950 Nobel Award Ceremony. [Complete List](#) . [Speech text](#). Total number of symbols (words): 5716. Diversity: 1868.

Rank	Symbol	Occurrences	Length	Rank	Symbol	Occurrences	Length
1	,	412	1	51	or	20	2
2	the	342	3	52	some	20	4
3	.	256	1	53	no	19	2
4	of	218	2	54	so	18	2
5	to	172	2	55	was	18	3
6	is	140	2	56	our	18	3
7	and	132	3	57	human	18	5
8	a	106	1	58	can	17	3
9	in	105	2	59	these	17	5
10	that	84	4	60	very	16	4
11	are	80	3	61	may	16	3
12	it	66	2	62	many	16	4
13	be	64	2	63	;	15	1
14	they	53	4	64	than	15	4
15	not	52	3	65	such	15	4
16	as	51	2	66	fear	15	4
17	which	46	5	67	motives	14	7
18	if	44	2	68	war	14	3
19	I	43	1	69	life	13	4
20	have	42	4	70	people	13	6
21	we	40	2	71	however	13	7
22	by	38	2	72	because	12	7
23	you	37	3	73	«	12	1
24	he	37	2	74	his	12	3
25	but	37	3	75	excitement	12	10
26	for	35	3	76	hate	12	4
27	will	34	4	77	most	12	4
28	their	33	5	78	your	12	4
29	'	32	1	79	great	12	5
30	with	32	4	80	an	12	2
31	from	30	4	81	think	11	5
32	power	30	5	82	become	11	6
33	this	29	4	83	been	11	4
34	when	28	4	84	motive	11	6
35	would	27	5	85	herd	11	4
36	more	27	4	86	much	11	4
37	one	27	3	87	out	10	3
38	there	27	5	88	should	10	6
39	who	26	3	89	could	10	5
40	has	26	3	90	those	10	5
41	them	25	4	91	politics	10	8
42	men	25	3	92	vanity	10	6
43	do	25	2	93	political	9	9
44	at	25	2	94	were	9	4
45	all	25	3	95	upon	9	4
46	what	25	4	96	desires	9	7
47	on	24	2	97	wish	8	4
48	other	24	5	98	?	8	1
49	love	24	4	99	man	8	3
50	had	22	3	100	desire	8	6

E. Symbols of two descriptions at the Fundamental Scale

Rank	Symbol	Occurrences	Length	Rank	Symbol	Occurrences	Length
201	boredom	4	7	301	preference	3	10
202	time	4	4	302	various	3	7
203	better	4	6	303	type	3	4
204	while	4	5	304	obvious	3	7
205	gambling	4	8	305	sometimes	3	9
206	serious	4	7	306	sank	3	4
207	long	4	4	307	away	3	4
208	found	4	5	308	cause	3	5
209	hand	4	4	309	end	3	3
210	old	4	3	310	killed	3	6
211	taken	4	5	311	innocent	3	8
212	members	4	7	312	believe	3	7
213	destructive	4	11	313	themselves	3	10
214	above	4	5	314	desired	3	7
215	within	4	6	315	step	3	4
216	enemies	4	7	316	wars	3	4
217	French	4	6	317	kind	3	4
218	way	4	3	318	where	3	5
219	communists	4	10	319	passions	3	8
220	effective	4	9	320	instinctive	3	11
221	sympathy	4	8	321	brothers	3	8
222	self	4	4	322	feeling	3	7
223	Nation	4	6	323	Russians	3	8
224	selfishness	4	11	324	enemy	3	5
225	moralists	4	9	325	ways	3	4
226	general	4	7	326	conflict	3	8
227	although	4	8	327	altruistic	3	10
228	politicians	4	11	328	against	3	7
229	since	4	5	329	operation	3	9
230	ideologies	4	10	330	fall	3	4
231	government	4	10	331	hunger	3	6
232	account	3	7	332	history	3	7
233	population	3	10	333	rivalry	3	7
234	South	3	5	334	current	2	7
235	North	3	5	335	theory	2	6
236	books	3	5	336	psychology	2	10
237	sort	3	4	337	facts	2	5
238	person	3	6	338	constitutional	2	14
239	between	3	7	339	began	2	5
240	cannot	3	6	340	right	2	5
241	politician	3	10	341	average	2	7
242	frequently	3	10	342	income	2	6
243	causes	3	6	343	want	2	4
244	action	3	6	344	tell	2	4
245	another	3	7	345	heard	2	5
246	far	3	3	346	questions	2	9
247	too	3	3	347	remote	2	6
248	wholly	3	6	348	scientific	2	10
249	duty	3	4	349	constantly	2	10
250	sense	3	5	350	thinking	2	8

E. Symbols of two descriptions at the Fundamental Scale

Rank	Symbol	Occurrences	Length	Rank	Symbol	Occurrences	Length
501	designed	2	8	551	century	2	7
502	deceive	2	7	552	feel	2	4
503	condemn	2	7	553	hatred	2	6
504	form	2	4	554	strange	2	7
505	appropriate	2	11	555	methods	2	7
506	feared	2	6	556	best	2	4
507	fellow	2	6	557	thoroughly	2	10
508	leads	2	5	558	produced	2	8
509	exciting	2	8	559	ill	2	3
510	provide	2	7	560	treated	2	7
511	rabbits	2	7	561	Western	2	7
512	impulse	2	7	562	countries	2	9
513	big	2	3	563	sum	2	3
514	contain	2	7	564	expensive	2	9
515	small	2	5	565	Germans	2	7
516	enmity	2	6	566	victors	2	7
517	actual	2	6	567	secured	2	7
518	member	2	6	568	advantages	2	10
519	mechanism	2	9	569	B	2	1
520	nations	2	7	570	large	2	5
521	regards	2	7	571	disguised	2	9
522	international	2	13	572	conclusion	2	10
523	discovered	2	10	573	intelligence	2	12
524	degree	2	6	574	ladies	2	6
525	says	2	4	575	economic	2	8
526	am	2	2	576	nor	2	3
527	line	2	4	577	mankind	2	7
528	Rhine	2	5	578	court	2	5
529	essential	2	9	579	civilized	2	9
530	danger	2	6	580	dance	2	5
531	TRUE	2	4	581	none	2	4
532	might	2	5	582	killing	2	7
533	regard	2	6	583	Royal	1	5
534	Mother	2	6	584	Highness	1	8
535	Nature	2	6	585	Gentlemen	1	9
536	cooperation	2	11	586	chosen	1	6
537	easily	2	6	587	subject	1	7
538	schools	2	7	588	lecture	1	7
539	turning	2	7	589	tonight	1	7
540	cruelty	2	7	590	discussions	1	11
541	everyday	2	8	591	insufficient	1	12
542	atom	2	4	592	statistics	1	10
543	bomb	2	4	593	organization	1	12
544	wicked	2	6	594	set	1	3
545	rival	2	5	595	forth	1	5
546	hating	2	6	596	minutely	1	8
547	burglars	2	8	597	difficulty	1	10
548	disapprove	2	10	598	finding	1	7
549	attitude	2	8	599	able	1	4
550	irreligious	2	11	600	ascertain	1	9

E. Symbols of two descriptions at the Fundamental Scale

Table E.2: Fundamental-scale profile of Bertrand Russell's speech given at the 1950 Nobel Award Ceremony.
[Complete profile](#). [Speech text](#). Total number of fundamental symbols: 25362. Diversity: 1247

Rank	Symbol	Probability	Occurrences	Length	Rank	Symbol	Probability	Occurrences	Length
1	∅	0.192475	5020	1	51	pr	0.000513	13	2
2	e	0.099270	2589	1	52	B	0.000512	13	1
3	t	0.077074	2010	1	53	fo	0.000512	13	2
4	n	0.050588	1319	1	54	.∅Th	0.000485	13	4
5	a	0.050490	1317	1	55	ot	0.000479	12	2
6	o	0.049389	1288	1	56	st	0.000477	12	2
7	i	0.049202	1283	1	57	ly	0.000475	12	2
8	s	0.047820	1247	1	58	'	0.000475	12	1
9	h	0.047428	1237	1	59	ur	0.000474	12	2
10	r	0.037977	990	1	60	ll	0.000473	12	2
11	d	0.017566	458	1	61	if	0.000472	12	2
12	l	0.017226	449	1	62	co	0.000471	12	2
13	f	0.015580	406	1	63	as	0.000471	12	2
14	c	0.015178	396	1	64	S	0.000471	12	1
15	w	0.013417	350	1	65	E	0.000469	12	1
16	m	0.010668	278	1	66	to	0.000439	11	2
17	y	0.009954	260	1	67	politic	0.000439	11	7
18	,	0.009707	253	1	68	F	0.000436	11	1
19	u	0.008401	219	1	69	ra	0.000433	11	2
20	p	0.007610	198	1	70	ca	0.000433	11	2
21	g	0.006424	168	1	71	∅f	0.000432	11	2
22	v	0.006262	163	1	72	ce	0.000430	11	2
23	b	0.004462	116	1	73	K	0.000428	11	1
24	.	0.004066	106	1	74	will	0.000425	11	4
25	l	0.001695	44	1	75	∅b	0.000399	10	2
26	k	0.001421	37	1	76	um	0.000398	10	2
27	nd	0.001258	33	2	77	em	0.000397	10	2
28	be	0.001149	30	2	78	M	0.000395	10	1
29	ma	0.000831	22	2	79	av	0.000395	10	2
30	of	0.000829	22	2	80	ev	0.000395	10	2
31	A	0.000827	22	1	81	su	0.000394	10	2
32	x	0.000826	22	1	82	ol	0.000394	10	2
33	T	0.000787	21	1	83	ver	0.000393	10	3
34	un	0.000747	19	2	84	se	0.000393	10	2
35	us	0.000713	19	2	85	whic	0.000390	10	4
36	.∅l	0.000668	17	3	86	woul	0.000363	9	4
37	by	0.000633	17	2	87	pp	0.000357	9	2
38	s,	0.000629	16	2	88	de	0.000356	9	2
39	mo	0.000627	16	2	89	im	0.000356	9	2
40	me	0.000626	16	2	90	ua	0.000355	9	2
41	ed	0.000599	16	2	91	ac	0.000355	9	2
42	ad	0.000592	15	2	92	op	0.000355	9	2
43	lo	0.000592	15	2	93	wi	0.000354	9	2
44	ve	0.000588	15	2	94	from	0.000354	9	4
45	om	0.000587	15	2	95	com	0.000354	9	3
46	∅d	0.000586	15	2	96	∅p	0.000353	9	2
47	W	0.000553	14	1	97	no	0.000353	9	2
48	ri	0.000551	14	2	98	hi	0.000353	9	2
49	ag	0.000514	13	2	99	so	0.000353	9	2
50	;	0.000513	13	1	100	ho	0.000352	9	2

E. Symbols of two descriptions at the Fundamental Scale

Rank	Symbol	Probability	Occurrences	Length	Rank	Symbol	Probability	Occurrences	Length
776	øgr	7.87E-05	2	3	1101	oli	3.95E-05	1	3
777	lec	7.87E-05	2	3	1102	requ	3.95E-05	1	4
778	lki	7.87E-05	2	3	1103	rable	3.95E-05	1	5
779	ødea	7.87E-05	2	4	1104	vanity	3.95E-05	1	6
780	?øAn	7.87E-05	2	4	1105	factio	3.95E-05	1	6
781	rimi	7.87E-05	2	4	1106	import	3.95E-05	1	6
782	day,	7.87E-05	2	4	1107	ps,	3.93E-05	1	3
783	joym	7.87E-05	2	4	1108	sn	3.93E-05	1	2
784	stsø	7.87E-05	2	4	1109	s,øhowever	3.93E-05	1	10
785	firs	7.87E-05	2	4	1110	oweve	3.93E-05	1	5
786	forø	7.87E-05	2	4	1111	dom	3.93E-05	1	3
787	notø	7.87E-05	2	4	1112	ldøb	3.93E-05	1	4
788	mora	7.87E-05	2	4	1113	øbet	3.93E-05	1	4
789	eøol	7.87E-05	2	4	1114	oøp	3.93E-05	1	3
790	hand	7.87E-05	2	4	1115	inø	3.93E-05	1	3
791	zedøm	7.87E-05	2	5	1116	døg	3.93E-05	1	3
792	løref	7.87E-05	2	5	1117	men,	3.93E-05	1	4
793	uchøa	7.87E-05	2	5	1118	,øga	3.93E-05	1	4
794	produc	7.87E-05	2	6	1119	døco	3.93E-05	1	4
795	inøcon	7.87E-05	2	6	1120	old	3.93E-05	1	3
796	lømake ¹ up	7.87E-05	2	9	1121	vil	3.93E-05	1	3
797	heødevilø	7.87E-05	2	9	1122	mak	3.93E-05	1	3
798	shouldøbe	7.87E-05	2	9	1123	joy	3.93E-05	1	3
799	y ¹ fiveømil	7.87E-05	2	10	1124	cia	3.93E-05	1	3
800	seriousness	7.87E-05	2	11	1125	day	3.93E-05	1	3
801	fromøboredom	7.87E-05	2	12	1126	nno	3.93E-05	1	3
802	not,øperhaps,	7.87E-05	2	13	1127	ate	3.93E-05	1	3
803	uch	7.86E-05	2	3	1128	ked	3.93E-05	1	3
804	ry	7.86E-05	2	2	1129	ist	3.93E-05	1	3
805	arø	7.84E-05	2	3	1130	aød	3.93E-05	1	3
806	'	7.84E-05	2	1	1131	e.ø	3.93E-05	1	3
807	llø	7.83E-05	2	3	1132	af	3.93E-05	1	2
808	rad	7.83E-05	2	3	1133	ç	3.93E-05	1	1
809	dr	7.83E-05	2	2	1134	i	3.93E-05	1	1
810	D	7.83E-05	2	1	1135	sc	3.93E-05	1	2
811	øun	7.79E-05	2	3	1136	be,	3.93E-05	1	3
812	wil	7.79E-05	2	3	1137	øif	3.93E-05	1	3
813	about	7.79E-05	2	5	1138	nøf	3.93E-05	1	3
814	y,øa	7.79E-05	2	4	1139	oot	3.93E-05	1	3
815	Gr	7.79E-05	2	2	1140	res	3.93E-05	1	3
816	oe	7.79E-05	2	2	1141	,øm	3.93E-05	1	3
817	lov	7.79E-05	2	3	1142	løw	3.93E-05	1	3
818	øMa	7.79E-05	2	3	1143	phy	3.93E-05	1	3
819	iew	7.79E-05	2	3	1144	øpo	3.93E-05	1	3
820	agr	7.79E-05	2	3	1145	yøp	3.93E-05	1	3
821	ivi	7.79E-05	2	3	1146	m,ø	3.93E-05	1	3
822	,ø«	7.79E-05	2	3	1147	tin	3.93E-05	1	3
823	.øO	7.79E-05	2	3	1148	ond	3.93E-05	1	3
824	esu	7.79E-05	2	3	1149	eci	3.93E-05	1	3
825	oøi	7.79E-05	2	3	1150	cook	3.93E-05	1	4

E. Symbols of two descriptions at the Fundamental Scale

Table E.3: Fundamental-scale profile for a MIDI version of Beethoven 9th Symphony, 4th movement. [Complete Profile](#). [Complete text](#). [Listen MIDI Version](#). Total number of fundamental symbols: 84645. Diversity: 2824.

