Norbert Kordek

On Some Quantitative Aspects of the Componential Structure of Chinese Characters

Norbert Kordek

On Some Quantitative Aspects of the Componential Structure of Chinese Characters

rys
wydawnictwo

Poznań 2013

Recenzent: prof. dr hab. Grażyna Demenko

Projekt okładki: Helena Oszmiańska

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

In this initial part of the book I would like to address some of the aspects of the writing process that cannot be quantified. By those aspects I mean the help I received from other people and my gratitude to them.

The person I am most indebted to is Professor Jerzy Bańczerowski, who has been inspiring me from the very moment I entered university – and that was a long time ago. I hope that this study, conceived on his ideas, is not going to disappoint him.

I owe special thanks to Piotr Woźnica, who wrote the Perl scripts, for patiently enduring my personality, constant demands, changes of mind, redesigns of algorithms, and generally unefficient handling of his time.

I would like to extend my gratitude to a number of colleagues from various institutions for their generosity and openness with sharing their resources and knowledge – particularly Jagoda Bruni from the University of Stuttgart, Richard Cook from the Wenlin Institute, Han Buxin from the Institute of Psychology, Chinese Academy of Sciences, William Strnad and Paweł Nowakowski of Adam Mickiewicz University.

I also would like to thank Szymon Grzelak for his continuing friendly support in various aspects of work on this volume, including solving technical problems and translation from Japanese.

Finally, but importantly, my family deserves my deepest gratitude for their patience, support and encouragement.

Preface

Despite the truth in F. Coulmas' words, who called writing "*the single most consequential technology ever invented*",[1] I did not intend to write a book exclusively on Chinese script. The initial concept was, in some ways, more ambitious than the final contents. It involved the quantitative analysis of all systems and levels of Mandarin Chinese that fit the type of analysis inspired by Bańczerowski's phonotactic theory,[2] including phonetics, phonology (both phonemic and syllabic systems), morphology and the script. I started with the chapter on Chinese script and it turned out to be a book. The complexity of Chinese script, the variety of research perspectives it offers, even in the very specific graphotactic framework, the number of issues that needed to be addressed and solved, and also the latent flexibility of Bańczerowski's theory[3] were all contributing factors resulting in a complete change of the concept. It is also not a book written entirely from the perspective of graphotactics – some other quantitative aspects of the Chinese witing system proved too attractive not to give them considerable attention, such as the problem of measuring the complexity of Chinese characters. As a result, as the title suggests, the graphotactics was presented as one of the possible quantitative approaches to the research of Chinese script. Nevertheless, the contents of the book should leave no doubt that graphotactics was intended to be the focus, and therefore, the subject which was explored the most thoroughly.

This book is not intended to be a theoretical study. Given the abundance of linguistic material and the space limitations, it seemed more reasonable to focus on designing an analysis of the corpus of characters and on presentation of the results. The resulting theoretical conseqences are, of course, discussed in respective sections of the book. This implies a non-theoretical character of this study, by which I mean it does not offer a more robust theoretical framework than Bańczerowski's original theory. All I did in the theoretical aspect of the study was to expand the very idea of this type of analysis beyond phonetics and phonology, and adjust the terminology to fit the requirements of the graphotactics, which included supplementing the framework with additional terms. The graphotactic theory is not presented in the axiomatic form, as originally intended – there still remains many unanalyzed aspects of Chinese graphotactics that should first be touched upon at the very least, before any theoretical generalizations are drawn up, especially in the axiomatic form.

Some more general theoretical issues pertaining to Chinese script were limited to the necessary minimum – by that I mean primarily to the general theory of writing

[1] Coulmas 2003: 1.

[2] See Chapter 3, the project was outlined in Kordek (2012).

[3] The proposal resulted in the analysis of the structure of Polish words in terms of letters of the alphabet by prof. P. Wierzchoń.

systems and semiotics of the Chinese script, the nature of Chinese characters with respect to speech, history and evolution, and also to traditional and etymological perspectives on the structure of Chinese characters. Those issues were extensively discussed elsewhere and I did not feel compelled to dwell on subjects to which I have nothing new to add, and that are not crucial to the problems on which the the study is focused. In this context it may seem that too much attention was devoted to the discussion of information processing related issues, but the simple fact is that contemporary research efforts concerning Chinese characters are focused on this very area. Surprisingly, or perhaps not, in recent years the most concrete and valuable contributions to the understanding of modern Chinese script were rendered not in the purely linguistic context, but by the information processing and computer technology related approaches. The evidence is obvious – the standards for Chinese character processing, standards of character components, ordering and typology of strokes, and also the databases of decomposed characters, the computer software, etc. It is not an exaggeration to say that the advances in information processing technology and the related research made the present study possible to carry out. Without progress in the collection, standardization, and annotation of East Asian scripts for the purpose of informaton processing and exchange, this project would have remain too daunting, if not impossible, to complete. It should be noted that those achievements usually are underemphasized in linguistic works. Technological progress in recent years has been staggering, but due to its extent and complexity, it has proven to be very confusing for the layman (or simple linguists, like myself). This is one of the reasons I devoted a substantial part of the book to the issues of Chinese character processing. This should not be misunderstood – this is still a linguistic work, but instead of repeating what can be found in almost every book on Chinese characters, I concentrate on investigating the potential sources of data best suited for graphotactic analysis. At the beginning, the graphotactic analysis promised fascinating results, if conducted properly, but its feasibility was not much more than a hopeful gamble. Given the complexity and novelty of the project, it is not surprising that there is only a limited body of research to draw upon, and the success also depended heavily on access to proper data. Both of those aspects happened to be greatly facilitated by all kind of by-products of collective efforts to make Chinese script a viable medium of communication in the information age.

There is at least one aspect that should have received more attention – the general introduction to quantitative linguistics. Due to the space restrictions, the discussion is limited to the presentation of concrete results pertaining to Chinese script.

The quantitative methods are still underutilized in the practice of linguistic research. They are underrepresented not only in comparison to the mainstream research output, but even more so in the domain of writing systems studies. Chinese script is relatively well studied, but despite the great importance of the componential structure of Chinese characters, relatively little is known from the quantitative perspective about the ways in which components combine. Those two facts plus the inspiring force of Bańczerowski's phonotactological theory are the reasons for writing this book. Although the analysis is couched in a specific theoretical framework, whenever it is possible theory-neutral language is used for the description of discussed problems. In the present volume the graphotactological theory was not developed to the extent that would allow a complete immersion of the discussion in theory-specific language and terminology. This in fact, as I mentioned before, is a result of a premeditated decision to focus on raw data analysis and interpretation of the results, leaving the formulation of a more robust theoretical concept of Chinese graphotactology, and the graphotactology in general, for the next step.

Apart from some questions of interest that I intentionally approached in an overtly simplistic manner, there are likely to be problems that I ignorantly sidestepped to address the issues on which I wished to focus. In cases like that criticism is more than welcomed.

The material is organized into 7 chapters, that now will be briefly introduced. **Chapter 1** presents background information on the traditional study of Chinese characters and most general issues pertaining to the subject of this book, including a short introduction of modern Chinese characterology, and a traditional approach to the componential structure of characters. The chapter is aimed at readers less familiar with problems related to Chinese script, including the last section in which I discuss the problem of the number of characters. This allows the most general quantitative aspect of the Chinese writing system to be put in perspective. **Chapter 2** is dedicated to the problem of research data that I discuss by presenting the available Chinese character sets. **Chapter 3** contains the theoretical preliminaries of the main research in this book. The discussed terminologies and concepts**,** in general**,** do not go beyond issues pertaining directly to the main research section in Chapter 7, but some of the directly unrelated problems are presented in more details, e.g. the extension of Bańczerowski's idea beyond the phonotactics. **Chapter 4** is focused on the modern approach to the structure of Chinese characters. The discussion includes the terminological issues, the types and levels of decomposition, the decomposition rules, the constituent types and component sets. The simplification of characters is also presented. **Chapter 5** introduces the most relevant models of Chinese character description from the perspective of Chinese graphotactics. The models are discussed with respect to their usefulness for graphotactic analysis, structure (with the focus on language of descriptive expressions), type and form of provided data, and also their purpose. The second part of the chapter discusses the grammatical models of Chinese script and research endeavors related to the main topic of this book. **Chapter 6** is a detailed presentation of different kinds of quantitative approaches to various aspects of Chinese script. It addresses the most common types of statistical studies involving character sets, as well as component and stroke levels of script. In the second part of the chapter I present the investigative results on the validity of two quantitative laws for Chinese script – the Zipf law and the Menzerath-Altmann hypothesis. I also attempt to examine in a more detailed way the validity of two general methods of measuring the complexity of script proposed by Altmann (2004) and Peust (2006). **Chapter 7** is a detailed presentation of the results of graphotactic analysis of Chinese script. In order to provide reasonably diversified results I examined two fundamentally different structural representations – the structural array of Chinese characters in the Cangjie Input Method and the traditional componential approach. Of the two analyses the former is intended as an auxiliary in approximating the Chinese writing system to alphabetic systems, whereas the latter is the proper graphotactic analysis. Additionally, for the sake of diversification and completeness, different character sets are analyzed – BIG5 as the largest homogenous set, two comparably large sets of traditional and simplified characters for comparative purposes, and the Unihan database (CJK Unified Characters) as the largest available character set. The analysis of different sets provided a large amount of interesting results – they are presented and interpreted as thoroughly as possible at this stage of Chinese graphotactical investigation.

I hope this book, beside demonstrating the validity of this type of research, will reveal at least some of the intricacy and internal logic of Chinese script and become one more piece of evidence testifying to the uniqueness of Chinese characters.

1. Preliminary considerations

The aim of the present book is to implement a portion of Bańczerowski's modified phonotactic theory to the analysis of Chinese script. The issues pertinent to the analysis presented in Chapter 7 will be discussed in the successive five chapters. Chapter 1 is intended as a brief introduction to the basic properties of Chinese script and as a presentation of the most general problems concerning the Chinese writing system. The scale of the planned analysis drastically restricts the introductory considerations, but a minimum of information sufficient to prepare even the readers unfamiliar with Chinese script is maintained.

1.1. On the nature of Chinese script

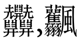
One of the most fascinating aspects of Chinese language is undoubtedly the script. This simple statement might raise a series of objections, especially from the struggling learners of Chinese or from fellow linguists who find different aspects of Chinese language more appealing. This is only to say that due to its unique features which include the traceability of development, structural and compositional complexity, an open-ended number of elements (characters), cultural load and artistic value, Chinese script offers research opportunities and presents challenges like no other contemporary writing system. This had been recognized by Chinese scholars in ancient times, which caused Western intellectuals to take an interest in Chinese script as early as during the first missionary contacts.[4] Chinese characters (漢字 *hànzi*) have been analyzed with different levels of intensity for at least a few centuries from the point of view of formation, etymology and evolution, structure, statistics, language planning policy and socio-cultural impact**.** The result is that Chinese script is probably the most extensively studied writing system that has ever existed and still remains a source of both popular interest and academic endeavours.

1.1.1. Terminology and scope

The term 'Chinese characters' (漢字 *hànzi*) referring to a research object is not as straightforward and unambiguous as it may seem. In its widest possible meaning it includes all forms of *hànzi* at all stages of the evolution of Chinese script.[5] The oracle-

[4] For more details see for example DeFrancis 1984b, Su 2002b, Unger 1990 & 2004, Boltz 1994.

[5] In this study the evolution of Chinese script is of marginal importance, it will be addressed briefly in some sections of the book.

bone depiction of a horse:‘ ’ and the equivalents – the small seal script: ‘ ’ and its modern regular script: ‘馬’ are all Chinese characters. Needless to say, both are different and ‘Chinese characters’ may refer to both or just one of them. The same can be said of: 𠔻, ,,𪚥 and 的, 一, 是, 不 – the difference being not the stunning complexity (43, 46, 47 and 64 strokes in the first group respectively) versus relative simplicity, but the fact that the first four are completely obscure, abandoned forms that no one uses. The first group is known only to a negligible number of the keenest charactereologists, and can be found only in the biggest existing dictionaries, while the second group are the four most frequently used characters in contemporary texts. There is another point in need of clarification. There is a problem with the ‘Chinese’ in ‘Chinese characters’. Beside *hànzi* in the parameters of interest there are Japanese *kanji* (and Korean *hanja* and Vietnamese *hán tự* for that matter). What about Japanese 国字 *kokuji*, Korean 韓國式漢字 *hanguksik hanja* and Vietnamese 𡨸喃 *chữ Nôm*?[6] There is yet another direction of inquiry – with regard to simplified (簡化字 *jiǎnhuàzì* or 簡體字 *jiǎntǐzì*) and traditional (正體字 *zhèngtǐzì* or 繁體字 *fántǐzì*) forms: is one interested in both or just one of them?[7] Finally, the treatment of variant forms of characters (異體字 *yìtǐzì*)[8] should also be addressed.

The historical change of form of Chinese script not only reflects the evolution of the writing system itself, but the changes in character use, their phonetic value and meaning, while mirroring the evolution of Chinese spoken language. This study is not concerned with the evolution of forms and historic types of script, although the etymology plays an important role in explaining the structure of modern characters, which in many instances cannot be ignored. The phonetic and semantic shifts together with the distinction between modern and ancient characters, and between literary and vernacular styles of written Chinese form a complex interplay between the characters:[9]

- use of a character is limited to either classical Chinese[10] (e.g. 曰 *yuē*, 玎 *dīng*, 駎 *zhòu*) or to modern Chinese (e.g. 氮 *dàn*, 甩 *shuǎi*, 粁 *qiān*);

[6] The term *chữ Nôm* denotes characters coined within the system of sinograms for the transcription of Vietnamese using the components of Chinese characters. In other words they are characters that are not borrowed from Chinese, despite looking like 漢字 *hànzi*, but are idiosyncratic to Vietnamese, a non-Chinese language.

[7] The choice between traditional and simplified forms used in the text of any book is a commitment, but one that comes naturally and is not a commitment based on the type of study. For reasons that do not even need to be explained the Chinese terms are glossed in traditional characters.

[8] For example: 夠 - 够 - 彀 (*gòu*). The problem of variant forms is too complex to even begin to discuss here, and will be explained in further sections of the book.

[9] Su 2001: 21-23. Some of the examples are borrowed from this work.

[10] The ambiguity and subtleties of the term will not be exploited here, as it is understood synonymously with the Chinese terms 古文 *gǔwén* ‘classical Chinese / ancient Chinese (script)’ and 文言文 *wényánwén* ‘classical literary Chinese’.

- a character is used both in classical and modern Chinese in a uniform manner (e.g. 馬 *mǎ*, 頭 *tóu*, 意 *yì*);
- a character is used both in classical and modern Chinese, but in each case differs in:
  - meaning (e.g. 謝 *xiè* – classical 'apologize' vs. modern 'thank', 脚 *jiǎo* – classical 'calf' vs. modern 'foot')
  - meaning and pronunciation (e.g. 听– classical *yǐn* 'smile' vs. modern *tīng* 'listen'.[11]

The seemingly straightforward term 'Chinese characters' turns out to be rather complex and the above questions are far from trivial. This study is ambitiously aimed at the totality of Chinese characters. As it can aleady be seen the term 'Chinese character' is used to designate the basic unit of Chinese script. The Chinese term '*hànzi*'(漢字) or the abbreviated 'character' will be used synonymously throughout the book. From the perspective of the problems outlined above, the term may refer to different subsets of the 'total' set. All of those synonymous terms will be used either in clear contexts that provide enough background for a proper understanding or the meaning will be provided explicitly. In practice the meaning will be dependent on the discussed or analyzed character set[12]. The referential range is restricted only by the exclusion of the ancient form of script, meaning that whatever is written in this book about Chinese characters, concerns only the modern forms[13] (with the exception of historical references).

1.1.2. Typology and relation to speech

In the Western theory of writing systems the Chinese script has always been a subject of debate on a general level. In theories that strive to create terms for the basic units of script so that the terms themselve indicate its characteristic features the Chinese case is complicated. It is generally agreed to classify Chinese script as logographic and the units of script as logographs, but, as for example DeFrancis (1984b) points out, for Chinese script the terminology constructed on this principle will always be questionable.[14] The terminology is not necessarily wrong, but it is never absolutely precise. The very often discussed topics in Western literature, including the characteristic features and typology of Chinese script and the relation of *hànzi* to the units of speech,

[11] The modern 听 is a simplified form of traditional 聽. The traditional form does not display this kind of phonetic and semantic duality. The disparity in this respect between traditional and simplified forms is not uncommon.

[12] See Chapter 2.

[13] See Section 1.2.

[14] DeFrancis 1984b: 71-73.

will be intentionally treated here as a secondary topic for the discussion and left out almost completely. The question of relation to the units of speech may bear some significance for statistical studies. Readers interested in the topic may refer to the vast literature discussing it from different angles in some selected notable works: Hill (1967), Künstler (1970), Trager (1974), Brice (1976), French (1976), Haas (1976a, 1976b and 1983), Baron (1981), Wang (1983), DeFrancis (1984b), Catach (1986), Stalph (1989), Mattingly (1992), Sampson (1985 and 1994), Boltz (1994), Yin & Rohsenow (1994), Nöth (1995), Chang (1996), Coulmas (1989 and 2003), Unger (1991 and 2004), Hannas (1997, 2003 and 2005) Rogers (1995 and 2005), Hyman (2006), Zhao & Baldauf (2008), Robinson (2009), and Han (2012).

In comparison, it is interesting to note how little discussion in Western literature is devoted to the terminology concerning the subunits of Chinese script,[15] with the term 'radical' typically being used indiscriminately. Fortunately, Chinese literature remedies this deficiency. The literature in Chinese is omitted from the above list simply because Chinese authors usually are not concerned with the problems of typology of writing systems or speech representation. In cases when these problems receive some attention they are typically treated as introductory and secondary concerns discussed in the West. Chinese authors (writing in Chinese) do not discuss the terms that should be used to designate the basic unit of Chinese script. Why should it occur to any Chinese characterologist to use any different term than 漢字 (or 汉字) *hànzi*? In this sense this book is written from the Chinese perspective. The general Chinese literature on *hànzi* is focused on the evolution of forms, etymology, structure and language planning. Some of the notable Chinese studies contributing to modern characterology and to the understanding of character structure are referred to throughout this book, particularly in Chapter 4.

1.1.3. Traditional characterology

Chinese characterology recognizes different domains and methodological approaches. This study is not going to provide a complete overview of them, but despite having a very specific and different goal, it is impossible to ignore some seemingly unrelated aspects of Chinese characterology. It is for that reason a brief overview of the history of 漢字 *hànzi* study is a necessary backdrop.

[15] There are of course exceptions, like Zhao & Baldauf (2008), Stalph (1989).

Traditional characterology (小學 *xiǎoxué*)[16] dates back to the Spring and Autumn Period in Chinese history (starting 770 BC) and prevailed until the last years of the Qing Dynasty. A detailed introduction here would be redundant, so suffice to say that due to social and political reasons the common feature of the traditional inquiries is their ancillary function in relation to 經學 *jīngxué*. The study of character pronunciation, form and meaning were subordinate to the study of the Confucian classics, which in turn constituted the subjects of imperial exams – the only way for promotion in the social hierarchy. As a result this pragmatic aspect of traditional characterology dominated and heavily suppressed inquiries into the very nature and structure of characters. The theoretical value of 小學 *xiǎoxué* does not go beyond the heritage of 說文解字 *Shuōwén Jiězì* (*SWJZ*). This may seem as an oversimplification, but it is to emphasize the theoretical value of Xǔ Shèn's work. It is true that during the Tang and Sui Dynasties a new academic subject 'Character Shape Study' was developed.[17] It is also a fact that scholarly efforts in every major dynasty had a significant contribution to sustaining the continuity and unity of Chinese script, as well as in standardization of character forms, meaning and pronunciation. However, not much was offered as far as the understanding of character structure or semiotic properties.[18] The most important work from the Tang Dynasty – 干祿字書 *Gānlù zìshū*[19] – was an orthographic dictionary of acceptable character forms for the purpose of taking the imperial exams where success or failure could depend on the miswriting of a single character. The 'Character Shape Study' was then the normative orthography for imperial officials. The Song Dynasty's 類篇 *lèipiān* and the Ming Dynasty's 康熙字典 *Kāngxī Zìdiǎn* were also dictionaries. Over the centuries every single important work on Chinese characters in dynastic China was either a dictionary or some sort of a collection of characters. The ancient studies of characters had their own logic, purpose and importance, different from modern characterology. For the purpose of this study this simplified picture of the traditional stage is sufficient and partially compensated by the fact that the importance and influence of the above mentioned 說文解字 *Shuōwén Jiězì* is given considerable attention in further sections.

漢字學 *hànzixué* 'Chinese characterology' is a modern term coined in China as a recognition of a new fully-fledged discipline that started to ripen at the advent of the 20th century. The impulse to shed the limitations of 小學 *xiǎoxué* was the discovery of oracle bones script and the collapse of traditional Chinese social and political order

[16] The term literally means 'minor learning' emphasizing its auxiliary status. It is called 'characterology' here from the modern perspective.

[17] Zhao & Baldauf 2008: 26-27.

[18] Those remarks refer only to the Chinese scholars. The studies done by the Catholic missionaries in China is a subject that deserves its own separate attention.

[19] 'Character Book for Official Posts'.

resulting in large scale language and social reforms. The archeological findings near Anyang were of course the direct reason for the rapid development of historical characterology, but almost simultaneously the political and cultural movements that were shaping the new post-imperial China had language reform as one of the priorities – the tradition of language reform and standardization simultaneous to political upheavals or as their immediate follow-ups dates back to the Qin Dynasty. These language reforms were seen as "*the major contribution to the country's unity, power consolidation and social stability*".[20] Chinese script, despite the growing attention from researchers and politicians, still had to wait a few decades to face a substantial change in the research approach, i.e. when the modern synchronic structural studies grew in prominence[21] and the whole discipline became more balanced.

1.2. Six categories (六書 *liùshū*)

The most commonly used classification of hanzi is still 許慎 Xŭ Shèn's six categories (六書 *liùshū*, lit. 'six graphs', 121 CE).[22] It is inevitable to critically introduce the six categories classification, despite its omnipresence in the literature. It's still the most handy method for demonstrating the most basic facts regarding the emergence, formation and construction of *hànzi*.

The typology based on six categories is a valid analytic tool (but not without shortcomings) for the early types of Chinese script. It was written as a study of one particular type of script – 小篆 *xiǎozhuàn* 'small seal script', but it has been a common practice to refer 六書 *liùshū* analysis to the other ancient forms of script – 甲骨文 *jiǎgŭwén* 'oracle-bone inscriptions' (Xŭ Shèn probably didn't even know of its existence), 金文 *jīnwén* **'**the bronze inscriptions', 大篆 *dàzhuàn* 'great seal script' and also the later forms (隸書 *lìshū* and the modern 楷書 *kǎishū*). The six categories typology is deficient in capturing the specifics of the modern writing system that was started with the Li-change 隸變 *lìbiàn* – the introduction of the clerical (official) script 隸書 *lìshū* and ultimately the regular script 楷書 *kǎishū*. Any classification of modern *hànzi* must take into account the evolutionary changes of shape, structure, composition, meaning and

[20] Zhao & Baldauf 2008: 26.

[21] See Section 1.3.

[22] The concept of 六書 *liùshū* is universally associated with Xŭ Shèn. In fact, the term has a longer history and was used by other authors, of whom 班固 Bān Gù (the author of 漢書 *Hànshū* 'The Book of Han') and 鄭玄 Zhèng Xuán (the commentator of 周禮 *Zhōulǐ* 'The Rites of Zhou' where the first mention of the six categories is found) are the most notable (Song & Jia 2003: 40-41). Xŭ Shèn's work was the most robust as a lexicographic treatise, and the most methodical and thorough. His work dealt with the Eastern Han writing system in a very detailed way.

pronunciation up to modern times, while 六書 captured an important but much earlier stage of evolution of Chinese characters, namely 小篆 *xiǎozhuàn*.

1. Pictographs (pictograms) 象形字 *xiàngxíngzì* was the earliest method of character formation employed during the earliest stages of the evolution of Chinese script. The form of the characters in this category was determined by the depicted object. It must be stressed that characters which fall into the category of pictographs were not truly pictographic even in the oracle-bone script form. For example, the oracle bone depiction of an eye: is undoubtedly pictographic in origin, but not in the semiotic and linguistic nature. Boltz observes that these inscriptions fail to satisfy even the basic requirement for pictographs – the realism of depiction which is necessary for a direct appeal to an object rather than a word.[23] *SWJZ* dealt with small seal script characters which are even less pictographic in depiction. An example of this is the character 'eye', the conventionalized form of which ' ' was greatly influenced by the shape of other small seal script characters, rather than by the oracle-bone inscriptions (not to that it varies from the depiction of an eye). The supposedly ideographic nature of Chinese characters was a subject of heated debate and is still a quite common conviction among laymen, but is no longer a subject of academic discussion.[24] There is one 'pictographic' feature of the pre-modern characters that sets them apart from forms of modern script (after Li-change 隸變) in a way relevant to this study. The continuous form renders them non-analyzable in terms of modern atomic structural units – the strokes – which means that the pre-modern pictographs are not decomposable at all. The counterparts in modern regular script are decomposable into conventional strokes: 364 characters out of a total 9,353 found in Xǔ Shèn's dictionary, or 4%) are pictographs.
2. Ideographs (ideograms, indicative characters) 指事字 *zhǐshìzì* symbolize abstract meanings. Characters belonging to this category are similar in origin to pictographs – the form is motivated by the ideas they denote – for example the ideograms symbolizing the idea of 'down' ' ',' ', '丅' in oracle-bones inscriptions, seal script and *Shuōwén Jiězì* respectively. The corresponding series for the ideograms for 'up': 丄 helps to illustrate the ideographicity of 指事字. Only about one hundred characters in *Shuōwén Jiězì* (1%) are ideograms.

[23] Boltz 1994: 31-33.

[24] This means 'ideographic' as opposed to 'representing the units of speech, instead of objects or ideas'. The terminology might be confusing, due to the existence of 'pictographic – ideographic' opposition which refers to different categories of 六書 *liùshū*. The pinnacle of the discussion on the very nature of Chinese writing system was probably the sharp exchange of arguments between H.G. Creel and P. Boodberg in the 1930's (see for example DeFrancis 1984b).

3. Compound ideographs (semantic compounds, logical aggregates, associative characters) 會意字 *huìyìzì* are composed of at least two characters to express abstract meanings. The following is an example of a character composed of two simple ideograms for ‘man’ and ‘tree’: 休 depicts a person leaning against a tree, meaning ‘to rest’. A frequently used device in the formation of ideograms is the repetition of components. For example, the character meaning ‘tree’ repeated two or three times: 林 ‘woods’, and 森 ‘forest’. These types of characters were an important leap forward in the Chinese writing system. Pictographs and simple ideograms represented mostly nouns, while the possibility of compounding made the 會意字 capable of much more flexible representation of verbs and adjectives.[25] About 1,200 characters in *Shuōwén Jiězì* (13%) are compound ideograms.
4. Picto-phonetic characters (phono-semantic characters) 形聲字 *xíngshēngzì* are complex characters composed of two parts – one indicating a semantic category and the second providing the phonetic information. For example, in the character 伴‘companion’ (伴 in regular script) the semantic (radical) 人 ‘man’ indicates that the whole character means some type of person, and the phonetic 半 (*bàn* in modern standard Mandarin) informs about the pronunciation. In principle, the idea of using the combination of existing characters to create a new one is the same as in the case of formation of compound ideograms. The important difference is the departure from the rule that all components of a complex character must contribute to its meaning. The role of the mechanism of combining semantic and phonetic elements in the evolution of characters into a full-fledged writing system simply cannot be overestimated. The first three categories would not be able to carry out the role of representing speech on their own. Therefore, with liberation from the purely semantic composition of characters the rules of formation gained flexibility necessary to transform the Chinese writing into a system capable of keeping up with the dynamics of speech. The importance of the picto-phonetic formation method has been growing over the centuries – only around 20% of oracle-bone characters were 形聲字 *xíngshēngzì*. This this type of chracter grew to 82% of seal script characters (7,697 in *SWJZ*), and exceeded over 90% of the contemporary inventory.

The remaining two categories pertain to the pragmatic aspects of character use, rather than to the formational aspects. In other words, characters belonging to these categories were formed in one of the four ways described above, but underwent a pragmatic category shift. In both cases Xǔ Shèn limited himself to the definitions and some examples.

[25] Yin & Rohsenow 1994: 19.

5. Notative characters (transformed cognates) 转轉字 *zhuǎnzhùzì* are described by Xǔ Shèn in a rather obscure and ambiguous way. Interpretations of Xǔ Shèn's criteria differ, but since the importance of the category of *zhuǎnzhùzì* is purely historical, it is enough to refer to just one representative explanation. Yin & Rohsenow (1994: 25-26) explain that related characters in this category have the same radical, are at least similar in meaning and pronunciation, and can be used for mutual explanation. As an example they give the characters 顶 *dǐng* and 颠 *diān* (in simplified regular script form, and in modern standard Mandarin pronunciation), which share the radical 页, a similar meaning ('top' and 'peak' respectively), and are similar phonetically. Notative characters are rare, and sources only give some examples instead of exhaustive lists.
6. Phonetic loans (borrowings, rebus characters) 假借字 *jiǎjièzì* are characters whose meaning is expanded or changed as they are used to represent newly coined words that lacked written representation. The borrowing is licensed purely by the phonetic value of the original word that the character represents, the meaning being irrelevant. This was a convenient method of updating the script, but since the borrowed characters belonged to a relatively small inventory of 'ideographic' categories (the first three on the above list) it also had serious limitations. The typical process of borrowing may be shown in the example of the character (亦 *yì* in standard modern Mandarin) with the original meaning 'armpit' that was chosen to represent the meaning 'also'. The character 亦 lost its original meaning and another character (picto-phonetic) was devised to represent it. Borrowing a character with a more concrete meaning to represent a word with a more abstract meaning and later coining a new picto-phonetic character for the original meaning as a disambiguation device are quite typical to the process of phonetic loaning. It is possible for the character to retain the original meaning, which is what happened in the case of (果 *guǒ*) which is borrowed to denote *guǒ* 'effect' in 如果 *rúguǒ*, and 結果 *jiéguǒ*. Nonetheless, the character also retained the original meaning 'fruit' .

The six categories system has its deficiencies – heterogeneity and ambiguity of criteria and overlapping of categories. However, unless one wants to analyze the *SWJZ* theory instead of using it as a research and illustratory tool, the shortcomings are absolutely insignificant. In most cases that is the role of SWJZ with respect to the modern form of Chinese script – a convenient prism through which to look at the most basic features of Chinese character formation and structure.[26]

The impact of *SWJZ* was so powerful that it overshadowed and determined the state of character study for almost two millennia. The discovery of the oracle-bone inscrip-

[26] *SWJZ* is far more important in etymological studies, but this aspect is beyond the focus of this study.

tions was one of the important factors that brought new challenges and gave an impulse for new paleontological and etymological studies, but still not much happened with regard to the scientific study of modern character structure. As it was already mentioned the typology of *hànzi* in *SWJZ* primarily addressed the problem of their creation (the first four categories), and secondarily the shifts in their use (the last two categories). 六書 *liùshū* offered some insight into character structure, but by no means was a structural classification. There is neither a difference between the structure of pictographs and ideographs, nor between the large number of compound ideographs[27] and picto-phonetic characters, like 休 (休) used as an example above. It has the most typical picto-phonetical component structure, which is a left-right component structure.

1.3. Modern characterology (漢字學 *hànzixué*)

In modern China the position of linguistics is quite unique. This fact stems directly from the uniqueness of the Chinese script. It was probably Zhou Youguang who first used the term 'modern Chinese characterology'.[28] Certainly he was the first to propose a systematic research program for the new discipline. In 1988 Zhu Dexi was still pondering whether 'Chinese characterology' was a valid academic subject of study.[29] Modern characterology emerged for a few independent reasons – the most general was the realization that without a proper study the Chinese script would become a burden that would hinder the advance of civilization, but would also cause problems in everyday life. East Asian countries were faced with the challenges of the modern information processing age that compelled interrelated efforts related to Chinese script to enable computer processing and the unification of encoding. That resulted in standard coded character sets and encoding standards. As Zhao and Baldauf rightly note: "*Perhaps no other country in the world but China has an IT industry so closely interrelated with its writing system and its study of linguistics.*"[30] Their study supports the conclusions that can be also be drawn from this book – the new perspective on the study of characters, new research possibilities, advancement in theoretical and practical approaches to Chinese script were fueled by the emergence of a new social order, by the necessity to keep up with change. The resultant progress, though, was not without a price. The linguistic contribution to Chinese information processing was essential in establishing the Chinese script as a proper tool of communication in the information age. In return,

[27] The multiplication of components (e.g. 森) is a device typical in the formation of compound ideographs, generally not used to create picto-phonetics.

