

**On Some Quantitative Aspects
of the Componential Structure
of Chinese Characters**

Norbert Kordek

**On Some Quantitative Aspects
of the Componential Structure
of Chinese Characters**



Poznań 2013

Recenzent: prof. dr hab. Grażyna Demenko

Projekt okładki: Helena Oszmiańska

Copyright by:
Norbert Kordek and Wydawnictwo Rys

Wydanie I, Poznań 2013

ISBN 978-83-63664-13-8

Wydanie:



Wydawnictwo Rys
ul. Różana 9/10
61-577 Poznań
tel./fax 048 61 833 16 03
kom. 0600 44 55 80
e-mail: rysstudio@o2.pl
www.wydawnictworys.com

Contents

Acknowledgements.....	11
Preface	13
1. Preliminary considerations	17
1.1. On the nature of Chinese script.....	17
1.1.1. Terminology and scope.....	17
1.1.2. Typology and relation to speech.....	19
1.1.3. Traditional characterology	20
1.2. Six categories (六書 <i>liùshū</i>).....	22
1.3. Modern characterology (漢字學 <i>hànzixué</i>).....	26
1.4. Number of Chinese characters.....	27
2. Chinese character sets.....	31
2.1. Noncoded Character Sets (NCSes)	32
2.1.1. Chinese noncoded character sets	32
2.1.1.1. China.....	32
2.1.1.2. Taiwan	33
2.1.2. Non-Chinese noncoded character sets.....	34
2.2. Coded Characters Sets (CCSes)	35
2.2.1. Chinese coded character sets.....	35
2.2.1.1. China.....	35
2.2.1.2. Taiwan	37
2.2.2. Non-Chinese and international CCSes	39
2.2.2.1. Unicode – Unihan.....	39
3. Theoretical preliminaries	42
3.1. Segmentotactology and segmentotactics	43
3.1.1. Introduction	43
3.1.2. Prerequisites for segmentotactic investigations.....	45
3.2. Phonotactics	45
3.2.1. Terminology.....	46
3.2.1.1. Tactophoneme	47

3.3. Beyond phonotactics.....	48
3.3.1. Orthotactics	49
3.3.2. Phonemotactics, syllabotactics and morphotactics	51
3.3.3. Graphotactics	52
3.3.3.1. Exemplary analysis.....	54
3.3.2.2. Levels of analysis.....	57
4. Structure of Chinese characters	59
4.1. Terminology	59
4.2. Hierarchy of constituents.....	61
4.3. Composition.....	61
4.4. Decomposition and component types.....	62
4.4.1. The decomposition	62
4.4.1.1. Decomposition rules.....	63
4.4.1.2. Decomposition structure	66
4.4.1.3. Functional categories of components.....	72
4.4.2. Component lists.....	74
4.4.3. 說文解字 <i>Shuōwén Jiězì</i> (SWJZ) and the modern components	78
4.4.4. Functional types of components	78
4.4.4.1. Modern ‘six categories’	79
4.5. Strokes	80
4.6. Simplification of characters	86
4.6.1. Extent of simplification	87
4.6.2. Simplification methods	88
4.6.3. Simplification of components.....	89
5. Models of Chinese character descriptions.....	91
5.1. Character description language (CDL) projects.....	91
5.1.1. Ideographic description sequence (IDS).....	91
5.1.1.1. Character Information Service Environment (CHISE)	93
5.1.1.2. Kawabata’s Kanji Database Project (KDP).....	95
5.1.2. Chinese Documents Processing Lab (CDP).....	99
5.1.3. Wenlin CDL	111

5.1.4. Summary	113
5.1.5. Other projects.....	113
5.1.5.1. Hanglyph CDL.....	114
5.1.5.2. Cjklib.....	115
5.1.5.3. Wikimedia Commons Chinese Characters Decomposition Project (CCDP)	115
5.2. Grammars	118
5.2.1. Distributional model I.....	118
5.2.2. Generative model	118
5.2.3. Distributional model II.....	120
5.3. Graphotactics related studies	125
5.3.1. Component combination database	125
5.3.2. Chinese orthography database.....	126
5.4. Psycholinguistics related studies	128
5.5. Mathematical models.....	128
6. Quantitative studies of Chinese script.....	131
6.1. Traditional statistical and quantitative studies	131
6.1.1. Frequency of characters	131
6.1.2. Frequency of components	134
6.1.3. Componential complexity of characters	140
6.1.4. Stroke statistics.....	142
6.1.4.1. Stroke number statistics	143
6.1.4.2. Stroke count.....	143
6.1.4.3. Stroke count and characters frequency.....	150
6.1.5. Quantitative properties of syllable-to-character mapping	152
6.2. Quantitative linguistic laws	154
6.2.1. Zipf's Law.....	154
6.2.2. Menzerath-Altmann Law.....	155
6.3. Script complexity.....	158
6.3.1. Compositional method.....	158
6.3.2. Intersectional method.....	159

6.3.3. The complexity of Chinese script	161
6.3.3.1. The compositional method and the Chinese script	162
6.3.3.2. The intersectional method and the Chinese script	165
6.3.3.3. The methods compared.....	168
6.3.3.4. Complexity reduction.....	175
6.3.4. Summary	178
7. Graphotactic analysis of Chinese script	179
7.1. Graphotactics of Cangjie input method	179
7.1.1. Introduction to the Cangjie input method (CIM)	179
7.1.2. The graphotactic analysis of Cangjie codes	187
7.1.2.1. CIM tactographemicity and t-graphotactemicity	189
7.1.2.2. T-efficiency	191
7.1.2.3. Tactographemic t-efficiency.....	192
7.1.2.4. CIM categorial graphotactemic efficiency.....	193
7.1.2.5. CIM graphemes dispersion	194
7.2. Graphotactics of Chinese script.....	200
7.2.1. Levels of analysis.....	201
7.2.2. The database	201
7.2.3. Big5 (CNS 11643 Plane 1 and 2).....	204
7.2.3.1. Big5 – immediate components	206
7.2.3.2. Big5 basic components.....	211
7.2.4. Comparative analysis.....	216
7.2.4.1 Immediate components	217
7.2.4.2. Basic components.....	221
7.2.5. Unihan	226
7.2.5.1. Immediate components	227
7.2.5.2. Unihan basic components	231
7.2.6. Summary	235
7.2.6.1. Tactographemicity and t-graphotactemicity.....	235
7.2.6.2. T-efficiency	241
7.2.6.3. Tactographemic t-efficiency.....	242

7.2.6.4. Categorical graphotactemic efficiency	243
7.2.7. Graphemic dispersion	244
7.2.7.1. Immediate components	244
7.2.7.2. Basic components.....	248
7.2.7.3. Summary	253
7.2.8. Complexity of graphotactemes in terms of graphemes.....	254
7.2.8.1. Immediate components	255
7.2.8.2. Basic components.....	257
7.2.8.3. Summary	259
7.2.9. Summary and concluding remarks	260
7.2.10. Perspectives	262
References.....	265
Appendix I – Chinese Documents Processing Lab (CDP) basic components list – ordered by frequency	281
Appendix II – Big5 character set.....	289
Appendix III – Unihan inventory of basic components	301