Rank	Symbol	Probability	Occurrences	Length	Rank	Symbol	Probability	Occurrences	Length
1	²	0.38332	32446	1	51	1	0.00287	243	1
2	x	0.03896	3298	1	52]	0.00278	235	1
3	Φ	0.03870	3276	1	53	5	0.00273	231	1
4	n	0.03320	2810	1	54	+	0.00267	226	1
5	@	0.01921	1626	1	55	[0.00265	224	1
6	³	0.01916	1622	1	56	A	0.00259	219	1
7	d	0.01769	1497	1	57	0	0.00261	221	1
8	9	0.01454	1231	1	58	!	0.00205	173	1
9	2	0.01359	1151	1	59	.	0.00189	160	1
10	?	0.01358	1149	1	60	8	0.00167	142	1
11	-	0.01321	1118	1	61	W	0.00160	135	1
12	é	0.01304	1104	1	62	P	0.00158	134	1
13	J	0.01221	1033	1	63		0.00150	127	1
14	E	0.01212	1026	1	64	D	0.00148	125	1
15	B	0.01108	938	1	65	L	0.00123	104	1
16	L	0.00997	844	1	66	¿	0.00117	99	1
17	Q	0.00979	828	1	67	Y	0.00115	97	1
18	N	0.00944	799	1	68	3	0.00107	91	1
19	/	0.00899	761	1	69	%	0.00100	84	1
20	6	0.00877	742	1	70	ÿX [↓] ↓ _↑ □	0.00087	73	7
21	;	0.00826	699	1	71	#	0.00081	69	1
22	C	0.00815	690	1	72	°@	0.00078	66	2
23	=	0.00801	678	1	73	w	0.00073	62	1
24	O	0.00773	654	1	74	½	0.00070	59	1
25	V	0.00748	633	1	75	\$	0.00064	54	1
26	K	0.00746	631	1	76	¶	0.00064	54	1
27	ã	0.00671	568	1	77	→	0.00063	53	1
28	'	0.00658	557	1	78	-	0.00057	48	1
29	4	0.00635	537	1	79		0.00057	48	1
30	¾	0.00600	508	1	80	ΦÿX [↓] ↓ _↑ □	0.00056	47	8
31	Z	0.00594	503	1	81	ú	0.00054	46	1
32	7	0.00590	499	1	82	ΦÿX [↓]	0.00054	46	4
33	G	0.00582	492	1	83	,	0.00054	46	1
34	I	0.00576	488	1	84	↓ _↑ □	0.00053	44	4
35	°	0.00517	438	1	85	¤	0.00050	43	1
36	&	0.00454	385	1	86	↓	0.00050	43	1
37	F	0.00426	360	1	87	4¾'4	0.00049	41	4
38	R	0.00410	347	1	88	f	0.00048	41	1
39	X	0.00409	346	1	89		0.00047	40	0
40	∏	0.00401	340	1	90	_	0.00047	40	1
41	*	0.00369	312	1	91	"	0.00047	40	1
42		0.00345	292	1	92	□	0.00046	39	1
43	S	0.00323	274	1	93	`	0.00044	37	1
44	T	0.00322	273	1	94	•	0.00044	37	1
45	:	0.00320	271	1	95	@³	0.00043	37	2
46	H	0.00318	269	1	96	°@³	0.00043	36	3
47	f	0.00302	255	1	97	Á	0.00042	36	1
48	M	0.00299	253	1	98	í	0.00042	36	1
49	U	0.00285	241	1	99	-¾'-	0.00038	32	4
50		0.00285	241	1	100	v	0.00038	32	1

E. Symbols of two descriptions at the Fundamental Scale

Rank	Symbol	Probability	Occurrences	Length	Rank	Symbol	Probability	Occurrences	Length
401	S?²G	0.00006	5	4	651	¿²ΦR¾	0.00004	3	5
402	²9Zn	0.00006	5	4	652	H?nM²	0.00004	3	5
403	d[]	0.00006	5	2	653	²Y¾²M¾nV	0.00004	3	8
404	²é	0.00006	5	2	654	"²,r.¾²"¾,	0.00004	3	10
405	†	0.00006	5	1	655	[]V	0.00004	3	2
406	²,	0.00006	5	2	656	[][0.00004	3	2
407	Φ8x	0.00006	5	3	657	[]3	0.00004	3	2
408	ãΦ	0.00006	5	2	658	[]N	0.00004	3	2
409	ΦB	0.00005	5	2	659	[]1	0.00004	3	2
410	Φ'	0.00005	5	2	660	[]7	0.00004	3	2
411	ã	0.00005	5	2	661	;x²7x	0.00004	3	5
412	X³	0.00005	5	2	662	Gxã[]S	0.00004	3	5
413	&³	0.00005	5	2	663	[];²²7	0.00004	3	5
414	#³	0.00005	5	2	664	%²Φ2x	0.00004	3	5
415	?²-?n	0.00005	4	5	665	&²ΦUx	0.00004	3	5
416	;	0.00005	4	2	666	¾²ΦEx	0.00004	3	5
417	1³2=	0.00005	4	4	667	²&xã[]	0.00004	3	5
418	L³²@	0.00005	4	4	668	²E²ΦS	0.00004	3	5
419	8³2D	0.00005	4	4	669	¾xã[]Q	0.00004	3	5
420	6³2B	0.00005	4	4	670	Φ]x²é	0.00004	3	5
421	édnN	0.00005	4	4	671	ãfx²²	0.00004	3	5
422	=³nU²	0.00005	4	5	672	Φ]x²Q	0.00004	3	5
423	;³nS²	0.00005	4	5	673	²Φ4x	0.00004	3	4
424	L²²C?	0.00005	4	5	674	xãfZ²	0.00004	3	5
425	dN²ΦLdL²ΦJdJ	0.00005	4	15	675	+xã[]Q	0.00004	3	5
426	E?	0.00005	4	2	676	ΦZx²]	0.00004	3	5
427	&²Φ¾	0.00005	4	4	677	2²Φ@x	0.00004	3	5
428	NxnN	0.00005	4	4	678	é/	0.00004	3	2
429	2-²Φ9-29²Φ--R	0.00005	4	14	679	62	0.00004	3	2
430	Sx	0.00005	4	2	680	72	0.00004	3	2
431	éx	0.00005	4	2	681	ãfl	0.00004	3	3
432	2Q	0.00005	4	2	682	7d²+d	0.00004	3	5
433	3²	0.00005	4	2	683	;x²@x	0.00004	3	5
434	//ãf	0.00005	4	4	684	O2²L2	0.00004	3	5
435	/²-/ãf9	0.00005	4	7	685	Gx²;x²V	0.00004	3	7
436	Q²ΦQxnQ	0.00005	4	8	686	;²é/²9/	0.00004	3	8
437	Q²²E/ãfQ	0.00005	4	8	687	Qx	0.00004	3	2
438]x²Bx²9xn]	0.00005	4	10	688	[]L	0.00004	3	2
439	7xn7²Φ-xnE	0.00005	4	10	689	[]J	0.00004	3	2
440	xã	0.00005	4	2	690	Φ-	0.00004	3	2
441	Φ]	0.00005	4	2	691	Bd	0.00004	3	2
442	@ã	0.00005	4	2	692	Y²²M²	0.00004	3	5
443	²ã .¾ã	0.00005	4	6	693	²;xnK	0.00004	3	5
444	.¾ãf	0.00005	4	4	694	²2xã[]	0.00004	3	5
445	K¾nT	0.00005	4	4	695	²/xnW	0.00004	3	5
446	Kãf5	0.00005	4	4	696	[]°@³Z	0.00004	3	5
447	VZ²JZ	0.00005	4	5	697	ΦZxnZ	0.00004	3	5
448	é?².²ãf	0.00005	4	7	698	/xãf]²	0.00004	3	6
449	?²:²ãf:	0.00005	4	7	699	:xn	0.00004	3	3
450	.¾²"¾,úJ	0.00005	4	8	700	C²²	0.00004	3	3

E. Symbols of two descriptions at the Fundamental Scale

Rank	Symbol	Probability	Occurrences	Length	Rank	Symbol	Probability	Occurrences	Length
2101	2d[]Z	0.00002	2	4	2771	4?4	0.00001	1	4
2102	[]V²[]	0.00002	2	4	2772	²4?	0.00001	1	4
2103	Xd²P	0.00002	2	4	2773	9d²-	0.00001	1	4
2104	!²²[]	0.00002	2	4	2774	-²Φ/	0.00001	1	4
2105	2dn&	0.00002	2	4	2775	4²Φ	0.00001	1	4
2106	4V²[]	0.00002	2	4	2776	²[]	0.00001	1	2
2107	E²[]	0.00002	2	4	2777	K²7Kn	0.00001	1	5
2108	²OKn	0.00002	2	4	2778	K²NK	0.00001	1	4
2109	Bd²	0.00002	2	4	2779	éK²7	0.00001	1	4
2110	°@³[]	0.00002	2	4	2780	xE	0.00001	1	2
2111	²ΦGK	0.00002	2	4	2781	QK²	0.00001	1	3
2112	:d²N	0.00002	2	4	2782	Xd	0.00001	1	2
2113	!dnU	0.00002	2	4	2783	OK	0.00001	1	2
2114	Ld²*	0.00002	2	4	2784	²	0.00001	1	2
2115	¼²²4	0.00002	2	4	2785	dx	0.00001	1	2
2116	EdSE	0.00002	2	4	2786	dI	0.00001	1	2
2117	Z² N²oN	0.00002	2	8	2787	Un	0.00001	1	2
2118	K²NK²EKã	0.00002	2	8	2788	ã*	0.00001	1	2
2119	²LKãfc²²	0.00002	2	8	2789	Jd	0.00001	1	2
2120	²4dxX²²P	0.00002	2	8	2790	¼	0.00001	1	1
2121	CK²LK²GK	0.00002	2	8	2791	Q	0.00001	1	2
2122	²;K²7Kn@	0.00002	2	8	2792	²/	0.00001	1	2
2123	[]X²ÁO²[] [0.00002	2	8	2793	dn[0.00001	1	3
2124	[] [²] O²¹-d	0.00002	2	9	2794	2dã	0.00001	1	3
2125	Φ=Kn=²²7²	0.00002	2	9	2795	J-	0.00001	1	2
2126	d²]dÉ²²Y&	0.00002	2	9	2796	ú@²	0.00001	1	3
2127	ãf=²Φ@Kãf	0.00002	2	9	2797	²¿³	0.00001	1	3
2128]²jZd²Qd²]	0.00002	2	10	2798	F	0.00001	1	2
2129	&d²2dI Q²[] V	0.00002	2	10	2799	[°	0.00001	1	2
2130	-,	0.00002	2	2	2800	[]°@	0.00001	1	3
2131	...F	0.00002	2	2	2801	4²Φ	0.00001	1	3
2132	²6	0.00002	2	2	2802	¿³	0.00001	1	2
2133	2³	0.00002	2	2	2803	ΦT	0.00001	1	2
2134	E-	0.00002	2	2	2804	§-	0.00001	1	2
2135	²A	0.00002	2	2	2805	-f	0.00001	1	2
2136	²¿	0.00002	2	2	2806	i	0.00001	1	1
2137	R	0.00002	2	2	2807	-,ú	0.00001	1	3
2138	²L	0.00002	2	2	2808	ã é	0.00001	1	3
2139	[]°	0.00002	2	2	2809	¿³ã	0.00001	1	3
2140	ã @	0.00002	2	3	2810	²²6	0.00001	1	3
2141	²²'	0.00002	2	3	2811	Φ@³	0.00001	1	3
2142	²j²	0.00002	2	3	2812	O-²§	0.00001	1	4
2143	³²L	0.00002	2	3	2813	...Fé²	0.00001	1	4
2144	,úé	0.00002	2	3	2814	³Φ	0.00001	1	2
2145	VE²	0.00002	2	3	2815	š	0.00001	1	1
2146	-²@	0.00002	2	3	2816	Ó	0.00001	1	1
2147	fV7	0.00002	2	3	2817	~	0.00001	1	1
2148	Φ@-	0.00002	2	3	2818		0.00001	1	1
2149	X°@	0.00002	2	3	2819	³#6	0.00001	1	3
2150	²ã#	0.00002	2	3	2820	²7³	0.00001	1	3

Appendix F

Language properties at different scales

- F.1 English Properties at different scales: [Table](#).
Total number of Texts: 128.
- F.2 Spanish Properties at different scales: [Table](#)
Total number of Texts: 72.
- F.3 Computer Programing Code Properties at different scales: [Table](#)
Total number of Codes: 37.
- F.4 MIDI Music Properties at different scales: [Table](#)
Total number of Pieces: 438

F. Properties of some languages at different scales

Table F.1: [Table](#).