[28] Zhou 1980.

[29] Zhu 1988: 1.

[30] Zhao & Baldauf 2008: 234.

the new technologies offered previously unthinkable research opportunities (the present study is an illustration of this), while at the same time the emerging standards of Chinese information processing probably irreversibly shaped and restricted, at least in part, present and future character studies. The main analytic part of this book is a good example. On one hand analysis on this scale would be simply impossible without the data provided by information processing technology, while on the other, the same technology restricts the way the research is conducted. The goals of establishing information processing standards are different than linguistic investigations. The data used in this study is not perfect, but it is the best available. It is difficult to imagine that anyone will ever make an attempt to work on a new camparable set of data for purely linguistic reasons, at least not without some new technology.

The importance of characterology beyond the discipline of linguistics is undeniable. This is especially true in reference to the IT sector, though it must be noted that in the end it was computer technology that influenced and changed Chinese characterology, not the other way around. The same technology that reshaped the study of characters enabled and gave a boost to statistical and quantitative investigations. Paradoxically the same technology also imposed certain convenient compromises. It is interesting to observe that any attempts at the grammatology of Chinese script ended at the turn of 1980s and 1990s which coincides with the emergence of the digital era.

Modern Chinese characterology is a balanced discipline – this study concentrates on synchronic structural research, but as the 'founding fathers' of 漢字學 *hànzixué* all point out that it should include several domains of study – including the traditional ones. Works of Zhou (1980), Gao & Fan (1985), Zhu (1988), Qiu (1988) show a very similar approach to the modern understanding of character study. A summary of the research domains may also serve as a summary of this section – Chinese characterology as a modern approach studies characters from all possible angles, which include the evolution, the structure, the function and the social aspects.

1.4. Number of Chinese characters

There is no one answer to the typical layman's question about the number of existing characters, but it happens to be not just a layman's problem. It is not possible to establish the absolute number of Chinese characters, yet it is important to give meaning to the numbers in the context of quantity of characters. In the further sections of this book quantitative data pertaining to different number of characters are quoted. Quantitative investigations are based on the corpuses consisting of a certain number of characters – interpretation of the results partially depends on the size of the investigated set. In order to put the numbers in perspective some basic information pertaining to this issue will be presented in this section.

Chinese literacy standards define the basic literacy levels as 1,500 characters for rural regions (farmers) and 2,000 for urban areas and office workers.[31] A middle school graduate is expected to know 3,500 characters,[32] and while the standard for college graduates are more difficult to estimate, 5,000 characters is probably a good approximation. The basic lists of frequently used characters contain: in China 3,500 characters, and in Taiwan 4,808.[33] A typical concise character dictionary contains more or less 7,000 entries. The corpuses of texts contain a different number of unique characters, depending on the size of the corpus itself and on the type of texts. The corpus-based statistics of the number of characters is much more telling than educational standards or contents of any dictionary.[34] There are many corpus studies that have produced different results. Table 1.1 lists some of notable examples of results based on research of non-specialized, non-technical texts:

Tab. 1.1 The number of characters in corpuses of different sizes

Total number of characters	Number of unique characters
1,807,398[35]	4,574
1,051,159[36]	4,667
21,600,000[37]	5,991
307,317,060[38]	9,711

Tab. 1.2[39] The number of characters in different corpuses

Corpus	Total number of characters	Number of unique characters
Classical Chinese	65,348,624	11,115
Modern Chinese:	193,504,018	9,933
Informative	106,254,415	8,954
Imaginative	87,249,603	8,435
Total	258,852,642	12,041

[31] Li 1988: 43. See also: http://www.accu.or.jp/litdbase/policy/chn/index.htm.

[32] This is the standard in the People's Republic of China.

[33] See Section 2.1.1. for details.

[34] In fact, both the lists of frequently used characters and the educational / literacy standards must be based on corpus studies.

[35] Su 2001: 36.

[36] Taiwan Ministry of Education: http://www.edu.tw/files/site_content/M0001/86news/ch2.html?open

[37] Su 2001: 34-35.

[38] http://ccl.pku.edu.cn:8080/ccl_corpus/CCL_CC_Sta_Xiandai.pdf.

[39] Da 2004: 6.

The corpus of social and natural sciences texts was gathered in the years 1977-1982, and it consisted of 7,754 unique characters (11,080,000 in total).[40] Da (2004) presented detailed results of his study that also included the classical texts (Tab. 1.2).

The large corpuses of diverse texts contain up to 10,000 unique characters, and as the numbers show, corpuses which are 10 times smaller contain up to 6,000. The highest number of 12,041 unique characters includes the 'classical' characters (not used in contemporary texts). The statistics for large corpuses are close in number to the deductive estimates in Yin & Rohsenow (1994) and Su (2001). Yin & Rohsenow (1994) very roughly estimate the number of all characters in current use as 10-20 thousand, including highly specialized terminology.[41] Su used the above mentioned corpus of social and natural sciences texts, to include the broadest spectrum of special terminology corpus based on data available at that time[42] (containing 7,754 characters). Additionally, he included two general frequency charts (containing 4,574 and 5,991 characters, also mentioned above), to calculate the number of modern Chinese characters as over 10,000.

The above data can be supplemented by the correlation of the number of frequency-ordered characters and the text coverage ratio (not to be confused with the understanding of texts), the details of which can be found in Section 6.1.1. Here it is sufficient to say that a knowledge of 3,000 characters guarantees the coverage of over 99% of written texts.[43] The discussion in this section should be enough to provide the right perspective on the character sets presented in the next chapter and on the corpuses used for graphotactic analysis. More details and different perspectives on the subject may be found in Zhou (1984), and Zhao and Zhang (2007).

[40] Su 2001: 38.

[41] Yin & Rohsenow 1994: 48-53.

[42] The book was first published in 1994.

[43] Su 2001: 35.

2. Chinese character sets

This study focuses on the quantitative properties of Chinese script. Consequently, in assuming this perspective the determination of a character inventory is of fundamental importance. The complex meaning of the term 'Chinese characters' makes the task nontrivial.[44] A *Chinese character* set is understood as a clearly defined (usually officially) standard collection of characters. Any non-standard collection of characters will be termed a *character inventory*. Character sets in this section are classified from the prespecive of computer information processing.[45] The terminology is borrowed from Lunde (2008), who recognizes two major categories: noncoded character sets (NCSes) and coded character sets (CCSes).

NCSes are sets of characters created for purposes different than information processing. Most typically, NCSes are an important part of the general education process, which includes language planning, teaching and literacy policy. In other words the design of NCSes is not related to computer processing technology. This is not to say that the two types of sets are unrelated, but this fact will not be discussed here. At this point suffice it to say that because of their purpose NCSes are usually smaller than CCSes, the former constituting the subsets of the latter.[46] Both types of character sets are relatively new inventions – NCSes are a result of organized efforts towards standardization and increasing the literacy rate, while CCSes are a natural product of the evolution of computer technology. The need for standardized sets of Chinese characters comes from the properties of the Chinese writing system itself. The sheer number of Chinese characters and their relation to the units of speech are potential factors impeding social, educational and technological progress. The character sets are the first important step toward resolving some of the important problems resident in the Chinese script.

Generally, character sets vary greatly based on national sets or standards. Character sets, both coded and non-coded, are aimed at different levels of literacy and different levels of the educational system or at standardization, consequently, they also distinguish different classes of characters (e.g. standard vs. variant or obsolete forms, modern vs. ancient, etc.). This chapter is devoted to the introduction of some of the most important character sets.

[44] See Chapter 1.

[45] The term 'character set' itself is used in information processing, instead of 'character list'.

[46] Lunde 2008: 79.

2.1. Noncoded Character Sets (NCSes)

As it was already mentioned, NCSes are functional sets of characters, usually designed for the purposes of education, language policy and planning. In contrast to some of the most important CCSes, the noncoded sets are country specific and are not subject of international standardization efforts. The NCSes in different countries (with the focus on China and Taiwan) will be briefly introduced in this chapter with an emphasis on the number and type of characters in each NCS examined.

The survey of standard character sets here has a very specific purpose – to facilitate the determination of the number and type of characters used for graphotactic analysis. The usefulness of a particular set depends on its size, type and the homogeneity of characters.

2.1.1. Chinese noncoded character sets

2.1.1.1. China

The historic aspect of designing character sets throughout the early years of the Republic of China and the People's Republic of Chna will be neglected here – the early attempts at making an inventory of characters more accessible, mostly by limiting and hierarchizing their number were extensively described by Su.[47]

There are three important official character sets in the People's Republic of China that were created for educational purposes and for the purposes of language policy and planning:[48]

- 现代汉语常用字表 *xiàndài hànyǔ chángyòngzìbiǎo* 'List of Frequently Used Characters'– 2,500 characters in the primary school curriculum;
- 现代汉语次常用字表 *xiàndài hànyǔ cìchángyòngzì biǎo* 'List of Secondary Frequently Used Characters' – an additional 1,000 characters in the middle school curriculum (the two lists are often jointly called 现代汉语常用字表 *xiàndài hànyǔ chángyòngzìbiǎo*);
- 现代汉语通用字表 *xiàndài hànyǔ tōngyòngzì biǎo* 'List of Commonly Used Characters' – 7,000 characters, including the 3,500 from the first two lists.

[47] Su 2001: 56-62.

[48] For example, Lunde 2008: 80-81.

Another notable list was published in 2009 – 通用规范汉字表 *tōngyòng guīfàn hànzì biǎo* ‘List of Commonly Used Standardized Chinese Characters’[49] that enumerates 8,300 characters. The list recognizes 3 levels of characters:

- level 1 – the 3,500 most frequently used characters (equivalent to the characters on 现代汉语常用字表 and 现代汉语次常用字表);
- level 2 – 3,000 less frequently used characters, together with level 1, constitute 6,500 characters used to satisfy the needs of printing and publishing standards;
- level 3 – 1,800 characters uses in surnames, geographical names, technical terminology and the classical characters used in the official Classical Chinese teaching materials for primary and middle school.

Six traditional characters and 51 variant forms were ‘reestablished’ as standard ones, and the forms of 44 characters were standardized. The purpose of this list is related to standardization, rather than to something pedagogical. The 通用规范汉字表 *tōngyòng guīfàn hànzì biǎo* is the largest consistent set of characters that does not contain traditional and variant forms, and for that reason it is the best choice for a graphotactic analysis of the characters used in the People’s Republic of China or for a comparative graphotactic analysis with the equivalent Taiwanese set.

The lists of simplified characters in the context of differences between the PRC and Taiwan should also be mentioned. An introduction of the simplification scheme can be found in Section 4.6.

2.1.1.2. Taiwan

Taiwan’s Ministry of Education issued four official list of characters that together totaled over 48,000 characters. The titles of the first two Taiwanese lists are analogous to the first two Mainland Chinese lists, but contain a much higher number of characters:[50]

- 常用國字標準字體表 *chángyòng guózì biāozhǔn zìtǐ biǎo* ‘List of Standard Forms of Frequently Used Characters’ – published in 1982, containing 4,808 characters;
- 次常用國字標準字體表 *cìchángyòng guózì biāozhǔn zìtǐ biǎo* ‘List of Standard Forms of Less Frequently Used Characters’ – published in 1982, containing 6,341 characters.

[49] http://www.china-language.gov.cn/doc/zb2009.pdf

[50] For example, Lunde 2008: 81-82. The numbers of characters on the lists differ slightly in different sources. For example, the Taiwan’s Ministry of Education online dictionary website provides very similar, yet different numbers (http://dict.variants.moe.edu.tw/ex.htm). The discrepancies are too small to discuss here. The actual number of characters used for graphotactic analysis will be discussed in Chapter 7.

The remaining two lists are collections of rarely used characters and variant forms:

- 罕用字體表 *hǎnyòng zìtǐ biǎo* 'List of Rarely Used Characters' – published in 1983, containing 18,480 characters;
- 異體國字字表 *yìtǐ guózì biǎo* 'List of Variant Form Characters' – published in 1984, containing 18,609 characters.

2.1.2. Non-Chinese noncoded character sets

The discussion of the character sets used outside the People's Republic of China and Taiwan will be very brief and limited to Japan and Korea. The use of Chinese characters in these countries, especially in Korea, is limited when compared to Mainland China and Taiwan. Due to the unification efforts and inclusion of Japanese and Korean 漢字 *hànzì* into the Unihan set, however, it seems reasonable to introduce the situation of Chinese characters in these countries, and provide character numbers to render a more complete general perspective.

The most notable of Japanese sets is the 常用漢字 *jōyō kanji* 'Frequently Used Chinese Characters' list, mentioned in different sections of this book. After the recent revision in 2010 the list contains 2,136 characters compulsory in the Japanese educational system – 1,006 are taught in primary school and the remaining 1,130 in secondary school. The second list – 人名用漢字 *jinmeiyō kanji* 'Characters Used in Personal Names', contains 861 characters beyond the *jōyō kanji*.[51] The third set – 表外漢字 *hyōgai kanji* 'Characters Outside the Chart' – is not a definitive official list of characters. As the title suggests all characters not listed in the first two lists fall into *hyōgai* category. The number of *hyōgai kanji* is difficult to estimate.[52]

The official Korean list of Chinese characters – 漢文教育用基礎漢字 *hanmun gyoyukyong gicho hanja* 'Basic Chinese Characters for Educational Purposes' contains 1,800 characters that are taught in middle and high school. Another official list – 人名用漢字 *inmyeongyong hanja* 'Characters Used in Personal Names' – contains 2,964 characters officialy approved for use in personal names.

[51] *Jinmeiyō kanji* may also refer to the joint set of *jōyō kanji*, which also can be used in personal names, and *jinmeiyō kanji*, for a total of 2,997 characters.

[52] Lunde (2008: 83) provides a number of 1,022 *hyōgai kanji* ('NLC Kanji'). On the other hand, the 日本漢字能力検定試験 *Nihon kanji nōryoku kentei shiken* 'Test of Japanese Kanji Aptitude', at the highest level evaluated, tests 6,000 *kanji*, which means the number of *hyōgai kanji* is closer to 3,000.

2.2. Coded Characters Sets (CCSes)

CCSes are a result of the pressing need to adapt the East Asian language societies to the challenges of global and local exchange, distribution and processing of information in the face of proliferation of computer technology. This section introduces the most important coded sets in the context of the main purpose this study. It concentrates on the number and type of characters in chosen Chinese locales and international sets, rather than on the history or compatibility and interchangeability between different CCSes. Those latter aspects are covered in many information processing-oriented sources, of which Lunde (2008) is probably the most comprehensive and referential.

The CCSes usually divide characters into at least two blocks labeled 'levels' or arrange them in a single block that is divided into planes. The characters in each level are arranged in rows and cells. Level 1 characters usually are the most frequently used characters, level 2 are less frequently used, etc.

2.2.1. Chinese coded character sets

2.2.1.1. China

The official coded character sets in the People's Republic of China (PRC) start with the abbreviation 'GB'.[53] Although it was Japan to first introduce a national coding standard, the CCSes in mainland China are probably most numerous, and at least a few of them are worth introducing here.

2.2.1.1.1. GB 2312-80 and GB/T 12345-90

The GB 2312-80 set standard was the first established in the PRC in 1981. It is listed here not because of its historic importance, but rather because of its 'twin' set – GB/T 12345-90, which is its traditional equivalent. The characters in GB 2312 are arranged in a 94×94 matrix. The Chinese characters are arranged in rows: in rows 16-55 (3,755 level 1 characters, arranged according to 拼音 *pīnyīn*); rows 56-87 (3,008 level 2 characters, arranged according to radicals and the number of remaining strokes); rows 88-89 are unassigned in GB 2312, but contain additional characters in GB/T.[54] The simplified set can be transformed into the traditional set with the use of 2,180 characters

[53] 'GB' stands for 国家标准 *guójiā biāozhǔn* 'National Standard'.

[54] Lunde 2008: 95 and 99-103, http://en.wikipedia.org/wiki/GB_2312.

(2,118 traditional equivalents and 62 of the additional characters in rows 88-89).[55] In most cases the positions of simplified and traditional equivalents are paired in exactly the same locations, which is convenient for the comparative analysis of both sets. The differences are documented and with proper treatment such analysis should not be problematic. The GB 2312 standard was extended and supplemented a few times, but the standards established by this method will not be discussed here. More details on the process of extending and supplementing the GB 2312 standard set can be found in Lunde (2008).

2.2.1.1.2. GB 13000.1-93 and GBK

The GB 13000.1-93 is a manifestation of the Chinese government's efforts to internationalize the standard of Chinese character coding. To put it simply, GB 13000.1 is the Chinese equivalent of the international standard ISO 10646.1-1993. The Chinese-specific part of the standard is designated 'GBK' (汉字内码扩展规范 *hànzì nèimă kuòzhăn guīfàn* 'Chinese Internal Code Specification'). GBK is a translation and extension of GB 2312-80, allowing the standard to cover all remaining characters in ISO 10646.1-1993. GBK specifies 21,003 Chinese characters, which is 101 more than in ISO 10646.[56]

2.2.1.1.3. GB 18030-2005

The evolution of international standards, especially the emergence and growing prominence of the Unicode, were reasons for establishing the newest incarnation of GB 2312 designated GB 18030-2005 (中文標準交換碼 *zhōngwén biāozhŭn jiāohuànmă* 'Chinese National Standard GB 18030-2005: Information technology—Chinese Coded Character Set'). Chinese authorities declared a subset of this standard as mandatory for supporting all the computer software sold in China. GB 18030 was established to accommodate the Unicode standard while remaining compatible with GB 2312 and GBK, and for that reason it may be viewed as a superset of all previous standards.[57] The number of characters covered by GB 18030 may be associated with Unicode standard (CJK Unified Ideographs) which is discussed in Section 2.2.2.1. Because of the association with the Unicode it supports both simplified and traditional characters.

[55] Lunde 2008: 101.

[56] Ibid., 104.

[57] http://en.wikipedia.org/wiki/GB_18030; Lunde 2008: 105-111.

2.2.1.2. Taiwan

There are two general features of Taiwanese (Republic of China) standards – they do not include simplified characters and they contain a relatively large number of characters. What is unique to Taiwan is the fact that the official national standard is not the most commonly used standard – the details are provided in the next two sections.

2.2.1.2.1. CNS 11643

The national standard of Taiwan (中文標準交換碼 *zhōngwén biāozhǔn jiāohuànmǎ* 'Chinese Standard Interchange Code') is the largest of the national standards in current use, enumerating 69,134 characters. There are two versions of the standard that need to be addressed – CNS 11643:1992 and CNS 11643:2007. For the sake of simplicity they will be introduced here in a unified way. CNS 11643 provides 13 occupied character planes, and a total number of 8,836 characters can be accommodated in each plane. The first seven planes are practically identical in CNS 11643:1992 and 11643:2007.[58] The remaining planes in the Tab. 2.1 represent the structure of CNS 11643:2007.[59] The character sets in each character plane are arranged according to the stroke count and radicals.

2.2.1.2.2. Big5 (五大 *wǔdà*)

Big5 is the other important standard, and it was the first on Taiwan when it was established in 1984 by the Institute for Information Industry of Taiwan. The first two planes of CNS 11643 enumerate 13,051 characters, which is identical to the number of characters in levels 1 and 2 of Big5.[60] CNS 11643 planes 1 and 2 are often described as a corrected version of BIG5.[61] The fact is that BIG5, at least in terms of software implementation, is the most widespread standard in Taiwan, and is used extensively in Hong Kong and Macau. From the perspective of the present study the encoding differences are irrelevant. What matters is the number and type of characters in a set. The consistent and identical selection of 13,051 characters for the first two levels of BIG5 and first two planes of CNS 11643 make them a natural selection for a graphotactic analysis of a traditional character set.

[58] http://www.cns11643.gov.tw/AIDB/encodings_en.do#encord1; Lunde 2008: 118-119.

[59] http://www.cns11643.gov.tw/AIDB/encodings_en.do#encord1; Lunde 2008: 115-120.

[60] Big5 encodes 13,053 characters, but due to a design error, two characters were duplicated.

[61] For example, so described in Lunde 2008: 115.

Tab. 2.1 Structure of CNS 11643 (the national standard)

Plane	Number of characters	Description
1	5,401	4,808 characters from the official frequently used character list (常用國字標準字體表 *chángyòng guózì biāozhǔn zìtǐ biǎo*), and an additional 593 characters (including 6 variant forms) frequently used in schools
2	7,650	6,330 characters from the official less frequently used characters (次常用國字標準字體表 *cìchángyòng guózì biāozhǔn zìtǐ biǎo*), 1,320 characters from the list of rarely used characters
3	6,148 + 128	rarely used characters and frequently used variant forms, coded by the EDPC (Electronic Data Processing Center 行政院主計處電子處理資料中心 *Xíngzhèngyuàn zhǔjìchù diànzi chǔlǐ zīliào zhōngxīn*) of the Executive Yuan
4	7,298	ISO 10646 characters from the on-line computerized Residency Information System (戶政用字 *hùzhèngyòngzì*) and other organizations, used in information technology
5	8,603	characters from the official list of rarely used characters (罕用字體表), excluding those accommodated in plane 2
6	6,388	variant forms
7	6,539	
10	8,836	variant forms[62]
11	3,698	
12	443	
13	763	
14	408	
15	6,831	

[62] Neither Lunde (2008) nor the official website (http://www.cns11643.gov.tw) provide detailed information on the sources of characters in planes 10-15. For the purposes of this study the convenient category of 'variant form' is sufficiently accurate. For simplicity's sake the above introduction ignores the earlier CNS 11643:1986 standard – its planes 14 and 15 were sources of characters scattered throughout the planes of later standards (see Lunde 2008: 119-120 for details).

2.2.2. Non-Chinese and international CCSes

2.2.2.1. Unicode – Unihan

The Unicode project is the most notable and successful international effort to accommodate the diversity of the world's scripts in a unified way. Unicode was developed by the Unicode Consortium in consultation with the ISO (International Organization for Standardization, more precisely ISO/IEC 10646). It is a character set that aims to provide a unique codepoint for every character of any script, independent from computer software and hardware. The largest subset of the Unicode, called CJK Unified Ideographs (Unihan),[63] pertains to the Chinese script used in the whole East Asia. The 'CJK' initials stand for China, Japan and Korea. Sometimes the abbreviation 'CJKV' is used to include Chinese characters used in Vietnam. The terminology used in the Unicode Standard, i.e. 'Han characters', 'CJK characters' or 'ideographs' is equivalent to the terms 'Chinese characters' and '*hànzi*' used throughout this book.

Tab. 2.2 CJK Unified Ideographs blocks

Block	Number of Characters	Description
CJK Unified Ideographs	20,902	Common
CJK Unified Ideographs, Extension A	6,682	Rare
CJK Unified Ideographs, Extension B	42,711	Rare, historic
CJK Unified Ideographs, Extension C	4,908	Rare, historic
CJK Unified Ideographs, Extension D	222	Uncommon, some in current use
CJK Compatibility Ideographs	268	Duplicates, unifiable variants, corporate characters
CJK Compatibility Ideographs, Supplement	478	Unifiable variants, not used with Ideographic Description Sequences (IDS)

[63] CJK Unified Ideographs (+ extensions) is a character set. Unihan is the name of a database containing CJK Unified Ideographs.

Development of the Unicode started in the beginning of the 1990s. Originally, the standard covered 20,902 CJK characters in the original block that was supplemented over the years in the form of official extensions. In the current version (6.2) it allocates 75,215 Han characters.[64] Table 2.2 shows the distribution and type of CJK characters in each block.[65] It should be noted that the original block contains both simplified and traditional characters.

A basic knowledge of the rules of the *hànzi* unification is important for understanding the content of the CJK subset of Unicode, and it is for that reason a more detailed introduction of the content design is indispensable. Detailed information on character unification is provided in Unicode 6.2.0 documentation – Chapter 12: East Asian Scripts. The introduction and examples below is based on Unicode 6.2.0 specifications:[66]

Rule 1 – Source Separation Rule. If two ideographs are distinct in a primary source standard,[67] then they are not unified.

The Unicode documentation gives an example of the various ununified forms of characters for 'sword':

剣劍剱劔劒釼

This rule was applied only to the characters in the original block.

Rule 2 – Noncognate Rule. In general, if two ideographs are unrelated in historical derivation (noncognate characters), then they are not unified.

The Unicode documentation gives an example of two graphically similar, but etymologically unrelated characters:

土≠士

Rule 3 – Abstract Shape Rule. By means of a two-level classification, the abstract shape of each ideograph is determined. Any two ideographs that possess the same abstract shape are then unified provided that their unification is not disallowed under either the Source Separation Rule or the Noncognate Rule.

The Abstract Shape Rule is based on an assumption that the typeface used as a surface manifestation of shape (e.g. computer display) is secondary to the underlying abstract shape. Fig. 2.1 shows the three-dimentional model of character shape representation that is used to determine the underlying shape of each character.

[64] Unicode 6.2.0, East Asian Scripts: 407 (http://www.unicode.org/versions/Unicode6.2.0/ch12.pdf).

[65] Based on Unicode 6.2.0, East Asian Scripts and Lunde 2008: 156.

[66] Ibid., 415-418.

[67] The source standards include some of the national standards introduced earlier in Section 2.2.

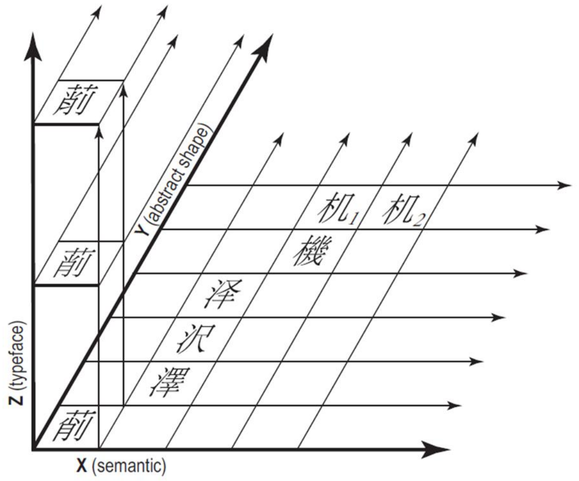


Fig. 2.1[68]

In other words the characters are graphically classified into two levels: the abstract shape (*Y*-axis) and the actual typeface (*Z*-axis). The diversity of forms on the *Z*-axis unfounded on the *Y*-axis is ignored in Unicode, which is to say that the characters are unified. To determine differences between an abstract shape and an actual shape, the structure and features of each character are analyzed in relation to the Ideographic Component Structure. This structure includes the number, type and relative positioning of the components, the structure of corresponding components, treatment of the character in a source character set, and a radical reference.[69] If characters differ in any of these respects, the characters are considered to possess different abstract shapes and are not unified. Fig. 2.2 ia an example of the Ideographic Component Structure, it may also serve as an introduction to character structure that is discussed in details in Chapter 4.

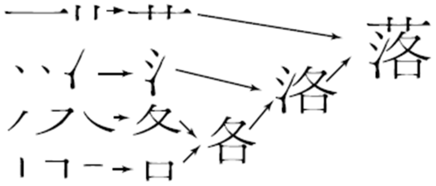


Fig. 2.2[70]

The abstract shape comparison may be illustrated in the following way:

[68] Unicode 6.2.0, East Asian Scripts: 415.

[69] Ibid., 416-417.

[70] Ibid., 417.

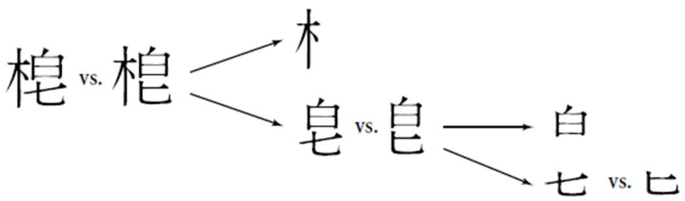


Fig. 2.3[71]

The potential candidates for unification are examined according to the aforementioned criteria and procedures. Table 2.3 shows examples of the characters that were not unified for different reasons.

Tab. 2.3[72] Distinct treatment of characters

Characters		Reason for distinct treatment
日	曰	Non-cognate characters
崖	厓	Different number of components
說	説	Distinct treatment in source sets
峰	峯	Different relative positions of components
擴	拡	Different components in the same relative composition
祕	秘	Different radicals[73]

The Unicode standard with respect to CJK characters openly assumes that there will always be a set of unencoded characters, simply because of the coinage of new characters.[74] To address this issue a system of unencoded character descriptions was devised. The system, referred to as the Ideographic Description Sequence (IDS) system, facilitates the interchange of text containing such characters. Because of the importance to the present study, the IDS system is introduced in greater detail in Chapter 5 and also mentioned in Chapter 4.

It is difficult to overestimate the importance of the Unicode Standard and the CJK Unified Ideographs set from the perspective of the present study, but also from a mostly general point of view. The Unicode Standard and the CJK Unified Ideograph set have not only removed barriers in information exchange and processing, but also have had a profound impact on the shape of modern characterology.

[71] Ibid.

[72] Ibid.

[73] This is the description of the actual unification process, there is no point to debate whether the radicals are a necessary part of the procedure.

[74] Unicode 6.2.0, East Asian Scripts: 423.

3. Theoretical preliminaries

This chapter presents the theoretical premises for the intended graphotactic analysis of Chinese script from the perspective of a more general segmentotactological framework.

3.1. Segmentotactology and segmentotactics

The first part of the chapter has two goals: to introduce the general idea of segmentotactology/segmentotactics, and to discuss its relation to the original theory of phonotactology/phonotactics as a source of inspiration for this book.

3.1.1. Introduction

This introduction is necessary to set straight the facts concerning the short history of segmentotactology and to give credit where credit is due. The origins of the present study date back to 2008 when prof. J. Bańczerowski gave a series of lectures on his new approach to phonotactology and phonotactics. The theoretical part was presented in the form of a rigorously formal axiomatic framework. The theoretical proposal had immediate research results in the analysis of the structure of Polish words in terms of letters of the alphabet by prof. P. Wierzchoń who conducted research on the Polish corpus of dictionary entries. The general idea behind the phonotactology was so capacious and flexible, as Bańczerowki and Wierzchoń had amply shown, it became self-evident that it could be used for inquiries into different lingual systems. That is precisely how and when this book was inspired.[75]

The axiomatic grid of the theory has never been published, but the ultimate form intended by its author – that is, segmentotactics, including graphotactics, required the theory to be constructed in a formal way.[76] The semi-formal introduction to phonotactology and phonotactics was provided in Bańczerowski (2009), in which the author shows its application to a fragment of Chinese phonotactics, based on *pīnyīn* transliterations of the *MDGB English-Chinese Dictionary* (CC-CEDICT) entries and confronts

[75] As it was mentioned in the preface, the project of extending the investigative range of Bańczerowski's concept beyond phonetics and phonology was presented in Kordek (2012). The reasons for restricting the analysis to graphotactics were also briefly presented in the preface, but in fact they are quite self-explanatory.

[76] Bańczerowski 2009: 9.

the results with the results obtained by Wierzchoń.[77] In the introduction to his research Bańczerowski provides a non-formal account of his concept. The present study adopts the same approach to the theoretical premises of the discussed problems. The phonotactological theory is treated here as a member of the class of segmentotactological theories that are based on similar principles and theoretical premises. The main proposal of this book outlined here and the demonstrated research practice in the last chapter may look like a substantial modification of the original framework, but in fact it is merely a natural follow-up to the ideas more or less explicitly expressed by Bańczerowski himself. For example, while commenting on the actual type of his analysis and the type of linguistic data he had at hand, he wrote:

> "Since neither of these dictionaries gives their entries in phonetic transcription, the exemplifications which will be adduced, reflect the graphotactic structure of these entries rather than the phonotactic."[78]

His humble comments on the results of analysis leave no doubts that Bańczerowski was fully aware of the real nature of the investigation and uncertain as to the actual prospects for this type of investigations:

> "The author is also fully aware of the approximate nature of the exemplifications being given. The unavailability of suitable phonotactic language material certainly weakened the value of these exemplifications. But nevertheless the proposed theory may turn out to be a source of inspirations which may result in more adequate elaborations of general and particular phonotactology. The author would also like to hope that the journey accomplished in the present, still imperfect, phonotactological vehicle, into the enormous expanse of words, will contribute to making at least one further small step towards a better understanding of the phonotactic reality of ethnic languages, a reality full of enigmas and surprises. However, if this hope is unfounded, that is, if the reader will get the impression of having wasted time on this article, then all that's left to do is to apologize for my misconceived approach to the reality in question."[79]

The evaluation of the analysis performed in this study depends on many factors, but one thing can be said with certainty – the failure would not be a result of misconceptions in the theory itself, but rather because of the misuse by the author.