English Properties at Different Scales									
Message Name	<i>L</i> = Length			<i>c</i> = Complexity			<i>[w]</i> = [words]		
	<i>d</i> = Specific Diversity			<i>[F.S.]</i> = [Fundamental Symbols]					
	<i>h</i> = entropy			<i>[chrs]</i> = [characters]			<i>[0-1]</i> = between 0 and 1		
	At Char Scale			At Fundamental Scale			At Word Scale		
	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>
	[chrs]	[0-1]	[0-1]	[F.S.]	[0-1]	[0-1]	[w]	[0-1]	[0-1]
1381.JohnBall.txt	1122	0.0294	0.830	899	0.121	0.715	227	0.515	0.914
1601.Hamlet.txt	645	0.0220	0.792	563	0.124	0.758	150	0.435	0.950
1588.QueenElizabethI.txt	1635	0.0636	0.815	1256	0.111	0.688	359	0.647	0.879
1601.QueenElizabethI.txt	5451	0.0097	0.744	4275	0.078	0.607	1140	0.340	0.865
1851.SojournerTruth.txt	1881	0.0271	0.748	1529	0.097	0.666	443	0.413	0.911
1877.ChiefJoseph.txt	770	0.0532	0.786	603	0.126	0.736	183	0.503	0.926
1901.MarkTwain.txt	2971	0.0162	0.755	2366	0.093	0.633	669	0.386	0.889
1923.BS.Eng.WilliamButlerYeats.txt	1689	0.0243	0.791	1363	0.096	0.691	320	0.522	0.920
1932.MargaretSanger.txt	6123	0.0085	0.737	4722	0.077	0.586	1162	0.343	0.847
1936.KingEdwardVIII.txt	2850	0.0147	0.788	2208	0.098	0.653	596	0.408	0.875
1938.BS.PearlBuck.txt	2458	0.0175	0.779	1911	0.094	0.660	520	0.379	0.893
1940.05.WinstonChurchill.txt	3530	0.0150	0.744	2910	0.077	0.646	703	0.415	0.873
1941.FranklinDRoosevelt.txt	3184	0.0173	0.747	2496	0.091	0.634	574	0.455	0.881
1942.MahatmaGandhi.txt	6106	0.0097	0.725	4724	0.073	0.604	1234	0.347	0.855
1944.DwightEisenhower.txt	1076	0.0418	0.788	885	0.122	0.718	208	0.577	0.925
1944.GeorgePatton.txt	3919	0.0138	0.744	3026	0.074	0.639	890	0.361	0.886
1946.WinstonChurchill.txt	6633	0.0089	0.725	5139	0.080	0.587	1285	0.388	0.850
1949.BS.Eng.WilliamFaulkner.txt	3016	0.0136	0.780	2414	0.088	0.646	622	0.399	0.884
1954.BS.Eng.ErnestHemingway	1808	0.0227	0.777	1452	0.101	0.673	367	0.499	0.919
1961.11.JohnFKennedy.txt	3596	0.0131	0.760	2816	0.091	0.629	680	0.465	0.892
1962.BS.Eng.JohnSteinbeck.txt	4925	0.0091	0.769	3852	0.084	0.616	952	0.404	0.859
1963.06.26.JohnFKennedy.txt	3146	0.0146	0.761	2452	0.082	0.633	665	0.358	0.875
1964.05.LyndonBJohnson.txt	5975	0.0095	0.729	4742	0.072	0.600	1168	0.368	0.848
1964.LadybirdJohnson.txt	4205	0.0150	0.719	3347	0.084	0.621	818	0.432	0.876
1964.MartinLutherKing.txt	6667	0.0081	0.738	5006	0.076	0.592	1266	0.393	0.862
1968.RobertFKennedy.txt	3025	0.0152	0.765	2353	0.080	0.636	629	0.315	0.898
1969.IndiraGhandi.txt	5279	0.0100	0.746	3962	0.078	0.608	1058	0.386	0.867
1969.ShirleyChisholm.txt	5102	0.0114	0.730	3983	0.078	0.602	967	0.396	0.867
1972.JaneFonda.txt	4053	0.0136	0.744	3095	0.091	0.620	799	0.432	0.873
1976.BS.Eng.SaulBellow.txt	1999	0.0255	0.758	1607	0.105	0.672	397	0.501	0.912
1981.RonaldReagan.txt	5877	0.0095	0.732	4552	0.074	0.599	1183	0.383	0.856
1983.BS.Eng.WilliamGolding.txt	1837	0.0256	0.772	1538	0.097	0.684	369	0.545	0.913
1986.BS.Eng.WoleSoyinka.txt	2529	0.0178	0.781	1972	0.099	0.666	482	0.508	0.892
1986.RonaldReagan.txt	3738	0.0158	0.732	3093	0.075	0.635	805	0.385	0.857
1991.BS.Eng.NadineGordimer.txt	2837	0.0176	0.756	2265	0.093	0.651	564	0.500	0.892
1992.BS.Eng.DerekWalcott.txt	611	0.0704	0.798	545	0.127	0.770	104	0.654	0.930
1993.BS.Eng.ToniMorrison.txt	1887	0.0228	0.783	1514	0.103	0.675	368	0.546	0.912
1993.MayaAngelou.txt	3660	0.0145	0.757	3000	0.082	0.644	794	0.392	0.835
1993.SarahBrady.txt	4409	0.0118	0.752	3555	0.069	0.618	969	0.332	0.869
1993.UrvashiVaid.txt	6545	0.0089	0.737	4895	0.076	0.595	1319	0.315	0.841

F. Properties of some languages at different scales

<i>Message Name</i>	At Char Scale			At Fundamental Scale			At Word Scale		
	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>
	[chrs]	[0-1]	[0-1]	[F.S.]	[0-1]	[0-1]	[w]	[0-1]	[0-1]
1994.NelsonMandela.txt	5181	0.0097	0.748	3930	0.043	0.633	1010	0.384	0.848
1995.BS.Eng.SeamusHeaney.tx	1508	0.0312	0.769	1225	0.112	0.689	287	0.561	0.915
1997.BillClinton.txt	6083	0.0087	0.741	4610	0.084	0.592	1303	0.322	0.845
1997.QueenElizabethII.txt	2172	0.0203	0.771	1795	0.088	0.665	449	0.461	0.900
2001.09.11.GeorgeWBush.txt	3522	0.0139	0.760	2777	0.086	0.637	673	0.443	0.882
2001.09.13.GeorgeWBush.txt	2922	0.0198	0.740	2484	0.078	0.654	550	0.456	0.874
2001.BS.Eng.VSNaipaul.txt	1665	0.0306	0.760	1365	0.113	0.679	348	0.500	0.899
2001.HalleBerry.txt	2840	0.0204	0.748	2303	0.089	0.654	649	0.337	0.849
2002.OprahWinfrey.txt	2685	0.0175	0.769	2061	0.099	0.644	609	0.373	0.865
2003.BethChapman.txt	4257	0.0148	0.720	3452	0.074	0.620	882	0.382	0.877
2003.BS.Eng.JMCoetzee.txt	1510	0.0364	0.757	1239	0.098	0.700	331	0.459	0.913
1606.LancelotAndrewes.txt	41451	0.0017	0.691	32985	0.040	0.503	9291	0.166	0.738
1833.ThomasBabington.txt	81977	0.0009	0.688	62980	0.035	0.487	15668	0.169	0.746
1849.LucretiaMott.txt	38756	0.0017	0.707	30664	0.043	0.509	7577	0.227	0.770
1851.ErnestineLRose.txt	39851	0.0016	0.711	32514	0.036	0.514	8301	0.196	0.764
1861.AbrahamLincoln.txt	20952	0.0027	0.722	16550	0.051	0.537	4007	0.254	0.808
1867.ElizabethCadyStanton.txt	29592	0.0022	0.705	23717	0.036	0.541	5862	0.253	0.784
1890.RusselConwell.txt	81989	0.0009	0.686	63660	0.034	0.483	17795	0.128	0.748
1892.FrancesEWHarper.txt	21988	0.0026	0.719	17224	0.051	0.537	4396	0.283	0.805
1906.MaryChurch.txt	8158	0.0072	0.722	6570	0.070	0.577	1558	0.375	0.852
1909.BS.SelmaLagerlof.txt	10046	0.0061	0.715	8424	0.059	0.568	2301	0.272	0.826
1915.AnnaHoward.txt	50806	0.0014	0.683	40013	0.036	0.496	10652	0.134	0.776
1916.CarrieChapman.txt	31123	0.0023	0.696	24697	0.047	0.521	6127	0.252	0.794
1916.HellenKeller.txt	13143	0.0046	0.724	10498	0.081	0.532	2562	0.335	0.829
1918.WoodrowWilson.txt	15039	0.0043	0.702	12279	0.050	0.555	2753	0.279	0.818
1920.CrystalEastman.txt	10557	0.0051	0.733	8326	0.071	0.564	2136	0.314	0.848
1923.JamesMonroe.txt	6485	0.0076	0.743	5151	0.067	0.596	1178	0.354	0.849
1923.NL.Eng.WilliamButlerYea	21120	0.0031	0.704	16724	0.053	0.539	4258	0.265	0.819
1925.MaryReynolds.txt	17911	0.0031	0.719	14404	0.059	0.535	4340	0.198	0.799
1930.NL.Eng.SinclairLewis.txt	29220	0.0023	0.705	24545	0.040	0.535	5708	0.282	0.799
1936.EleanorRoosvelt.txt	9186	0.0063	0.710	7082	0.062	0.573	1968	0.233	0.830
1938.NL.PearlBuck.txt	50855	0.0013	0.698	41104	0.033	0.507	10270	0.178	0.767
1940.06.A.WinstonChurchill.tx	19584	0.0035	0.707	15511	0.052	0.545	3784	0.282	0.822
1940.06.B.WinstonChurchill.tx	25152	0.0025	0.715	20006	0.049	0.529	4909	0.242	0.802
1941.HaroldIckes.txt	12131	0.0046	0.736	9806	0.060	0.568	2449	0.295	0.822
1947.GeorgeCMarshall.txt	8669	0.0058	0.746	6843	0.071	0.576	1608	0.363	0.843
1947.HarryTruman.txt	13420	0.0045	0.720	11008	0.056	0.557	2459	0.292	0.821
1948.BS.Eng.ThomasEliot.txt	7381	0.0078	0.724	5864	0.082	0.582	1467	0.344	0.845
1950.MargaretChase.txt	9313	0.0056	0.749	7284	0.071	0.578	1717	0.327	0.845
1950.NL.Eng.BertrandRussell.t	32621	0.0021	0.705	25362	0.049	0.522	5716	0.327	0.821
1953.DwightEisenhower.txt	14887	0.0042	0.709	11441	0.059	0.549	2910	0.285	0.811
1953.NelsonMandela.txt	27937	0.0025	0.703	22128	0.046	0.530	4967	0.289	0.801
1957.MartinLutherKing.txt	39237	0.0018	0.700	31528	0.039	0.510	7953	0.159	0.780
1959.RichardFeynman.txt	39621	0.0018	0.693	31335	0.042	0.511	8218	0.159	0.786

F. Properties of some languages at different scales

<i>Message Name</i>	At Char Scale			At Fundamental Scale			At Word Scale		
	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>
	[chrs]	[0-1]	[0-1]	[F.S.]	[0-1]	[0-1]	[w]	[0-1]	[0-1]
1961.01.JohnFKennedy.txt	7433	0.0070	0.739	6039	0.070	0.588	1521	0.348	0.852
1961.04.JohnFKennedy.txt	8697	0.0071	0.708	6936	0.068	0.568	1715	0.353	0.845
1961.05.JohnFKennedy.txt	35545	0.0019	0.703	27984	0.048	0.517	6588	0.233	0.799
1962.09.JohnFKennedy.txt	11652	0.0057	0.697	9748	0.061	0.554	2441	0.308	0.827
1962.10.JohnFKennedy.txt	14787	0.0045	0.708	11833	0.056	0.552	2772	0.293	0.829
1962.12.MalcomX.txt	80830	0.0009	0.697	62446	0.034	0.479	17561	0.095	0.757
1963.06.10.JohnFKennedy.txt	18539	0.0035	0.699	14797	0.056	0.545	3680	0.277	0.815
1963.09.20.JohnFKennedy.txt	20998	0.0031	0.706	16571	0.055	0.534	3988	0.273	0.804
1963.MartinLutherKing.txt	8526	0.0063	0.736	6873	0.067	0.586	1731	0.304	0.837
1964.04.MalcomX.txt	15616	0.0044	0.706	11543	0.074	0.537	3381	0.198	0.817
1964.NelsonMandela.txt	63224	0.0012	0.699	50334	0.034	0.498	11935	0.180	0.767
1965.03.LyndonBJohnson.txt	20217	0.0031	0.716	16523	0.049	0.544	4169	0.235	0.805
1965.04.LyndonBJohnson.txt	6171	0.0088	0.733	4919	0.069	0.598	1286	0.326	0.849
1967.MartinLutherKing.txt	37609	0.0018	0.702	30603	0.043	0.514	7366	0.237	0.794
1968.MartinLutherKing.txt	23357	0.0028	0.711	18085	0.051	0.535	5119	0.195	0.793
1969.RichardNixon.txt	26380	0.0027	0.700	20360	0.050	0.524	5070	0.218	0.805
1972.RichardNixon.txt	25132	0.0030	0.690	19868	0.042	0.527	5406	0.172	0.795
1974.RichardNixon.txt	9735	0.0056	0.730	7773	0.064	0.570	1959	0.274	0.833
1976.NL.Eng.SaulBellow.txt	28967	0.0025	0.699	23205	0.046	0.527	5639	0.266	0.799
1979.MargaretThatcher.txt	17079	0.0037	0.715	13812	0.057	0.541	3219	0.312	0.820
1982.RonaldReagan.txt	26695	0.0027	0.695	21482	0.046	0.531	5052	0.277	0.807
1983.NL.Eng.WilliamGolding.tx	24742	0.0026	0.703	20114	0.052	0.534	5150	0.267	0.812
1983.RonaldReagan.txt	26631	0.0025	0.708	21269	0.052	0.528	5236	0.242	0.814
1986.NL.Eng.WoleSoyinka.txt	48990	0.0015	0.695	39377	0.038	0.507	9034	0.280	0.778
1987.RonaldReagan.txt	15753	0.0043	0.705	12717	0.057	0.553	3170	0.296	0.822
1988.AnnRichards.txt	15083	0.0043	0.712	11992	0.057	0.558	3126	0.278	0.827
1991.GeorgeBush.txt	8716	0.0064	0.739	7152	0.069	0.584	1782	0.328	0.844
1991.NL.Eng.NadineGordimer.t	22521	0.0032	0.694	18005	0.052	0.533	4386	0.286	0.802
1992.NL.Eng.DerekWalcott.txt	37759	0.0018	0.702	29694	0.046	0.512	7407	0.266	0.774
1993.NL.Eng.ToniMorrison.txt	17471	0.0035	0.719	13811	0.059	0.549	3492	0.294	0.813
1995.ErikaJong.txt	12131	0.0051	0.717	10030	0.049	0.569	2401	0.252	0.832
1995.HillaryClinton.txt	12878	0.0048	0.714	10546	0.053	0.567	2487	0.289	0.823
1995.NL.Eng.SeamusHeaney.tx	36355	0.0020	0.689	28865	0.045	0.517	7054	0.270	0.787
1997.EarlOfSpencer.txt	6509	0.0078	0.741	5284	0.073	0.598	1327	0.383	0.857
1997.NancyBirdsall.txt	13010	0.0052	0.717	10323	0.055	0.569	2312	0.279	0.833
1997.PrincessDiana.txt	8558	0.0072	0.717	6942	0.071	0.576	1759	0.343	0.849
1999.AnitaRoddick.txt	9966	0.0059	0.722	7987	0.072	0.568	2040	0.313	0.843
2000.CondoleezzaRice.txt	7551	0.0079	0.729	5972	0.073	0.591	1511	0.341	0.854
2000.CourtneyLove.txt	38575	0.0019	0.697	32153	0.041	0.513	8344	0.196	0.799
2001.NL.Eng.VSNaipaul.txt	30520	0.0024	0.697	23708	0.047	0.524	6327	0.194	0.788
2003.NL.Eng.JMCoetzee.txt	20992	0.0028	0.708	16858	0.052	0.531	4593	0.241	0.793
2005.NL.Eng.HaroldPinter.txt	29213	0.0025	0.702	23586	0.041	0.539	5833	0.255	0.803
2005.SteveJobs.txt	12200	0.0056	0.708	9360	0.065	0.568	2615	0.273	0.832
2007.NL.Eng.DorisLessing.txt	26956	0.0024	0.706	21783	0.046	0.524	5898	0.212	0.793

Table F.2: [Table](#)

Spanish Properties at Different Scales									
	<i>L</i> = Length			<i>c</i> = Complexity			<i>[w]</i> = [words]		
	<i>d</i> = Specific Diversity			<i>[F.S.]</i> = [Fundamental Symbols]					
	<i>h</i> = entropy			<i>[chrs]</i> = [characters]			<i>[0-1]</i> = between 0 and 1		
Message Name	At Char Scale			At Fundamental Scale			At Word Scale		
	<i>L</i> [chrs]	<i>d</i> [0-1]	<i>h</i> [0-1]	<i>L</i> [F.S.]	<i>d</i> [0-1]	<i>h</i> [0-1]	<i>L</i> [w]	<i>d</i> [0-1]	<i>h</i> [0-1]
1805.Simon Bolivar.txt	2480	0.0226	0.734	2118	0.0817	0.660	462	0.4978	0.878
1813.Simon Bolivar.txt	4127	0.0119	0.749	3306	0.0811	0.612	739	0.4493	0.864
1830.Simon Bolivar.txt	1120	0.0402	0.768	923	0.1073	0.709	201	0.6020	0.930
1931.Manuel Azana.txt	1617	0.0291	0.763	1282	0.0983	0.675	297	0.5118	0.906
1936.Dolores Ibarruri.txt	3069	0.0199	0.733	2204	0.0867	0.626	641	0.3276	0.863
1936.Jose Buenaventura Durru	3672	0.0155	0.726	2977	0.0820	0.605	690	0.4420	0.877
1938.Dolores Ibarruri.txt	4273	0.0133	0.727	3468	0.0819	0.597	774	0.4109	0.846
1945.Juan Domingo Perón.txt	5861	0.0096	0.726	4760	0.0666	0.596	1192	0.3649	0.866
1959.Fulgencio Batista.txt	466	0.0751	0.806	416	0.1202	0.772	85	0.6824	0.947
1967.BS.Esp.MiguelAngelAsturi	4237	0.0123	0.745	3412	0.0780	0.614	804	0.4216	0.845
1971.BS.Esp.PabloNeruda.txt	2326	0.0181	0.773	1789	0.0900	0.669	468	0.4466	0.859
1973.Bando Nro 5.txt	4601	0.0130	0.717	3711	0.0738	0.599	801	0.4569	0.860
1973.Salvador Allende.txt	3809	0.0139	0.748	2938	0.0841	0.622	700	0.4486	0.868
1976.Jorge Videla.txt	3380	0.0148	0.753	2556	0.0876	0.623	604	0.4371	0.875
1977.BS.Esp.VicenteAleixandre	1265	0.0348	0.775	1059	0.1020	0.698	241	0.5685	0.917
1978.Juan Carlos I.txt	5507	0.0096	0.737	4345	0.0769	0.584	973	0.4224	0.848
1982.BS.Esp.GabrielGarciaMar	2738	0.0175	0.751	2188	0.0905	0.641	522	0.4808	0.876
1982.Leopoldo Galtieri.txt	694	0.0634	0.778	634	0.1025	0.754	119	0.6387	0.934
1990.BS.Esp.OctavioPaz.txt	3345	0.0158	0.740	2619	0.0909	0.625	613	0.4633	0.878
2004.Pilar Manjón.txt	1149	0.0409	0.768	950	0.1095	0.702	209	0.5646	0.917
2008.J. L. Rodriguez Zapatero.t	2549	0.0177	0.767	2005	0.0853	0.647	449	0.4543	0.886
2008.Julio Cobos.txt	1443	0.0340	0.755	1210	0.0901	0.688	280	0.4929	0.907
2010.BS.Esp.MarioVargasLlosa	2179	0.0225	0.752	1756	0.0928	0.654	424	0.4811	0.888
2010.Raúl Castro.txt	1415	0.0339	0.765	1158	0.1071	0.695	260	0.5577	0.912
2010.Sebastian Pinera Echeniq	2203	0.0213	0.757	1718	0.0902	0.648	432	0.4005	0.890
1868.CarlosMCespedes.txt	8081	0.0063	0.736	6361	0.0629	0.582	1457	0.4056	0.836
1819.Simon Bolivar.txt	63674	0.0011	0.696	50374	0.0340	0.488	11502	0.2286	0.751
1912.Emiliano Zapata.txt	14493	0.0041	0.715	11946	0.0483	0.550	2590	0.3610	0.811
1917.Emiliano Zapata.txt	9001	0.0059	0.732	6967	0.0693	0.563	1619	0.4033	0.826
1918.Emiliano Zapata.txt	8025	0.0070	0.724	6515	0.0597	0.584	1438	0.4124	0.830
1933.JAntonioPrimoDeRivera.t	16896	0.0039	0.692	13498	0.0528	0.528	3190	0.3047	0.803
1946.Jorge Eliecer Gaitan.txt	18953	0.0034	0.704	14620	0.0561	0.531	3544	0.2782	0.811
1952.Eva Perón.txt	5672	0.0100	0.721	4383	0.0719	0.588	1124	0.3060	0.839
1959.Fidel Castro.txt	15237	0.0050	0.702	11936	0.0567	0.540	2892	0.2950	0.810
1964.Ernesto Che Guevara.txt	40987	0.0020	0.672	32534	0.0432	0.497	7172	0.2665	0.779
1967.Ernesto Che Guevara.txt	33029	0.0023	0.681	26129	0.0460	0.508	5870	0.2891	0.788
1967.Fidel Castro.txt	30129	0.0023	0.693	23837	0.0424	0.516	5519	0.2232	0.788
1967.NL.Esp.MiguelAngelAsturi	26424	0.0030	0.674	21555	0.0463	0.513	4901	0.3128	0.787
1970.Salvador Allende.txt	11048	0.0056	0.709	8734	0.0600	0.563	1865	0.3850	0.834
1971.Pablo Neruda.txt	19893	0.0031	0.704	15712	0.0538	0.532	3683	0.3503	0.806

F. Properties of some languages at different scales

<i>Message Name</i>	At Char Scale			At Fundamental Scale			At Word Scale		
	<i>L</i> [chrs]	<i>d</i> [0-1]	<i>h</i> [0-1]	<i>L</i> [F.S.]	<i>d</i> [0-1]	<i>h</i> [0-1]	<i>L</i> [w]	<i>d</i> [0-1]	<i>h</i> [0-1]
1972.Salvador Allende.txt	42804	0.0015	0.706	33483	0.0429	0.501	7417	0.2694	0.778
1973.Augusto Pinochet.txt	23950	0.0025	0.714	18621	0.0491	0.527	4193	0.3146	0.797
1977.NL.Esp.VicenteAleixandre	12379	0.0056	0.695	10068	0.0609	0.552	2379	0.3611	0.818
1979.Adolfo Suarez.txt	79333	0.0009	0.679	61701	0.0349	0.476	13201	0.2120	0.751
1979.Fidel Castro.txt	74583	0.0011	0.669	59493	0.0320	0.476	12838	0.2078	0.743
1981.Adolfo Suarez.txt	7346	0.0065	0.751	5531	0.0674	0.574	1348	0.3116	0.842
1981.Roberto Eduardo Viola.txt	23067	0.0029	0.698	18209	0.0499	0.525	3823	0.3369	0.799
1982.Gabriel Garcia Marquez.t	11419	0.0061	0.693	9358	0.0550	0.555	2095	0.4086	0.831
1982.Felipe González.txt	38382	0.0019	0.681	30636	0.0416	0.499	6592	0.2758	0.782
1983.Raul Alfonsín.txt	18599	0.0034	0.704	14833	0.0501	0.538	3309	0.2950	0.805
1987.Camilo Jose Cela.txt	8301	0.0076	0.714	6554	0.0647	0.575	1591	0.3903	0.830
1989.Carlos Saul Menem.txt	6450	0.0085	0.732	4966	0.0763	0.580	1199	0.3369	0.845
1989.NL.Esp.CamiloJoseCela.t	33979	0.0022	0.676	27214	0.0419	0.509	6293	0.2867	0.777
1990.NL.Esp.OctavioPaz.txt	25831	0.0029	0.685	20968	0.0460	0.518	4836	0.3002	0.788
1992.Rafael Caldera.txt	14167	0.0045	0.700	11091	0.0572	0.539	2504	0.3323	0.810
1996.Jose Maria Aznar.txt	29982	0.0022	0.699	24043	0.0403	0.522	5071	0.2727	0.782
1999.Hugo Chavez.txt	66784	0.0013	0.667	53936	0.0329	0.482	12768	0.1912	0.760
2000.Vicente Fox.txt	42804	0.0015	0.706	33483	0.0429	0.501	7417	0.2694	0.778
2001.Fernando de la Rúa.txt	6342	0.0090	0.723	4926	0.0717	0.588	1129	0.3862	0.853
2005.Daniel Ortega.txt	40751	0.0019	0.689	32819	0.0260	0.530	7651	0.1979	0.778
2006.Alvaro Uribe.txt	26323	0.0026	0.695	21129	0.0475	0.522	4555	0.3407	0.776
2006.Evo Morales.txt	18755	0.0042	0.680	14748	0.0500	0.527	3393	0.2626	0.812
2006.Gaston Acurio.txt	24311	0.0029	0.689	19835	0.0437	0.525	4360	0.2927	0.803
2006.Hugo Chavez.txt	18043	0.0040	0.697	14287	0.0538	0.536	3353	0.2827	0.808
2007.Cristina Kirchner.txt	27524	0.0027	0.692	21156	0.0487	0.506	5008	0.2452	0.795
2007.Daniel Ortega.txt	18653	0.0039	0.687	14823	0.0487	0.533	3373	0.2541	0.805
2010.NL.Esp.MarioVargasLlosa	37797	0.0021	0.677	30184	0.0433	0.507	7034	0.3149	0.763
CamiloJoseCela.LaColmena.Nc	8041	0.0077	0.710	6548	0.0640	0.573	1623	0.3672	0.829
GabrielGMarquez.DicursoCarta	7397	0.0088	0.706	5891	0.0689	0.576	1443	0.4012	0.844
GabrielGMarquez.MejorOficioE	16483	0.0036	0.712	12920	0.0566	0.537	2949	0.3591	0.808
MarioVargasLlosa.DiscursoBue	10772	0.0061	0.700	8661	0.0637	0.555	1986	0.3912	0.820
OctavioPaz.DiscursoZacatecas.	11767	0.0048	0.718	0	0.0576	0.551	2238	0.3177	0.810

Table F.3: [Table](#)

Programing Code Properties at Different Scales									
	<i>L</i> = Length			<i>c</i> = Complexity			<i>[w]</i> = [words]		
	<i>d</i> = Specific Diversity			<i>[F.S.]</i> = [Fundamental Symbols]					
	<i>h</i> = entropy			<i>[chrs]</i> = [characters]			<i>[0-1]</i> = between 0 and 1		
	At Char Scale			At Fundamental Scale			At Word Scale		
<i>Message Name</i>	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>	<i>L</i>	<i>d</i>	<i>h</i>
	[chrs]	[0-1]	[0-1]	[F.S.]	[0-1]	[0-1]	[w]	[0-1]	[0-1]
BoolFunctWithMultiplexerLogic.C.t	3369	0.0258	0.7191	2751	0.0712	0.634	1112	0.1457	0.794
ChainedScatterTable.CSharp.txt	793	0.0668	0.785	605	0.1306	0.727	201	0.2289	0.890
CopyFolderNContent.CSharp.txt	986	0.0517	0.811	687	0.1223	0.743	203	0.2463	0.910
ExtendedEuclidean.C.txt	200	0.1650	0.704	187	0.1872	0.679	88	0.2841	0.904
Factorial.CSharp.txt	138	0.2174	0.837	111	0.2793	0.802	40	0.5500	0.961
FibonacciNumbers.CSharp.txt	229	0.1485	0.809	174	0.2299	0.751	64	0.4375	0.924
GameOfLife.C.txt	729	0.0782	0.763	499	0.1403	0.647	247	0.1862	0.893
HanoiTowers.Java.txt	2055	0.0345	0.761	1526	0.0858	0.683	512	0.1816	0.846
HeapSort.CSharp.txt	768	0.0625	0.759	524	0.1221	0.658	261	0.1801	0.901
HeapSort.Java.txt	1110	0.0514	0.746	808	0.1002	0.663	340	0.1765	0.854
InsertAfterBefore.CSharp.txt	611	0.0655	0.813	393	0.1552	0.731	141	0.2624	0.935
IsPrime.C.txt	572	0.1049	0.766	431	0.1740	0.734	162	0.3519	0.905
Levenberg.MathLab.txt	1728	0.0324	0.733	1214	0.1112	0.625	579	0.1658	0.826
MathLab.Fr.MathLab.txt	5579	0.0133	0.704	4098	0.0639	0.591	1723	0.1207	0.790
MatrixLUDecomp.CSharp.txt	1183	0.0482	0.730	884	0.0894	0.633	420	0.1262	0.858
MatrixLUDecomp.Python.txt	2084	0.0331	0.739	1610	0.0981	0.604	702	0.1624	0.783
MetaWords.FormsAnsClasses.CSha	6069	0.0129	0.774	4863	0.0506	0.651	1341	0.1081	0.826
ModularInverse.C.txt	220	0.1727	0.693	218	0.1743	0.689	95	0.3368	0.906
PartDifEqtnsLaplaceEq.MathLab.tx	2113	0.0289	0.685	1574	0.0807	0.585	843	0.1163	0.780
PartDifEqtnsWaveEqtn.MathLab.tx	670	0.0821	0.732	569	0.1318	0.682	249	0.2731	0.854
PermutationAlgorithm.Csharp.txt	2499	0.0256	0.725	1591	0.0886	0.614	825	0.1079	0.839
PermutationAlgorithm.Java.txt	5913	0.0108	0.766	3024	0.0612	0.582	1305	0.0743	0.776
Polinom.CSharp.txt	396	0.1111	0.812	226	0.1947	0.755	92	0.3696	0.917
QuadraticPrograming.CSharp.txt	2078	0.0322	0.785	934	0.1167	0.645	485	0.1464	0.847
QuickSort.CSharp.txt	1204	0.0507	0.744	759	0.1120	0.648	376	0.1516	0.898
SnakeGame.C.txt	4833	0.0166	0.734	4415	0.0392	0.661	1545	0.1003	0.804
Sumation.CSharp.txt	360	0.1000	0.849	188	0.1915	0.746	71	0.3521	0.895
BlowfishEncryption.C.txt	22732	0.0040	0.741	17075	0.0362	0.628	4808	0.2552	0.674
FiniteElements.MathLab.txt	8128	0.0095	0.700	6703	0.0513	0.574	2802	0.1056	0.732
FTPFunctions.CSharp.txt	34172	0.0026	0.758	21721	0.0335	0.571	7410	0.0421	0.714
MathLab.programa2.MathLab.txt	25429	0.0040	0.652	16954	0.0363	0.487	9346	0.0289	0.686
MathLab.Taller.MathLab.txt	9531	0.0069	0.761	4405	0.0661	0.560	2166	0.0559	0.741
MatrixFuncs.CSharp.txt	16318	0.0053	0.691	11930	0.0370	0.532	5801	0.0336	0.742
MetaWordsMainForm.CSharp.txt	186597	0.0006	0.748	124312	0.0229	0.515	41392	0.0272	0.648
NetPlex.Classes.CSharp.txt	86696	0.0010	0.777	53863	0.0322	0.535	19976	0.0327	0.680
NetPlex.Forms.CSharp.txt	347738	0.0003	0.774	226256	0.0209	0.502	69994	0.0212	0.637
NetPlexMainForm.CSharp.txt	191198	0.0005	0.763	139473	0.0219	0.509	40258	0.0305	0.633
Sociodynamica.Forms.txt	9994	0.0080	0.762	6872	0.0534	0.623	2498	0.1201	0.759
Sociodynamica.Module1.txt	44370	0.0018	0.742	22334	0.0365	0.529	10263	0.0284	0.665
Sociodynamica.Module2.txt	27903	0.0032	0.738	16273	0.0482	0.523	7932	0.0538	0.706
Sociodynamica.Module3.txt	11890	0.0067	0.744	7673	0.0508	0.574	3599	0.0622	0.763
ViscomSoft.ScannerActivex.CSharp.	31157	0.0031	0.748	23600	0.0363	0.572	6483	0.0956	0.684
WebSite.Inmogal.php.txt	75224	0.0014	0.705	42260	0.0363	0.493	19279	0.0339	0.631
WebSite.RistEuropa.Html.txt	47289	0.0021	0.692	25804	0.0350	0.494	11715	0.0430	0.595
WebSite.TiempoReal.Html.txt	33799	0.0026	0.709	17610	0.0474	0.506	7509	0.0752	0.587