[77] Bańczerowski 2009. Wierzchoń's analysis was not published independently.

[78] Ibid., 9.

[79] Ibid., 22.

3.1.2. Prerequisites for segmentotactic investigations

Following Bańczerowski's suggestion[80] a distinction should be made between:

- segmentotactology (subdiscipline of linguistics, a class of linguistic theories);
- segmentotactics (subject matter of segmentotactology, comprised of all segmentotactic objects and relations).

The subject matter of segmentotactology may be briefly defined as a word grammar – understood as a calculus that in research practice produces relevant results by means of computational analysis of different levels of representation.

An analysis of this type requires a certain type of data to be available for computational processing. Bańczerowski lists four conditions for a database to be considered suitable for phonotactological analysis:[81]

(i) it should be sufficiently representative of the vocabulary of a given language;
(ii) the entries should be solely words (not including syntagms composed of more than one word);
(iii) it should be accessible in an electronic form;
(iv) the word-entries should be given in phonetic transcription.

Confronted with the reality of Chinese electronic dictionaries, such conditions turn out to be rather demanding and present the most challenging task in conducting the research. For example, the *MDGB English-Chinese Dictionary* (CC-CEDICT) that was used by Bańczerowski contains numerous syntagm-entries, which means that it does not satisfy condition (iv). Due to the lack of appropriate digital databases, certain types of the segmentotactological analyses must be limited to theoretical considerations. Fortunately, this is not the case with the segmentotactical investigation of Chinese characters.

3.2. Phonotactics

In this section the original terminological setting of phonotactic theory is introduced in an abridged version. The terminology in question was presented in similar form in Kordek (2012).

[80] Ibid., 8.

[81] Ibid., 9.

3.2.1. Terminology

The terminology used in the remainder of this book needs to be defined here for a few obvious reasons. First of all, the uniqueness of Bańczerowski's concept is directly reflected in the terms he uses. Secondly, the expanded framework, including the graphotactic component, will utilize analogons of the terms coined for the need of phonotactics. All definitions here are quoted after Bańczerowski. An ***utterance*** is "*a spatio-temporal physical object, individual and concrete, produced hinc et nunc by a definite speaker in a definite time and space... In a certain sense an utterance is a linear object consisting of phonical substance, having its beginning, duration and termination in time, and immediately preceded and succeeded by pauses.*"[82] A ***vocabulon*** (*actual word*) is a "*maximal unit of linear, that is, sequential, ordering of an utterance. Putting it differently, the linear structure of an utterance may be imagined as a sequence consisting of vocabulons as always linearly continuous and relatively easily distinguishable units within utterances.*"[83] A ***phonaton*** is "*any subvocabulonic part or segment of various size, provided it is linguistically relevant. Each phonaton is also as individual and concrete as its corresponding vocabulon, and it is always a linearly continuous unit. Needless to say, every vocabulon will be treated as a particular kind of phonaton.*"[84] A ***phonon*** is a minimal phonaton; this term is similar to ***sound*** or ***actual phone***, but is preferred for technical reasons.[85] A ***phone*** is "*a set of all those phonons which are homophonous with a given phonon*".[86] A ***vocable*** is "*a set of all those vocabulons which are homophonous and homosignificative with a given vocabulon*".[87] The term ***word*** would be ambiguous in this terminological setting. The definitions so far form a preliminary phonotactic setting, one that allows the definitions of the remaining phonotactic terms that will have direct analogons in other domains of segmentotactics. A ***phonotacteme*** is a phonetic representation of a linear structure of a vocable – a sequence of phones which are constituants of a given vocable.[88]

So far no new types of linguistic segments or units have been defined. The new terms were coined for the sake of precision and for technical reasons to avoid ambiguity. At this point, however, the introduction of theory-specific terms is necessary.

[82] Ibid., 10.

[83] Ibid.

[84] Ibid.

[85] Ibid.

[86] Ibid.

[87] Ibid.

[88] Ibid.

3.2.1.1. Tactophoneme

Vocables consist of sequences of phones; a different way of putting it is to say that certain sets of phones sequentialize (tactify) in the vocables. A ***tactophoneme*** will be conceived as a set of phones that tactify in a phonotacteme. For purely illustrative purposes it is most practical to avail to an example based on the letters of alphabet which is also a method used by Bańczerowski (2009). For English tactophoneme {*A, R, T*}[89] which is a set of three 'phones' (letters, in fact), out of all possible permutations seven result in phonotactemes representing the English vocabulons: *ART*, *TAR*, *RAT*, *TARA*, *TART*, *TARTAR*, *TATAR*.

The properties of a tactophoneme may be described in terms of:[90]

(i) ***phonicity*** – the number of phones which are its elements;

(ii) ***phonotactemic range*** – the set of all phonotactemes generated out of it;

(iii) ***phonotactemicity*** (phonotactemic load) – the number of all phonotactemes generated out of it.

The characteristics of the tactophoneme in the above example are as follows:

(i) phonicity: 3.

(ii) phonotactemic range: {ART, TAR, RAT, TARA, TART, TARTAR, TATAR} phonotactemicity: 7.

Other important phonotactic properties are described by:

(i) ***tactophonemic dispersion*** – the set of all tactophonemes to which a given phone belongs;

(ii) ***tactophonemic dispersion number*** – the number of all tactophonemes to which a given phone belongs;

(iii) ***phonotactemic dispersion*** – the set of all phonotactemes in which a given phone occurs;

(iv) ***phonotactemic dispersion number*** – the number of all phonotactemes in which a given phone occurs.[91]

Another relevant property of tactophonemes is described by their ***phonotactemic efficiency*** – the ratio between the phonotactemicity and the phonicity of a given tactophoneme.[92] The phonotactemic efficiency of the exemplary tactophoneme {*A, K, T*} equals 2.33 (its phonotactemicity is 7, and its phonicity is 3).[93] The notion of phonotactemic efficiency may also be understood as the ratio between the number of all phono-

[89] English serves here as a better example than Chinese.

[90] Bańczerowski 2009: 13.

[91] Ibid., 14.

[92] Ibid., 15.

[93] A more detailed exemplary analysis will be presented in the section on graphotactics.

tactemes and the number of all tactophonemes.[94] The last terms introduced here are related to a ***tactophonome***, which is defined as a set of equiphonous tactophonemes, i.e. comprised of the same number of phones or having the same phonicity.[95] Derivative terms include ***tactophonemicity***, which is defined as a tactophonome related to phonotactemicity and phonotactemic efficiency.[96] The notion of tactophonome will prove useful and important in the analysis performed in Chapter 7.

The discussion in this section does not cover the full extent of Bańczerowski's phonotactics – it is limited to those theoretical aspects that are pertinent to the actual graphotactic analysis. There are at least two aspects of Bańczerowski's proposal related to the tactophonome that are omitted – tactophonomic phone-basis, and tactophonomic equiphony and disphony.[97] These two issues have a different status in relation to graphotactic analysis – the former can be rather easily implemented, but is left out due to the space limitations; the latter is much complicated as it involves different properties than the analyzed ones. The properties of Chinese characters make equiphonic/disphonic analysis even more complex.

3.3. Beyond phonotactics

As it was already mentioned, Bańczerowski was well aware of the fact that he was exemplifying his phonotactical framework with an inquiry into a different level of language structure that he informally termed 'graphotactic'.[98] This section is devoted to the domains of the extended phonotactic framework, but focuses mainly on the theoretical premises of Chinese graphotactics. Orthotactics is set apart from graphotactics. The distinction is justified in Section 3.2.1. Phonemotactics, syllabotactics and morphemotactics are given an extremely brief treatment which is limited only to the basic terminology illustrating the similarity of all domains of segmentotactics. More detailed treatment showing the perspectives of the segmentotactic investigations in Chinese, but without an actual analysis, can be found in Kordek (2012).

[94] Bańczerowski 2009: 15.

[95] Ibid., 16.

[96] Ibid.

[97] Ibid., 20-21.

[98] Ibid, 17. For this level we use a different term, for reasons to be explained in the sucessive sections.

3.3.1. Orthotactics

The understanding of the term ‘graphotactics’ in the present study is slightly different than in Bańczerowski (2009). He used it in the context of the representation of Chinese words in 拼音 *pīnyīn* transliteration. In Kordek (2012), it was tentatively proposed to distinguish between ‘orthotactics’ and ‘graphotactics’, with the former pertaining to the structure of words in terms of units of alphabetical scripts, and the latter pertaining to the structure of Chinese characters and probably other non-alphabetic scripts.[99] The reason for this term becomes apparent when we are confronted with the diversity of the writing systems of world languages. It is probably justified to assume that in the case of languages using alphabetical writing systems the two terms could be used synonymously, since it seems difficult to associate them with two different levels of tactical analysis in those languages. In alphabetical scripts there is no other relevant graphical level other than orthography. However, the same cannot be said of languages with non-alphabetical writing systems, such as Chinese. The graphic aspect is inherently associated with the Chinese script; on the other hand it is not immediately obvious what ‘orthography’ means in reference to Mandarin Chinese. The units that tactify into the written representations of words in the two types of writing systems are of a very different nature. The letter type units in alphabetical systems more or less directly reflect the phonetic or phonemic properties of a vocable, while in the case of the Chinese logographic script the internal structure of individual characters is not restricted by such properties of vocables. In other words, if this terminological distinction is to be accepted, orthotactics would pertain to writing systems dependent on the phonetic and phonological properties of a given language, especially the alphabetic systems, while ***graphotactics***[100] would refer to systems with a different setup of relations between the speech and the script. Due to the lack of investigation of other scripts, at this point this distinction can only be claimed to pertain to Chinese script.

The orthotactics of Chinese script is not then a direct inquiry into the writing system, but instead into its alphabetical transliteration. The proposed terminology is analogous to the phonotactical case. The introduction in this section is intended as an illustration and is limited only to the basic terms. The section on graphotactics, as pertaining to the main topic of the study, is more detailed and thorough.

The ***orthotacteme*** will be the linear representation of vocables in terms of letters. The ***tactorthoneme*** will be conceived as a set of letters that tactify in an orthotacteme. The following terms relate to a tactorthoneme:

(i) ***orthocity*** – the number of letters which are its elements;

[99] This section is in large part an expanded version of the considaraions in Kordek (2012: 112-113).

[100] ‘Graphemotactics’ also comes to mind as an alternative term, that was in fact used in Kordek (2012).

(ii) ***orthotactemic range*** – the set of all orthotactemes generated out of it;

(iii) ***orthotactemicity*** (orthotactemic load): the number of all orthotactemes generated out of it;

(iv) ***orthotactemic dispersion*** – the set of all orthotactemes in which a given letter occurs;

(v) ***orthotactemic dispersion number*** – the number of all orthotactemes in which a given letter occurs;

(vi) ***orthotactemic efficiency*** – the ratio between the orthotactemicity and the orthocity of a given tactorthoneme.

As already mentioned, in the case of analysis of Chinese script the orthotactic analysis is an inquiry into the transliteration system. The results presented by Bańczerowski (2009), based on the 拼音 *pīnyīn* transliteration, reflect the relevant properties of Chinese. For example, the orthotactemic efficiency is expected to be lower than in Polish. The reason for this is the syllable and word structure of Chinese and the related issue of the syllable-morpheme-word correspondence.[101] The typical Chinese word consists of two syllables. Every syllable is subject to rigorous restrictions on its linear structure. Typically, only one permutation of the elements of a tactophoneme is allowed (the same is true for tactorthonemes). For example, the tactorthoneme {*R,E,N*} tactifies into one orthotacteme only: {*REN*} ('man'). The only theoretical possibility of increasing the orthotactemic efficiency of most Chinese tactorthonemes is the existence of a vocable consisting of a duplicated syllable – {*RENREN*} ('people'), as is the case in this particular example. In the case of tactorthonemes that can tactify into bisyllabic vocables, for example {*S*, *H*, *I*, *H*, *E*}, the typical efficiency equals one, with the exception of cases where there exist orthotactemes representing the vocables with reversed syllabic linear order. In the above example the orthotactemic efficiency equals 2, since both orthotactemes *SHIHE* and *HESHI* (both meaning 'suit, suitable', among other things) represent Chinese vocables. The restrictions on the linear order of syllables and the related small number of syllables in Chinese are the main factors which reduce phonotactemic and orthotactemic efficiency. On the other hand the possibility of syllable duplication and permutations in the syllabic linear order – a phenomenon non-existent in Polish – increase the efficiency. In extreme cases the efficiency may increase to values not seen in Polish:
{N, A, I}: {NAI, NIAN, NAINAI, NIANNIAN, NINA, NANI, NAINA, NA'NAI, NAN'AI, AINAN, NANAI, NAINAN, NANNAI, NINIAN, NIANNI, AINAI, NAIAI, AINA, NAAI, AINIAN, NIANAI, NI'AN'AI}.

Intuitively, out of the properties having an opposite effect on efficiency, the number of syllables and the restrictions on linear order within the syllable are expected to dom-

[101] These properties have significance for every type of tactical analysis of Mandarin Chinese, not only orthotactical.

inate the tactical properties of Chinese vocables. This intuition is confirmed by the results obtained by Bańczerowski. The orthophonemic efficiency of Polish is 1.36 while that of Chinese is only 1.11.[102]

3.3.2. Phonemotactics, syllabotactics and morphotactics

The section on remaining domains of (Chinese) segmentotactology will be limited to the introduction of the terminology. There are a few reasons for such restrictions. First of all, a more detailed introduction of topics not directly related to the main subject of the book is not possible due to the space limitations. Secondly, also because of the secondary importance of the remaining domains for the present study, the introduction could not bring anything new compared to Kordek (2012). This section is intended only as an illustration of the flexibility of the original phonotactic concept. The terminology in question briefly illustrates the segmentotactic investigations of phonemic, syllabic, and morphemic levels of language.

Phonemotactics is understood as a segmentotactical analysis of vocables in terms of phonemes. ***Phonemotacteme*** is the linear representation of vocables in terms of phonemes. ***Tactophoneme*** is conceived as a set of phonemes that tactify in a phonemotacteme. ***Phonemicity***, ***phonemotactemic range***, ***phonemotactemicity***, ***phonemotactemic dispersion***, ***phonemotactemic dispersion number***, ***phonemotactemic efficiency***, etc., will be defined analogously to the phonotactic counterparts.

Syllabotactics is understood as a segmentotactical analysis of vocables in terms of syllables. ***Syllabotacteme*** is the linear representation of vocables in terms of syllables. ***Tactosyllable*** is a set of syllables that tactify in a syllabotacteme. ***Syllabocity***, ***syllabotactemic range***, ***syllabotactemicity***, ***syllabotactemic dispersion***, ***syllabotactemic dispersion number***, ***syllabotactemic efficiency***, etc., will be defined analogously to the phonotactic counterparts.

Morphotactics is understood as a segmentotactical analysis of vocables in terms of morphemes. Morphotacteme is the linear representation of vocables in terms of morphemes. ***Tactomorpheme*** is conceived as a set of morphemes that tactify in a morphotacteme. ***Morphemicity***, ***morphotactemic range***, ***morphotactemicity***, ***morphotactemic dispersion***, ***morphotactemic dispersion number***, ***morphotactemic efficiency***, etc., will be defined analogously to the phonotactic counterparts.

[102] Bańczerowski 2009: 15-16 and 18-19.

3.3.3. Graphotactics

Probably the most unique tactical analysis in Mandarin Chinese refers to one of its most unique systems – the script. The terminology and theoretical premises of graphotactics are discussed from the perspective of Chinese script, while the problem of universality of the proposal is not addressed here. The complexity of the Chinese writing system presents the problem of determining the most basic concepts of the tactical analysis of characters. It is provisionally proposed that a graphotactic counterpart of phone be the ***grapheme*** – a component part of a character at any layer of decomposition,[103] excluding the strokes. This understanding of the term is convenient for representing the structure of units of Chinese script, but it should be noted that some notable definitions offer different perspectives. The grapheme is usually understood analogously to phoneme – as an abstract minimal unit of script[104] represented by allographs (glyphs). Coulmas (2003) offers no reference of the term related to the Chinese script. Rogers equates 'grapheme' with 'character'[105] and refers to components as 'ligatures'.[106] Köhler's (2008) general idea of a grapheme is in concord with the interpretation adopted in this study, but it requires an element of a script to play a role in the representation of units of speech, as well as either phonetic or semantic representation:[107]

> "A grapheme is any graphical sign which, on its own, represents in at least one context a portion of linguistic material. Hence, the letter <c> is a grapheme regardless of the fact that it appears also in sequence with <h> for another sound. On the other hand, diacritics such as accents would not be considered as graphemes but as parts of complex graphemes because they do not represent any sound, sound combination, word, or meaning. They are rather distinctive features which serve to differentiate graphemes. Sequences such as <ch> will then be considered as syntagmas."

The above definition is clearly aimed at different types of script.

The terminological ambiguities do not change the fact that it is clear what 'grapheme' should mean in graphotactic theory. Grapheme is synonymous with 'component', a definition of which was provided by Zhao & Baldauf (2008) and is quoted in Section 4.1.

[103] See Chapter 4. It should be noted that the databases used in graphotactic analysis of Chinese script contain character entries representing Chinese signary, rather than vocables representing vocabulary.

[104] Coulmas 2003: 36, Rogers 2005: 10-11.

[105] Rogers 2005: 26.

[106] Ibid., 39.

[107] Köhler 2008: 4.

The remaining basic graphotactic terminologies are analogous to proposed in the previous sections: The ***graphotacteme*** is the spatial representation of vocables in terms of graphemes. The ***tactographeme*** is conceived as a set of graphemes that tactify in a graphotacteme. The following terms relate to the tactographeme:

(i) ***graphemicity*** – the number of graphemes which are its elements;
(ii) ***graphotactemic range*** – the set of all graphotactemes generated out of it;
(iii) ***graphotactemicity*** (***graphotactemic load***) – the number of all graphotactemes generated out of it;
(iv) ***tactographemic dispersion*** – the set of all tactographemes to which a given grapheme belongs;
(v) ***tactographemic dispersion number*** – the number of all tactographemes to which a given grapheme belongs;
(vi) ***graphotactemic dispersion*** – the set of all graphotactemes to which a given grapheme belongs;
(vii) ***graphotactemic dispersion number*** – the number of all graphotactemes to which a given grapheme belongs;
(viii) ***graphotactemic efficiency*** – the ratio between the graphotactemicity and the graphemicity of a given tactographeme.

The dispersion may be understood as a distribution of components (graphemes) between the graphotactemic units – tactographemes and graphotactemes, hence two types of dispersion.

The average efficiency of the tactographemes is not expected to be high, since the majority will have efficiency equal to 1. This is due to the fact that in most cases the same set of components makes up only a single character; however, the character formation rules allow for variations in the spatial arrangement of components resulting in different characters, as well as for the recurrence of components, which is another important mechanism of character formation. The following examples of tactographemes and their graphotactemic range illustrate these properties:[108]

{木} : {木 *mù* 'tree'，林 *lín* 'woods'，森 *sēn* 'forest'};
{一, 日}: {旦 *dàn* 'dawn'，亘 *gèn* 'continuous'};
{一, 亅}: {丁 *dīng* 'cubes',亍 *chù* 'footstep'};
{句, 多}: {够 *gòu* 'enough', 夠 *gòu* 'enough'};
{木, 日}: {杲 *gǎo* 'bright', 杳 *yǎo* 'obscure, dim'};
{一, 大}: {天 *tiān* 'heaven', 'day', 夫 *fū* 'man'}.

[108] Examples are taken from Kordek 2012: 117.

The examples show different major formational strategies that increase the efficiency of Chinese graphotactemes. The first three examples exemplify the recurrence of elements, the fourth ({句, 多}), is an instance of linear rearrangement, and the last two are examples of spatial rearrangement.

The basics of a graphotactic analysis will be shown based on the example of above tactographemes and graphotactemes.

3.3.3.1. Exemplary analysis

The basics of a graphotactic analysis will be shown based on the example of the above sets of tactographemes and graphotactemes. This is done to introduce additional terminology and facilitate an understanding of proper analysis on a large scale performed in Chapter 7.

3.3.3.1.1. Graphemicity related analysis

This section exemplifies the basic types of graphotactic investigations of Chinese script. The form in which the examples are presented may differ from the actual analysis. Due to the tiny size of the sample set of tactographemes and graphotactemes, the presentation of the results in the form of lists poses no problem. On the other hand, a presentation in form of diagrams might seem excessive; this is the exact opposite of the analysis in Chapter 7.

Tab. 3.1 Graphemicity and graphotactemicity of tactographemes

Tactographeme	Graphemicity	Graphotactemicity
{木}	1	3
{一, 日}	2	2
{一, 亅}	2	2
{句, 多}	2	2
{木, 日}	2	2
{一, 大}	2	2

The first set of data shown in Tab. 3.1 concerns the graphemicity of each tactographeme.

Another important notion is the graphotactemic efficiency of tactographemes which is a measure of their generative power. It may be applied to the individual tactographemes, to the subset of tactographemes or to the whole tactographemic system. In the exemplary set there are 6 tactographemes and 13 graphotactems, which means that the average graphotactemic efficiency of the whole system is 2.17. It seems that in normal analysis, due to the number of elements, the efficiencies for individual tactographemes would not be listed, but for the exemplary set it can be done without sacrificing too much space:

Tab. 3.2 Graphotactemic range and efficiency of tactographemes

Tactographeme	Graphotactemic range	Graphotatemic efficiency
{木}	{木, 林, 森}	3
{一, 日}	{旦, 亘}	2
{一, 亅}	{丁, 亍}	2
{句, 多}	{够, 夠}	2
{木, 日}	{杲, 杳}	2
{一, 大}	{天, 夫}	2

In the actual analysis – when large corpuses of data are at play – it is utterly impractical to list individually both the graphotectemic efficiencies, but any type of individual data, e.g. the graphemicity for every single tactographeme. It is more convenient, and more significant from an analytical perspective, to classify the tactographemes with the same graphemicity into families – ***tactographons***. In other words, tactographons are the sets (families) of tactographemes with identical graphemicity. The graphemicity of tactographons will be used as a name for respective families (t-families). In the exemplary set there are two tactographons:

1: {{木}}

2: {{一, 日}, {一, 亅}, {句, 多}, {木, 日}, {一, 大}}.

Tactographemicity is the number of tactographems of which a given tactographon consists. ***T-graphotactemicity*** (to distinguish it from graphotactemicity) is the number of graphotactemes generated out of a given tactographon – in other words, t-graphotactemicity is the number of graphotactemes generated out of all tactographemes with a certain graphemicity. Also graphotactemic efficiency can be calculated for every tactographon (t-efficiency). The tactographemicity, t-graphotactemicity and t-efficiency in the exemplary set are summarized in Tab. 3.3.

Tab. 3.3 Properties of tactograhons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	1	3	3
2	5	10	2

3.3.3.1.2. Dispersion related analysis

Dispersion pertains to the distributional properties of graphemes. There are two types of dispersion: graphotactemic and tactographemic. Both can be expressed either in sets of elements (range) or in the number of elements (number).

There are 6 graphemes in the exemplary set: {木, 一,日, 句, 多, 大}. Their dispersional properties are summarized in Tab. 3.4:

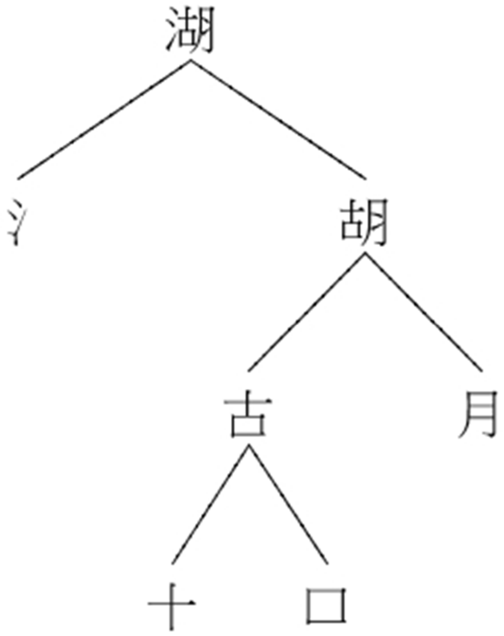
Tab. 3.4 Dispersion of graphemes

Grapheme	Graphotactemic dispersion (range)	Tactographemic dispersion (range)	Graphotactemic dispersion number	Tactographemic dispersion number
木	{木, 林, 森, 杲, 杳}	{{木} , {木, 日}}	5	2
一	{旦, 亘, 丁, 亍, 天, 夫}	{{一, 日}, {一, 亅}, {一, 大}}	6	3
日	{旦, 亘, 杲, 杳}	{{一, 日}, {木, 日 }}	4	2
句	{够, 夠}	{{句, 多}}	2	1
多	{够, 夠}	{{句, 多}}	2	1
大	{天, 夫}	{{一, 大}}	2	1
		Average:	3.5	1.7

The dispersion range is impractical to present in the case of large sets of graphemes and graphotactemes. Even in alphabetic scripts where the number of graphemes is relatively small, tens of thousands of graphotactemes are typically (depending on the average graphotactemic efficiency of tactographemes) not much smaller in number than the tactographemes which would have to be listed. The dispersion range in most cases would be used only for referencial and control reasons. The dispersion numbers on the other hand can and should be presented in cases of small number of graphemes – the analysis of Cangjie codes in Chapter 7 is a good example. However, in the case of Chinese script, where the graphemes are counted at least in the hundreds, it is still impractical to work on the dispersion numbers of the individual graphemes. Most significant information provided by the dispersion data is the average dispersion and the standard central tendency and dispersion measures, such as the mean, the median, the range, the variance and the standard deviation.

3.3.2.2. Levels of analysis

The exemplification of graphotactic analysis in the previous section does not take into account the complexity of the internal structure of Chinese characters. This problem will be discussed in details in the next chapter. This section is only a brief introduction to the problem. The following example is the decomposition of the character 湖 *hú* 'a lake':[109]



The components (graphemes) in the decomposition tree (氵 'water', 胡 *hǔ* 'beard', 古 *gǔ* 'ancient', 月 *ròu* 'flesh', 十 *shí* 'ten', 口 *kǒu* 'mouth') are spread on 3 different levels. The first branching is a decomposition into immediate components (氵 and 胡). The left node contains a non-decomposable component (氵). The right node component decomposes into two more basic components (古 and 月), one of which (古) can be further analyzed into two even smaller constituents (十 and 口).[110] This recalls the phrasal structure of a sentence represented by X-bar syntactic trees distinguishing between the intermediary and true phrasal components. The trees are a convenient way of presenting the constituent structure of *hànzi*, but the form of representation is not directly relevant to graphotactics. It is not immediately clear whether graphemes in all nodes and levels are a valid subject of graphotactic analysis. The components revealed by the first and last branchings, i.e. immediate and non-decomposable constituents, seem like the natural units of interest. At this point the problem of the graphotactical status of the intermediate components (i.e. 古, 十, 月) and individual strokes remains

[109] The examples of character decomposition in this section are taken from Kordek (2012).

[110] The atomic level of decomposition (strokes) is not shown in the example. It is discussed in the next chapter.

unaddressed. These problems, along with other related issues pertaining to the structure and decomposition of characters, graphotactically relevant constituent types, etc., are discussed in the next chapter.

4. Structure of Chinese characters

The literature on the subject of Chinese script does not lack studies on the structure of Chinese characters, which display all degrees of extensiveness, thoroughness and focus on different structural properties. The aim of this chapter is not to present a comprehensive, state-of-the-art overview of theoretical and descriptive research in the field, but instead it is intended to focus on the issues relevant to the graphotactic analysis of Chinese script with a brief introduction of the more general aspects. For more extensive readings on the structure of Chinese characters one should refer to Wang (1983), Qiu (1988), Stalph (1989), Fan (1990), Yin & Rohsenow (1994), Fei (1997), Shen & Shen (1998), Gao (1999), Wang (1999b), Wang et al. (2001), Su (2002), Wang (2002), Song & Jia (2003), Lü (2004), Chuang & Teng (2009), and Chen et al. (2011). The Chinese coding standards for information processing in Mainland China (GF 3001-1997, GF2001-2001) and Taiwan (CNS 11643-2, CNS 11643-3), as well as the international standards (Unicode and ISO) must also be mentioned.

4.1. Terminology

From the perspective of constituent structure a distinction should be made between simple and complex characters – the former decompose directly into strokes while the latter into components differing in the degree of complexity. The English terminology referring to the constituent parts of Chinese characters is not unambiguous. The main controversy concerns the use of the term 'radical'. It is used as an equivalent of two different Chinese terms: 部首 *bùshǒu* 'indexing component' and 邊旁 *biānpáng* 'radical'[111]. The former pertains to the indexing function of certain components used for the ordering of and searching for characters in character sets and dictionaries, whereas the later is related to the traditional formative parts of complex characters. The referential official standard of components (GF 3001-1997 – 'Chinese Character Component Standard of GB 13000.1 Character Set for Information Processing') defines neither 部首 *bùshǒu*, nor 邊旁 *biānpáng*, and in the context of structural and graphical decomposition of characters in GF 3001-1997, only the term 部件 *bùjiàn* 'component' is used; the English language equivalents involving the use of 'radicals' are specified in the GF 0012-2009 standard – 'Specification for Identifying Indexing Components of GB 130001. Chinese Character Set', which covers the indexing properties of character constituents. The 'radical' in the context of componential structure should be under-

[111] The English translations are in accordance with the GF 0012-2009 standard. Other equivalents used in the literature in English are 'indexing radical' for 部首 *bùshǒu* and 'side component' or simply 'component' for 邊旁 *biānpáng*.

stood as a ‘side component’ of a picto-phonetic character, i.e. either the semantic or phonetic part.[112]

The problem of indexing characters by their parts has no relevance for graphotactic analysis and for that reason, to avoid any terminological confusion, the term ‘radical’ will be avoided here and the term ‘component’ or the more general term ‘constituent’ will be used instead. The GF 3001-1997 standard mentioned above contains the definitions of some basic terms that are important from the graphotactic perspective – the central term being 汉字部件 *hànzì bùjiàn* “Chinese character component”. A component is rather vaguely defined as a “*unit in the structure of character having the constituent function*”.[113] Derivative terms are also defined: 成字部件 *chéngzì bùjiàn* ‘free component’, 非成字部件 *fēichéngzì bùjiàn* ‘bound component’[114], 基础部件 *jīchǔ bùjiàn* ‘basic component’, and 合成部件 *héchéng bùjiàn* ‘compound component’. The present study employs the neutral term ‘component’ in the sense of 部件 *bùjiàn* in GF 3001-1997. A much clearer definition, however, explaining the difference between components and radicals, was formulated by Zhao & Baldauf:[115]

> “Components are a new concept, born out of the need for designing schemes for computer typing, and hence, is a flexible term. In addition to strokes and radicals, there is a need to reconstruct characters into more maneuverable units,… The component is purely a graphological composing unit, qualitatively between strokes and simple characters, with an emphasis on position in constructing the character regardless of its phonetic and semantic functions. Therefore, the component is essentially different from the radical in that the radical is either semantically or phonetically rational, but the component is not. It is based on the strokes, but normally smaller and simpler than a radical.”

In the context of a graphotactic framework the term ‘grapheme’ will be used synonymously.

[112] For example, Fan 1990: 105.

[113] GF 3001-1997: 2.

[114] Ibid. The English equivalents of the Chinese terms proposed in the document can be misleading. 成字部件 are the components that can function as standalone characters while 非成字部件 are bound forms that are only allowed as parts of characters.

[115] Zhao & Baldauf 2008: 14.

4.2. Hierarchy of constituents

The structural descriptions of Chinese script usually deal with the structure of characters in a bottom-to-top approach – by starting at the most atomic level of smallest elements, then proceeding to the intermediate components and fnally arriving to the characters level. In the following sections the constituent structure is introduced from the top to bottom perspective that serves better the ultimate purpose of this book, but this is merely a technical issue.

From the point of view of the compositional properties there are two distinctive types of characters – simple and compound (complex).[116] The simple characters are formed by a single component, the compound characters by at least two components. The internal structure of characters is multilayered – the complex ones are always decomposable into components, components decompose into strokes. The decomposition may have a different depth depending on the degree of complexity of the most complex component. The complexity of components, that is, the depth of their decomposition, is a criterion for their classification.The simple characters are formed by one component and their depth of decomposition into components is null.[117] The different types of constituents will be introduced in the next sections.

4.3. Composition

The components of characters are arranged in one of the conventional ways. The composition of components is not linear in nature, contrary to most alphabetic scripts. Characters differ in terms of the type, number and spatial arrangement of components. The number of possible structures may vary according to the degree of detail. For reasons that will become apparent later, the introduction of the structural composition types will be based on the twelve Ideographic Description Characters (IDC) that are part of the Unicode standard.[118] IDCs are graphic descriptions of the internal composition of compound characters (Tab. 4.1).

For the sake of simplicity the number of compositional categories represented by the IDCs is a compromise between economy and the level of detail. It is possible to differentiate between lesser or greater numbers of structure types, but this would result in structure types that are either too general or unnecessarily complicated. The problems of the composition types and their representation are also discussed in the section

[116] For an example see Xiao (1994).

[117] As simple as it sounds, the practical implementation of this definition proves to be difficult. This problem has no direct importance to the main purpose of this book and it will not be pursued in more detail. The problem was thoroughly discussed by Xiao (1994).

[118] http://www.unicode.org/charts/PDF/U2FF0.pdf

on character description languages. It should be noted that the composition types bear no direct significance for quantitative analysis itself.

Tab. 4.1 Ideographic Description Characters

Structure type	IDC	Examples
left to right	⿰	乣搈褑
above to below	⿱	褭名盲
left to middle and right	⿲	粥衍讎
above to middle and below	⿳	象賞嶲
full surround	⿴	井卪因
surround from above	⿵	夃风太
surround from below	⿶	凶鼎斗
surround from left	⿷	𠥃匠𠥓
surround from upper left	⿸	在塺嫠
surround from upper right	⿹	与𤤃鸟
surround from lower left	⿺	乪尳匙
overlaid	⿻	七中乘

4.4. Decomposition and component types

This section is devoted to the detailed introduction of the rules of character decomposition and classification of components from the general and graphotactic perspectives.

4.4.1. Decomposition

Decomposition plays a crucial role in determining the composition of constituents, and more importantly, in establishing component sets for individual characters and for the whole writing system. In other words, decomposition pertains to the most basic problem of Chinese graphotactics, i.e. its basic units of analysis.

4.4.1.1. Decomposition rules

The rules of decomposition are usually provided in the form of general guidelines and there is generally a lack a detailed description of the procedure. It is usually assumed that the procedure of decomposing characters into components is intuitive, and therefore, there is no need to formalize or algorithmize every aspect of it. Typically, only controversial issues (like the treatment of variant forms of components and graphically similar components) are given more attention.

The componential structure of characters is not always unambiguous. In fact, the general rules must leave a certain degree of indeterminancy that should be resolved at the level of basic components. The problem may be exemplified by two of the components of the character ‘醫’ – ‘殳’ and ‘矢’.[119] There are two types of possible decomposition criteria – etymological and structural.[120] From the etymological perspective there is no reason to decompose ‘矢’, despite the structural features suggesting the presence of two immediate constituents – ‘𠂉’ and ‘大’, both of which are attested to in a number of other characters.[121] Decomposition in modern characterology somewhat relies on structural principles. The discussed component is a clear example of that. Unfortunately, matters are far more complicated than the given example. A further decomposition of ‘大’ is structurally motivated, yet it is not uncontroversial. Out of four referential character databases containing the structural componential information, 2 decompose ‘大’ into ‘人’ and ‘一’ (Kawabata’s IDS database and Wenlin 4.1); 2 treat it as a basic component (CDP and CHISE);[122] and Fan does not decompose ‘矢’ at all.[123] Since no explanations are offered, it can only be inferred that the non-decomposability comes from the inference of etymology that treats ‘大’ as a non-decomposable character. The component ‘殳’ presents yet another problem of resemblance to etymologically unrelated elements – sometimes the upper components are equated with the character 几 *jī*, instead of being treated as a separate component ‘⺇’. Another example of analyzing graphically similar components is provided by the components ‘土’ and ‘士’. They may be treated in dfferent ways:

- the Kawabata’s database descriptions are identical for both characters, assigning them two components: ‘十’ and ‘一’;
- Wenlin 4.1 treats ‘土’ as a basic component, ‘士’ is decomposed;
- in CDP and CHISE both are basic components.

The component ‘疋’, as in ‘疑’, receives basic interpretation in Kawabata’s IDS, and in both the CDP and the CHISE. However, Wenlin 4.1 splits the component into ‘乛’ and ‘龰’. This is the result of a different treatment of variant forms of components. Etymologically ‘龰’ is a graphic variant of ‘止’ and Wenlin lists such allographic forms separately. This is yet another example of etymological inference. Despite the long his-

[119] Chuang & Teng 2009: 25.

[120] Fan 1990: 103; Chuang & Teng 2009: 22.

[121] The Wenlin 4.1 database shows 565 characters containing ‘大’ and 180 containing ‘𠂉’.

[122] The referential databases will be introduced in further sections of the book.

[123] Fan 1990: 103.

tory of the evolution and reforms of Chinese script, a large portion of the constituents of Chinese characters is etymologically motivated. It can be claimed that the structural features prevail as a decompositional criteria, but it cannot be said that etymological considerations are systemically removed from the componential analysis of characters. The examples so far have illustrated the influence of etymological considerations on the depth of analysis, or, in other words, on the inventory of basic components. In some cases structural and etymological criteria render contradictory analysis. For example, '旗' structurally is decomposed as [[方][𬀩[𠂉][其]], while etymological analysis shows different components – '㫃' and '其'.[124] This is also an example of the 'external' vs. 'internal' analysis discussed in sections 4.4.1.3. and 4.4.4 As a rule of a thumb, the deeper the decomposition level, the more probable the conflict of the two decomposition motivators is. Fan (1990) also points out the possible inference of stroke order on the decomposition. He gives an example of the character '区' that irrespective of the stroke order can be decomposed into '匚' and '㐅', and into '一', '㐅' and '𠃊' , in view of the stroke order.[125]

Chinese characters were not created by one person at a certain point in time using strict and formal principles of composition. They have their history, etymology and semiotic motivation. In other words, it makes sense to treat certain elements of script in a way that cannot be motivated by purely structural criteria.[126] In fact, any complex component may be treated as a decomposable compound, but for etymological reasons and for the sake of tradition there are elements that are treated as non-decomposable regardless of how complex their structure is. For example, the character 殳 for historical reasons is considered a radical, but structure-wise it can be decomposed into two more basic components. This, however, must have a profound effect on any component-based formal descriptions, which either fail to capture the 'spirit' of Chinese characters, or are burdened with inherent indeterminacy. The different rules of decomposition, or simply the sets of components are usually a result of purpose driven commitments, and therefore, it should not be claimed that one set of rules is better than another. The procedure for isolating the set of components (graphemes) used in graphotactic analysis will be described in Chapter 7. In this section some of the existing decomposition schemes will be presented.

[124] The example is borrowed from T. Kawabata's website: http://kanji-database.sourceforge.net/ids/ids-analysis.html?lang=en and Su 2001: 92.

[125] Fan 1990: 103-104. The problem of stroke order is not discussed in this book.

[126] There were attempts at strictly formal (structural) analysis (Rankin 1965, Rankin et al. 1966, 1970), which will be briefly introduced in further sections.

The inventory of components used in the T. Kawabata database[127] is based on the GF 3001-1997 standard as a basic set and the CDP components as a supplementary set. It is for that reason both sets will be discussed in some detail in this study. The rules of decomposition for GF 3001-1997 are introduced below, the CDP components are refered to in different sections of this study.

In an ideal situation a set of precise and unambiguous rules leads to a set of components, or, in other words, an inventory of components is the result of a decomposition of a set of characters with the use of precise rules and instructions. The People's Republic of China official standard for character components (GF 3001-1997) provides very few details on the very procedure of decomposition. In this respect the standard is an instruction for the use of a list of 560 components, rather than a description of the decomposition procedure. The only direct explanation of the decomposition procedure leading to the formulation of the list of 560 components in the standard is a very general statement: "*The 'List of Basic Components' was established after summarizing, categorizing and computing the results of the decomposition of every single of the 20,092 characters in the GB 13000.1 standard set*".[128] From a practical perspective the instructions for decomposing characters with the use of an existing list explains the contents of the database of character decompositions, but the exact criteria for isolating the components on the list are still unclear. Despite that, the instructions in the standard are important for understanding the contents of GF 3001-1997 based character decomposition databases:[129]

1. Mutually separated or connected elements can be isolated, for example:
   明→ 日, 月
   名→ 夕, 口
   韭→ 非, 一
   crossing elements cannot be isolated (decompose directly into strokes), for example:
   串 is not decomposable into 中, 中
   东 is not decomposable into 七, 小
   small number of overlapping elements can be isolated when such decomposition does not influence the structure and the number of strokes, for example:
   幾→ 幺, 幺, 戈, 人 (人 (overlaps with (戈)
   孝→ 耂, 子 (子 (overlaps with (耂)

[127] The KDP database is used as a basic source of componential data in graphotactic analysis (see Section 5.1.1.2.).

[128] Ibid., 3.

[129] Ibid., 4-5.

2. Whenever it is possible, characters should be decomposed in an etymologically motivated way[130] ("*original disassembly*"). In cases when the motivated decomposition is not possible, or is contradictory to structural and graphical criteria, characters should be decomposed in accordance with structural and graphical criteria. For example:
   赤→ 土, 小 (motivated)
   亦→ 亠, 小 (unmotivated)
   虎→ 虍, 几 (etymology contradicted by graphical form – unmotivated)
   東→ (etymology (日, 木) contradicted by overlapping form – not decomposed),
   example of a multilayed analysis:
   给→ 纟, 合 (1st layer – motivated)
   合→ 亼, 口 (2nd layer – motivated)
   亼→ 人, 一 (3rd layer – unmotivated).
3. The compositional variants of components are treated as one component. For example: 亻= 人, 飠 = 食.
4. The listed components should be treated as basic and should not be further decomposed.
5. The listed components can be used to assemble characters, and should not be used to assemble intermediate non-character structures. For example: 自, 田 and 丌 can be used to assemble 鼻, while 自 and 田 should not be used to assemble 畠.

4.4.1.2. Decomposition structure

Only structurally simple characters do not decompose into components. The structurally complex characters are composed of hierarchical tiers of constituents.The first tier consists of the immediate constituents, and the last tier consists of the basic components. The tiers between the first and the last consist of intermediate components, or intermediate and basic components. In other words only the status of the last tier elements is predetermined to be basic components. The number of tiers and type of components in each tier, except for the last one, depends on the individual characters. The Ideographic Description Characters introduced in the previous section carry infor-

[130] Decomposition of characters in an etymologically motivated way often involves a multilayered etymological analysis of the characters (GF 3001-1997: 4).

mation about the number of immediate constituents of a character (1st tier components) and the type of their spatial composition.[131]

The seemingly simple question of the number of components of a given character should entail, more or less explicitly, the type of elements in question. The possible answer might refer to the number of immediate constituents, basic components, or the overall number of components on all levels of decomposition, including the intermediate components. The immediate constituents represent the most natural degree of awareness of componential structure for the average users of Chinese script. Components on deeper levels of decomposition require theoretical models or at least extensive knowledge of the component system. In this study the problem of the number of components will be limited to the immediate and basic components. This restriction will be manifested in the graphotactic analysis of Chinese script presented in Chapter 7.

The examination of the componential structure of characters shows at least three types of potentially relevant components from the graphotactic perspective:

- basic;
- immediate;
- intermediate.

These categories are interrelated, only basic and immediate ones are disjointed**.** As already mentioned, the basic components (基礎部件 *jīchǔ bùjiàn*) are non-decomposable into components other than individual strokes. All components at the lowest level of decomposition are basic, but basic components may appear at any level of decomposition – their only defining feature is the decomposability directly into strokes. 'Immediacy' of components is a functional category relative to the level of decomposition, and unless indicated otherwise the term 'immediate constituent' refers to the character level. Of course, all non-basic (complex) components have at least two immediate constituents. The category of immediate constituents is not equal to the category of complex (compound) components. Intermediate components are the immediate non-basic components of another component (not a character).

From the perspective of systemic functions there are two types of components:

- free forms;
- bound forms.

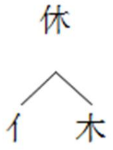
Some of the components also function as standalone characters (free forms), others cannot occur as standalone characters (bound forms). A general tendency, rather than an absolute rule, is that the more complex a component is (the more components there are which comprise it) and/or the closer it is to being the immediate (1st tier) constituent of a character, the more probable that it is a free form. The multilayered composition of characters and different types of components are best shown by examples. The

[131] IDCs are used this way in the Unicode's CJK Unified Ideograph descriptions of characters, but their use is by no means limited to the immediate constituents.

following examples with short explanations illustrate the increasing complexity of characters in terms of compositional depth in different notations. The notations include the component's tier structure, an IC analysis tree and bracketed string representation:

(1) 休

1st tier components: 亻, 木;



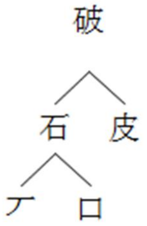
[休[亻][木]]

The character is decomposed directly into basic components – there is one tier of components, meaning the immediate constituents of the character are its basic components. One of the components can function as a standalone character ('木'); '亻' is a bound form – a distributional variant of the full character '人'.

(2) 破

1st tier components: 石, 皮

2nd tier components: 丆, 口



[破[石[丆][口][皮]]

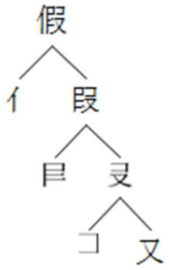
Two tiers of decomposition, 4 components overall: 3 basic, 1 immediate compound component ('石'). One of the immediate constituents is decomposable into two basic components. One of the four components is a bound form ('丆').

(3) 假

1st tier components: 亻, 叚

2nd tier components: 𡰪, 叏

3rd tier comonents: コ, 又

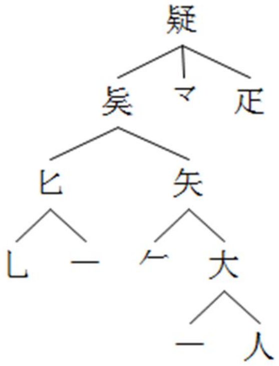


[假 [亻][叚][戸][𠬝 [コ][又]]

Three tiers of decomposition, 6 components overall: 4 basic (‘亻’, ‘戸’, ‘コ’, ‘又’), 1 intermediate (‘𠬝’), 1 immediate compound component (‘叚’) decomposing into two 2nd tier components: a basic ‘戸’ and an intermediate ‘𠬝’, an immediate constituent of ‘叚’, which further decomposes into two basic components. Only one of the six components can function as a standalone character (‘又’).

(4) 疑

1st tier components: 𠤕, マ, 𤴔
2nd tier components: 匕, 矢
3rd tier comonents: 乚, 一, 𠂉, 大
4th tier components: 人, 一

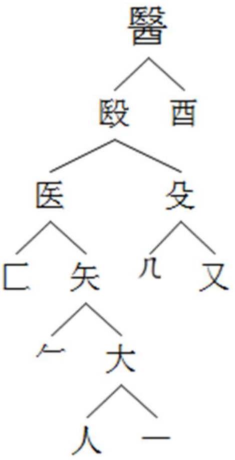


[疑[𠤕[匕[乚][一]][矢[𠂉][大[人][一]]]][マ][𤴔]]

Four tiers of decomposition, 11 components overall: 7 basic (‘マ’, ‘𤴔’, ‘乚’, ‘一’, ‘𠂉’, ‘人’, ‘一’), 3 intermediate (‘匕’, ‘矢’, ‘大’) and 1 immediate compound component (‘𠤕’) decomposing into two intermediate components; 4 of the components are bound forms(‘𠤕’, ‘マ’,‘乚’, ‘𠂉’).

(5) 醫

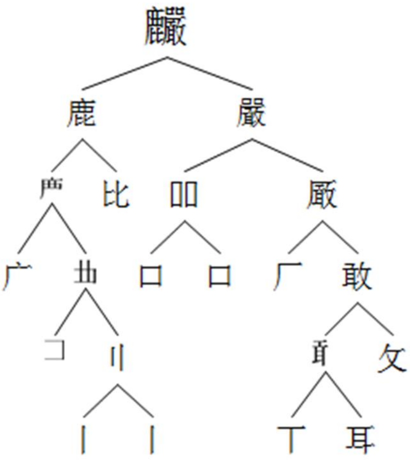
1st tier components: 殹, 酉
2nd tier components: 医, 殳
3rd tier comonents: 匚, 矢, 几, 又
4th tier components: 𠂉, 大
5th tier components: 人, 一



[醫[殹[医[匚][矢[𠂉][大[人][一]]]][殳[𠘧][又]]][酉]]
Five tiers of decomposition, 12 components overall: 7 basic (‘酉’, ‘匚’, ‘𠘧’, ‘又’, ‘𠂉’, ‘人’, ‘一’), 4 intermediate (‘医’, ‘殳’, ‘矢’, ‘大’) and 1 immediate compound component (‘殹’); 5 of the components are bound forms[132](‘殹’, ’ 医’, ’匚’, ’𠘧’, ‘𠂉’).

(6) 䴕

1st tier components: 鹿, 嚴
2nd tier components: 𢉖, 比, 吅, 厳
3rd tier comonents: 广, 丗, 口, 口, 厂, 敢
4th tier components: コ, 𠁣, 𦔮 , 攵
5th tier components: 丨, 丨, 丁, 耳



[䴕[鹿[𢉖[广][丗[コ][𠁣 [丨][丨]]]][比]][嚴[吅[口][口]][厳[厂][敢[𦔮 [丁][耳]][攵]]]]]

[132] This example also illustrates the fact that any assumptions regarding Chinese characters are relative to a particular set of characters. The listed components are bound forms in a set of comtemporary traditional characters. ‘医’ *yī* is a simplified equivalent of ‘醫’, and ‘殹’ is an ancient characters (source: Wenlin 4.1 database). The problem of graphically similar components is briefly mentioned in the next section. Here it should be noted that ‘𠘧’ is different than ‘几’ *jī* and ‘匚’ (radical 22 in 康熙 *Kāngxī* system) is different than radical 23 ‘匸’ *xì*’ (source: Wenlin 4.1 database).

5 tiers of decomposition, 20 components overall, 11 basic ('比', '广', '口', '口', '厂', 'コ', '攵', '丨', '丨', '丅', '耳'), 7 intermediate ('⿸厂廿', '吅', '厰', '丗', '敢', '〢', '⿱丅耳'), 2 immediate compound components ('鹿', '嚴'); the number of free form components is particularly difficult to determine. 7 ('口' is occurring twice) components undoubtedly are free forms ('鹿', '嚴', '比', '口', '敢', '耳'), while the status of 7 components is ambiguous and dependent on the type of characters they refer to:

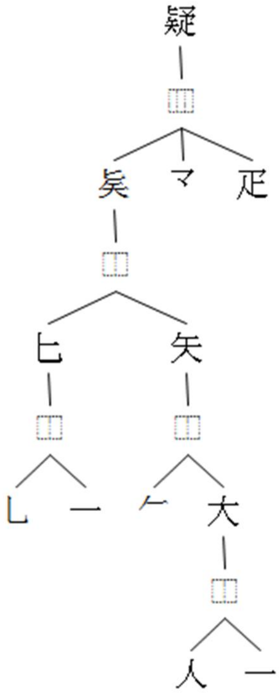
- '吅' is an archaic character;
- '厰' in Wenlin 4.1, no meaning or dictionary entry is provided, and there is only a reference to classical dictionaries;
- '广' is obsolete (Wenlin 4.1) in the traditional characters set, and is a simplified equivalent of '廣' *guǎng*;
- 〢 is a Suzhou numeral 'two';
- '攵' is radical 66 (*pū*), and has the same etymology as the character '攴'*pū* 'beat, strike';
- '丅' is a variant of 下 *xià*;
- '厂' is a radical with no character entry, and therefore its treatment is unambiguous.

The character 麣 *yán* itself is a CNS 11643-1992 plane 3, row 66 character, which means it is a very rarely used character or a rare variant form. Wenlin 4.1 lists it as a variant of ⿱嚴鹿, which is not listed in CNS 11643-1992 at all, nor is it provided a meaning or dictionary entry. It stands to reason that the character's (free form component) status should be granted to the other variant forms also, no matter how rare. The problem is that there is no clear-cut definition of variant forms as opposed to obsolete characters. Intuitively, '丅', in the same way as '厂', can probably be rendered as an obsolete form, but it is not obvious how it affects its status in respect to the discussed categories of components.[133] Even the components that correspond in form to the simplified equivalents of traditional characters are unequivocal candidates for dismissal as free forms. The decomposed character is a traditional one, but it does not have a simplified equivalent – it is rare enough that it was not included in the simplification scheme. Suzhou numerals probably should not be considered Chinese characters. Having the same etymology as a full character is not enough to consider an element as having the same status, as in the case with '攵' and '攴'. There are no easy answers for the problems outlined here, and the solutions will not be addressed any further. Since in this study determining the status of components is not the main purpose of decomposition, at this point it is sufficient to point out the 'identity' problems of some components.

[133] The status of components in East Asian scripts complicates matters even more. '丅', for example, is included on the Japanese 表外漢字 *hyōgai kanji* list, and for that reason from the perspective of CJK Unified Ideographs, the character should be considered a free form.

The above notations provide information on the constituency of a character, but do not contain information on the composition of elements, in which they differ fundamentally from the IDCs. It is possible to integrate the IDCs into any of the above notations, for example:

疑
1st tier components: ⿲𠤕, マ, 疋
2nd tier components: ⿰匕, 矢
3rd tier comonents: ⿰乚, 一, ⿰𠂉, 大
4th tier components: ⿰人, 一



[疑[⿲[𠤕[⿰[匕[⿰[乚][一]]][矢[⿰[𠂉][大 [⿰[人][一]]]]]]][マ][疋]]]

4.4.1.3. Functional categories of components

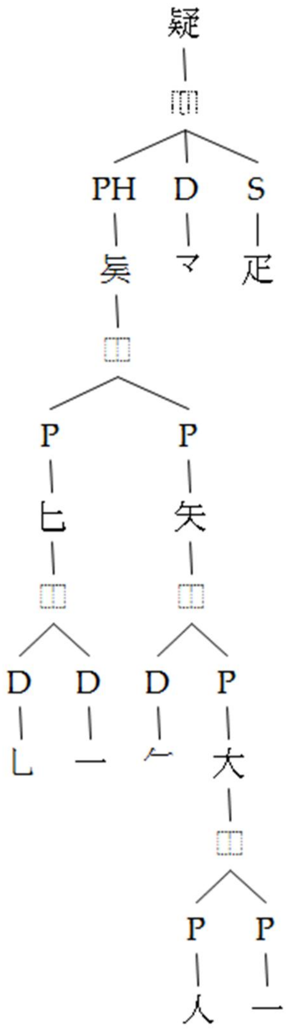
The componential and compositional information in the decompositional model can be supplemented by the functional description of components. Su associates the functions of components with the form, pronunciation and meaning of characters and refers to the functional classification of units of character composition as a study of the 'internal structure' of characters, as opposed to the 'external structure' that is concerned with structural decomposition into components.[134] Different functional classifications of components may differ in detail, but the general idea is based on the

[134] Su 2001: 75 and 92-93.

same premises. For example, Chuang & Teng[135] list the following functions of components:

- pictorial (表形 *biǎoxíng*) – P, components that display properties similar to pictographic characters (象形字 *xiàngxíngzì*);
- signific (表義 *biǎoyì*) – S, components that contribute to the meaning;
- phonetic (示音 *shìyīn*) – PH, components that contribute to the pronunciation;
- diacritic (標示 *biāoshì*) – D, bound form components that distinguish characters;
- substitutive (替代 *tìdài*) – ST, components that substitute other components in certain structures. For example, '大' in the character '奠' is a substitute for '丌'.

The classification proposed by Chuang & Teng has certain limitations, at least without clear and detailed rules, and its application to all levels of decomposition raises some doubts. The analysis below is proposed just for the sake of illustrating the inclusion of functional categories into the componential analysis:



[135] Chuang & Teng 2009: 23-24.

[疑 [⿰ [PH [㝵 [⿱ [P [匕 [⿻ [D [乚]][D [一]]]]][P [矢 [⿱ [D [𠂉]][P [大 [⿻ [P [人]][P [一]]]]]]]]]]][D [龴]][S [疋]]]]

The alternative proposal of functional classification proposed by Su (2001) will be presented in the section on the modern classifiation of characters.

4.4.2. Component lists

The list of basic components differ according to the different standards. The sources of differences and principles of constructing an inventory of components are explained extensively in Chuang & Teng[136] upon whose study the discussion in this section is largely based, including all of the component lists. Given the importance of component lists, for the main purpose of this study, it is necessary to address the problem in a comprehensive way as much as possible. There are two Chinese official standards for defining and listing the basic components – the Chinese GF3001-1997 (Information-processing Components of Chinese Characters with GB13000.1 Character Specification – 信息处理用 GB13000.1 字符集汉字部件规范) and the Taiwanese CNS 11643-2 (Basic Components for Chinese Characters and Their Properties – 中文字基礎部件及部件屬性). The former is based on a character set specified in GB13000.1 containing 20,902 characters, the latter on planes 1 and 2 of CNS 11643 containing total 13,051 characters. GF3001-1997, published in 1997, specifies 560 basic components, while CNS 11643-2, published in 2007, specifies 517 basic components. There are a few sources of the discrepancies between the two sets and the differences are not limited just to the number of basic components. The common part of the two sets contains 391 components. This portion of basic components is exhaustively listed below:[137]

一, 丨, 丿, 丶, ㇏, 乙, 乚, 亅, 𠃌, ㇉, ⺄, 𠃊, 乛, ㇇, 乀, 二, 匕, 𠤎, 匚, 匸, 十, 厂, 丂, 丁, 七, 𠀁, 丅, 与, 𠂇, 冂, 𠘨, 凵, 卜, ⺊, ⺆, 人, 亻, 儿, 入, 八, 几, 勹, 𠥃, 乂, 𠂉, 力, 丬, 𠂆, ⺈, 𠂊, 龴, 𠘧, 七, 亠, 冖, 氵, 丷, 冫, 刂, 刀, 刂, 力, 卩, 㔾, ⺋, 𠃜, 厶, 又, 𠬛, 乃, 九, 了, 丩, コ, 𠀃, 乜, 巜, 土, 士, 大, 寸, 尢, 工, 干, 廾, 弋, 卄, 丈, 于, 才, 丌, 丰, 龵, ⺧, 巨, ユ, 口, 囗, 山, 巾, 少, 夂, 夊, 夕, 尸, 彡, 彳, 彑, 久, 㐅, 乇, 千, 宀, 广, 丬, 女, 子, 孑, 小, ⺌, 中, 屮, 巛, 川, 川, 己, 已, 幺, 廴, 弓, 彑, 彐, ⺕, 阝, 巳, 巳, 也, 𠂔, 丑, 乡, 孓, 巳, 丸, 戈, 木, 朩, 歹, 𡯁, 牙, 犬, 犭, 王, ⺹, 丏, 不, 井, 卅, 夫, 屯, 廿, 市, 丑, 旡, 主, 朮, 五, 爿, 日, 曰, 止, ⺪, 中, 內, 冃, 冃, 毌, 罒, ⺆, 戶, 手, 龵, 扌, 攵, 斤, 月, 夕, 毛, 氏, 气, 爪, 爫, 巨, 父, 片, 牛, 生, ⺗, 𧘇, 𠬝, 丰, 丹, 及, 心, ⺗, 忄, 文, 火,

[136] Chuang & Teng 2009: 25-30.

[137] Ibid., 25-26.

巛, 廴, 之, 㐅, 尢, 戸, 毋, 毌, 母, 水, 氵, 氺, 爿, 丑, 尹, 巴, 弔, 夬, 肀, 尸, 丯, 瓦, 甘, 石, 示, 礻, 世, 本, 未, 末, 夫, 𡗗, 市, 戊, 戉, 田, 皿, 目, 内, 罒, 电, 甩, 且, 冉, 冊, 凹, 凸, 史, 央, 由, 甲, 申, 兄, 㠯, 册, 业, 瓜, 白, 禾, 朩, 乒, 丘, 乍, 乎, 疒, 立, 必, 永, 癶, 巨, 皮, 矛, 弗, 民, 夬, 丱, 聿, 业, 而, 耳, 臣, 襾, 西, 覀, 吏, 夷, 朿, 甚, 卍, 乐, 戋, 亜, 肉, 月, 虍, 虫, 曲, 曳, 且, 冎, 冎, 竹, ⺮, 缶, 耒, 自, 臼, 𦥑, 舟, 年, 豸, 韦, 缶, 㐅, 镸, 米, 羊, 𦍌, ⺶, 衣, 衤, 州, 糸, 糹, 聿, 𦘒, 艮, 艮, 豕, 車, 酉, 更, 柬, 求, 甫, 匝, 見, 貝, 里, 串, 㡀, 豸, 身, 采, 我, 鳥, 䍃, 𡘓, 言, 亥, 長, 镸, 雨, 事, 東, 疌, 豕, 叀, 直, 亞, 門, 果, 金, 非, 秉, 臾, 承, 革, 禺, 食, 食, 垂, 禹, 重, 為, 飛, 鬥, 鬼, 莫, 畢, 𠀍, 黑, 黽, 熏.

One of the important sources of discrepancies is the theoretical assumption concerning the treatment of graphical variant forms of components. As already mentioned some components in certain distributional contexts change their shapes as a type of graphical accommodation. The variant forms may be treated as identical to the basic form or as separate entities. GF3001 assumes the former option, CNS 11643-2 the latter. The components listed below in parentheses are listed separately in CNS 11643-2 and are treated as the alloghraphs of the basic forms in GF3001.1997:[138]

丿（丿）, 乚（𠃊）[139], 几（几）, 匕（匕）, 𠆢（人）, 儿（儿）, 入（入）, 八（八）, 几（几）, 乂（乂）, 又（又）, 九（九）, 土（土）, 大（大）, 工（工）, 久（久）, 丷（丷）, 女（女）, 小（小）, 巳（巳）, 木（木）, 木（木）, 犬（犬）, 王（王）, 夫（夫）, 屯（屯）, 旡（旡）, 止（止）, 牛（牛）, 文（文）, 火（火）, 尢（尢）, 水（水）, 本（本）, 电（电）, 且（且）, 瓜（瓜）, 生（生）, 禾（禾）, 禾（禾）, 丘（丘）, 立（立）, 皮（皮）, 业（业）, 耳（耳）, 至（至）, 朿（朿）, 耒（耒）, 舟（舟）, 衣（衣）, 更（更）, 柬（柬）, 求（求）, 見（見）, 里（里）, 采（采）, 雨（雨）, 金（金）, 非（非）, 垂（垂）, 重（重）, 熏（熏）, 丂（丂）, 与（与）, 夬（夬）.

Both standards differ in the treatment of particular characters and components. There are elements that are listed as basic components in CNS 11643-2, which are decomposed in GF3001-1997 and vice versa. For example, the character ‘妻’ is a basic element in CNS 11643-2; in GF3001-1997 it is decomposed into ‘肀’ and ‘女’. The character ‘象’ is a basic component in GF3001-1997; in CNS 11643 it is decomposed into ‘⺈’, ‘罒’ and ‘豕’. The different treatments result in different inventories of basic elements.[140]

[138] Ibid., 26.

[139] In GF3001 ‘乚’ and ’𠃊’ are treated as the same component, even though from the etymological perspective the resemblance is accidental.

[140] The examples are borrowed from Chuang & Teng (2009: 26).

45 basic components specific to CNS 11643-2 are listed below (the corresponding compositionally derived components in GF3001-1997 are provided in parentheses, if they exist):[141]

一, 巾, 乇, 生, 用, 耳, 丽, 韭, 首, 齒, 䜌, 勿（丿丿）, 甩（乇）, 至（𠫓）, 以（𠄌）, 互（彑）, 丙（内）, 亥（夕）, 羽（刁）, 重（重）, 堇（重）, 妻（⺕）, 典（曲）, 雈（主）, 隹（主）, 無（卌）, 卑（甶）, 隶（肀）, 面（目）, 馬（⾺）, 烏（⿃）, 兼（𠦃）, 鬲（丫）, 帶（卌）, 曹（曲）, 凷（声）, 棄（枼）, 婁（串）, 庸（肀）, 壺（亞）, 單（甲）, 鼎（爿）, 齊（月, 氏）, 龍（㠯）, 羲（𢦏）.