Table F.4: [Table](#)

MIDI Music properties At Fundamental Scale					
<i>N</i> = Length		<i>[F.S.]</i> = [Fundamental Symbols]			
<i>d</i> = Specific Diversity		<i>[chrs]</i> = [characters]			
<i>h</i> = entropy		<i>[w]</i> = [words]		<i>[0-1]</i> = between 0 and 1	
At Fundamental Scale					
<i>Period/Style</i>	Pieces	<i>Composers</i>	<i>N</i> [F.S.]	<i>d</i> [0-1]	<i>h</i> [0-1]
Total	438	>71			
Medieval		12	182839	0.0143	0.6281
Reinainssance		10	308629	0.0215	0.5952
Baroque		8	1396651	0.0222	0.5386
Classical		7	2409305	0.0222	0.5343
Romantic		13	4809201	0.0202	0.5182
Impressionistic		4	1363204	0.0244	0.5139
20th Century		8	1455986	0.0243	0.5220
Chinese		Several	474214	0.0304	0.5414
Hindu-Raga		Several	109055	0.0409	0.6220
Movie Themes		Several	400105	0.0349	0.5781
Rock		5	619413	0.0287	0.4968
Venezuelan		>20	894249	0.0273	0.4549

Appendix G

MIDI music properties. Musicnet

Properties of text codes obtained texts from *MIDI* music files can be seen in the link indicated below.

Table G.1: [MIDI MUSIC PROPERTIES.](#)

<http://www.gfebres.com/F0IndexFrame/F132Body/F132BodyPublications/MusicComplexityModels/MusicNet.Tree/MusicNet.htm>

MusicNet.												
Class	Type	Period/Style	Region	Genre			Spec. diversity		Entropy		2nd Ord. Ent.	
					Composers	Pieces	Ave.	Std.Dev.	Ave.	Std.Dev.	Ave.	Std.Dev.
Total					71	453						
Western	Academic	Medieval			12	40	0.062	0.026	0.649	0.048	0.949	0.037
		Reinainssance			10	31	0.048	0.016	0.622	0.037	0.935	0.041
		Baroque			8	55	0.039	0.013	0.581	0.057	0.911	0.050
		Classical			7	45	0.040	0.019	0.566	0.059	0.896	0.049
		Romantic			13	89	0.049	0.021	0.602	0.068	0.914	0.061
		Impressionistic			4	34	0.050	0.015	0.582	0.052	0.921	0.044
		20th Century			8	35	0.052	0.017	0.559	0.057	0.888	0.062
	Traditional		Venezuelan Tradition.	>20	56	0.049	0.014	0.540	0.056	0.929	0.036	
	Popular / Contemp.			Movie Themes		18	0.048	0.010	0.615	0.051	0.934	0.033
				Rock	5	24	0.041	0.010	0.585	0.043	0.919	0.045
			Jazz									
			Regie Tecno									
Asian	Traditional		Hindu-Ra, Raga	Several	14	0.083	0.019	0.697	0.061	0.974	0.026	
			Chinese	Several	12	0.048	0.015	0.582	0.038	0.915	0.046	

Appendix H

Numerical data of the symbol
frequency profiles for *MIDI* music

H. Numerical of some languages at different scales

Table H.1: Numerical data for symbol probability for different types of music. Symbols determined by the fundamental scale method

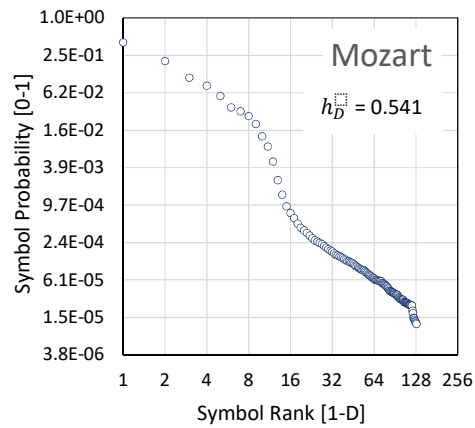
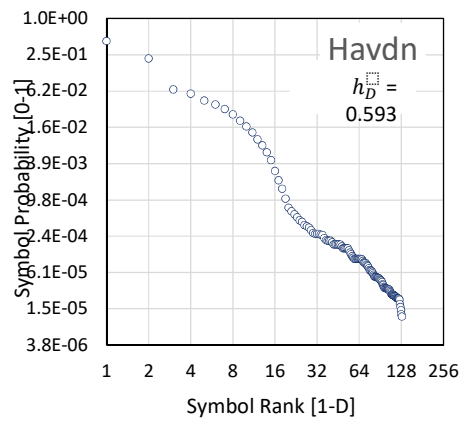
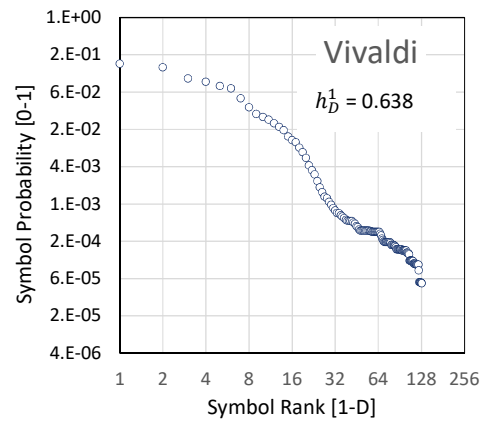
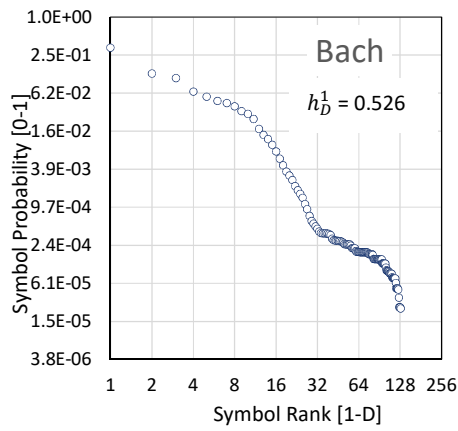
Symbol probability for several types of music. Observation scale = Diversity = 129 Symbols. From r=1 to r=62												
rnk.	Medieval	Renaiss.	Baroque	Classic	Romant.	Impres.	20th Cty.	Chinese	Raga	V.Thems.	Rock	Venez.
1	5.96E-01	6.96E-01	8.87E-01	8.92E-01	9.00E-01	8.92E-01	9.05E-01	8.25E-01	5.85E-01	7.82E-01	8.63E-01	9.02E-01
2	1.52E-01	1.35E-01	3.63E-02	1.31E-02	1.12E-02	2.69E-02	2.45E-02	6.29E-02	1.37E-01	1.04E-01	4.32E-02	1.83E-02
3	7.92E-02	5.16E-02	7.92E-03	8.52E-03	7.58E-03	1.43E-02	7.00E-03	9.77E-03	6.85E-02	1.10E-02	8.38E-03	1.01E-02
4	4.62E-02	1.80E-02	5.29E-03	6.87E-03	5.86E-03	6.53E-03	5.11E-03	6.19E-03	5.25E-02	5.88E-03	5.13E-03	7.17E-03
5	2.75E-02	1.00E-02	4.10E-03	5.77E-03	4.82E-03	4.73E-03	4.13E-03	5.05E-03	2.79E-02	4.31E-03	4.05E-03	5.56E-03
6	1.61E-02	7.38E-03	3.49E-03	4.99E-03	4.11E-03	3.79E-03	3.44E-03	4.46E-03	1.01E-02	3.57E-03	3.46E-03	4.72E-03
7	1.04E-02	5.62E-03	2.98E-03	4.40E-03	3.63E-03	3.32E-03	2.91E-03	3.94E-03	5.69E-03	3.15E-03	3.03E-03	4.03E-03
8	6.74E-03	4.48E-03	2.63E-03	3.99E-03	3.26E-03	3.05E-03	2.62E-03	3.47E-03	4.38E-03	2.91E-03	2.68E-03	3.32E-03
9	5.33E-03	3.80E-03	2.36E-03	3.62E-03	2.95E-03	2.79E-03	2.43E-03	3.15E-03	3.60E-03	2.61E-03	2.37E-03	2.88E-03
10	4.20E-03	3.40E-03	2.10E-03	3.36E-03	2.67E-03	2.38E-03	2.17E-03	3.06E-03	3.46E-03	2.42E-03	2.17E-03	2.62E-03
11	3.75E-03	2.92E-03	1.91E-03	2.92E-03	2.40E-03	2.13E-03	1.95E-03	2.83E-03	3.27E-03	2.20E-03	2.01E-03	2.33E-03
12	2.99E-03	2.67E-03	1.82E-03	2.60E-03	2.18E-03	1.87E-03	1.80E-03	2.77E-03	2.98E-03	2.02E-03	1.90E-03	2.13E-03
13	2.69E-03	2.54E-03	1.65E-03	2.38E-03	2.03E-03	1.72E-03	1.68E-03	2.68E-03	2.87E-03	1.94E-03	1.78E-03	1.94E-03
14	2.61E-03	2.33E-03	1.54E-03	2.33E-03	1.92E-03	1.62E-03	1.52E-03	2.56E-03	2.61E-03	1.87E-03	1.67E-03	1.81E-03
15	2.51E-03	2.23E-03	1.45E-03	2.23E-03	1.80E-03	1.51E-03	1.43E-03	2.51E-03	2.51E-03	1.80E-03	1.55E-03	1.73E-03
16	2.05E-03	2.11E-03	1.40E-03	2.07E-03	1.67E-03	1.36E-03	1.35E-03	2.45E-03	2.31E-03	1.69E-03	1.47E-03	1.67E-03
17	1.88E-03	1.93E-03	1.31E-03	1.95E-03	1.55E-03	1.24E-03	1.28E-03	2.31E-03	2.25E-03	1.68E-03	1.43E-03	1.59E-03
18	1.63E-03	1.85E-03	1.27E-03	1.83E-03	1.45E-03	1.20E-03	1.21E-03	2.18E-03	2.08E-03	1.61E-03	1.37E-03	1.53E-03
19	1.50E-03	1.80E-03	1.25E-03	1.68E-03	1.38E-03	1.18E-03	1.12E-03	1.99E-03	2.05E-03	1.54E-03	1.31E-03	1.37E-03
20	1.39E-03	1.64E-03	1.21E-03	1.50E-03	1.33E-03	1.14E-03	1.03E-03	1.85E-03	2.00E-03	1.48E-03	1.29E-03	1.26E-03
21	1.30E-03	1.56E-03	1.14E-03	1.35E-03	1.25E-03	1.12E-03	9.74E-04	1.70E-03	1.86E-03	1.44E-03	1.25E-03	1.15E-03
22	1.22E-03	1.52E-03	1.08E-03	1.24E-03	1.17E-03	1.10E-03	9.43E-04	1.61E-03	1.85E-03	1.41E-03	1.17E-03	1.02E-03
23	1.09E-03	1.47E-03	1.01E-03	1.18E-03	1.12E-03	1.01E-03	8.99E-04	1.53E-03	1.72E-03	1.36E-03	1.13E-03	9.50E-04
24	9.73E-04	1.43E-03	9.53E-04	1.12E-03	1.06E-03	9.15E-04	8.58E-04	1.42E-03	1.71E-03	1.30E-03	1.09E-03	9.01E-04
25	9.70E-04	1.35E-03	8.92E-04	1.05E-03	1.02E-03	8.50E-04	8.26E-04	1.31E-03	1.62E-03	1.23E-03	1.07E-03	8.80E-04
26	8.85E-04	1.27E-03	8.40E-04	9.83E-04	9.75E-04	7.91E-04	7.94E-04	1.28E-03	1.62E-03	1.18E-03	1.05E-03	8.60E-04
27	8.29E-04	1.20E-03	7.95E-04	9.39E-04	9.31E-04	7.71E-04	7.52E-04	1.26E-03	1.54E-03	1.14E-03	1.01E-03	8.39E-04
28	7.93E-04	1.14E-03	7.51E-04	9.12E-04	8.95E-04	7.29E-04	7.24E-04	1.24E-03	1.53E-03	1.10E-03	9.92E-04	7.83E-04
29	7.53E-04	1.11E-03	7.09E-04	8.85E-04	8.55E-04	6.63E-04	6.96E-04	1.20E-03	1.45E-03	1.07E-03	9.42E-04	7.02E-04
30	7.04E-04	1.04E-03	6.70E-04	8.47E-04	8.21E-04	6.28E-04	6.53E-04	1.11E-03	1.44E-03	1.04E-03	9.16E-04	6.45E-04
31	6.65E-04	1.02E-03	6.44E-04	8.08E-04	7.93E-04	6.08E-04	6.20E-04	1.06E-03	1.39E-03	1.02E-03	9.01E-04	6.21E-04
32	6.32E-04	1.00E-03	6.23E-04	7.51E-04	7.69E-04	5.95E-04	5.89E-04	1.00E-03	1.36E-03	1.01E-03	8.79E-04	5.79E-04
33	6.07E-04	9.96E-04	6.11E-04	6.93E-04	7.41E-04	5.72E-04	5.55E-04	9.42E-04	1.32E-03	1.00E-03	8.49E-04	5.38E-04
34	5.89E-04	9.78E-04	5.89E-04	6.42E-04	7.09E-04	5.43E-04	5.18E-04	8.78E-04	1.32E-03	9.78E-04	7.99E-04	5.15E-04
35	5.70E-04	9.31E-04	5.65E-04	6.23E-04	6.85E-04	5.08E-04	4.98E-04	8.47E-04	1.32E-03	9.61E-04	7.79E-04	4.85E-04
36	5.30E-04	8.97E-04	5.47E-04	6.16E-04	6.54E-04	4.75E-04	4.82E-04	8.34E-04	1.25E-03	9.61E-04	7.69E-04	4.51E-04
37	5.14E-04	8.23E-04	5.33E-04	6.08E-04	6.22E-04	4.46E-04	4.67E-04	8.28E-04	1.22E-03	9.44E-04	7.51E-04	4.27E-04
38	4.86E-04	7.82E-04	5.15E-04	5.93E-04	5.95E-04	4.25E-04	4.61E-04	8.10E-04	1.15E-03	9.22E-04	7.34E-04	4.12E-04
39	4.74E-04	7.46E-04	5.01E-04	5.76E-04	5.65E-04	4.09E-04	4.46E-04	7.88E-04	1.15E-03	9.04E-04	7.28E-04	3.98E-04
40	4.65E-04	7.04E-04	4.77E-04	5.56E-04	5.44E-04	3.93E-04	4.31E-04	7.81E-04	1.12E-03	8.79E-04	7.22E-04	3.79E-04
41	4.65E-04	6.85E-04	4.67E-04	5.19E-04	5.25E-04	3.78E-04	4.22E-04	7.70E-04	1.11E-03	8.60E-04	7.14E-04	3.50E-04
42	4.63E-04	6.79E-04	4.58E-04	4.81E-04	5.16E-04	3.64E-04	4.14E-04	7.32E-04	1.08E-03	8.40E-04	7.10E-04	3.22E-04
43	4.44E-04	6.78E-04	4.48E-04	4.59E-04	5.02E-04	3.47E-04	4.02E-04	6.95E-04	1.05E-03	8.25E-04	7.04E-04	3.07E-04
44	4.35E-04	6.36E-04	4.32E-04	4.42E-04	4.83E-04	3.28E-04	3.86E-04	6.59E-04	1.04E-03	7.99E-04	6.94E-04	2.97E-04
45	4.26E-04	6.27E-04	4.25E-04	4.23E-04	4.70E-04	3.15E-04	3.77E-04	6.28E-04	1.01E-03	7.91E-04	6.77E-04	2.82E-04
46	4.18E-04	5.99E-04	4.16E-04	4.06E-04	4.58E-04	3.08E-04	3.68E-04	5.93E-04	9.88E-04	7.79E-04	6.55E-04	2.68E-04
47	3.93E-04	5.74E-04	4.07E-04	3.83E-04	4.47E-04	3.01E-04	3.55E-04	5.76E-04	9.21E-04	7.55E-04	6.39E-04	2.60E-04
48	3.93E-04	5.44E-04	3.97E-04	3.65E-04	4.33E-04	2.93E-04	3.37E-04	5.63E-04	8.91E-04	7.31E-04	6.24E-04	2.51E-04
49	3.89E-04	5.18E-04	3.92E-04	3.49E-04	4.21E-04	2.85E-04	3.28E-04	5.40E-04	8.79E-04	7.21E-04	6.05E-04	2.40E-04
50	3.84E-04	5.04E-04	3.85E-04	3.30E-04	4.10E-04	2.76E-04	3.21E-04	5.22E-04	8.65E-04	7.14E-04	5.80E-04	2.23E-04
51	3.66E-04	4.91E-04	3.79E-04	3.16E-04	4.03E-04	2.64E-04	3.07E-04	5.21E-04	8.59E-04	7.12E-04	5.53E-04	2.13E-04
52	3.61E-04	4.63E-04	3.71E-04	3.10E-04	3.98E-04	2.50E-04	2.99E-04	5.14E-04	8.46E-04	6.94E-04	5.38E-04	2.08E-04
53	3.58E-04	4.46E-04	3.59E-04	3.05E-04	3.90E-04	2.42E-04	2.91E-04	5.09E-04	8.41E-04	6.76E-04	5.21E-04	2.02E-04
54	3.51E-04	4.28E-04	3.47E-04	2.99E-04	3.82E-04	2.36E-04	2.87E-04	5.00E-04	8.07E-04	6.53E-04	5.07E-04	1.99E-04
55	3.32E-04	3.97E-04	3.32E-04	2.94E-04	3.73E-04	2.30E-04	2.81E-04	4.88E-04	8.02E-04	6.44E-04	4.97E-04	1.94E-04
56	3.20E-04	3.76E-04	3.25E-04	2.87E-04	3.62E-04	2.24E-04	2.75E-04	4.83E-04	7.97E-04	6.24E-04	4.81E-04	1.88E-04
57	3.16E-04	3.56E-04	3.16E-04	2.81E-04	3.52E-04	2.16E-04	2.66E-04	4.65E-04	7.89E-04	6.05E-04	4.64E-04	1.75E-04
58	3.07E-04	3.54E-04	3.12E-04	2.74E-04	3.43E-04	2.09E-04	2.59E-04	4.58E-04	7.89E-04	5.93E-04	4.55E-04	1.67E-04
59	2.96E-04	3.41E-04	3.06E-04	2.63E-04	3.36E-04	2.03E-04	2.52E-04	4.43E-04	7.74E-04	5.76E-04	4.45E-04	1.62E-04
60	2.95E-04	3.39E-04	3.01E-04	2.51E-04	3.23E-04	1.96E-04	2.46E-04	4.38E-04	7.70E-04	5.70E-04	4.37E-04	1.54E-04
61	2.93E-04	3.30E-04	2.92E-04	2.45E-04	3.13E-04	1.89E-04	2.45E-04	4.30E-04	7.68E-04	5.67E-04	4.28E-04	1.49E-04
62	2.87E-04	3.22E-04	2.85E-04	2.39E-04	3.04E-04	1.83E-04	2.34E-04	4.20E-04	7.65E-04	5.53E-04	4.22E-04	1.40E-04

H. Numerical data of the symbol frequency profiles for MIDI music

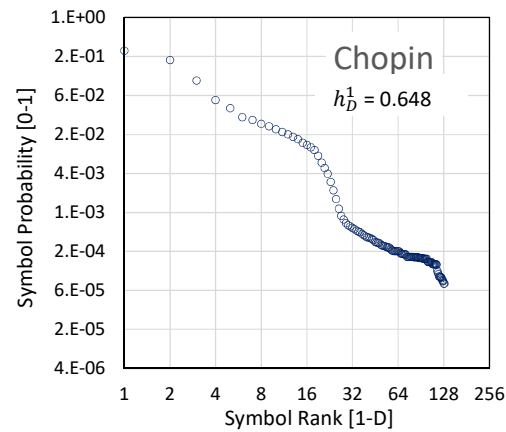
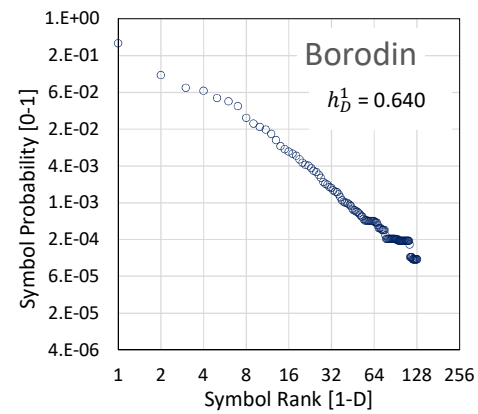
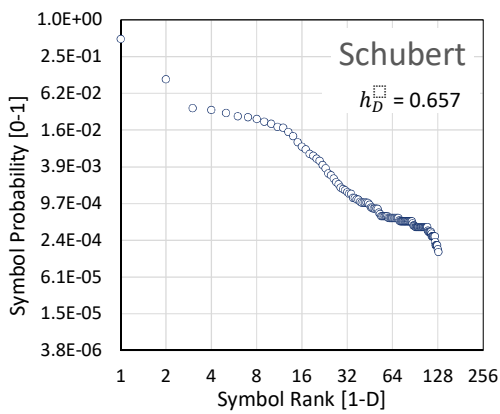
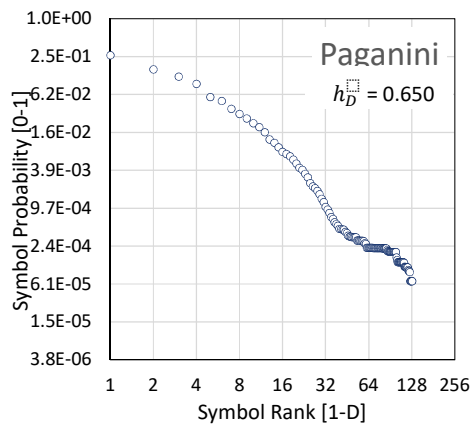
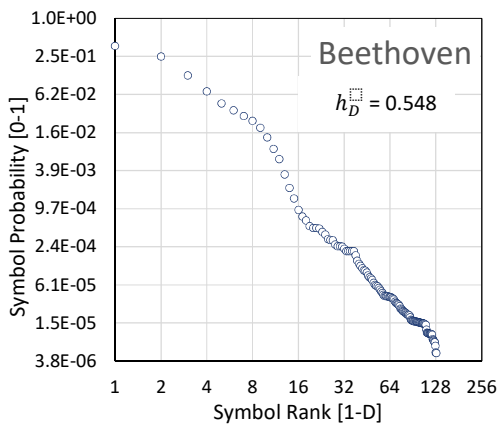
Symbol probability for several types of music. Observation scale = Diversity = 129 Symbols. From r=63 to r=129												
rnk.	Medieval	Renais.	Baroque	Classic	Romant.	Impres.	20th Cty.	Chinese	Raga	V.Thems.	Rock	Venez.
63	2.87E-04	3.16E-04	2.80E-04	2.35E-04	2.94E-04	1.78E-04	2.21E-04	3.99E-04	7.52E-04	5.43E-04	4.16E-04	1.36E-04
64	2.87E-04	3.05E-04	2.71E-04	2.31E-04	2.85E-04	1.73E-04	2.15E-04	3.73E-04	7.46E-04	5.24E-04	4.13E-04	1.29E-04
65	2.86E-04	2.98E-04	2.62E-04	2.24E-04	2.75E-04	1.66E-04	2.09E-04	3.53E-04	7.45E-04	5.19E-04	4.04E-04	1.21E-04
66	2.83E-04	2.96E-04	2.60E-04	2.15E-04	2.67E-04	1.60E-04	2.04E-04	3.39E-04	7.44E-04	5.15E-04	3.90E-04	1.13E-04
67	2.82E-04	2.93E-04	2.55E-04	2.08E-04	2.61E-04	1.56E-04	2.00E-04	3.29E-04	7.23E-04	5.07E-04	3.85E-04	1.12E-04
68	2.81E-04	2.82E-04	2.47E-04	2.03E-04	2.58E-04	1.53E-04	1.93E-04	3.29E-04	7.23E-04	4.98E-04	3.78E-04	1.11E-04
69	2.70E-04	2.74E-04	2.41E-04	1.96E-04	2.54E-04	1.51E-04	1.88E-04	3.22E-04	6.98E-04	4.87E-04	3.70E-04	1.10E-04
70	2.58E-04	2.62E-04	2.35E-04	1.91E-04	2.49E-04	1.49E-04	1.80E-04	3.10E-04	6.97E-04	4.83E-04	3.68E-04	1.08E-04
71	2.43E-04	2.54E-04	2.30E-04	1.88E-04	2.45E-04	1.48E-04	1.70E-04	3.05E-04	6.96E-04	4.82E-04	3.67E-04	1.07E-04
72	2.41E-04	2.53E-04	2.28E-04	1.84E-04	2.39E-04	1.47E-04	1.63E-04	3.02E-04	6.96E-04	4.79E-04	3.64E-04	1.05E-04
73	2.40E-04	2.41E-04	2.20E-04	1.81E-04	2.35E-04	1.45E-04	1.56E-04	3.02E-04	6.95E-04	4.76E-04	3.60E-04	1.03E-04
74	2.34E-04	2.29E-04	2.14E-04	1.76E-04	2.30E-04	1.44E-04	1.51E-04	2.98E-04	6.76E-04	4.73E-04	3.55E-04	1.02E-04
75	2.27E-04	2.23E-04	2.09E-04	1.70E-04	2.24E-04	1.41E-04	1.48E-04	2.89E-04	6.75E-04	4.69E-04	3.46E-04	9.84E-05
76	2.19E-04	2.08E-04	2.06E-04	1.62E-04	2.19E-04	1.40E-04	1.46E-04	2.88E-04	6.62E-04	4.61E-04	3.39E-04	9.51E-05
77	2.10E-04	1.99E-04	2.03E-04	1.58E-04	2.15E-04	1.39E-04	1.46E-04	2.88E-04	6.55E-04	4.50E-04	3.34E-04	9.19E-05
78	1.99E-04	1.95E-04	2.00E-04	1.53E-04	2.12E-04	1.36E-04	1.45E-04	2.85E-04	6.53E-04	4.44E-04	3.32E-04	8.92E-05
79	1.95E-04	1.87E-04	1.98E-04	1.50E-04	2.09E-04	1.31E-04	1.44E-04	2.84E-04	6.34E-04	4.44E-04	3.28E-04	8.35E-05
80	1.91E-04	1.87E-04	1.93E-04	1.46E-04	2.07E-04	1.26E-04	1.43E-04	2.81E-04	6.28E-04	4.36E-04	3.23E-04	7.84E-05
81	1.89E-04	1.85E-04	1.91E-04	1.42E-04	2.04E-04	1.21E-04	1.42E-04	2.80E-04	6.28E-04	4.33E-04	3.21E-04	7.48E-05
82	1.86E-04	1.80E-04	1.85E-04	1.38E-04	2.01E-04	1.16E-04	1.41E-04	2.79E-04	6.28E-04	4.31E-04	3.14E-04	7.09E-05
83	1.80E-04	1.78E-04	1.77E-04	1.35E-04	1.97E-04	1.14E-04	1.40E-04	2.77E-04	6.28E-04	4.27E-04	3.11E-04	6.56E-05
84	1.77E-04	1.77E-04	1.69E-04	1.32E-04	1.93E-04	1.13E-04	1.37E-04	2.66E-04	6.28E-04	4.23E-04	3.09E-04	6.37E-05
85	1.76E-04	1.76E-04	1.64E-04	1.30E-04	1.88E-04	1.08E-04	1.35E-04	2.58E-04	6.28E-04	4.18E-04	3.04E-04	5.98E-05
86	1.74E-04	1.76E-04	1.61E-04	1.27E-04	1.84E-04	1.05E-04	1.33E-04	2.58E-04	6.28E-04	4.13E-04	2.98E-04	5.72E-05
87	1.69E-04	1.70E-04	1.59E-04	1.25E-04	1.80E-04	1.02E-04	1.29E-04	2.55E-04	6.28E-04	4.04E-04	2.97E-04	5.63E-05
88	1.65E-04	1.69E-04	1.56E-04	1.23E-04	1.77E-04	1.01E-04	1.28E-04	2.48E-04	6.28E-04	4.02E-04	2.93E-04	5.56E-05
89	1.63E-04	1.68E-04	1.55E-04	1.20E-04	1.74E-04	9.96E-05	1.26E-04	2.46E-04	6.28E-04	3.96E-04	2.78E-04	5.45E-05
90	1.62E-04	1.66E-04	1.52E-04	1.19E-04	1.72E-04	9.72E-05	1.24E-04	2.39E-04	6.28E-04	3.89E-04	2.66E-04	5.36E-05
91	1.59E-04	1.65E-04	1.52E-04	1.17E-04	1.69E-04	9.48E-05	1.21E-04	2.29E-04	6.28E-04	3.88E-04	2.62E-04	5.28E-05
92	1.59E-04	1.62E-04	1.50E-04	1.14E-04	1.67E-04	9.20E-05	1.11E-04	2.23E-04	6.28E-04	3.87E-04	2.59E-04	5.22E-05
93	1.58E-04	1.62E-04	1.46E-04	1.08E-04	1.62E-04	8.61E-05	1.02E-04	2.22E-04	6.28E-04	3.78E-04	2.57E-04	5.17E-05
94	1.49E-04	1.61E-04	1.42E-04	1.04E-04	1.59E-04	8.10E-05	9.62E-05	2.21E-04	6.28E-04	3.71E-04	2.55E-04	5.06E-05
95	1.49E-04	1.60E-04	1.40E-04	1.02E-04	1.54E-04	7.63E-05	8.45E-05	2.20E-04	6.28E-04	3.64E-04	2.53E-04	4.97E-05