There are 68 GF3001-1997 specific basic components, 33 of which have compositionally related basic components in CNS 11643 (provided in parentheses):[142]

乁, 玉, 四, 丆, 兀, 习, 亍, 方, 圭, 亡, 四, 上, 柬, 万, 𠂤, 兆, 而, 巨, 下, 予, 勹, 象, 斥, 三, ⺕, 氺, 㓁, 丫, 尺, 八, 夊, 丏, 矢, 冂, 丿丿（勿）, 乇（甩）, 𠫓（至）, 𠄌（以）, 彑（互）, 内（丙）, 夕（亥）, 刁（羽）, 重（重, 堇）, ⺕（妻）, 曲（典）, 主（隹, 雈）, 𠂉（雈）, 卌（無）, 甶（卑）, 肀（卑）, 目（面）, ⾺（馬）, ⿃（烏）, 𠦃（兼）, 丫（鬲）, 卌（帶）, 曲（曹）, 声（凷）, 枼（棄）, 串（婁）, 肀（庸）, 亞（壺）, 甲（單）, 爿（鼎）, 月, 氏（齊）, 㠯（龍）, 𢦏（羲）.

Another set of discrepancies is created by the differences in the standardized shapes of components. CNS 11643-2 contains 15 unique basic components that differ in shape, and they are listed below, with exemplary characters in parentheses (in a succession: CNS 11643-2, GB130001/GF3001):[143]

匕（化化）, 丰（丰丰）, 犮（拔拔）, 夕（炙炙）, **龜**（龜龜）, 儿（奎奎）, 𠫓（充充）, 凡（恐恐）, 龶（害害）, 艹（艾艾）, 龵（邦邦）, 丯（彗彗）, 冖（牽牽）, 冊（冊冊）, 臼（叟叟）.

From the perspective of GF3001-1997 there are 12 unique components listed below, with exemplary characters in parentheses (in succession: GB130001/GF3001, CNS 11643):[144]

[141] Chuang & Teng 2009: 27.

[142] Ibid., 28.

[143] Ibid.

[144] Ibid.

月（前前），止（延延），尒（鰥鰥），于（刊刊），彐（彗彗），户（肩肩），臼（叟叟），乑（聚聚），戢（櫼櫼），册（册册），乛（敢敢），𠀎（華華）.

Finally, the sheer number of characters the both standards are based on is reflected in the number of basic components. GF3001-1997 contains simplified basic components, CNS 11643 does not. Some of the non-simplified components are contained only in GF3001-1997, beause they are parts of characters listed only in GB13000.1. Basic components of both types are listed below:[145]

钅，纟，乡，贝，讠，鸟，车，刂，⺉，马，门，饣，𡗗，见，长，丿丨，龙，丬，𭕄，东，小，东，为，韦，卌，攴，𠃓，戋，乌，乆，乐，𠂇，飞，专，书，乂，𠃊，卄，巛，肃，丰，户，凸，夹，丑，丄，丣，丣，丣，亚，亜，兜，黾，与，冄，先，甲，爲，亩，丗，円，戋，巨，朩，匝，吏，龙，两，市，丑，央，双，龜，乁，无，事，黽，亐，乘，垂，垂，亞，乄，个，夕，壴，西.

The picture is completed with archaic basic components in each standard: ‘弗’ and ‘吏’ in CNS 11643-2 and ‘冊’ in GF3001-1997.[146]

A few other sets of basic components will be addressed further in this study:

- designed at Academia Sinica Chinese Documents Processing Lab as a part of the CDP project, containing 441 basic components, the above introduction is in fact a part of its design - it will be discussed in some detail in Section 5.1.2.;
- Stalph’s set of 485 minimal graphemes, obtained by an analysis of minimal pairs of *jōyō kanji* characters (Section 5.2.3.);
- two related sets retrieved by a recursive analysis of the Ideographic Description Sequences[147] of the CJK Unified Ideographs set:
    - T. Kawabata IDS descriptions: 593 basic components (Section 5.1.1.2. and 7.2.5.2.);
    - CHISE IDS descriptions: 667 basic components (Section 5.1.1.1.);
- retrieved by a recursive analysis of the subsets (especially the Big5 set) of the Ideographic Description Sequences of the CJK Unified Ideographs (Chapter 7).

[145] Ibid., 29.

[146] Ibid.

[147] See Section 5.1.1.

4.4.3. 說文解字 *Shuōwén Jiězì* (SWJZ) and the modern components

The lingering influence of SWJZ often clouds the relationship between the structure of modern Chinese characters in regular script and that of small seal script meticulously analyzed by Xǔ Shèn. As Cook points out,[148] the characteorology is often equated with the study of ancient scripts (small seal and earlier), even if characterologists do not confuse the point of reference and meaning of their studies. It is more popularly believed that the components of regular script components are confused with the components found in SWJZ. The fact that the ancient script characters are discussed with the use of the regular script does not help to resolve the confusion. This section is aimed at pinpointing the differences other than mere shapes between the SWJZ component inventory and the modern one discussed above. The types of differences are similar to those between any two different sets of components (e.g. GF3001 and CNS 11643-2 discussed above); i.e. some components are basic in one set, decomposable in another and vice versa. The examples are borrowed from the Chuang & Teng study:[149]

- some components are basic in SWJZ, and decompose in modern script. For example, the small seal character ‘易’ is a non-decomposable depiction of a chameleon or a lizard – the regular script counterpart ‘易’ decomposes into ‘日’ and ‘勿’, while the small seal character ‘癸’ is a non-decomposable depiction of a water flowing from four directions ‘癸’. The regular script counterpart decomposes into ‘癶’ and ‘天’.
- some components are basic in modern script and decompose in SWJZ. For example, the regular script character ‘束’ is a basic component. The small seal counterpart ‘束’ decomposes into ‘木’ and ‘口’. The regular script character ‘東’ is a basic component, while the small seal counterpart ‘東’ decomposes into ‘日’ and ‘木’.

4.4.4. Functional types of components

The classification of components according to their ‘inner’ function (pertaining to semantic and phonetic properties) in modern regular script characters proposed in Chuang & Teng (2009) was very briefly introduced in Section … The authors did not elaborate on the proposed classes in much detail. An alternative proposal, simpler and more intuitive to apply, was put forward in Li & Kang (1993), Kang (1993) and Su

[148] Cook 2003: 22-23.

[149] Chuang & Teng 2009: 24-25.

(2001). The general idea in Chuang & Teng (2009) is similar to that of these authors, but there are only three functional ('internal') types of components called 字符 *zìfú*:

- significs (意符 *yìfú*) – components contributing to the meaning of a character, e.g. '火' *huǒ* 'fire' in: 燈 *dēng* 'lamp', 燒 *shāo* 'roast', 炎 *yán* 'scorching hot', 灼 *huó* 'burn';
- phonetics (音符 *yīnfú*) – components contributing to the pronunciation of a character,[150] e.g. '黃' *huáng* in: 璜 *huáng*, 磺 *huáng*, 簧 *huáng*, 熿 *huáng*, 獚 *huáng*, 癀 *huáng*;
- symbolics (記號 *jìhào*) – components having a purely symbolic function, contributing to neither the meaning nor the pronunciation of the whole character, e.g. in '火' *huǒ* 'fire' in: 炫 *xuàn* 'show off', 煩 *fán* 'vexed'; '黃' *huáng* 'yellow' in: 橫 *héng* 'horizontal', 廣 *guǎng* 'wide'.

4.4.4.1. Modern 'six categories'

The modern take on the structural classification of characters in terms of functional classes of components is modeled on the SWJZ's 'six categories'. The number of distinguished classes is the same as in SWJZ, and the general criteria are also similar. The six classes are distinguished on the basis of functions of the immediate constituents. The fundamental difference compared to SWJZ is the reference to the modern regular script (楷書 *kǎishū*), instead of to the small seal script (小篆 *xiǎozhuàn*):[151]

- ideographic (會意字 *huìyìzì*) – defined in the same way as the equivalent class in SWJZ, e.g. 析 *xī* 'cut up, analyse' ('wood' + 'axe'), 休 *xiū* 'rest' ('person' + 'tree'), 磊 *lěi* 'rampart' (3 x 'stone');
- picto-phonetic (形聲字 *xíngshēngzì*) – also defined in the same way as in SWJZ, e.g.癀 *huáng* 'jaundice' (疒 'illness' + 黃 *huáng*), 娶 *qǔ* 'marry a woman' (女 'woman' + 取 *qǔ* 'take'), 甥 *shēng* 'nephew' (男 'male' + 生 *shēng*);
- signific-symbolic (半意符半記號 *bànyìfúbànjìhàozì* 'half signific, half symbolic characters') – characters having 意符 and 記號 as immediate components, e.g. 燒 *shāo* 'roast' (火 'fire' + 堯 *yáo*), 咬 *yǎo* 'bite' (口 'mouth' + 交 *jiāo*). Due to the evolutionary changes in pronunciation, independent from the evolution of script, many characters classified in SWJZ as picto-phonetic underwent a category shift to become 半意符半記號;
- phono-symbolic (半音符半記號 *bànyīnfúbànjìhàozì* 'half phonetic, half symbolic characters') – characters having 音符 and 記號 as immediate components,

[150] Su (2001: 93) does not require the tone of the syllable represented by the whole character to be identical to the tone of the syllable representing the 音符 *yīnfú* component.

[151] Ibid., 94-101.

e.g. 炫 *xuàn* ‘show off’ (‘火’ ‘fire’ + 玄 *xuán*), 派 *pài* ‘faction’ (氵’water’ + 𠂢*pài*), 笨 *bèn* ‘stupid’ (竹 ‘bamboo’ + 本 *běn*). Similarly to the previous category, many picto-phonetic characters in SWJZ underwent independent evolution, in this case, an evolution of meaning to become 半音符半記號;

- simple symbolic characters (獨體記號字 *dútǐjìhàozì*) – non-decomposable characters comprised of one 記號 component, e.g. 子 *zǐ* ‘child’, 矢 *shǐ* ‘arrow’, 舟 *zhōu* ‘boat’. Important sources of 獨體記號字 are the characters classified in SWJZ as pictographs (象形字 *xiàngxíngzì*) and phonetic loans (假借字 *jiǎjièzì*);
- complex symbolic characters (合體記號字 *hétǐjìhàozì*) – characters having only 記號 as immediate components, e.g. 特 *tè* (牛 *niú* ‘bull’ + 寺 *sì* ‘temple’), 頭 *tóu* ‘head’ (豆 *dòu* ‘bean’ + 頁 *yè* ‘page’), 穌 *sū* ‘revive’ (魚 *yú* ‘fish’ + 禾 *hé* ‘grain’). This type evolved mostly from small seal script picto-phonetic characters whose 意符 and 音符 components became 記號. Also, the ancient pictographs and ideographs are possible sources of 合體記號字.

In contrast to Xǔ Shèn’s classification, all six categories are homogenous, meaning that they are devoted to the structural properties of characters. The classification is rather coarse and taken literally it could not classify the non-ideographic characters with more than two immediate components. It is not difficult though to look past the SWJZ-modelled classes and incorporate the totality of characters into the modern classification of characters based on the internal functions of components.

The topic of character classification according to their inner structure is not the focus of this study, but this brief introduction of the modern classification of characters serves the practical purpose of demonstrating three things – the extent of evolutionary changes in the script, the limitations of the traditional classification, and the modern structure of characters from the functional perspective. A more detailed account of these problems can be found in Qian (1990).

4.5. Strokes

In regular script (楷書 *kǎishū*) the atomic units of constituency are the strokes.[152] There are two one-stroke characters and one-stroke basic components, but the levels of analysis should not be confused.

Even at the atomic level of character structure there are theoretical options of segmentation that result in different stroke systems. The decomposition of individual

[152] The ancient characters (from oracle bone inscriptions to small seal script) usually are not decomposed into conventional strokes or strokes in a narrow sense (Su 2001: 65). There are, however, notable exceptions. For example, Cook (2004) in his extensive study of SWJZ provides the ordering of SWJZ radicals by the count order of strokes in a broad sense. Also see Gao (1999): 11; Song & Jia (2003); and Huang (2006).

characters into strokes is not disputable, at least in a lexicographic approach – for the purpose of sorting and ordering the characters fine distinctions between types of strokes are not necessary. Strokes are simply the parts of a character written in a conventional way without the writing instrument breaking contact with the writing surface; that is to say, strokes are the continuous parts of characters from the perspective of the writing process.

Classifications of strokes for the purpose of fine structural decomposition of characters usually have a few features in common:

- distinguish between basic and combining (subordinated) strokes;
- inclusion of certain stroke types;
- some strokes types are always classified as the same type (basic or combining).

The traditional views on the inventory of strokes are supplemented by the advancements in information processing, thus creating a rather complicated situation.

From the perspective of traditional characterology there are 5 basic stroke types used for classificatory purposes: horizontal 橫 *héng* (一), vertical 豎 *shù* (丨), slant 撇 *piě* (丿), dot 點 *diǎn* (丶), and bend 折 *zhé* (乛). This set is a general class system into which all other strokes are placed. This stroke classification system is also used in stroke-based input methods,[153] and may be used as a primary or secondary criterion for the ordering of characters and components. The five categories are the basic stroke features that abstractly describe the graphical and structural properties of strokes and are used for categorizing strokes. A set of basic strokes is a basis for decomposition and/or generation of compound strokes.

Tab. 4.2 Basic strokes

Chinese name	Description	Shape	Symbol	Examples
橫 héng	horizontal	一	H	三，上
豎 shù	vertical	丨	S	丰，非
撇 piě	slant	丿	P	少，千
點 diǎn	dot	丶	D	為，血
捺 nà	right falling	㇏	N	尺，史
提 tí	rising	㇀	T	孑，刁

The basic shapes of individual strokes, which are the minimal units of character decomposition, are categorized in different ways in different systems that may or may not include secondary basic shapes. For the sake of simplicity the system chosen for

[153] This refers to both Wubi (五筆字型輸入法 *wǔbǐ zìxíng shūrùfǎ*) and Wubihua (五笔画输入法 *wǔ bǐhuà shūrùfǎ*) methods.

presentation here is a ‘distilled’ version of a few notable systems.[154] It is presented in the form of three tables containing the lists of basic strokes (Tab. 4.2), combining features (Tab. 4.3) and compound strokes (Tab. 4.4).

The second component necessary for generating coumpound strokes are the combining features. A combining feature is not a structural part of a stroke, but rather it defines the spatial orientation or direction of one basic stroke in relation to another with which it is connected. Additionally, a combining feature can be directly connected with the shape of a basic stroke, or changes in shape of that basic stroke.

Tab. 4.3 Combining features of strokes

Chinese name	Description	Symbol
折 zhé	bend	Z
鉤 gōu	hook	G
彎 wān	curve	W
左/右 zuǒ/yòu	left/right	Z/Y
扁 biǎn	flat	B

Compound strokes result from the combination of basic strokes. Their composition in relation to each other is described with the use of combining features, as it is shown in Tab. 4.4.

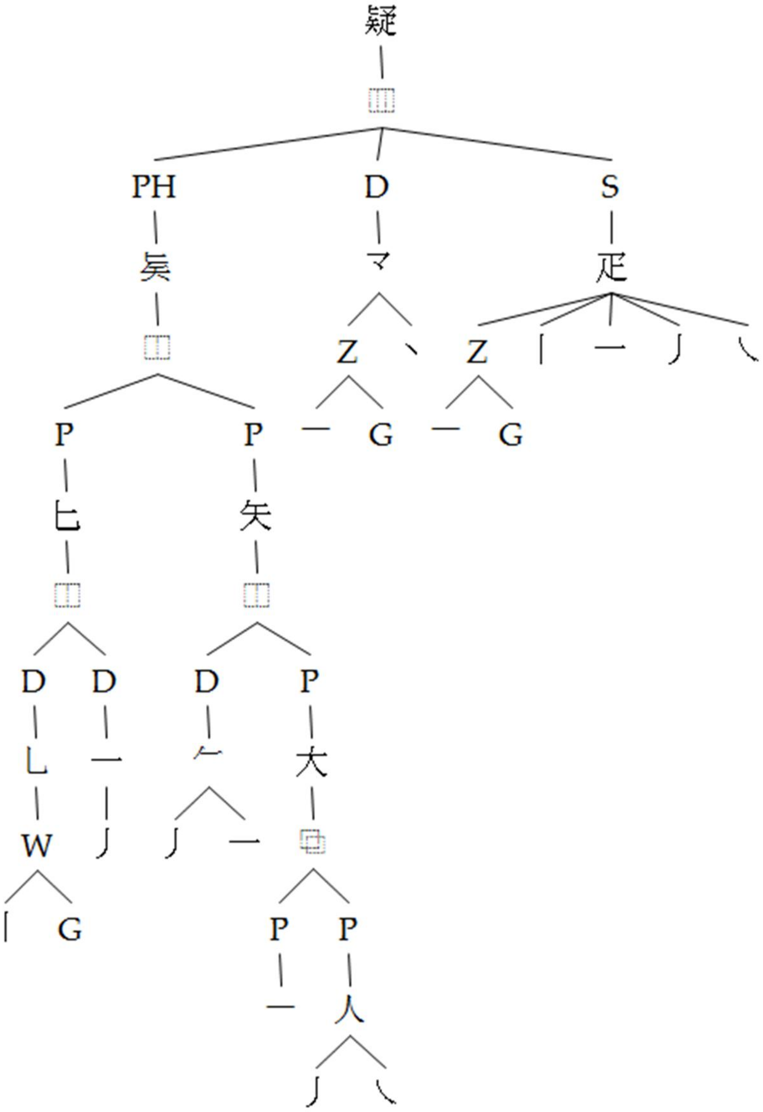
The combining features are in parentheses. The count of elements outside the parentheses is also the count of basic elements in a compound stroke. There is one exception to the principle that combining features are not structural parts of strokes – the ‘hook’ is always manifested structurally, and it is not classified as a basic stroke, because it is a ‘bound’ stroke, meaning it does not occurr in isolation. For that reason, in the system presented here, it cannot be said that compound strokes are composed of basic strokes only, which is the same reason the ‘hook’ is not in parentheses. The Latin alphabet notation is more problematic – some compound strokes are composed of the same basic strokes written in the same order. In such cases, at least one combining feature should be preserved in the notation to keep the distinction (e.g. SG-WSG, SH-SHZ-SWH). For the sake of consistency and simplicity *zhé* is the default combining feature not indicated in the alphabetic notation; in other cases the combining features should be indicated by a corresponding letter.

[154] Su 2001: 69; Sun 2006; GF 2001-2001; CNS 11643-3; http://zh.wikipedia.org/zh/%E7%AC%94%E7%94%BB.

Tab. 4.4 Compound strokes

Stroke description	Stroke shape	Symbol	Examples
橫 (折) (豎)	㇕	HS	口，回
橫(折)撇	㇇	HP	又，水
橫鉤	㇖	HG	定，了
豎鉤	㇚	SG	小，事
豎(折)橫	㇗	SH	互，山
豎(彎)橫(左)	┘	SWHZ	肅
豎(彎)橫(右)	㇄	SWH	忙，四
豎(折)提	㇙	ST	民，良
撇(折)橫	㇜	PH	公，累
撇(折)點	㇛	PD	災，經
撇鉤	ノ	PG	乂
(彎)豎鉤	㇁	WSG	狗，家
捺鉤	㇂	NG	戰，我
(扁)捺鉤	㇃	BNG	必，心
橫(折)豎(折)橫	㇅	HSH	凹
橫(折)豎(彎)橫	㇍	HSWH	躲，殳
橫(折)豎(折)提	㇊	HST	鳩
橫(折)豎鉤	㇆(𠃌)	HSG	刁，月
橫(折)捺鉤	㇈	HNG	風，飛
豎(折)橫(折)豎	㇞	SHS	吳，亞
豎(折)橫(折)撇	ㄣ	SHP	㐆
豎(彎)橫鉤	㇟	SWHG	已，記
橫(折)豎(折)橫(折)豎	㇎	HSHS	凸
橫(折)豎(折)橫(折)撇	㇋	HSHP	及
橫(折)豎(彎)橫鉤	乙	HSWHG	乞
橫(折)撇(折)(彎)豎鉤	㇌	HPWSG	都，隊
豎(折)橫(折)豎鉤	㇉	SHSG	號，弓
橫(折)豎(折)橫(折)豎鉤	㇡	HSHSG	扔，孕

The relationship of basic and compound strokes is analogous to the relationaship between basic and complex components. The decomposition of characters into strokes is also multilayered, but the number of layers is limited to two. In terms of stroke-constituents there are no intermediate layers of decomposition. Stroke components are either basic (basic strokes) or complex (compound strokes) – characters that are composed only of basic strokes have one layer of decomposition (immediate stroke-components), while characters composed of at least one compound stroke have two layers.

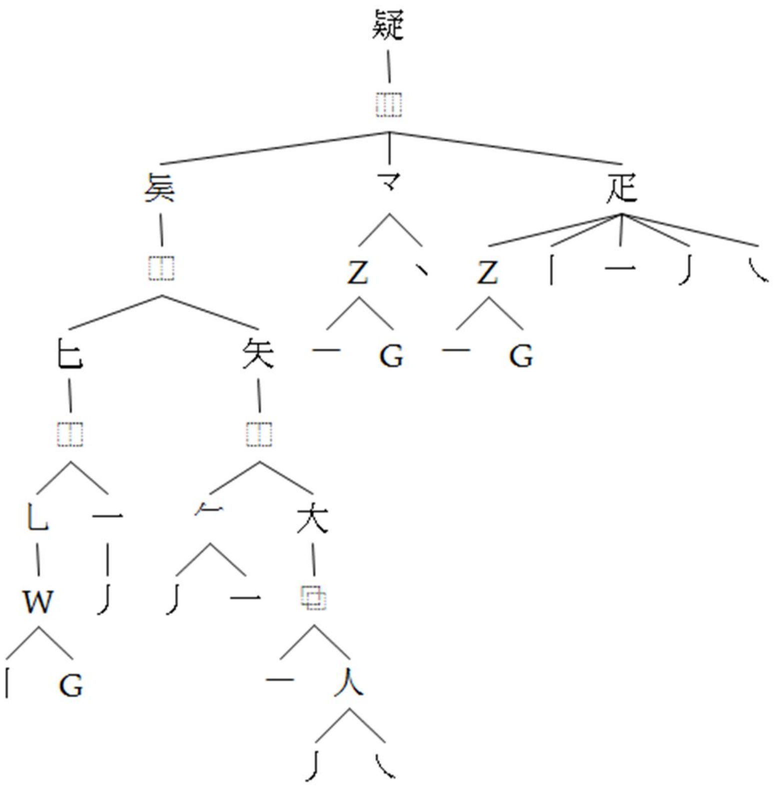


[疑[⿲[PH[𠤕[⿰[P[匕[⿰[D[乚[W[丨][G]]]][D [一[丿]]]]]][P[矢[⿰[D[𠂉[丿][一]]][P[大[⿻[P[一]][P[人[丿][㇏]]]]]]]]]]]][D[マ[Z[一][G]][丶]]][S [𤴔[Z[一][G]][丨][一][丿][㇏]]]]]

The above represenation is the complete decomposition of the character '疑', but because it contains heterogeneous units, a few additional comments are necessary. To some degree the representation can be compared to other graphic explanations, including the morphological, syllabic, phonemic, and phonetic levels in the syntactic IC-analysis tree, or in the X-bar notation. Units of Chinese script have their own specifics and all analogies to the levels of speech have severe limitations. In this case it seems reasonable to include the seemingly heterogeneous levels of analysis, but relevant for a purpose. Following Su's terminology[155] it can be reasonably argued that the representations of character component structure refer to either external or internal composi-

[155] Su 2001: 75 and 92-93.

tion. The internal composition that includes etymological and functional properties of components is not subjects of interest for this study any more than absolutely necessary – the representation of internal composition should contain the component structure from the etymological perspective and functional information. The stroke components should be excluded as the stroke level pertains to the purely graphical, i.e. that which pertains to external analysis. Etymological analysis dates back to the seal script, if not further, which is not decomposable into strokes. External analysis that is based purely on graphical criteria should only be concerned with properties relevant to the graphical structure of characters. In some cases, when internal and external analyses render different componential structures, it is impossible to mix the two types of representations. To sum up, from the perspective of external structure the full representation of the decomposition of the character '疑' has the following form:



[疑[⿲[𠤕[⿰[匕[⿰[乚[W[丨][G]]][一[丿]]]][矢[⿱[𠂉[丿][一]][大[⿻[一][人[丿][㇏]]]]]]]][マ[Z[一][G]][丶]][𤴔[Z[一][G]][丨][一][丿][㇏]]]]

Another problem is created by changing the shape of components. In the above example, the last stroke (lower right corner) of the component '矢' is in fact the 'dot' stroke '丶'. The analysis shows the decomposition of the character '矢'. The handling of cases like this should correspond to the treatment of variant forms of components.

The Unicode Consortium introduced its own system of strokes and their description.[156] In principle, it is similar to the one proposed above. Due to the limitation of space it will not be introduced here, but suffice it to say that in the most recent 6.2 version the Unicode system consists of 36 strokes. This system is potentially important in the context of graphotactic analysis of Chinese script. At this point of development the coding of Chinese characters does not contain stroke information. Moreover, to the best knowledge of the author, there is no project currently underway that is aimed at providing information on strokes to IDS descriptions of characters (T. Kawabata, CHISE). The IDS descriptions, however, have reached the stage where introducing the stroke components of each character contained in the CJK Unified Ideographs set requires relatively little effort, and it is conceivable that the Unicode Consortium will eventually mandate storing this type of metadata. It is even possible that the whole procedure could be limited to the assignment of component strokes to a set of a few hundred basic components, and then the IDS descriptions could be used to extrapolate the component strokes to any set of characters. The limitations of this study do not allow the conduct of the graphotactic analysis in terms of strokes, but this is a possibility that will certainly be explored in the future.

Another insight into the stroke component of Chinese characters is offered by the Wenlin CDL[157]. This system is designed for constructing the entirety of Chinese characters from strokes. Bishop & Cook (2003) estimate that "*less than fifty stroke types is sufficient for the construction of practically all characters in a modern printed style*",[158] which is probably a good approximation of the number of finely distinguished stroke elements at the atomic level of the structure of *hànzi*.

4.6. Simplification of characters

The simplified forms have existed at all stages of the evolution of Chinese script, as 'vulgar' and unofficial forms of the more complex characters promoted by authorities.[159] On the other hand, simplification is, to some degree, inherent in standardization.[160] The current situation in Chinese-writing countries with regard to simplification standards is rather complicated. While not completely irrelevant to the subject of this study, simplification standards will be introduced only with regard to China and Taiwan by discussing facts that are helpful to describe the extent of the difference between the two sets of characters. The ongoing contentious debate over character sim-

[156] http://www.unicode.org/charts/PDF/U31C0.pdf

[157] See Section 5.1.3.

[158] Bishop & Cook 2003: 2.

[159] Zhao & Baldauf 2008: 30.

[160] An extensive study of the standardization of Chinese script can be found in Zhao & Baldauf (2008).

plification, with its strong political overtones, is ignored here completely, as well as general research issues related to the problem of simplification (historical, cultural, and social aspects, literacy, and writing acquisition).[161] In this section of the book the discussion of the problem will be limited to the basic facts related to the number of simplified characters and to the the classification of simplification mechanisms. The quantitative aspects of simplification are discussed in more detail in Chapters 6 and 7.

4.6.1. Extent of simplification

The official list of characters simplified as a result of the 1956 simplification scheme (简化字总表 *jiǎnhuàzì zǒngbiǎo* 'General List of Simplified Characters'), was published in 1964, and consists of 2,235 characters (2,249 items)[162] which are internally classified into three sublists:

– the first list contains 350 items – characters that can only be used as standalone characters, i.e. characters that are not used as components of other characters, e.g. 处, which is is a simplified equivalent of 處 . However, a character having 處 as a side component must retain it in the unchanged form: 擄 *jù* 'evidence'.

– the second list contains 146 items that are used as components. The list is further subcategorized into two lists:
  - There are 132 elements that can be used as standalone characters and components of other characters, e.g. 带 ,which is a simplified equivalent of the standalone character 帶 *dài* 'belt' or 'carry'. The character is also an equivalent of the same element used as a side component 滞, which is the simplified form of 滯 *zhì* 'stagnant'.
  - There are 14 elements that are used only as side components, e.g. 讠, which replaces the traditional 言, and is used only as a side component, as in 说 (說) 'speak'. The standalone character 言 *yán* 'speech' is not simplified. This part lists side components, not characters.

– the third list contains 1,753 items – characters simplified as a result of applying the simplified side components on the second list, e.g. 趙 - 赵, 漢 - 汉, 話 -话.

The 'General List of Simplified Characters' is a part of a national standard and may give the impression of an exhaustive list, but referencing the footnotes of the 1986 ver-

[161] Those aspects are thoroughly covered in Zhao & Baldauf (2008); DeFrancis (1984b); Taylor & Taylor (1995) and in Chen (2004).

[162] The last version was published in 1986. The full list is easily accesible online, at: http://www.zsjy.gov.cn/yywz/yypg/gfwj/17.htm.

sion of the list should be made to appreciate the real extent of simplification. It is explicitly stated that the third list is not intended as an exhaustive set for the set of all characters and that for practical reasons it does not need to be. The standard character set for simplification was determined by the 1962 edition of the 新华字典 *Xīnhuá Zìdiǎn* 'New China Dictionary', for which about 8,000 characters were collected. It is explicitly stated in the 'General List of Simplified Characters' that the mechanism of simplification of the 1,753 characters may be extended, if needed, to other characters as well. The extension of the simplification method to the CJK Unified Ideographs set would increase the number dramatically. For example, in Kawabata's IDS database there are 1,756 elements containing just the '金' element. This number is not equivalent to the number of characters, but it is very close. The Wenlin database shows 1,165 characters with this component. The Wenlin database approximately reflects the scale of the total numerical increase of the simplified characters.[163]

Standardization of character forms, number and order of strokes was carried out alongside the simplification of characters, but these issues will not be discussed here in any greater detail than has already been done.

4.6.2. Simplification methods

Strictly from the purely graphical perspective, simplification of characters can be described as a replacement of a character or its part by a graphically simpler one. For the purpose of this study an extensive survey of the simplification mechanisms is not necessary, but a brief introduction is more than justified. The types of graphic changes in the process of simplification are pertinent to the topic of this book, but it is sufficient to present the basic mechanisms without excessive details.

The methods of simplification are motivated by heterogeneous criteria, and can be classified into the following categories:[164]

Substitution

Graphical (unmotivated):

- component replacement – a simpler component replaces a more complex one in a character structure, e.g. 趙 - 赵, 漢 - 汉;
- a part replacing the whole – a part (not necessarily a component) of a character is chosen to represent the whole character, e.g. 兒 - 儿, 術 - 术, 奪 - 夺;

[163] It should be taken into account that '金' is a very frequent component.

[164] Yin & Rohsenow 1994: 107-112; Zhao & Baldauf 2008: 45-46.

- outline preservation – only a general shape of a character is preserved, while other elements are deleted,[165] e.g. 齊 - 齐, 變 - 变.

Phonetic

- a simpler component replaces a more complex phonetic component with the same or similar pronunciation,[166] e.g. 種 - 种, 殲 - 歼 *jiān*;
- a simpler character replaces a more complex character, one which is fully or partially homonymic, e.g. , 後 - 后, 隻 - 只.

Stylistic

- using a simpler character of a different writing style (typically grass or running style characters to replace a more complex one, e.g. 長 - 长, 為 - 为.

Creation

- coining a simpler character to replace a more complex one is the least common method, and typically picto-phonetic and compound ideograph methods are used, e.g. 雙 - 双, 體 - 体.[167]

4.6.3. Simplification of components

In the official set of 2,235 simplified characters there are 41 systematically simplified components, including 14 side components mentioned in Section 4.6.1. The full list is provided in Table 4.5.[168]

Chuang & Teng (2009) present the full statistics of simplified component sets with references to the CDP set. According to them the characters in the official simplified set are composed from a set totaling 1,122 simplified components: 367 basic and 755 compound. 326 of the basic components are common with the BIG5 (CDP) inventory.[169] 41 components listed in the table are the unique simplified basic components.

[165] In a sense, it is the opposite of the 'part replacing whole' mechanism.

[166] This is a special case of component replacement.

[167] In most cases, as Yin & Rohsenow (1994: 111) point out, the simplified forms were not newly created, but characters already in unofficial use, some even for centuries.

[168] Ibid., 73.

[169] Chuang & Teng 2009: 72.