96	1.47E-04	1.60E-04	1.34E-04	9.98E-05	1.50E-04	7.40E-05	7.95E-05	2.18E-04	6.28E-04	3.63E-04	2.51E-04	4.93E-05
97	1.47E-04	1.57E-04	1.29E-04	9.79E-05	1.48E-04	7.25E-05	7.68E-05	2.18E-04	6.28E-04	3.47E-04	2.48E-04	4.88E-05
98	1.47E-04	1.50E-04	1.21E-04	9.55E-05	1.45E-04	7.04E-05	7.57E-05	2.17E-04	6.28E-04	3.91E-04	3.40E-04	4.85E-05
99	1.44E-04	1.48E-04	1.14E-04	9.33E-05	1.41E-04	6.87E-05	7.43E-05	2.17E-04	6.28E-04	3.89E-04	3.34E-04	4.72E-05
100	1.42E-04	1.45E-04	1.04E-04	8.67E-05	1.34E-04	6.61E-05	7.28E-05	2.13E-04	6.28E-04	3.85E-04	3.28E-04	4.59E-05
101	1.42E-04	1.39E-04	1.00E-04	7.93E-05	1.28E-04	6.30E-05	7.15E-05	2.13E-04	6.28E-04	3.81E-04	3.28E-04	4.05E-05
102	1.41E-04	1.31E-04	9.44E-05	7.56E-05	1.25E-04	5.92E-05	6.90E-05	2.08E-04	6.28E-04	3.76E-04	3.18E-04	3.64E-05
103	1.30E-04	1.26E-04	8.71E-05	7.37E-05	1.22E-04	5.60E-05	6.51E-05	1.85E-04	6.28E-04	3.62E-04	3.15E-04	3.44E-05
104	1.26E-04	1.17E-04	8.18E-05	7.23E-05	1.19E-04	5.27E-05	6.11E-05	1.60E-04	6.28E-04	3.55E-04	3.10E-04	3.34E-05
105	1.25E-04	1.10E-04	7.87E-05	7.06E-05	1.16E-04	5.08E-05	5.76E-05	1.56E-04	6.28E-04	3.34E-04	3.05E-04	3.24E-05
106	1.22E-04	1.02E-04	7.67E-05	6.92E-05	1.11E-04	4.89E-05	5.28E-05	1.54E-04	6.28E-04	3.28E-04	3.04E-04	3.15E-05
107	1.20E-04	9.69E-05	7.47E-05	6.85E-05	1.06E-04	4.57E-05	5.16E-05	1.53E-04	6.28E-04	3.14E-04	2.95E-04	3.13E-05
108	1.19E-04	9.34E-05	7.18E-05	6.74E-05	1.01E-04	4.14E-05	5.12E-05	1.51E-04	6.28E-04	3.12E-04	2.94E-04	3.08E-05
109	1.16E-04	9.16E-05	6.78E-05	6.59E-05	9.85E-05	3.85E-05	5.09E-05	1.50E-04	6.28E-04	2.84E-04	2.90E-04	2.84E-05
110	1.14E-04	8.90E-05	6.10E-05	6.52E-05	9.49E-05	3.54E-05	5.01E-05	1.50E-04	6.28E-04	2.82E-04	2.89E-04	2.54E-05
111	1.11E-04	8.67E-05	5.65E-05	6.39E-05	9.16E-05	3.43E-05	4.81E-05	1.49E-04	6.28E-04	2.31E-04	2.88E-04	1.67E-05
112	1.09E-04	8.48E-05	5.22E-05	6.29E-05	8.85E-05	3.39E-05	4.43E-05	1.49E-04	6.28E-04	2.19E-04	2.87E-04	1.64E-05
113	1.06E-04	8.14E-05	4.75E-05	6.22E-05	8.62E-05	3.35E-05	4.25E-05	1.47E-04	6.28E-04	2.19E-04	2.86E-04	1.58E-05
114	1.02E-04	7.70E-05	4.43E-05	6.11E-05	8.49E-05	3.29E-05	4.13E-05	1.47E-04	6.28E-04	2.13E-04	2.85E-04	1.51E-05
115	1.01E-04	7.28E-05	4.21E-05	5.90E-05	8.32E-05	3.24E-05	4.04E-05	1.46E-04	6.28E-04	2.13E-04	2.85E-04	1.47E-05
116	9.85E-05	6.97E-05	4.06E-05	5.32E-05	8.12E-05	3.19E-05	4.00E-05	1.46E-04	6.28E-04	2.13E-04	2.82E-04	1.45E-05
117	9.72E-05	6.90E-05	3.91E-05	4.64E-05	7.88E-05	3.13E-05	3.96E-05	1.44E-04	6.28E-04	2.13E-04	2.72E-04	1.42E-05
118	9.09E-05	6.81E-05	3.82E-05	4.25E-05	7.53E-05	3.07E-05	3.92E-05	1.43E-04	6.28E-04	2.10E-04	2.60E-04	1.37E-05
119	8.41E-05	6.61E-05	3.62E-05	4.12E-05	6.98E-05	3.02E-05	3.87E-05	1.43E-04	6.28E-04	2.10E-04	2.57E-04	1.30E-05
120	8.12E-05	6.49E-05	3.18E-05	4.05E-05	6.55E-05	2.93E-05	3.85E-05	1.42E-04	6.28E-04	2.07E-04	2.51E-04	1.28E-05
121	7.91E-05	6.36E-05	3.00E-05	4.00E-05	6.24E-05	2.84E-05	3.84E-05	1.40E-04	6.28E-04	1.99E-04	2.39E-04	1.26E-05
122	7.80E-05	6.01E-05	2.90E-05	3.93E-05	6.03E-05	2.77E-05	3.80E-05	1.39E-04	6.28E-04	1.98E-04	2.24E-04	1.23E-05
123	7.28E-05	5.95E-05	2.82E-05	3.84E-05	5.81E-05	2.71E-05	3.76E-05	1.30E-04	6.28E-04	1.94E-04	2.09E-04	1.17E-05
124	6.71E-05	5.88E-05	2.75E-05	3.77E-05	5.33E-05	2.69E-05	3.42E-05	1.11E-04	6.28E-04	1.94E-04	1.96E-04	1.06E-05
125	6.19E-05	5.79E-05	2.61E-05	3.68E-05	4.43E-05	2.67E-05	2.49E-05	1.07E-04	6.28E-04	1.14E-04	1.77E-04	9.94E-06
126	5.86E-05	5.71E-05	2.15E-05	3.50E-05	4.12E-05	2.63E-05	2.12E-05	8.77E-05	6.28E-04	1.10E-04	1.60E-04	9.19E-06
127	5.30E-05	5.17E-05	1.90E-05	3.25E-05	3.90E-05	2.39E-05	2.00E-05	7.54E-05	6.28E-04	1.06E-04	1.51E-04	8.17E-06
128	4.81E-05	3.54E-05	1.52E-05	2.26E-05	3.52E-05	1.60E-05	1.94E-05	7.36E-05	6.28E-04	1.05E-04	1.44E-04	7.15E-06
129	3.75E-05	3.08E-05	1.25E-05	1.95E-05	2.53E-05	1.36E-05	1.90E-05	7.15E-05	6.28E-04	9.99E-05	1.37E-04	5.91E-06

Appendix I

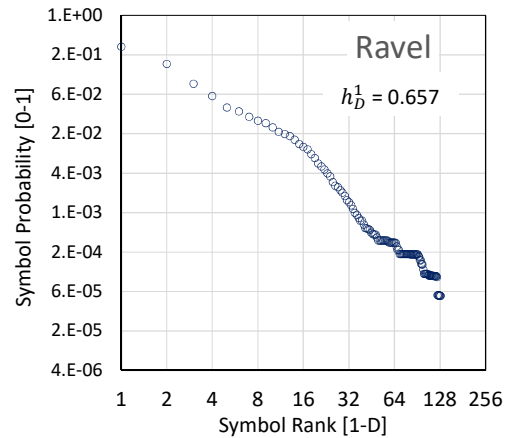
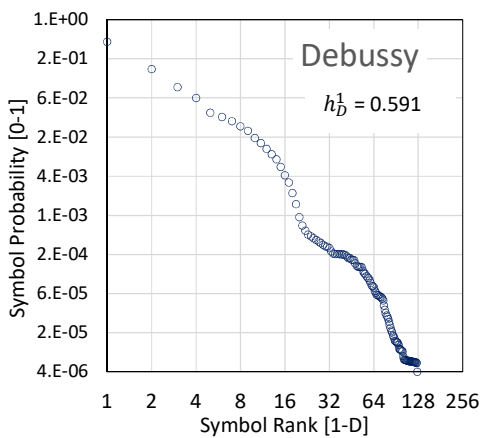
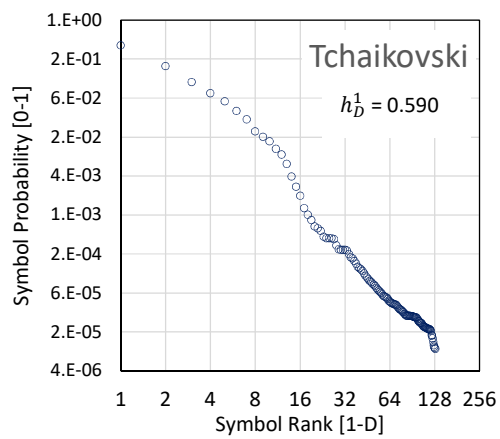
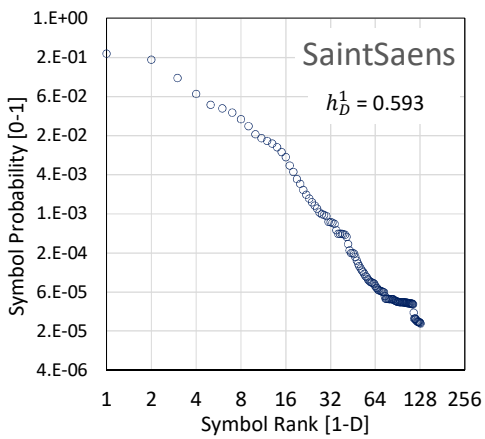
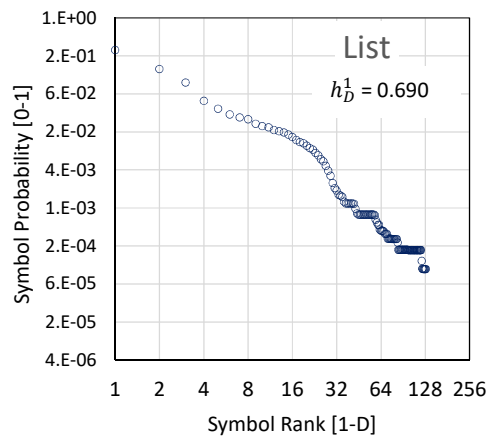
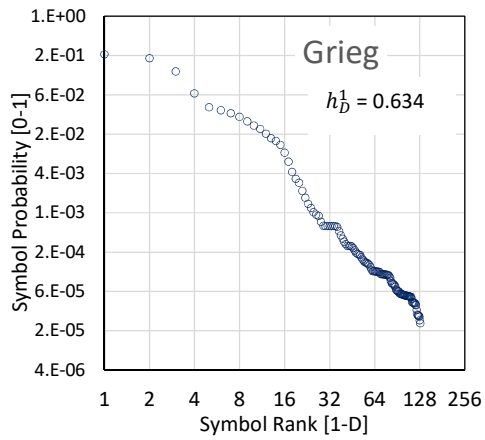
Symbol probability profiles of the music by composer



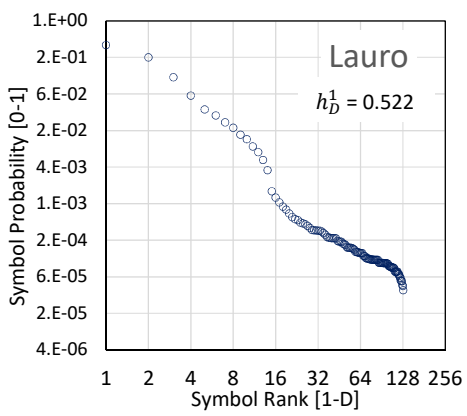
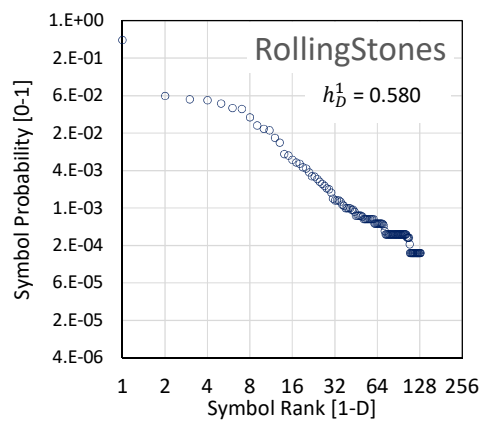
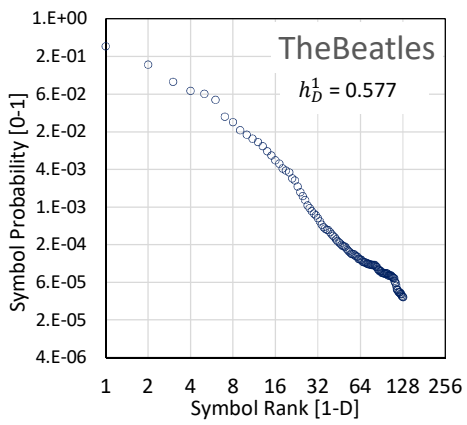
I. Symbol probability profiles of the music by composer



I. Symbol probability profiles of the music by composer



I. Symbol probability profiles of the music by composer



Appendix J

Music styles by composer, in the space (*specific diversity, entropy, 2nd order entropy*), ($\mathbf{d}, \mathbf{h}_D, \mathbf{h}_{D^{[2]}}$).

J. Music styles by composer, in the space spec. diversity, entropy, 2nd order entropy

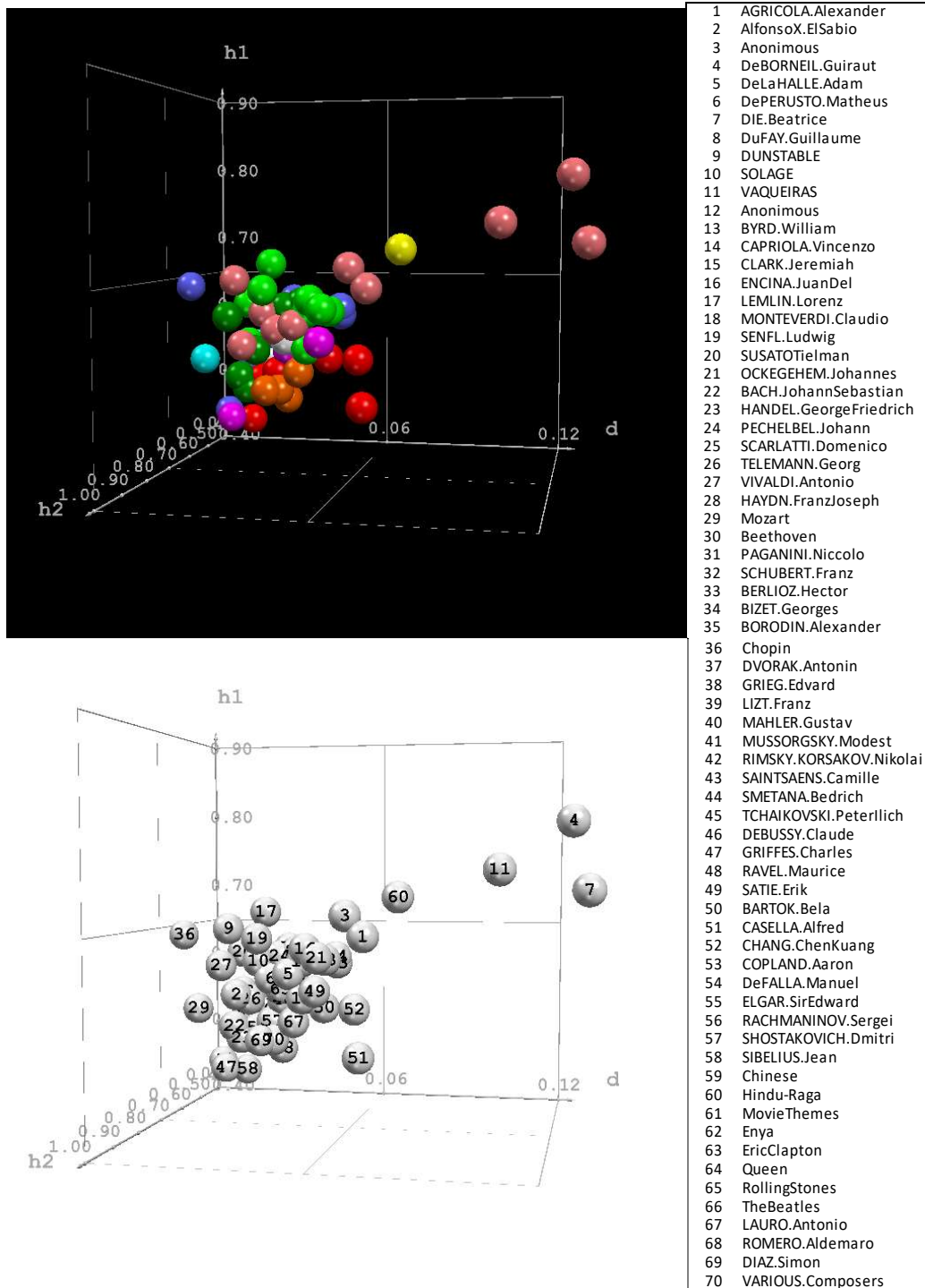


Figure J.1: A view of music style by composers represented in the space: $d, h, h^{[2]}$

J. Music styles by composer in the space (spec. diversity, entropy, 2nd order entropy)

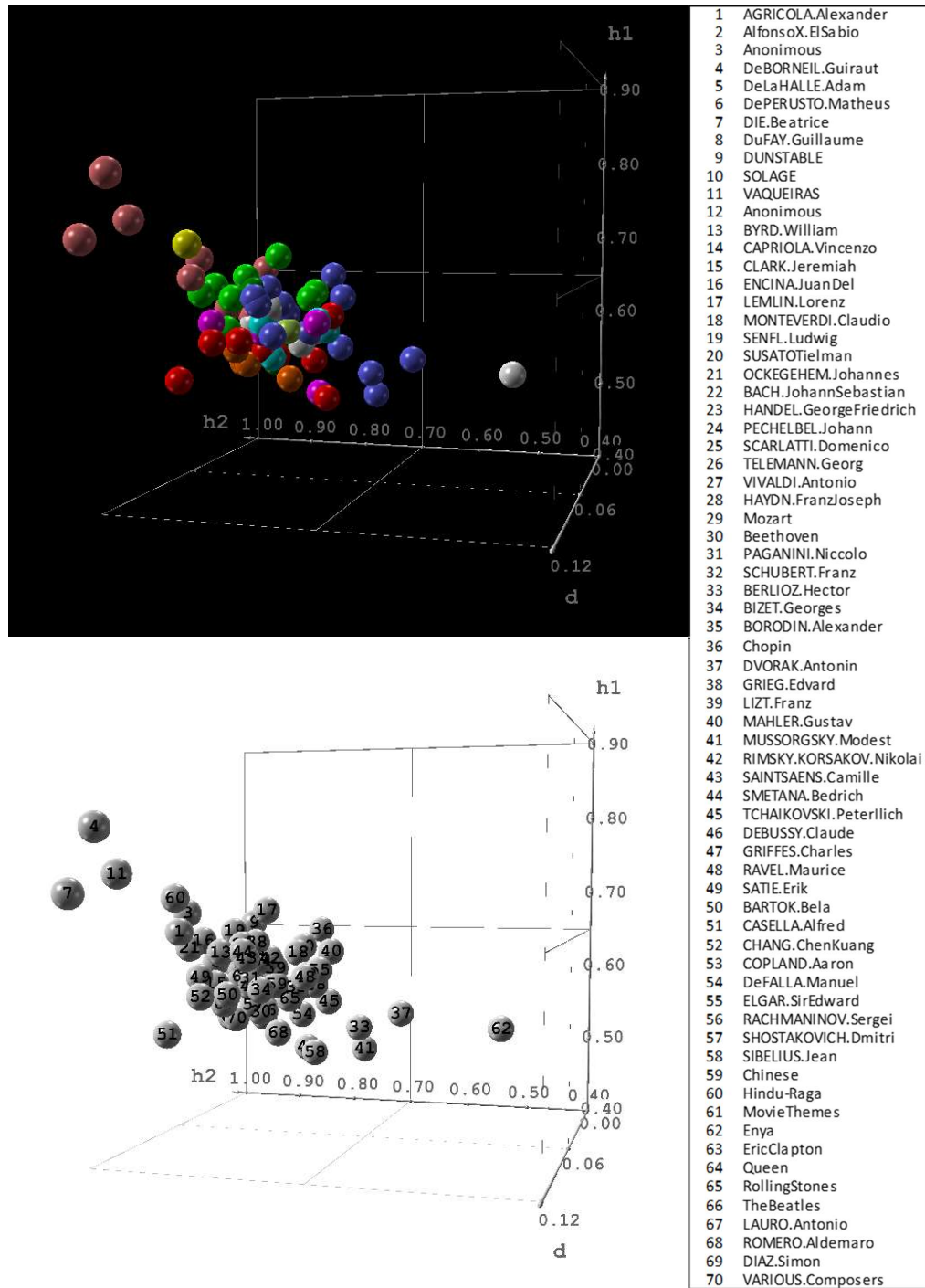


Figure J.2: A view of music style by composers represented in the space: $d, h, h^{[2]}$

J. Music styles by composer, in the space spec. diversity, entropy, 2nd order entropy

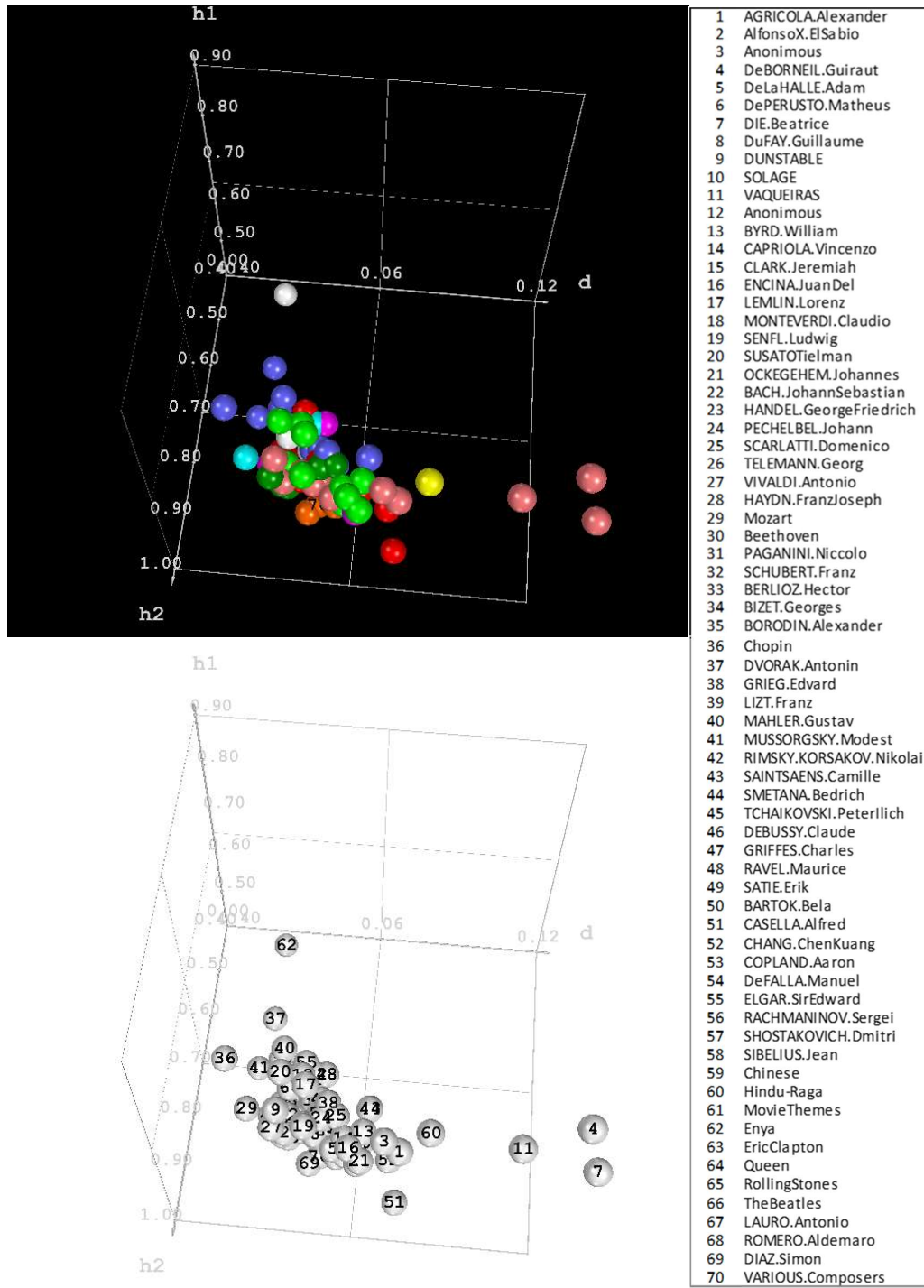


Figure J.3: A view of music style by composers represented in the space: $d, h, h^{[2]}$