Tab. 4.5 The list of simplified components

No.	Component	Example	No.	Component	Example	No.	Component	Example
1	〢	坚	15	𠂉	农	29	见	现觅觃
2	刂	师	16	彐	当寻归	30	长	张
3	丿丨	啸	17	𠃓	场疡觞	31	鸟	岛
4	讠	计狱	18	飞	飞	32	为	伪
5	𠂇	泼	19	习	习	33	书	书
6	ス	劲	20	纟	级辫辔	34	戋	栈笺盏
7	马	马	21	𡗗	寿	35	龙	拢宠垄
8	乡	丝	22	韦	伟围苇	36	东	冻岽鸫
9	戈	烧	23	专	传	37	东	练
10	丬	壮寝	24	卌	带	38	钅	衔针
11	饣	饭	25	车	军库轮	39	乐	烁
12	乌	乌	26	小	变	40	甲	单
13	门	们问	27	内	锅纳	41	两	满俩魉
14	⺍	学应	28	贝	贪则坝			

5. Models of Chinese character descriptions

The discussion in this chapter concentates on the survey of the most notable proposals of more or less formal descriptions of the structure of Chinese characters. The chapter is divided into three main sections devoted to the languages of character description, graphematical treatments of Chinese script, and approaches similar to the graphotactic framework.

5.1. Character Description Language (CDL)[170] projects

Character Description Languages (CDL) are linguistic systems aimed at describing Chinese characters using regular expressions constructed in accordance with more or less formal grammars. Three major CDL projects will be introduced in the following sections. The final subsection presents a related project that for various reasons is less significant to the discussion in this study.

5.1.1. Ideographic Description Sequence (IDS)

The Unicode Standard includes a staggering number of more than 75 thousand characters in the CJK Unified Ideographs block, although there are many characters that remain unencoded. As it was already mentioned in Section 2.2.2, that the Unicode developers remedied this problem with an ideographic description sequence (IDS), which is a syntactic device for characters description, aimed primarily to represent the unencoded *hànzi* by the means of 12 ideographic description characters.[171] In other words IDS works on the premises that all characters may be broken down into more primitive parts, all of which are encoded, and that there are regularities in the structural formation of characters that can be captured by a small number of syntactic expressions.

The purposeful design of the IDS system comes with a price. The 12 IDCs allow a description of all possible structural arrangements, but only in the sense that they render the desired character, and in some cases the descriptions are merely graphical

[170] The term 'character description language' was in fact used for one specific language created by Tom Bishop at the Wenlin Institute. The term is very handy and will be used in this book as a generic term for any language serving a similar purpose. The original CDL will be referred to as 'Wenlin CDL'.

[171] IDCs were introduced in Section 4.3. The reference materials for this section can be found in the official Unicode documentation, Chapter 12 on the East Asian scripts, at:
http://www.unicode.org/versions /Unicode6.2.0/ ch12.pdf.

translations of the correct interpretations. IDS lacks the descriptors of three-partite (other than ⿲ and ⿳) and four-partite structures. It is possible to represent the structure of 森 as ⿱木林 , which is the actual IDS description of this character. Such a rendering, however, is forced by the limitations of the system. For example, the repertoire of IDCs allows the representation of the character 𣊫 as ⿰昌昌, or as ⿱昍昍, and analogously, 𠈌 as ⿱从从, or as⿰仌仌.[172] A quick survey of various sources reveals different explanations for the layout of exemplary characters. The Wenlin database decomposes 𣊫 into four 日 without any structural descriptor. 𠈌 is decomposed as ⸬人 (etymologically, where ‘⸬’ is a CDP descriptor[173]) and 仌仌 in Wenlin CDL components,[174] while the Chinese Text Project website describes the structure of 𠈌 as top-bottom (= ⿱从从).[175] Typically the correct representation of characters like the discussed pair is the CDP descriptor ‘⸬’, which IDS lacks.

Although IDSes use a formal language to represent the structure of characters, they are not a formal way of encoding items into the Unicode Standard – “*ideographic descriptions are more akin to the English phrase “an ‘e’ with an acute accent on it” than to the character sequence <U+0065, U+0301>*”.[176]

From the formal perspective the syntax used by IDS is known as the Backus-Naur Form (BNF) notation technique. BNF is a tool for describing the syntax of a context-free grammar. A full description of the IDS grammar (G) may be found in the ‘IRG Principles and Procedures’:[177]

Let G = {Σ, N, P, S}, where…

- Σ: the set of terminal symbols including all coded radicals, coded ideographs, and the 12 IDCs.
- N: the set of 5 non-terminal symbols

N = {IDS, IDS1, Binary_Symbol, Ternary_Symbol, CDC[178]}

- S = {IDS}, which is the start symbol of the grammar
- P: a set of rewrite rules

The following is the set of rewriting rules P:

[172] The actual IDSs for the exemplary characters found in Kawabata’s IDS database are: 𣊫⿰昌昌, 𠈌⿱从从⿰仌仌(alternative structures).

[173] See Section 5.1.3.

[174] See Section 5.1.3.

[175] http://ctext.org/dictionary.pl?if=en&char=%F0%A0%88%8C.

[176] http://www.unicode.org/versions/Unicode6.2.0/ch12.pdf.

[177] http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg34/IRGN1646Confirmed.doc.
The Ideographic Rapporteur Group (IRG), previously called the CJK Joint Research Group, is an advisory committee that is in fact directing the development (e.g. controlling character additions, maintaining the standard) of the CJK Unified Ideographs.

[178] Stands for Character Description Components.

- IDS::=<Binary_Symbol><IDS1><IDS1>|<Ternary_Symbol><IDS1><IDS1><IDS1>
- <IDS1> :: = <IDS> | <CDC>
- <CDC> :: = coded_ideograph | coded_radical | coded_component
- <Binary_Symbol> :: I = ⿰|⿱ | ⿴ | ⿵ | ⿶ | ⿷ | ⿸ | ⿹ | ⿺ |
- <Ternary_Symbol> :: = ⿲ | ⿳

(The IDCs are not the part of the Character Description Components.)

IDS is particularly important for the present study not because of its design features, but for the amount of graphotactic data that is available through it. The IDS descriptions of CJK Unified Ideographs will be the basis for the graphotactic analysis of the selected sets of characters. The data on the relative arrangement of components (IDCs) is not exploited in this study, but it is a tempting perspective for distributional studies of character components.

This section focuses on introducing two projects involving the IDS descriptions of Chinese characters. Both are related to the Unihan database, but differ in relevant details. Both are candidates as a source of graphotactic information contained in IDS descriptions. The information on the component inventory used in the IDSes in both cases is meager, at best. Fortunately, this is not a significant problem, since the component sets can be extracted automatically. It is reasonable to assume that the basic set of components used in both databases is the GF 3001-1997 standard, because that standard is used in the Unicode.[179] Relevant differences regarding the component sets are discussed below.

5.1.1.1. Character Information Service Environment (CHISE)

The Character Information Service Environment is one of the largest open-source projects aiming to resolve the problems with information processing of different types of scripts.[180] The project consists of several sub-projects, including those focused on Chinese character processing. CHISE does not operate on characters defined as the code points, but rather on a prescribed set of features assigned to each character. The features include structural, phonetic and semantic information, and also CCS code points used to facilitate the information exchange in environments with the most widespread coded characters sets – Unicode and ISO.[181] For the purposes of the present project the IDS part of CHISE is most relevant. The character descriptions contain

[179] Zhang 2008.

[180] http://www.chise.org/.

[181] Morioka 2008: 148.

IDS data of the CJK Unified Ideographs coded characters set, and reference to the Unicode code points. Samples of IDS are introduced below to exemplify the format of CHISE structural descriptions:

U+4EA7	产	⿱⿱亠丷厂
U+4EA8	亨	⿱⿱亠口了
U+4EA9	亩	⿱亠田
U+4EAA	亪	⿱亦乁
U+4EAB	享	⿱⿱亠口子
U+4EAC	京	⿱⿱亠口小
U+4EAD	亭	⿱⿳亠口冖丁
U+4EAE	亮	⿱⿳亠口冖几
U+4EAF	亯	⿱⿱亠口日
U+4EB0	亰	⿱亠&CDP-8CED;

U+65D4	旔	⿰方建
U+65D5	旕	⿱於⿰口匕
U+65D6	旖	⿸⿰方𠂉奇
U+65D7	旗	⿸⿰方𠂉其

U-0002303F	𣀿	⿰⿱⿰⿳夕⿱一夕巾⿳夕⿱一夕巾火攵
U-00023040	𣁀	⿰⿱咸鬵攴
U-00023041	𣁁	⿱&GT-00458;文
U-00023042	𣁂	⿱&GT-K00305;&GT-17008;
U-00023043	𣁃	⿱文⿻了&GT-K00059;

The CHISE-IDS database contains 74,568 characters from the CJK Unified Ideographs basic block and extensions A, B, C and D, including the 214 indexing radicals. The CHISE-IDS uses components from a few different inventories of constituents, and this is one of the practical reasons for selecting the alternative KDP database for the graphotactic analysis over CHISE (the reasons are discussed in more details in the next section). Apart from the basic set of CJK components (appearing in graphical form), CHISE-IDS also uses CDP,[182] CBETA and *Konjaku Mojikyo* (今昔文字鏡)[183] components. This makes the results of graphotactic analysis substantially more difficult (but not impossible) to interpret. Nonetheless, what really prevents CHISE-IDS-based anal-

[182] See Section 5.1.2.

[183] Morioka 2008: 156.

ysis at this point is the fact that only CDP components are well-documented, with the remaining sets being rather obscure in this respect.[184]

5.1.1.2. Kawabata's Kanji Database Project (KDP)[185]

The content of the KDP database is very similar to that of the CHISE-IDS – the essential part consists of the Unihan set of characters (CJK Unified Ideographs basic block and extensions A, B, C and D). The raw unedited database contains 76,066 items, not all of which are suitable to be included in the graphotactic analysis. The exclusion of compatibility ideographs (relevant for information processing only), supplemental radicals, and 695 CDP components (used for decomposition, and assumed to be the non-character items) results in a total of 74,810 items. Further refinement of the database is probably possible, but not necessary from a statistical point of view.

The edited KDP database is the primary basis for all graphotactic investigations designed in this study (except for the Cangjie analysis – see Section 7.1.). Chapter 7 presents the result of an analysis of the whole KDP and its three selected subsets. Potentially there is a very large number of subsets that could be investigated, but there seems to be a limited number of possibilities that substantially contribute to the understanding of Chinese script. The main discussion of the analyzed sets and their relation to KDP is continued in Chapter 7. At this point it is important to stress that the edited version of the KDP database, containing a minimalized number of non-character items, is a basic source of graphotactic data. The immediate components of Chinese characters are extracted directly from KDP; the procedure for extracting basic components is described in Chapter 7.

The KDP database in very similar to CHISE in terms of the character inventory, but differs in some relevant respects from the IDS in terms of content. From the graphotactic perspective the most relevant difference is the inventory of components. To illustrate the differences between KDP and CHISE, the examples chosen to represent the KDP database are exactly the same as those in the CHISE section:

U+4EA7	产	⿱&CDP-8BAE;厂
U+4EA8	亨	⿳亠口了
U+4EA9	亩	⿱亠田
U+4EAA	亦	⿱亦乀
U+4EAB	享	⿳亠口子
U+4EAC	京	⿳亠口小

[184] The author did not succeed in gathering more details.

[185] http://kanji-database.sourceforge.net/.

U+4EAD	亭	⿱&CDP-8C4D; 丁	
U+4EAE	亮	⿱&CDP-8C4D; 几 [G]	⿱&CDP-8C4D; 儿 [TJK]
U+4EAF	亯	⿳亠口日	
U+4EB0	京	⿳亠日小	
U+65D4	旔	⿰方建	
U+65D5	旕	⿱於叱	
U+65D6	旖	⿰方⿱𠂉奇	
U+65D7	旗	⿰方&CDP-8CFC;	
U+2303F	𣀿	⿰⿱⿰⿳夕歹巾⿳夕歹巾火攵	
U+23040	𣁀	⿰⿱咸鬲攴	
U+23041	𣁁	⿱亽文	
U+23042	𣁂	⿱小文	
U+23043	𣁃	⿱文⿻了八	

This small sample does not show all the details of how the IDS descriptions in the two sets differ (Tab. 5.1), but identifies all relevant types of discrepancies, which can be divided into three provisional categories:

– different set of components: 𣁁𣁂𣁃;

– different structure (IDCs): 亭享京;

– different components and structure: 产亭亮亯京旕旖旗𣀿.

In cases of alternative decompositions that vary according to different locales (亮 in the sample set) the KDP descriptions provide the sources with divergent treatment. The type of components is strongly related to the type of structure (represented by the IDCs), hence the most numerous category of differences in IDS descriptions involves both aspects – structure and inventory. The sample character 产 is a good illustration of the influence of the components set on the structure rendered in the IDCs. Choosing '亠' as an immediate component, instead of '&CDP-8BAE'; ('亠丷'), determines the entire component set for this character ({亠,丷, 厂} instead of {亠丷, 厂}) and limits the structural setup. As a result, there are three immediate components in an above-to-below setup which excludes the '⿱' structure (the only possible option in the KDP IDS description of this character) leaving two possible treatments of three-element above-to-below setups: '⿱⿱' and '⿳'. In this case, as in many other similar setups, the choice between the two representations seems to be arbitrary. The differences in structural representations only are not relevant for the present study. As an example, below are discrepancies in the descriptions of the sample character 旖:

CHISE-IDS: ⿸⿰方𠂉奇   KDP: ⿰方⿱𠂉奇

Tab. 5.1 CHISE-IDS and KDP IDS descriptions

Character	CHISE-IDS		KDP	
	IDC and number of components		IDC and number of components	
产	⿱⿱	3	⿱	2
亨	⿱⿱	3	⿳	3
亩	⿱	2	⿱	2
亪	⿱	2	⿱	2
享	⿱⿱	3	⿳	3
京	⿱⿱	3	⿳	3
亭	⿱⿳	4	⿱	2
亮	⿱⿳	4	⿱	2/2[186]
亯	⿱⿱	3	⿳	3
亰	⿱	2	⿳	3
旔	⿰	2	⿰	2
旕	⿱X⿰XX	3	⿱	2
旖	⿸⿰	3	⿰X⿱XX	3
旗	⿸⿰	3	⿰	2
斅	⿰⿱⿰⿳X⿱XXX⿳X⿱XXXXX	10	⿰⿱⿰⿳XXX⿳XXXXX	8
斆	⿰⿱	3	⿰⿱	3
文	⿱	2	⿱	2
斈	⿱	2	⿱	2
斉	⿱X⿻XX	3	⿱X⿻XX	3

‘X’ indicates the placement of a single component in structures that otherwise would be ambiguous.

Ultimately, the discrepancies have no effect on the result of the graphotactic analysis, because they involve structural representations, while the component sets in both cases are identical. It was already mentioned that the analytical part of this study is not concerned with the spatial arrangements of components in general, and their IDC representations in particular. In other words, only the differences involving constituent sets of any type (immediate components in the case of IDS descriptions) have any relevance for the graphotactic analysis.

Since the quantitative exploration of IDC representations is a viable and quite natural research perspective, the problems or arbitrariness and disparity of representation discussed above must be addressed before continuing any further – at least a detailed account of the rules and criteria used in the database of choice should be provided. The integration of IDC data into the graphotactic analysis is not as straightforward as it

[186] Different component sets in different locales.

may seem – the structural arrangement data provided in the IDC part of IDS are more or less equivalent to the linear arrangement of letters in alphabetic scripts, i.e. the part that is intentionally ignored in the graphotactic framework. The IDC part of IDS should rather be considered a complement of the graphotacic analysis in some larger scale project – for example, a grammar for Chinese script. The IDCs might turn out to be indispensable in the investigation of tactographonic equigraphy and disgraphy,[187] but at this point that is pure speculation. It can only be stated that the two largest databases providing the data necessary for graphotactic analysis of Chinese script differ significantly with respect to the IDC representations of structural arrangement of immediate components and the componential representation of characters. The KDP was chosen over the CHISE because the KDP seems to be more suitable; a cursory survey of the IDS descriptions in both databases leaves that impression, which is by no means a conclusion based on hard quantitative evidence. It is, however, abundantly clear that the choice of KDP immediate components would leave a narrower margin for arbitrary choices from the IDC representations.

Only three of the sample characters (亩, 旔, ⿱人义) are assigned the same IDS descriptions in the discussed databases. The significance of the source of discrepancies should not be ignored, and their extent is yet to be estimated. Nonetheless, given the focus of this study and the space limitations, a discussion of differences between CHISE-IDS and KDP here must be restricted to the most relevant reasons that led to the choice of KDP over CHISE-IDS. The reasons for choosing the KDP database as an analytic basis can be summarized in a few points:

- the components inventory in KDP is more homogenous than in the case of CHISE-IDS, and it includes the basic inventory which is assumed to be the GF3001-1997 standard and the CDP components of different kinds (both basic and compound);
- the CDP components are well documented and relatively easy to identify and interpret;
- the KDP decomposition criteria are stated more explicitly as 'physical';[188]
- KDC IDS descriptions account for the locale-specific differences in decomposition by indicating, when neccessary, the source of a given structure.

Some more details on the handling of IDS descriptions in KDP are provided in Chapter 7.

[187] Bańczerowski 2009: 21. Also see Section 3.2.1.1.

[188] http://kanji-database.sourceforge.net. There is an alternative IDS database available on T. Kawabata's website with the IDS descriptions based on 'semantic' decomposition. This database is significantly smaller (over 18,000 characters); a semantically motivated decomposition is of secondary importance to this study.

5.1.2. Chinese Documents Processing Lab (CDP)

漢字構形資料庫 *hànzigòuxíng zīliàokù* is a project developed at the Chinese Document Processing Lab, Institute of Information Technology, Academia Sinica, in Taiwan.[189] The project was commenced in 1998 with the collection of 13,051 BIG5 characters with their CDP CDL descriptions. The CDP database has been growing over the years, and as of 17 April 2011, it contained the staggering figure of 165,653 characters, which is the largest inventory of Chinese characters of which the author of this book is aware. It must be noted that CDP is a database of all types of script, hence the overwhelming number. A closer look at the CDP character inventory shows, predictably, the relatively standard number (still probably the highest)[190] of regular script (楷書 *kǎishū*)characters:[191]

Tab. 5.2 Contents of the CDP database

Regular script characters	91,510	165,653
Collected variant form characters from 漢語大字典 *Great Dictionary of Chinese Characters*	12,208	
說文解字詁林 *A Forest of Glosses on the Shuowen Jiezi*	11,100	
金文編 Bronze inscription characters	22,729	
楚系簡帛文字編 Chu Silk Manuscripts and Bamboo Slips Characters	37, 614	
殷墟甲骨刻辭類纂 Oracle bone inscription characters	2,700	

On the downside, the CDP database has a private format and is accessed through the *cdphanzi* software package (2.7 is the current version). This means that the componential descriptions cannot be easily accessed in a way required by the theoretical framework of this book. CDP uses a set of 441 simple components that is based on the Big5 character set; 382 of them are the primary components, 59 are secondary vari-

[189] http://cdp.sinica.edu.tw/cdphanzi/.

[190] The *Konjaku Mojikyo* (今昔文字鏡) database, having collected 90,000 Chinese characters is a close competior. See http://www.meijigakuin.ac.jp/~pmjs/archive/2000/mojikyo.html; and http://www.baike.com/wiki/%E4% BB% 8A%E6%98%94%E6%96%87%E5%AD%97%E9%95%9C.

[191] http://cdp.sinica.edu.tw/cdphanzi/documents/history1010417.pdf.

ants.[192] The total number of Big5 basic components is 2,297; 441 simple and 1,856 compound ones.[193]

The character description language in CDP operates on components and descriptions for its structural arrangement. There are 13 graphic descriptors of the spatial arrangement:[194]

Tab. 5.3 CDP descriptors

Descriptor function	Graphic Representation	Explanation	Example of CDP CDL expression
Type of spatial arrangement of components	⿰	left-right component composition	弧 = 弓⿰瓜
	⿱	top-bottom component composition	岔 = 分⿱山
	⿴	outside-inside component composition	圁 = 囗⿴言
Writing order of components	形⃞ ⊡	two elements indicating the beginning and the end of the sequence of components that are arranged in the actual writing order	解 = 形⃞角刀牛⊡
Type of spatial arrangement of recurring component and number of recurrences	⦂	2 vertical recurrences	吕= ⦂口
	⁝	3 vertical recurrences	三 = ⁝一
	∘∘	2 horizontal recurrences	林 = ∘∘木
	∘∘∘	3 horizontal recurrences	⿲吉吉吉 = ∘∘∘吉
	∴	3 recurrences in triangular arrangement	森 = ∴木
	∘∘∘∘	4 horizontal recurrences	
	⁞	4 vertical recurrences	
	∷	4 recurrences in square arrangement	⿱牪牪=∷牛

[192] For example, 忄 and ⺗, variants of 心.

[193] Chuang & Teng 2009.

[194] http://proj1.sinica.edu.tw/~cdp/service/documents/T960419.pdf.

The three categories of descriptors (the first column) form five classes of structures. Each of the spatial arrangement types forms a separate class. In order to bring into focus some important issues it is best to avail to some more examples of the CDP description language:

1. 闟 = 門⟁∴下
2. 覉 = 亠⩓目⩓死
3. 嚞 = ∘∘∘吉
4. 灪= 氵⟑∘∘屏 刂

All descriptions are accurate and precise, meaning that they provide unambiguous information on the character's componential composition. More problematic cases will be handled in this section.

One of the main objectives of the CDP project is the facilitating of creation of missing forms of characters. The CDP CDL is probably the most flexible component-based CDL, and yet it is subject to substantial limitations. It is not always possible to render a unique structural description of a character. In other words, a CDP CDL expression may render more than one character form. Even more numerous are the opposite cases, where a character may be described by more than one expression. This is not necessarily a flaw in CDL. A certain degree of indetermination is inherent in Chinese characters as a writing system with long history of evolution and reforms. The reasons for this structural indetermination may be summed up in the following points:

- the selective evolutionary changes in character and component shapes;
- the adaptation of the shapes of components and strokes to different distributional contexts;
- ambiguous status of components in particular characters (cases other than evolutionary irregularities);
- structural similarities between different components;
- conflicting etymological and structural motivations.

The CDL description for the character 解 in the table above only specifies the order of writing, and does not provide full information on the spatial arrangement of {刀, 牛}. More general rules for character production (distributional properties of components) exclude two of the three possible arrangements – 刀 on the bottom and 牛 on the top, and 牛 on the left and 刀 on the right. This still leaves two possibilities – 刀 on the left and 牛 on the right, and the actual composition, 刀 on the top and 牛 on the bottom.

The formula 解 = 彫角⟑刀⩓牛⊡ gives precise structural information on the composition of the elements, but the graphic descriptors are only allowed to bind from among the 2,297 basic components that form the lexicon of CDP CDL. The expression in the table describes {刀, 牛} as separate components. They do not constitute a compound, which means that this element does not belong to the lexicon. This is con-

firmed by the fact that even a large database shows 0 characters with 刀 and 牛 in the top-bottom structure as a component. Another formula that complies to the 'component-only binding' rule is possible:
解 = 形角刀△牛⊡.

This expression seems to be essentially in accordance with the salient features of the language and is more precise than the expression in the table. The character 豐 may serve as another example. Neither of the possible expressions, 形山oo丰豆⊡ or 形山丰丰豆⊡, provide sufficient information for a complete reconstruction. The ambiguity of some of the CDP CDL expressions is a serious limitation for its use in computer applications, but nevertheless, it is a very useful and convenient as an visual information descriptor. The CDP CDL is a description language, and as such, lacks some of the formal features of a grammar for characters (see section on the grammars of characters). As a result, instead of assigning components to distributional classes that unambiguously determine component compositional properties in different contexts, it presumes the native writer's competence that renders '解' as the only possible arrangement of the '角刀牛' sequence.

The sources of problems with coherent and uniform structural description of characters are also addressed in different sections of this book (especially in the section on character structure). The CDP CDL is a good exemplification of the indeterminacy of character composition and component inventory caused by the ambiguous status of some components and the arbitrariness of the inventory.

Given the importance of the CDP database it is reasonable to introduce here a full list of the 441 CDP simple components. The CDP database is a potential source of alternative graphotactic analysis of the BIG5 and larger character sets. The CDP component list also renders the best comparative background for the inventory constructed as a result of present research.[195] Table 5.4 is an exhaustive list of simple components arranged according to the stroke count.[196] The reference numbers of components are provided in the second column, and the number of occurrences in the set of characters is shown in column 4, while the frequency of components is shown in column 5. The exact methodology is explained in Chuang & Teng (2009: 40), and in a less detailed way, in Section 4.4.

[195] See Chapter 7.

[196] Chuang & Teng 2009: 34-39

No.	Group	Component	Examples	No.	Group	Component	Examples
1	1	一	是天同旦	41		儿	四罕
2	2	㇀	刁	42	35	入	全兩糴仚
3	3	丨	在條引叟芈	43	36	乂	艾刈
4	4	丿	向少呂胤	44		㐅	學希凶爻
5	5	丶	主凡兔卵勺	45	37	儿	兒虎兆坴
6	6	㇏	尺	46	38	几	處凡飢冗
7	7	乙	乾乞壓	47		几	朵
8		乚	孔	48	39	⺇	夙
9	8	亅	丁了予	49	40	力	万別
10	9	𠃌	成局司幻	50	41	𠂊	你
11	10	㇉	吳	51		⺈	免
12	11	⺄	訊虱	52	42	勹	包句
13	12	㇄	亡陋曷	53	43	𠃋	留派旅兜印
14	13	㇖	慶予壽疋	54	44	七	叱
15		㇇	了候今	55	45	亠	六商夜京
16	14	乀	尐	56	46	⺉	班
17	15	二	些次元仁貳	57	47	冫	冷弱
18	16	十	什南早古	58	48	丷	南商弟帝幸
19	17	丅	丏	59	49	⺀	於冬
20	18	厂	原產反岸詹	60	50	冖	學軍受勞帝
21	19	𠂇	有在	61	51	コ	假巨卣
22	20	𠤎	能此尼旨罾	62	52	ユ	侯
23		ヒ	它	63	53	卩	報命印卯
24	21	匚	匠	64		㔾	範犯卷宛厄
25	22	匸	區	65	54	丩	叫收赳
26	23	七	切皂柒	66	55	凵	出匈屆凶禽
27	24	丁	可打頂亭	67	56	刀	分切絕召賴
28	25	丂	巧兮亟攷	68		刂	到
29	26	与	焉与	69	57	力	動加勞辦務
30	27	卜	下外赴鳪	70	58	又	受友取隻反
31		⺊	上桌竊	71		㇇	祭
32	28	冂	同南向剛炯	72	59	乃	仍秀孕
33		⺆	角	73	60	厶	能公參私允
34	29	⼌	囧	74	61	マ	令甬
35	30	𠂉	無傷族鹽乞	75	62	九	究軌旭尻厹
36	31	丿丨	介冥丌氘	76	63	乜	乜
37	32	𠂆	派遞后盾梔	77	64	了	亨釘
38	33	人	以幾閃企卒	78	65	巜	粼
39		亻	他候	79	66	干	幸岸幹汗刊
40	34	八	分六匹睿扒	80	67	于	宇迂盂吁

81	68	丰	那半奉丮	123	110	弓	發張弱躬彎粥
82	69	土	在去地社疆	124	111	⺕	急尋雪帚
83	70	士	志穀壯王敖	125	112	丑	韋
84	71	工	江空功式左貢	126	113	屮	出弢艸茻
85	72	大	天因奇夾枎	127	114	也	他拖匜迆
86	73	尢	拋尬尷	128	115	女	要好委威瀛
87	74	廾	開算奔戒葬	129	116	廴	建
88	75	丌	畀	130	117	𧘇	辰畏
89	76	丈	仗	131	118	小	你少京叔齋
90	77	㐄	降桀韋	132		⺌	當尚肖
91	78	ㅌ	虐	133	119	子	好學孔孟囝
92	79	寸	將專付冠刌	134		孑	孑
93	80	弋	代式鳶芅隿	135	120	彑	彙
94	81	厶	育流充棄	136		彐	彝
95	82	才	材閉鼒	137	121	阝	那院
96	83	丬	北燕	138	122	乡	鄉雍
97	84	少	步歲賓	139	123	孓	孓
98	85	口	可和問叫各	140	124	幺	麼後樂幼茲幻
99	86	囗	國	141	125	巛	經腦災巡韋
100	87	山	島微岸仙峽	142		川	順訓夼氚
101	88	巾	市布帽飾帥匝	143		川	流荒侃
102	89	巾	爾	144	126	王	全現班弄匡閏
103	90	乇	托	145	127	井	耕刱丼
104	91	毛	托宅	146	128	夫	規扶麩芙
105	92	千	乖阡芊	147	129	主	責
106	93	彳	得	148		丰	害
107	94	彡	形須參彥彪	149	130	耂	考
108	95	㐅	匆囪	150	131	丐	鈣
109	96	夂	各處隆贛	151	132	廿	度燕堇
110	97	夊	後愛致	152	133	龷	共備展散
111	98	久	畝疚灸羑	153	134	木	校新條林樂查
112	99	𠂎	卯	154		朩	余
113	100	夕	多外名夢矽	155	135	朮	述痲
114	101	丸	執芄尥	156	136	巿	沛旆芾
115	102	凡	恐贏	157	137	卅	卅
116	103	广	應	158	138	不	否杯罘鴀
117	104	冖	家	159	139	犬	類哭默莽戾倏
118	105	䒑	前業善兼朔岡	160		犭	狗
119	106	尸	局屋刷殿辟	161	140	歹	死列夙
120	107	己	記改忌巹	162		歺	餐
121	108	巳	起包巷熙祀	163	141	五	吾伍
122	109	已	已	164	142	屯	頓純囤窀迍

165	143	旡	既蠶潛	207	176	文	產蚊斑虔斐閔斌
166	144	戈	或幾找划盞	208	177	火	勞燈滅秋灰災炎
167	145	牙	呀穿雅迓	209		灬	然盡顯
168	146	艹	花	210	178	⺌	率函脊兆
169	147	卝	苟	211	179	辶	這
170	148	止	此步歸武歷企	212	180	之	芝
171		龰	是從徙	213	181	冘	沈鴆尣髧
172	149	日	是時間者普晶	214	182	心	想愛感寧恥
173	150	曰	書勗汩昳	215		⺗	恭
174	151	⺜	曼	216		忄	情
175		冃	冒冑	217	183	冖	羣
176	152	冄	衰	218	184	肀	唐
177	153	中	忠衷沖霻	219	185	户	倉
178	154	罒	象	220	186	⺕	爭
179	155	冋	雋	221	187	夬	快英
180	156	內	納芮氝	222	188	弔	弟佛第盄
181	157	手	看拿掌摩	223	189	爿	將藏寢妝
182		龵	拜	224	190	丑	紐羞
183		扌	把	225	191	巴	把爸爬疤弝
184	158	攵	教條變務肇煞鼇	226	192	尸	聲眉
185	159	毛	尾秏毯氅	227	193	凷	鹿
186	160	气	氣汽芞	228	194	尹	君伊笋
187	161	牛	解件牢犀犛	229	195	毋	毒
188		⺧	先告	230		毌	貫虜
189	162	龶	邦蚌耆夆	231		母	每姆
190		丰	豐彗	232	196	水	冰泉尿穎盥沓
191	163	片	牌洴	233		⺡	法衍
192	164	斤	所近質匠欣斧	234		⺢	暴漆黎藤
193	165	爪	爬抓笊	235		氺	壞屬鰥
194		爫	受	236	197	夫	春
195		𠂎	印毿	237	197	瓦	瓶瓷
196	166	戶	所房妒	238	198	𠀎	賽襄菁
197	167	父	交爸蚊	239	199	未	業妹釐寐
198	168	月	有前明服朋閒	240		末	抹茉
199		夕	望	241	200	示	款票尉佘
200	169	氏	紙昏痻	242		礻	社
201	170	丹	坍旃彤	243	201	甘	某甜邯疳
202	171	小	亦赤	244	202	世	葉屜貰痲泄
203	172	𧘇	表	245	203	本	笨鉢梱
204	173	𠂢	派旅	246	204	巿	囊
205	174	勿	物易忽刎囫	247	205	丙	病柄昺邴
206	175	及	級岌	248	206	石	研磨岩拓礱磊

249	207	卌	卌	291	246	𦘒	肅
250	208	犮	髮拔魃	292	247	𡗗	庚
251	209	戊	茂戌	293	248	𠨬	假霞
252		戉	越	294	249	弗	費佛剃氟茀
253	210	以	似苡	295	250	民	眠敃苠
254	211	目	看相省眼眉臭	296	251	皮	被頗疲簸髮
255	212	且	姐助宜疽	297	252	丱	聯
256	213	田	當界細畫福略甸	298	253	癶	發
257	214	由	油迪笛寅胄	299	254	矛	柔茅矜楙袤
258	215	甲	鴨押閘匣厴	300	255	卍	卍
259	216	申	神暢氰	301	256	耳	聲取聞聶弭
260	217	皿	盡醘齍	302	257	甘	其
261	218	罒	罷	303	258	臣	宦臥嚚臧挋
262	219	史	駛	304	259	襾	要
263	220	央	英映盎	305		西	晒垔茜氥
264	221	兄	免	306		両	両
265	222	冉	再髯聃	307	260	吏	使
266	223	冊	刪柵	308	261	朿	策棗刺棘
267		冊	狦	309	262	而	需耐耍洏
268	224	冊	侖	310	263	亙	恆
269	225	业	並業	311	264	至	到屋室姪臷
270	226	电	奄	312	265	戋	或
271		电	電	313	266	夷	姨痍荑
272	227	内	離亂竊	314	267	虍	虎
273	228	凹	兇	315	268	且	直具
274	229	凸	凸	316	269	夬	訣
275	230	𠂤	官師耜	317	270	曲	農蛐髷
276	231	生	產星性隆甥甦眚	318	271	虫	強雖蟲蜂蛋融
277	232	乍	作怎窄厏	319	272	曳	洩
278	233	禾	和乘委穎穌囷	320	273	业	虛
279	234	𠭊	段	321	274	冎	骨咼
280	235	丘	兵邱蚯	322	275	肉	腐臠胾肭
281	236	白	的原樂怕皆皇	323		月	能育臉胡
282	237	瓜	狐瓣瓞窳蠡	324		夕	然將祭炙
283	238	乎	呼虖	325	276	缶	寶搖缺陶鬱罄
284	239	用	備佣甬	326	277	耒	耕誄
285	240	甩	甩	327	278	年	哖
286	241	𠂔	姊笫趀	328	279	𢁺	制
287	242	疒	病	329	280	缶	卸
288	243	立	位童站笠	330	281	竹	竹
289	244	必	秘瑟謐邲閟	331		⺮	第
290	245	永	泳昶羕	332	282	自	息咱郎

333	283	臼	兒舀臿閻	374	317	𤣥	牽
334		𦥑	學與叟盥寠	375	318	㡀	敝
335	284	舟	船俯	376	319	鼠	鼠獵
336	285	㐆	殷	377	320	長	張萇饕
337	286	㐅	鹵𡿺	378		镸	套肆
338	287	𧰨	象𧰼	379	321	𤔔	敢
339	288	衣	裝依裁哀	380	322	亞	惡啞氬
340		衤	裡	381	323	叀	專
341	289	亥	孩刻氦	382	324	東	陳菄鶇
342	290	米	氣料迷菊糞粟	383	325	事	剚倳
343	291	羊	洋群氧羸羍	384	326	雨	電漏
344		𦍌	美	385	327	丽	麗
345	292	州	洲𨚵	386	328	𦣻	憂
346	293	聿	建筆律盡	387	329	豖	啄瘃冢剢
347		𦘒	書	388	330	果	課顆巢裹贏
348	294	艮	很良退痕茛	389	331	疌	捷
349		艮	即朗	390	332	典	典
350	295	亜	亞	391	333	門	們問菛
351	296	羽	習翻扇翁翅翍	392	334	妻	妻
352	297	糸	緊絲徽辮	393	335	非	匪靠排悲靡剕
353		糹	給	394	336	無	撫蕪廡甒
354	298	𦣝	姬頤宧	395	337	秉	秉
355	299	車	軍連輕陣輿庫	396	338	臾	庾腴萸斞
356	300	甫	補葡圃簠專鵏	397	339	隹	進難隻霍截
357	301	更	便甦郠	398	340	卑	脾鼙庳
358	302	束	速刺柬辣	399	341	金	錢鑒淦
359	303	酉	酒醫酸釁	400	342	隶	康隸逮
360	304	豕	家猪逐豚蟸	401	343	承	承
361	305	求	球救裘毬逑莍	402	344	韭	齏虀韰韱
362	306	里	裡童野厘	403	345	革	鞋鞏緙鞗
363	307	串	患賗	404	346	面	麵緬靨靦䩄
364		弗	弗	405	347	禺	萬遇愚偶顒
365	308	見	現覺覞	406	348	垂	睡郵箠厜
366	309	貝	員則贏貳狽賫	407	349	重	種動衝董
367	310	我	義俄鵝	408	350	禹	齲萭鄅
368	311	𣶒	肅淵	409	351	食	養飧
369	312	身	射鯓	410		飠	飯
370	313	鳥	島	411	352	首	道馗馘艏
371	314	釆	番釋粵竊	412	353	為	偽寪鄬
372	315	豸	貓綌	413	354	飛	騛
373	316	言	這說信譬獄謄	414	355	鬲	隔融鬻鬳

415	356	馬	媽驚騎闖騰馵
416	357	鬥	鬧
417	358	烏	烏嗚鄔
418	359	鬼	愧魔魁嵬
419	360	兼	廉歉謙蒹
420	361	隺	確鶴篧
421	362	堇	勤謹廑
422		重	董
423	363	莫	難漢
424	364	帶	滯蔕遰
425	365	曹	遭糟
426	366	棄	棄
427	367	畢	嗶篳鵯
428	368	婁	數屢樓簍
429	369	庸	傭鄘
430	370	壺	壺
431	371	鼎	鼐濎
432	372	斲	斲
433	373	單	戰彈闡
434	374	黽	蠅鼇鄳
435	375	黑	點黨墨黴嘿
436	376	熏	勳薰醺
437	377	齊	濟劑齋薺
438	378	齒	齡齧嚙
439	379	龍	籠襲隴龐
440	380	羲	犧
441	381	龜	鬮

Tab. 5.4 CDP basic components

The CDP database stores a huge amount of information valuable to the graphotactic framework. The next table is an example of how multilayered information on a single character (顙) and its components is stored in the database:[197]

Tab. 5.5 CDP database structure

Character	Structure type	Structure type and reference number		Component order	Simple component order	Simple components by stroke-count
顙	桑⩕頁	⩕	1	桑頁	又又又木一自八	一八又又又木自
桑	叒⟁木	⟁	2	叒木	又又又木	又又又木
叒	∴又	∴	5	∴又	又又又	又又又
木	木		0	木	木	木
又	又		0	又	又	又
頁	百⟁八	⟁	2	百八	一自八	一八自
百	一⟁自	⟁	2	一自	一自	一自
自	自		0	自	自	自
八	八		0	八	八	八

[197] Chuang & Teng 2009: 91.

Every character is hierarchically linked to the information on the structure of its subparts to the simple component level, which contains more detailed information than that contained in the IDS descriptions.[198]

Table 5.6 exemplifies CDP character representation table:[199]

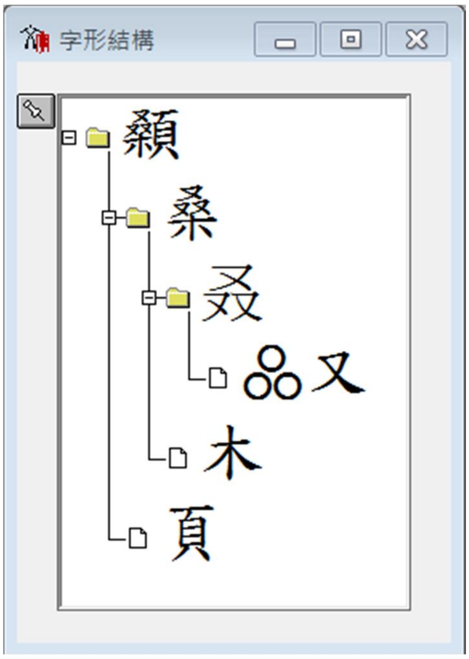
Tab. 5.6

	Character	Structure type		Component order	Simple component order	Simple component stroke-count order
1	繠	⩕	1	桑頁	又又又木一自八	一八又又又木自
	掰	⩕	1	手分手	手八刀手	八刀手手
	簠	⩓	2	⺮甫皿	⺮甫皿	皿⺮甫
	鬵	⩓	2	○○旡鬲	旡旡鬲	旡旡鬲
3	㔶	⟁	3	广鼻	广自田丌	广丌田自
	龥	⟁	3	龠口	人冊一口口口	一人口口口冊
4	邍	形⊡	4	辶彔备	辶夂田彑氺	夂彑辶氺田
	豐	形⊡	4	山○○丰豆	山丰丰一口丷丄	一口山丷丰丰
5	虤	○○	5	○○虎	虍儿虍儿	儿儿虍虍
	戔	8	5	8戈	戈戈	戈戈
	贔	∴	5	∴貝	貝貝貝	貝貝貝

The graphic interface of the CDP database (*cdphanzi*) does not directly display all the above information. For example, the character 繠 is decomposed in the following way:

[198] It is possible to extract the equivalent details from the IDS expressions by a recursive analysis of the whole database.

[199] Ibid., 90. The example set was modified due to technical reasons.



The CDP database has been integrated into the CHISE-IDS[200] and the KDC, but since few details are available on this process it is not clear how the integration was carried out and what the motivations were. It can be inferred from the end result that the CDP components were used for the decomposition in instances when the basic set of components was insufficient. The extent of the integration will be discussed in greater detail in Chapter 7.

The CDL database must be considered one of the most referential and comprehensive sources of information on Chinese characters – it stores information on archaic forms of Chinese script (the oracle bone inscriptions, the bronze script, and the seal script)[201] and an exhaustive inventory of regular script characters and their semantic, phonetic and structural descriptions. For technical reasons related to data accessibility, however, it was not chosen as the primary data source for the project detailed in this book. In other words, the CDP is not the basis for the graphotactic analysis of Chinese characters for practical reasons only, this despite the fact that it remains one of the the best available database of Chinese characters for non-quantitative research. Additionally, the CDP CDL is probably the best description language for non-computer applications. In the future it will be interesting to compare the results of CDP-based graphotactic analysis with the one based on IDS descriptions.

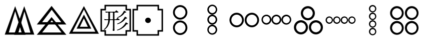
[200] Morioka 2008: 156.

[201] One of the main purposes behind the CDP database is the preservation of the ancient heritage of Chinese script.

5.1.3.Wenlin CDL

The Wenlin CDL (WL CDL) is a part of a commercial project (Wenlin Software for Learning Chinese) developed at the Wenlin Institute. The humble name of the software may suggest it is merely another learning aid, but it also is a powerful research tool and one of the most extensive souces of information on Chinese characters. In the current version Wenlin covers the whole Unihan portion of Unicode 6.1 and even goes beyond it, resulting in a total of 84,044 characters.[202] The introduction here concentrates on the features related to the componential structure representation in the Wenlin database and to the WL CDL.

Tab. 5.7 Features of different CDLs

CDL	Language features	
	Descriptors	Constituents
CDP		441 basic components 1,856 compound components
IDS	⿰⿱⿲⿳⿴⿵⿶⿷⿸⿹⿺⿻	Nominally 560 basic components[203]
WL	Cartesian coordinate system	System of 39 basic strokes[204]

Compared to CDP CDL and IDS, the most notable difference in the WL CDL design is its actual use in the rendering of characters from CDL descriptions – while CDP and IDS structural descriptors refer to the mental representations of character structure, i.e. they are intended to give the user an idea about the shape and structure of a described character without the computer rendering a capability, the WL CDL is a language interpretable for the computer generation of characters. It operates on the Cartesian coordinate system providing unambiguous mathematical descriptions, rather than on the descriptors of spatial arrangement.[205] The other relevant difference is the inventory of basic constituents – WL CDL operates on a set of the basic strokes. Table 5.7 summarizes the relevant differences in the descriptions of the CDLs discussed so far.

The WL CDL approaches the structure of characters from a very practical perspective. Essentialy it uses the strokes as a character description, but also predefined com-

[202] http://www.wenlin.com/cdl/#stat.

[203] This is based on the fact that the GB 130001 standard is a translation of ISO 10646 (identical to the basic block of CJK Unified Ideographs – see Section 2.2.1.) The purpose of the table is to provide the type of CDL expressions, not the exact number.

[204] http://www.wenlin.com/cdl/cdl_strokes_2004_05_23.pdf.

[205] Cook 2003: 106.

ponents are used in the descriptions as well. In both cases it provides accurate grid coordinates positioning the constituents in the structure. The expressions contain data on the constituents and their relative positioning expressed in grid coordinates. For example, the WL CDL description of the character ‘疑’ has the following form:

```
!cdl
<cdl char='疑' uni='7591'>
    <comp char='' uni='e5c7' points='0,0 52,128' />
    <comp char='龴' uni='9fb4' points='68,4 122,42' />
    <comp char='疋' uni='758b' points='48,48 128,128' />
</cdl>
```

The description specifies three components and provides two anchor points for each of them that position a rectangular referential frame of the component in realtion to the structure of character. The CDL descriptions are intended for interpretation by a computer, and therefore, they are not transparent to humans, at least not without training and effort. For that reason the visible database entries are not WL CDL expressions, but rather the WL CDL components are listed without the grid coordinates. Additionally, in a separate section, the CDP and IDC descriptors are used to indicate the relative positioning of the components. The descriptor section often contains components different than the WL CDL components – in the case of the exemplary character the relevant database entry provides a description of this in the following way:

“匕矢龴疋△; a total of 3 CDL comp elements (V=0): (龴疋) ”.

Despite WL CDL being the most accomplished language in terms of the descriptive precision and character processing, access to the componential information stored in the Wenlin database is restricted in a way similar to the CDP database. As a result, its relevance to graphotactic analysis is very limited. It must, however, be acknowledged that the Wenlin database has played an essential role in this present study as a reference source for gathering and verifying information on Chinese characters.[206]

5.1.4. Summary

All of the CDL projects introduced in the above sections are significant in relation to the main focus of this book, though at the same time, it must be said that all have vexing limitations stemming from the intended purposes of their design or from the format of the data. On paper the CDP CDL seems to be the most promising alternative as a basis for the graphotactic analysis of Chinese script, but that depends on the acces-

[206] The author also wishes to express gratitude for the personal help from dr Richard Cook of the Wenlin Institute who pointed out the alternative sources for componential description of characters.

sibility of the data. At this point the Chinese Document Processing Project is an indispensable source of reference materials for character decomposition and construction of the component inventory, and it provides a referential comparative background for the results of the graphotactic analysis. All CDL projects introduced in this section borrow from the CDP component inventory to supplement more basic sets (IDS projects) or for auxiliary purposes (Wenlin). The IDS projects, especially Kawabata's Kanji Database Project, made the graphotactic analysis possible by providing the required data. The Wenlin database has proven to be a useful accessory in providing and verifying information on Chinese characters. Furthermore, the Wenlin website[207] provided extensive documentation and reference materials, which included valuable theoretical insights.[208]

5.1.5. Other projects

The efforts that include the description of the componential structure of characters are not limited to those discussed in the previous sections. The projects introduced here include the most notable ones that for different reasons have no direct influence on this study, but due to their nature and content are related to the graphotactic framework. The introductions will be extremely brief, not only due to the space limitations, but also due to the fact that in most cases the relevant documentation is very limited.

5.1.5.1. Hanglyph CDL

Hanglyph[209] is a computer oriented CDL project aimed at the generation of graphical forms of characters from CDL descriptions.[210] In this respect it is quite similar to the Wenlin CDL, in the use of strokes as base constituents. Hanglyph operates on a system of 41 strokes and 5 'operators', which is equivalent in function to IDCs and CDP graphic descriptors:

- top-bottom;
- left-right;
- enclosing;

[207] www.wenlin.com.

[208] Some insights on different CDLs can be found in Lin & Song 2007; Haralambous 2011; and Xue & Gu 2012.

[209] http://www.hanglyph.com/en/hanglyph-index.shtml.

[210] Yiu & Wong 2003.

- partially enclosing (with seven possible directions indicated by numbers);
- cross.[211]

For refinement of the relative topography of constituents four relations are used:[212]

- dimension – specifying the relative dimensions of constituents, operating on four Boolean relations (less than, more than, not less than, and not more than);
- alignment – specifying the location of the constituents with the use of five descriptors (at top, at bottom, at left, at right, and centered), or with the descriptors in combination (for example, at bottom left);
- touching – specifying whether the constituents are in contact, through the use of two descriptors (touching and not touching);
- scale – adjusting the size of the rendered character.

For reasons of operational economy of the system, frequently occurring combination of strokes (equivalent to components) are coded as a system of macros, thus simplifying the CDL expressions.[213]

Hanglyph CDL is treated here marginally for two reasons – it is stroke-based and, more importantly, it has never been implemented. Since the status of the project, at least according to the website updates, has not changed in a long time, it is difficult to determine whether Hanglyph CDL is still being developed or if it has been abandoned. It looks promising as an unambiguous character rendering language. At the present, possibilities of its application for other purposes, particularly applications involving character component structure, remain undetermined.

5.1.5.2. Cjklib

Cjklib is a library of CJK characters implemented in the Python programming environment, and possesses some functionalities that include pronunciation, radicals, components and stroke decomposition.[214] The project website provides only a few details directly related to graphotactic analysis. The character set used in cjklib is closely related, if not identical, to the Unihan database. Based on this information it can be assumed that the decomposition data contained in the library are related to Kawabata's KDP IDS. The cjklib support features are unrelated to graphotactics, but since the author of this book failed to implement the cjklib package, the introduction here will be limited to the few general statements. Also, in this case it is not evident if the project is

[211] Ibid., 87-88.

[212] Ibid., 88.

[213] Ibid., 87-89.

[214] http://code.google.com/p/cjklib/

still being developed. The functionalities related to the componential data listed in the project description are:[215]

- decomposition of characters into components;
- component tree with structural information;
- component search with equivalent forms.

These functionalities suggest a similarity to the Wenlin database, but with a cruder user interface. Cjklib is a project worth following as it has all the features of a powerful non-commercial research tool.

5.1.5.3. Wikimedia Commons Chinese Characters Decomposition Project (CCDP)

Wikimedia Commons Chinese character decomposition project[216] is an open-source project based on a collaboration of the Internet community. After the latest update it covers 20,902 characters which is equivalent to ISO 10646 / CJK Unified Ideographs (basic block). The CCDP decomposition data format makes them directly available for graphotactic analysis. The sample entry in the database is shown below (the upper numeric line is for explanatory purposes only):[217]

1	2	3	4	5	6	7	8	9	10	11
欒	23	吕	絲言	19	?	木	4	?	VFD	木

1 – character (Unicode order)
2 – stroke count
3 – type of composition
4 – components in the first part of a character
5 – stroke count in the first part of a character
6 – first part verification check ('?' means unverified, empty means verified)
7 – components in the second part of a character
8 – stroke count in the second part of a character
9 – second part of the verification check
10 – Cangjie code[218]
11 – radical

[215] http://code.google.com/p/cjklib/wiki/Features

[216] http://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition

[217] Based on the legend available on the website: http://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition

[218] See Section 7.1.

The type and form of the information contained in the CCDP database is rather unusual. Compared to the previous version of the database, a tendency toward standardization can be observed, but some unique solutions are still part of the database format. The most notable feature is the unique set of structural descriptors. There are 12 descriptors used in the CCDP system:[219]

- 一 – non-decomposable character;
- 吅 – left to right structure;
- 吕 – above to below structure;
- 回 – full surround;
- 咒 – vertical structure, recurrent element in the top part;
- 弼 – horizontal composition of three, the third being the repetition of the first;
- 品 – recurrence of three elements;
- 叕 – recurrence of four elements;
- 冖 – above to below structure, separated by ‘冖’;
- + – “graphical superposition or addition”;
- ? – “unclear, seems compound but ...”;
- * – “atypical”[220] above to below structure.

The selection of the represented structures strikes the observer as rather non-standard. In general the system works, but in many cases in an unintuitive, ambiguous and artificial way. It is not difficult to notice that some of the descriptors are used to indicate different kinds of problems with the representing structures of some characters. Descriptions like ‘unclear’ and ‘atypical’ offer no insights into the composition. Also, it is ambiguous as to what the justification is for the very specific ‘咒’ and ‘弼’ descriptors – in case of the former there is a more general descriptor ‘吕’, while the latter lacks such an equivalent (there is no equivalent of ‘⿲’). There is no descriptor representing three element vertical structures (⿳), but instead a separator ‘冖’ is used to indicate a specific type of ⿳ composition. The structural descriptions are not overtly relevant to the present study, but they directly affect the component sets into which characters are decomposed. As a result, not only are the types of compositions different than the standard 12 Unicode types (represented by the IDCs), but also the constituency differs. The upper part of the exemplary character ‘欒’ in the KDP database is decomposed into three components: 糸, 言, 木, and in CCDP into two components: 絲, 言. On the immediate components level it results in two different tactographemes: {糸, 言, 木} and {絲, 言, 木} correspondingly. The tactographemes consisting of the basic components are identical in both cases. Further examples illustrate the specificity of CCDP decompositions (the equivalent KDP decompositions are given in parentheses):

[219] http://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition

[220] Parentheses indicate direct quotations from the website description.

(i)	宿	11	吕	宀	3	佰	(⿱宀佰)
(ii)	窉	12	吕	宀	3	爿丙	(⿱宀⿰爿丙)
(iii)	寡	14	吕	宀	3	丆且分	(⿳宀直分)
(iv)	寢	14	吕	宀	3	爿?冖又	(⿱宀⿰爿𠬶)
(v)	徵	14	吅	彳攵	7	?	(⿲彳⿱山币攵)

A brief examination of the decompositions shows that in two of the examples, (i) and (ii), the immediate component sets are identical (basic component sets also), while in the remaining two examples, the decompositions differ on both levels.[221] In examples (iv) and (v), the compositions are incomplete – there is an unaccounted for part of the structures. A more general overview of the two databases indicates that the decompositions on the level of immediate components differ significantly. The discrepancies on the basic component level are less apparent, and their extent can be determined in a quantifiable way, though it must be noted that this remains outside the focus of this study.

The Chinese Character Decomposition Project at the present state of development presents substantial difficulties for a direct use in graphotactical analysis for a few main reasons:

- the ratio of ‘unverified’ and ‘unclear’ decompositions is quite high;
- a considerable number of decompositions are incomplete;
- the unique approach to the structure of characters calls for a careful examination of the CCDP component system to understand the significance of the results.[222]

5.2. Grammars

The projects introduced so far concentrated on the various types of descriptions of Chinese characters. This section focuses on the more systemic approaches that treat the Chinese script as a system possessing the characteristic features of a grammar. Grammars either describe or generate expressions. Natural languages constitute truly open systems, whereas the Chinese script is a quasi open-ended system. In other words a sentence need not be decreed as correct before it is even uttered; it is enough for it to be in accordance with the rules of grammar and the characters.

[221] The basic components are extracted through a recursive analysis of the KDP database.

[222] The nonstandard solutions in character decomposition is not a problem *per se* - it would not make sense to duplicate the existing, more extensive decomposition databases (CHISE, KDP).

5.2.1. Distributional model I

One of the earliest attempts to apply the current linguistic methods to construct a model of Chinese script was Rankin's dissertation (1965) and two collaborative strudies (Rankin et al. 1966 and 1970). The resulting model was heavily criticized by Wang (1983) and Stalph (1989) for the purely mechanical application of segmentation methods without any reference to semantic and functional factors that produced no valuable results, and for a model with no explanatory power.[223] Rankin's model is evidence of the futility of an ill-considered application of linguistic methods without recoursing to the specifics of the subject matter. The model concentrated on establishing the segmentation procedure allowing the extraction of components and on defining the components of Chinese characters. It is in his attempts at defining the components where Rankin failed.

5.2.2. Generative model

J. Ch. Wang (1983) approached the system of Chinese script from the perspective of competence and character production. Wang's model mimics the standard model of generative grammar – it consists of three parts:

- Base component:
    - Inherent features assignment rules;
    - Component amalgamation rules;
- Transformational component;
- Writing order assignment component.[224]

The base component is responsible for generating the proto-forms of characters; the transformational rules of the transformational component arrange the composition of components in the internal structure and modify the shapes of components (if neccessary) to produce the surface form of a character; the writing order assignment component is responsible for the stroke order.[225]

In the context of the generative model of grammar there is an important difference between sentences and characters – the former form a infinite set and their production is only limited by the rules of grammar, whereas the latter differ significantly. The number of characters is open-ended, but in a very limited way – the production of new

[223] Wang 1983: 61; Stalph 1989: 40-42.

[224] Wang 1983: 89.

[225] Ibid., 89.

well-formed characters is a subject to substantial restrictions – each such new form must be accepted in a process that has nothing to do with the grammatical model. In other words at any given moment it is possible to produce a list containing all characters, or at least close to it. The inherent features of characters forced Wang to adopt a rather artificial assumption that his model, as a model of competence, is a grammar generating characters that a person does not know – he admits explicitly that:

> "In the lexicon of a native writer of Chinese there must exist a list of all the actual characters he knows. Since the total number of characters a person knows at any given time forms a finite set, and their shape and structure are largely determined by convention, it would be counterintuitive to assume that these characters are generated anew from the character formatives each time they are used. Even the most regular group of characters, namely the phonetic compounds, will have to listed."[226]

It seems also that Wang puts too much trust in the average writer's competence. It is true that a large portion of characters is easily and unambigiously decomposed into component parts even by learners of Chinese script that mastered a certain number of characters. It is also a fact, however, that a large number of characters are ambiguous in this respect.[227]

All grammars of natural languages face the problem of irregularities, idiosyncratic cases and exceptions to the rules and it cannot be expected that the generative model of Chinese character production will account for every single case. Wang's major model flaw is pertinent to the present study – the negligence of addressing the problem of the component inventory. Wang does not provide a list of components and addresses the issue by merely by introducing some general guidelines for the decomposition of characters. Wang simply assumes the existence of such an inventory, which would be equivalent to the lexical component in the early development of generative grammar. Failing to list the components causes all sorts of problems, even for generating of highly regular characters.[228] It seems to be an understatement to suggest that total immersion in the generative model made Wang's analysis artificial. It appears that abandoning the idea of modeling the writer's competence would help with at least some of the issues.

[226] Ibid., 90.

[227] Some of the issues are discussed in Section 5.1. It must be noted that Wang is aware of the problems (1983: 73-75) and claims that they are resolved by the underlying regularities (1983: 75).

[228] The failings of Wang's model in this respect are described in some detail in Stalph (1989: 42-48).

5.2.3. Distributional model II

Probably the most recent intentional attempt at formulating the grammar of Chinese script is the work of J. Stalph (1989) that challenged Wang's statement that the corpus-based approach to the analysis of Chinese characters is undesirable, inadequate and impossible to carry out.[229] Stalph's work deserves a detailed discussion, but due to the different focus of the present study and the space limitations it will be restricted only to an introduction of the basic ideas and Stalph's most important results. His understanding of the grammar of script is best illustrated by his own words:

> "Die Graphematik hat mithin in bezug auf die graphische Struktur der Kanji eine Elementenbestimmung durchzuführen, ein Inventar dieser Elemente zu erstellen und die graphotaktischen Bedingungen zu beschreiben, die zu Komposition regelkonformer, graphisch wohlgeformter Einheiten führen."[230]

The foundation of the analysis is the Japanese 'Frequently Used Chinese Characters' set (常用漢字 *jōyō kanji*) that contained 1,945 characters before the 2010 revision. Stalph, through his contastive graphical analysis of minimal pairs of *kanji*, came up with a list of 485 graphemes[231] that are the minimal constituents of characters. This aspect of analysis is the most pertinent to the present study. The sheer number of analyzed characters seems rather modest compared to the Unihan database, or even to the Big5 set, but one must realize that Stalph had no automatic methods at his disposal, and this fact makes the achievement of his analysis simply more impressive. The choice of the corpus was a practical one – it needed to be of a manageable size and contain frequently used characters – conditions that the *jōyō kanji* fulfilled. Stalph's inventory of components is based on sound principles of decomposition,[232] and the number of elements is within the range of other established sets of components.[233] It is difficult to stipulate how the substantial increase of the analyzed minimal pairs influences the number of graphemes. The number of components is slightly lower than in the official Chinese and Taiwanese standards based on much larger character sets and slightly higher than in the CDP inventory that is constructed on similar principles. It is not

[229] Wang 1983: 67-68.

[230] Stalph 1989: 29.

[231] Idid., 81-115.

[232] Ibid., 69-72.

[233] CNS 11643-2, GF3001-1997 and Chinese Documents Processing Lab (CDP) – see Sections 4.4. and 5.1.2.

obvious what the exact nature of the quantitative relations are between the number of characters and the number of components. As a general tendency, the increase of characters must be correlated with the increase in the number of components, but at some point some components may be replaced by a smaller number of more basic ones. Stalph's inventory of basic components cannot serve as a direct reference to the results of graphotactic analysis designed for this study also for some other reasons than incompatibility in size of the initial character sets. The inventory is not based exclusively on the *jōyō kanji*, and in Stalph's own words: *"Die Analyse geht mithin in nicht wenigen Fällen weit über das Corpus der Jōyōkanji hinaus."*[234] For that reason it is actually difficult to determine the actual size of the initial set of characters. Also, *jōyō kanji* contains characters that are idiosyncratic to Japan and, more importantly, the rules for isolating the components are not identical – a brief examination of Stalph's inventory shows that the IDS decomposes many of its components into more basic elements. Still, from the perspective of the grammar of Sino-Japanese script it is a remarkable and important work.

Every isolated component is described by providing the following information:[235]

94 五	reference number component
五悟語	list of characters
3 (0.15%); KKWJ 322.	number of characters (%); external reference
Distr.: Gruppe I (– links, – unten)	distributional class[236]

Tab. 5.8 Stalph's distributional classes[237]

Class	Features	Examples	Class	Features	Examples
A	+ free	工口耳	J	– left, – above	及互本
B	– below	丘申尚	K	– horizontal	一屮 龶
C	– above	了石金	L	– vertical, – right	亻彳冫
D	– right	矛卵采	M	– vertical, – left	乎印東
E	– left	人下免	N	– horizontal, – below	宀 ⺾ 雨
F	– vertical	半身夜	O	– horizontal, – above	儿 ⺀ 小
G	– right, – below	乂北夕	P	– horizontal, – vertical	凸承喪
H	– right, – above	冫 匹歹	Q	+ enclosing	冂匚口
I	– left, – below	士天四			

[234] Ibid., 68.

[235] Ibid., 87.

[236] See Tab. 5.8.

[237] Ibid., 116-119.

Stalph provides extensive statistical data regarding *jōyō kanji*:[238]

- frequency of components;
- average number of strokes;
- frequency of characters correlated with the internal structure;
- frequency of structures;
- complexity of characters;
- correlation of complexity and structural type.

The statistical and quantitative studies of Chinese script are discussed in details in the next chapter, but since the last two types of data are most pertinent to the graphotactic analysis discussed in Chapter 7, it is justified to present the data here:

Tab. 5.9 Complexity of *jōyō kanji* in terms of number of components[239]

Number of components	Kanji	%
1	250	12.58
2	803	41.29
3	570	29.31
4	258	13.26
5	58	2.98
6	6	0.31
Total	1,945	100

Tab. 5.10 Complexity of *jōyō kanji* in terms of the number of components correlated with the structure types[240]

Number of components	Types of structure				Total
	□	⿰	⿱	⿸	
1	250	-	-	-	250
2	-	477	219	107	803
3	-	350	162	58	570
4	-	162	70	26	258
5	-	40	16	2	58
6	-	4	1	1	6
Total	250	1,033	468	194	1,945

[238] Some of the data refer to more extensive character sets.

[239] Stalph 1989: 120.

[240] Ibid., 128.

The determination of the inventory of basic components and the classes of distribution was necessary to formulate the rules of grammaticality and well-formedness for *kanji* which are the lasting value of Stalph's analysis. The rules are introduced below in a slightly modified order:[241]

1) *Kanji* consist of a single component ("*grapheme*") or a combination of at most 8 components.
2) The distributional properties of components render the following structures agrammatical (for typographical reasons the letters are arranged horizontally – the first letter in a sequence represents the top part):

horizontal structures:	XD	XG	XH	XK	XL	XN	XO
	EX	IX	JX	KX	MX	NX	OX
vertical structures:	XB	XF	XG	XI	XL	XM	XN
	CX	FX	HX	JX	LX	MX	OX

where 'X' represents any class.

The following 64 horizontal and 64 vertical structures are well-formed (again, for typographical reasons the vertical structures are represented by horizontally arranged letters):

horizontal:

AA	AB	AC	AE	AF	AI	AJ	AM
BA	BB	BC	BE	BF	BI	BJ	BM
CA	CB	CC	CE	CF	CI	CJ	CM
DA	DB	DC	DE	DF	DI	DJ	DM
FA	FB	FC	FE	FF	FI	FJ	FM
GA	GB	GC	GE	GF	GI	GJ	GM
HA	HB	HC	HE	HF	HI	HJ	HM
LA	LB	LC	LE	LF	LI	LJ	LM

vertical:

AA	AC	AD	AE	AH	AJ	AK	AO
BA	BC	BD	BE	BH	BJ	BK	BO
DA	DC	DD	DE	DH	DJ	DK	DO
EA	EC	ED	EE	EH	EJ	EK	EO
GA	GC	GD	GE	GH	GJ	GK	GO

[241] Ibid., 132-138.

IA	IC	ID	IE	IH	IJ	IK	IO
KA	KC	KD	KE	KH	KJ	KK	KO
NA	NC	ND	NE	NH	NJ	NK	NO

3) Structures with more than 4 components arranged horizontally are agrammatical.
4) Structures with more than 4 components arranged vertically or embedded in an enclosing structure are agrammatical.
5) Multicomponential structures that exceed 36 strokes are agrammatical.
6) Free forms (single component characters) and well-formed multicomponential characters may bemultiplied – doubled, tripled and quadrupled, given that other rules are observed.
7) Duplications and quadruplications may assume the function of an enclosing structure.[242]
8) Enclosing structures (‘*kamae*’) may be divided into two categories: pure *kamae* and structures formed by the components primarily belonging to other distributional classes.[243]
9) P class components either appear as standalone characters or their distribution is confined to single graphical compounds.

The well-formedness rules should be evaluated against a specific set of characters, components and distributional classes. In this respect the Stalph’s grammar seems to be viable. The validity of Stalph’s graphotactic analysis was tested against the Menzerath-Altmann Law[244] with positive results, which is additional evidence supporting the legitimacy of corpus-based decomposition.

5.3. Graphotactics related studies

The studies involving the combinability of components of Chinese characters are scarce, and the analyses that focus precisely on the quantitative aspect of components combination are even rarer. The term ‘graphotactics’ was used in the context of Chinese characters by Stalph (1989), but its understanding was considerably different than in this book. There are very few works that have approached the subject in a way similar to the present study, in both method or research goal. The most notable three are the studies of Han (1994 and 1995), Chuang & Teng (2009) and Chen et al. (2011). Han’s study, as well as Chen’s (et al.) will be briefly introduced below. Some of the graphotactics related results presented in Chuang & Teng (2009) were discussed in

[242] This is rather a technical point.

[243] This, again, is a model-specific rule, and therefore, more details discussed by Stalph are left out here.

[244] See Section 6.2.2.

section 5.1.2., those more directly related to the graphotactic analysis will be presented in Chapter 7.

5.3.1. Component combination database

Han's analysis was concerned with the frequency of components of Chinese characters and the frequency of component combinations from the perspective cognitive psychology studies on letters and the frequency effect of groups of letters. Based on the assumed correspondence of component combinations in Chinese characters to the combinations of alphabetic script letters. He analyzed the composition of 6,763 characters in the GB2312-80 set using an inventory of 567 components ('the database of Chinese constituents').[245] As a result Han was able to show a correlation between the number and frequency of components – the distribution exhibited an uneven pattern with most components having a low frequency of occurrence in characters.[246] More interestingly his studies were also concerned with the combination of components occurring in characters. The general idea is similar to the notion of the tactographeme, but the actual analysis is based on a different type of unit. Han's 'combination of components' always involves two components, regardless of the complexity of a character to which the combining components belong. He gives an example of component combinations in the character 部 consisting of three components: 立, 口, 阝 – the character is assigned three combinations of components ({立, 口}, {立, 阝}, {口, 阝}),[247] instead of one set of components (tactographeme). Han's analysis featured 7,583 combinations ('component combination database'[248]) in the GB 2312-80 character set. The distribution of combinations of components revealed an uneven pattern – the increase in frequency was correlated with the decrease in the number of combinations. Unfortunately, in the 1994 and 1995 papers Han outlined only a general summary of the analysis and examples of the detailed results of his research.

5.3.2. Chinese orthography database

The study of Chen et al. (2011) also recognizes the relevance of componential structure as an important variable in the psychological processes associated with the acqui-

[245] The inventory was based on a computer analysis published in 1988 on the frequency and information processing dictionary of Chinese characters – 汉字信息字典 (Han 1994: 148).

[246] Han 1994: 148-149.

[247] Ibid., 149; Han 1995: 27.

[248] Han 1994: 148.

sition and recognition of characters. Their efforts concentrated on designing and creating a knowledge database of Chinese orthography. The primary concern was to facilitate both the process of teaching and learning traditional Chinese characters and the reaserch on the structure of characters. The authors selected a set of 6,097 frequently used characters that was based on the BIG5 level 1 (5,401 characters) and a set designed by the Chinese Knowledge and Information Processing Group of Academia Sinica (5,656 characters). Based on their selected character set, 439 components were isolated, and 246 of which correspond to the standalone characters, while the remaining 193 function only as components.[249] The authors of the discussed study endeavored to:[250]

- establish a database of the componential structure of traditional Chinese characters;
- analyze the properties and frequencies of components;
- establish a statistical relationship of component occurence with character structural types;
- detail the structural properties of the complex characters involving side components (邊旁 *biānpáng*).

The structural information is encoded by means of 11 descriptors, similar in function to the Unicode's IDCs and CDP descriptors. For practicality's sake, there are no special graphical symbols representing different types of structure, but instead, standard symbols easily accessible on any computer system are used:[251]

Tab. 5.11 Examples of structural descriptons

Structure type	Symbol	Expression example
Standalone character	X	內 = X
Top-bottom	–	貢 = – (工, 貝)
Left-right	|	烤 = | (火,考)
Full enclosure	0	困 = 0 (囗, 木)
Left-top enclosure	/	仄 = / (厂, 人)
Right-top enclosure	\	或 = \ (戈, – (口, 一))
Left-bottom enclosure	L	超 = L (走, – (刀, 口))
Top-square enclosure	^	凰 = ^ (几, – (白, 王))
Bottom-square enclosure	V	凶 = V (凵, 乂))
Left-square enclosure	<	匪 = < (匚, 非))
Left-right spread	T	夾 = T (大, 人, 人)

[249] Chen et al. 2011: 272.

[250] Ibid., 271.

[251] Ibid., 274-275.

The orthography database stores many kinds of data regarding the structural properties of characters and components. For example, each component is assigned a list of characters generated out of it reflecting the generative complexity. This is illustrated below with he example of the 牛*niú* ‘cow’ component:[252]

牛 30 件朱 (侏株殊珠茱硃蛛誅跦銖姝洙) 牟 (眸) 牢牽 (縴) 犁犀 (墀遲) 解
(懈邂蟹廨) 犖犛

There is no point in quoting here all the statistical data found in the study – identical or similar information obtained from the larger sets of characters are provided in different sections of this book. It should be noted that the project introduced in Chen et al. (2011) is probably the most complete database on the structural properties of characters and their components, including all kinds of quantitative information concerning the frequency of components and correlation with the structure types. The authors are primarily concerned with the applications facilitating writing acquisition, but a database in this format can be accessory to many types of Chinese script analysis. The purposeful design of the database results in its most serious limitation – its size. 6,097 frequently used characters is enough for educational and didactic purposes, but the suitability for large-scale research is restricted.

5.4. Psycholinguistics related studies

The present study is not concerned with the psycholinguistic status of components and componential structure of characters, but as it was mentioned in the previous sections, the componential analysis of Chinese characters is often motivated by psycholinguistic investigations. In fact, research on components and componential structures is probably most common in the psycholinguistic context. Apart from the sparse studies motivated by the results of psycholinguistic research, there are many more purely psycholinguistic studies on the acquisition of writing, reading, recognition of Chinese characters and the psycholinugistics of Chinese language processing in general. These studies involve, to some extent, the discussion of the componential structure and role of components in the mental processing of Chinese script. The following are selected examples: Yin (1991), Matsunaga (1994), Feldman & Siok (1997 and 1999), Taft & Zhu (1997), Wang et al. (1999), Ding et al. (2004), Xing et al. (2004), Tan et al. (2005a), Tan et al. (2005b), Perfetti et al. (2006), Lo et al. (2007), Wang & Yang (2008), Bi et al. (2009), Chen & Yeh (2009), Li (2009), and Vanderschot (2011). Chinese ortography is

[252] Ibid., 284.

valuable in this respect, because it offers perspectives and opportunities unavailable in any other type of script.[253]

5.5. Mathematical models

Another approach that should be mentioned here is the mathematical modeling of the structure of Chinese characters. Mathematics can be an effective modeling tool for different purposes, usually only indirectly related to purely linguistic considerations. The typical applications of mathematical tools for the description of Chinese characters pertain to information processing, more precisely to optical recognition of character technology (OCR, including the handwiring), automatic segmentation of characters, and representation of the structure of characters in computer systems. Studies of this type, often involving very spphisticated mathematical apparatuses, are important primarily for their practical value, though often enough they also offer some theoretical insights relevant to a linguistic perspective. Some of the notable studies using this approach include: Fujimura & Kagaya (1969), Stallings (1975), Thompson (1980), Lai et al. (1996), Iwanowski (2004), Liu (2008), Liu et al. (2010), and Jin et al. (2012), just to name a few.

[253] Packard 2000: 3.

6. Quantitative studies of Chinese script

6.1. Traditional statistical and quantitative studies

The properties of Chinese script make it a natural subject for statistical and quantitative analysis. In fact, statistical information regarding the structural and pragmatic aspects of Chinese characters has a fundamental importance for language policy, planning and teaching. In general, it is not uncommon to apply statistical methods to language-related fields, especially in language teaching and testing, but what sets the Chinese language apart is the extent to which statistics are used, and most of all, the relevance of statistical methods applied to the writing system. In the case of languages with some sort of alphabetic writing system, the statistical information on the elements of script (e.g. letter frequency lists) is of secondary importance at best. By contrast, in the case of Chinese, script-related inquiries are at least as important as statistical studies that pertain to other aspects of the language. This section briefly introduces the major types of script-related statistical studies of Chinese.

6.1.1. Frequency of characters

Frequency lists of Chinese characters are one of the most common forms of statistical research on the subject. They play an important role in educational designs, thus facilitating the efforts to increase the literacy rate among the Chinese and enhance the efficiency of the teaching process. Being entirely dependent on the corpus type and size, the lists may differ slightly, but the contents are rather stable. Table 6.1 contains a list of the 20 most-frequently used characters:

1. Leeds University corpus of Internet Chinese,[254]
2. Jun Da's website on Chinese text computing,[255]
3. Beijing University Center for Chinese Linguistics[256]
4. Taiwan Ministry of Education,[257]
5. Kanji Character Frequency List:[258]

[254] 281,660,631 characters – http://corpus.leeds.ac.uk/frqc/i-zh-char.num.

[255] 258,852,642 characters – http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=MO and Da 2004.

[256] 307,317,060 characters – http://ccl.pku.edu.cn:8080/ccl_corpus/.

[257] 1,051,159 characters – http://www.edu.tw/files/site_content/M0001/86news/ch2.html?open.

[258] More than 23,000,000 characters – Chikamatsu et al. 2000.

Tab. 6.1 20 most frequently used characters

Rank	1.	2.	3.	4.	5.
1	的	的	的	的	日
2	一	一	一	一	一
3	是	是	是	是	十
4	我	不	了	有	二
5	了	了	在	在	人
6	不	在	人	人	大
7	在	人	不	不	年
8	有	有	国	大	会
9	人	我	有	中	国
10	个	他	中	為	三
11	这	这	他	以	本
12	他	个	这	國	長
13	上	们	我	會	中
14	大	中	和	上	五
15	来	来	大	了	出
16	到	上	个	我	事
17	中	大	上	年	社
18	们	为	为	時	是
19	就	和	年	來	者
20	说	国	地	這	月

This section is not intended to discuss in detail the sources of discrepancies between the characters on different lists, but a few may be pointed out immediately:

- different sizes of corpuses;
- different source of corpuses;
- different set of characters;
- different languages – Chinese (1.-4.), Japanese (5.);
- different inclusive years of the sources surveyed.

Another frequency-related set of statistics refers to accumulative frequency of characters in relation to the readability of the body of texts. This is an indispensable tool for a precise compilation of the frequency based character lists which are the basis for educational policy regarding literacy acquisition and character teaching. Table 6.2 is a summary of the first computer assisted statistical study of corpus of simplified character texts; it shows the accumulative frequencies of graded character sets, additionally correlated with the totals and average number of strokes for each set:[259]

[259] Su 2001: 35.

Tab. 6.2 Accumulative frequency of characters

Graded Character Sets	Character ranks	Accumulative Frequency (%)	Total Number of Strokes	Average Number of Strokes
I. Grade (most frequently used characters)	1-500	77.419	3,622	7.244
II. Grade (frequently used characters)	501-1000	90.819	4,355	8.710
III. Grade (secondary frequently used characters)	1001-1500	95.898	4,840	9.680
IV. Grade (scarcely used characters)	1501-3000	99.597	15,655	10.437
V. Grade (rare characters)	3001-5991	100.000	23,682	11.599
Total:	5991	100.000	63,154	10.541

Accumulative frequency is a textual coverage ratio, meaning that the knowledge of 500 characters allows a person to read 77% of texts. The extracted data on the relation of the frequency of characters sets with the textual coverage ratio and the average number of strokes can be transposed onto a diagram. This graphically illustrates the rapidly decreasing gain in the text coverage ratio and steadily increasing average number of strokes. This is shown in Fig. 6.1.

It is a stunning realization that nearly 3,000 characters is necessary to cover less than 0.5% of the texts, while the first 3,000 frequency-ranked characters cover more than 99.5%. Viewed from the language teaching perspective, the drastically dropping ratio of learning effort to reading efficiency is an important factor in teaching process designs and literacy assessments. Li (1988) in a detailed comparative study of 3 frequency lists based on different corpuses (in terms of size and type of texts) and the official lists of frequently used characters compiled by the Chinese governmental institutions came to a conclusion that in fact only the first 2,500 characters have a frequency high enough (accumulating to 99% of the text coverage) to be taught during primary school.

The increasing average number of strokes with the increasing character ranks is not surprising, it can statistically be explained by reformulation of a corollary of Zipf's law corollary: shorter words are used more frequently, meaning simpler (in terms of number of strokes) characters are statistically used more frequently. The statistics of the number of strokes will be covered in Section 6.1.4.

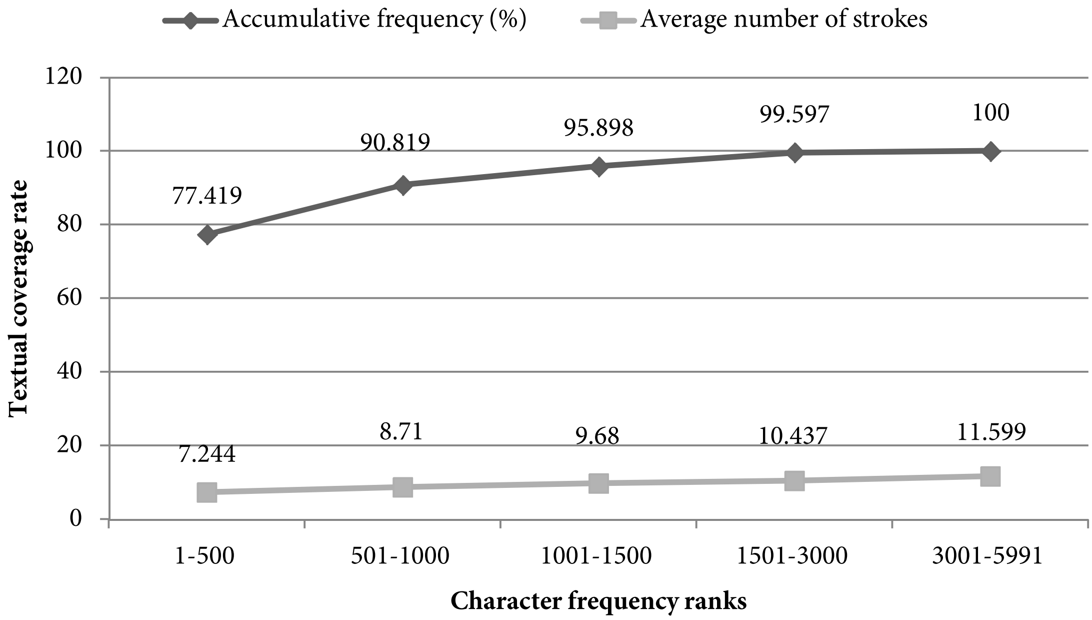


Fig. 6.1 Correlation of average number of strokes and textual coverage rate with the frequency ordered characters

6.1.2. Frequency of components

The traditional treatment of the frequency of components is usually understood as a frequency of radicals. Since the notion of 'radicals' was almost completely ignored in this study[260] it might be a good opportunity to present some quantitative data pertaining to the indexing radicals (康熙 *Kāngxī* radicals). The table below was compiled by the Taiwan Ministry of Education, based on a corpus of 4,667 characters. The radicals are listed in the traditional order.

Tab. 6.3 Frequency of 康熙 *Kāngxī* radicals

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
1	一	01	18	0.38	35,842	3.4097
2	丨	01	3	0.06	6,085	0.5788
3	丶	01	4	0.08	2,553	0.2428
4	丿	01	10	0.21	4,590	0.4366
5	乙	01	7	0.14	5,469	0.5202
6	亅	01	3	0.06	7,459	0.7095

[260] See Section 4.1.

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
7	二	02	10	0.21	5,879	0.5592
8	亠	02	9	0.19	2,090	0.1988
9	人	02	188	4.02	69,884	6.6482
10	儿	02	16	0.34	5,510	0.5241
11	入	02	4	0.08	6,273	0.5967
12	八	02	11	0.23	10,300	0.9798
13	冂	02	6	0.12	3,428	0.3261
14	冖	02	6	0.12	161	0.0153
15	冫	02	10	0.21	614	0.0584
16	几	02	4	0.08	108	0.0102
17	凵	02	5	0.10	4,344	0.4132
18	刀	02	46	0.98	18,435	1.7537
19	力	02	27	0.57	11,374	1.0820
20	勹	02	9	0.19	1,004	0.0955
21	匕	02	4	0.08	2,953	0.2809
22	匚	02	6	0.12	253	0.0240
23	匸	02	4	0.08	1,514	0.1440
24	十	02	14	0.29	7,135	0.6787
25	卜	02	4	0.08	735	0.0699
26	卩	02	9	0.19	2,369	0.2253
27	厂	02	8	0.17	1,240	0.1179
28	厶	02	2	0.04	2,214	0.2106
29	又	02	11	0.23	7,562	0.7193
30	口	03	239	5.12	42,495	4.0426
31	囗	03	23	0.49	13,787	1.3116
32	土	03	85	1.82	24,394	2.3206
33	士	03	5	0.10	884	0.0840
34	夂	03	1	0.02	1	0.0000
35	夊	03	1	0.02	121	0.0115
36	夕	03	8	0.17	6,260	0.5955
37	大	03	25	0.53	11,395	1.0840
38	女	03	91	1.94	11,743	1.1171
39	子	03	23	0.49	7,638	0.7266
40	宀	03	53	1.13	15,780	1.5012
41	寸	03	11	0.23	8,076	0.7682
42	小	03	4	0.08	3,183	0.3028
43	尢	03	5	0.10	3,459	0.3290
44	尸	03	24	0.51	4,435	0.4219
45	屮	03	1	0.02	16	0.0015

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
46	山	03	42	0.89	1,652	0.1571
47	巛	03	3	0.06	268	0.0254
48	工	03	6	0.12	3,019	0.2872
49	己	03	5	0.10	2,944	0.2800
50	巾	03	27	0.57	8,315	0.7910
51	干	03	6	0.12	6,027	0.5733
52	幺	03	4	0.08	938	0.0892
53	广	03	36	0.77	7,193	0.6842
54	廴	03	3	0.06	1,508	0.1434
55	廾	03	5	0.10	301	0.0286
56	弋	03	1	0.02	1,355	0.1289
57	弓	03	18	0.38	2,712	0.2580
58	彐	03	2	0.04	27	0.0025
59	彡	03	7	0.14	2,250	0.2140
60	彳	03	31	0.66	11,786	1.1212
61	心	04	158	3.38	22,020	2.0948
62	戈	04	15	0.32	11,464	1.0906
63	戶	04	6	0.12	4,028	0.3831
64	手	04	225	4.82	26,193	2.4918
65	支	04	1	0.02	530	0.0504
66	攴	04	29	0.62	10,210	0.9713
67	文	04	5	0.10	2,080	0.1978
68	斗	04	4	0.08	836	0.0795
69	斤	04	7	0.14	4,037	0.3840
70	方	04	9	0.19	7,506	0.7140
71	无	04	1	0.02	156	0.0148
72	日	04	74	1.58	25,198	2.3971
73	曰	04	9	0.19	8,742	0.8316
74	月	04	12	0.25	15,264	1.4521
75	木	04	204	4.37	29,336	2.7908
76	欠	04	16	0.34	3,105	0.2953
77	止	04	10	0.21	5,516	0.5247
78	歹	04	11	0.23	759	0.0722
79	殳	04	8	0.17	868	0.0825
80	毋	04	5	0.10	1,747	0.1661
81	比	04	1	0.02	1,225	0.1165
82	毛	04	4	0.08	300	0.0285
83	氏	04	3	0.06	2,456	0.2336
84	气	04	9	0.19	1,061	0.1009

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
85	水	04	272	5.82	30,050	2.8587
86	火	04	82	1.75	13,987	1.3306
87	爪	04	4	0.08	611	0.0581
88	父	04	4	0.08	575	0.0547
89	爻	04	2	0.04	551	0.0524
90	爿	04	1	0.02	66	0.0062
91	片	04	4	0.08	1,466	0.1394
92	牙	04	1	0.02	104	0.0098
93	牛	04	15	0.32	3,183	0.3028
94	犬	4	39	0.83	2,423	0.2305
95	玄	05	2	0.04	791	0.0752
96	玉	05	66	1.41	8,763	0.8336
97	瓜	05	5	0.10	73	0.0069
98	瓦	05	5	0.10	197	0.0187
99	甘	05	3	0.06	553	0.0526
100	生	05	4	0.08	5,862	0.5576
101	用	05	5	0.10	2,684	0.2553
102	田	05	26	0.55	7,966	0.7578
103	疋	05	2	0.04	304	0.0289
104	疒	05	52	1.11	2,153	0.2048
105	癶	05	3	0.06	3,373	0.3208
106	白	05	10	0.21	38,012	3.6161
107	皮	05	2	0.04	406	0.0386
108	皿	05	19	0.40	1,986	0.1889
109	目	05	58	1.24	8,585	0.8167
110	矛	05	2	0.04	34	0.0032
111	矢	05	7	0.14	1,372	0.1305
112	石	05	48	1.02	3,556	0.3382
113	示	05	32	0.68	4,847	0.4611
114	禸	05	3	0.06	1,205	0.1146
115	禾	05	40	0.85	7,040	0.6697
116	穴	05	24	0.51	2,222	0.2113
117	立	05	7	0.14	2,721	0.2588
118	竹	06	68	1.45	9,009	0.8570
119	米	06	27	0.57	1,401	0.1332
120	糸	06	118	2.52	22,606	2.1505
121	缶	06	6	0.12	303	0.0288
122	网	06	14	0.29	1,340	0.1274
123	羊	06	14	0.29	3,748	0.3565

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
124	羽	06	18	0.38	771	0.0733
125	老	06	4	0.08	3,885	0.3695
126	而	06	3	0.06	3,610	0.3434
127	耒	06	5	0.10	99	0.0094
128	耳	06	19	0.40	3,023	0.2875
129	聿	06	4	0.08	80	0.0076
130	肉	06	90	1.92	10,778	1.0253
131	臣	06	4	0.08	356	0.0338
132	自	06	2	0.04	2,948	0.2804
133	至	06	4	0.08	2,138	0.2033
134	臼	06	5	0.10	4,460	0.4242
135	舌	06	5	0.10	194	0.0184
136	舛	06	2	0.04	185	0.0176
137	舟	06	14	0.29	1,082	0.1029
138	艮	06	2	0.04	339	0.0322
139	色	06	2	0.04	967	0.0919
140	艸	06	180	3.85	11,845	1.1268
141	虍	06	9	0.19	1,982	0.1885
142	虫	06	69	1.47	1,662	0.1581
143	血	06	2	0.04	346	0.0329
144	行	06	8	0.17	5,076	0.4828
145	衣	06	48	1.02	7,108	0.6762
146	襾	06	3	0.06	4,738	0.4507
147	見	07	11	0.23	5,088	0.4840
148	角	07	4	0.08	1,499	0.1426
149	言	07	125	2.67	27,335	2.6004
150	谷	07	2	0.04	93	0.0088
151	豆	07	8	0.17	338	0.0321
152	豕	07	6	0.12	896	0.0852
153	豸	07	4	0.08	164	0.0156
154	貝	07	53	1.13	11,036	1.0498
155	赤	07	3	0.06	96	0.0091
156	走	07	12	0.25	3,555	0.3381
157	足	07	52	1.11	4,186	0.3982
158	身	07	5	0.10	1,115	0.1060
159	車	07	34	0.72	6,453	0.6138
160	辛	07	8	0.17	1,182	0.1124
161	辰	07	3	0.06	356	0.0338
162	辵	07	85	1.82	26,162	2.4888

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
163	邑	07	24	0.51	6,809	0.6477
164	酉	07	28	0.59	2,445	0.2326
165	釆	07	4	0.08	254	0.0241
166	里	07	5	0.10	3,176	0.3021
167	金	08	113	2.42	6,722	0.6394
168	長	08	1	0.02	2,554	0.2429
169	門	08	28	0.59	7,244	0.6891
170	阜	08	44	0.94	9,245	0.8795
171	隶	08	1	0.02	16	0.0015
172	隹	08	19	0.40	4,061	0.3863
173	雨	08	26	0.55	5,127	0.4877
174	青	08	4	0.08	541	0.0514
175	非	08	3	0.0	971	0.0923
176	面	09	3	0.06	2,056	0.1955
177	革	09	14	0.29	335	0.0318
178	韋	09	3	0.06	171	0.0162
179	韭	09	1	0.02	1	0.0000
180	音	09	5	0.10	1,679	0.1597
181	頁	09	33	0.70	8,252	0.7850
182	風	09	7	0.14	982	0.0934
183	飛	09	1	0.02	348	0.0331
184	食	09	32	0.68	2,960	0.2815
185	首	09	1	0.02	478	0.0454
186	香	09	3	0.06	678	0.0645
187	馬	10	38	0.81	1,856	0.1765
188	骨	10	11	0.23	2,151	0.2046
189	高	10	1	0.02	2,336	0.2222
190	髟	10	7	0.14	312	0.0296
191	鬥	10	4	0.08	187	0.0177
192	鬯	10	1	0.02	44	0.0041
193	鬲	10	0	0.00	0	0.0000
194	鬼	10	8	0.17	332	0.0315
195	魚	11	24	0.51	609	0.0579
196	鳥	11	31	0.66	498	0.0473
197	鹵	11	4	0.08	79	0.0075
198	鹿	11	6	0.12	317	0.0301
199	麥	11	3	0.06	281	0.0267
200	麻	11	3	0.06	1,310	0.1246
201	黃	12	1	0.02	364	0.0346

No.	Radical	Number of strokes	Number of characters	%	Number of occurences	%
202	黍	12	3	0.06	124	0.0117
203	黑	12	11	0.23	3,113	0.2961
204	黹	12	0	0.00	0	0.0000
205	黽	13	0	0.00	0	0.0000
206	鼎	13	2	0.04	47	0.0044
207	鼓	13	1	0.02	118	0.0112
208	鼠	13	1	0.02	35	0.0033
209	鼻	14	1	0.02	45	0.0042
210	齊	14	2	0.04	77	0.0073
211	齒	15	6	0.12	194	0.0184
212	龍	16	2	0.04	255	0.0242
213	龜	16	1	0.02	40	0.0038
214	龠	17	0	0.00	0	0.0000

Some notable studies referring to frequency of components from different perspectives and in different contexts were already mentioned in previous sections: Stalph (1989), Han (1994 and 1995), Teng & Chuang (2009), and Chen et al. (2011). Appendix I contains the frequency ordered list of CDP components.

6.1.3. Componential complexity of characters

The number of constituents is a basic measure of the complexity of characters. The typical measure in this context is the number of strokes (discussed in the next section) that also has a practical function of ordering characters. Quantitative studies of Chinese script in terms of the number of components are much rarer, probably because their practical uses are more limited. An example of a study of this kind can be found in Su (2001). The summarized results of an analysis of a corpus of texts consisting of 21,656,578 characters, containing 7,785 unique *hànzi* are presented in Tab. 6.4.

The graphical representation of the data in Fig. 6.2 is similar to a Gaussian curve. It is difficult to directly compare the results with similar results rendered by the graphotactic analysis in Chapter 7. Su provides no information on the type of components used for the decomposition. The author of this book did not succeed in obtaining a copy of the original source of the data. Direct comparison with the results presented in Section 7.2.8 show considerable differences, but it is reasonable to infer from the data in Tab. 6.4 that the two analyses relate to different types of components. The analysis presented by Su can neither be directly compared to immediate components, nor to basic components extracted from the IDS descriptions.

Tab. 6.4[261] Correlation of complexity with frequency

Number of components	Number of characters	% of all characters	Number of occurences	Frequency of occurences (%)
1	323	4.149	5,611,317	25.910
2	2,650	34.040	10,191,803	47.061
3	3,139	40.321	4,652,330	21.482
4	1,276	16.391	1,046,913	4.834
5	323	4.149	142,005	0.656
6	70	0.899	11,192	0.052
7	3	0.038	1,017	0.005
8	1	0.013	1	0
Total	7,785	100.000	21,656,578	100.000

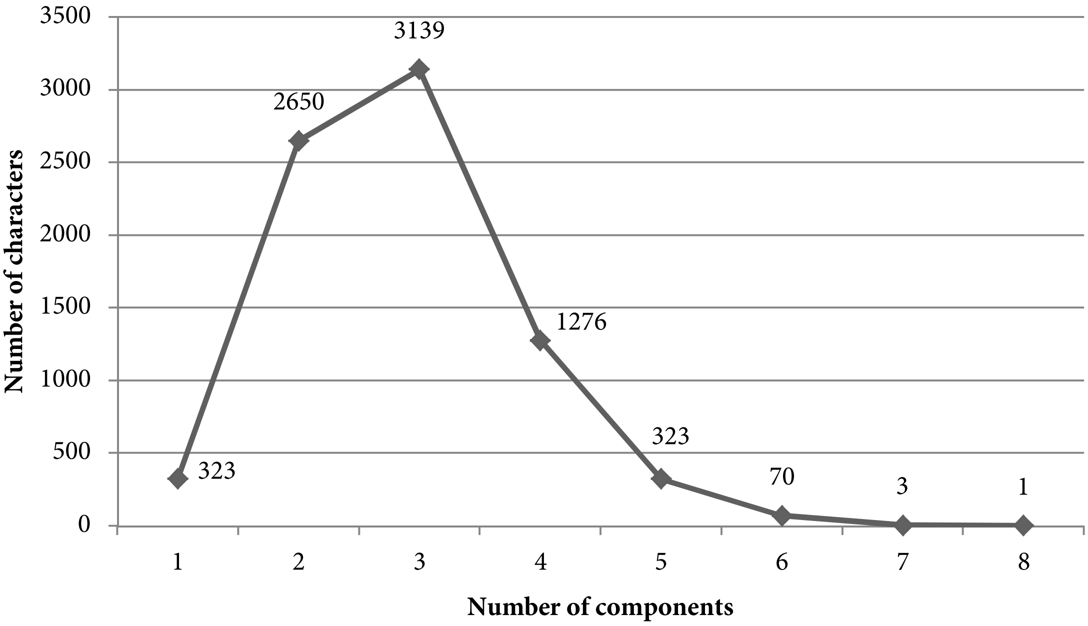


Fig. 6.2 Correlation of number of characters with componential complexity

The data provided by Su concerning the quantitative relations between the number of components, the number of characters, and their frequency is reformatted and presented below for a better illustration:

[261] Su 2001: 88. The data are taken from 汉字信息字典 *hànzi xìnxīzìdiǎn* ‘Dictionary of Chinese Character Information’, published in 1988.

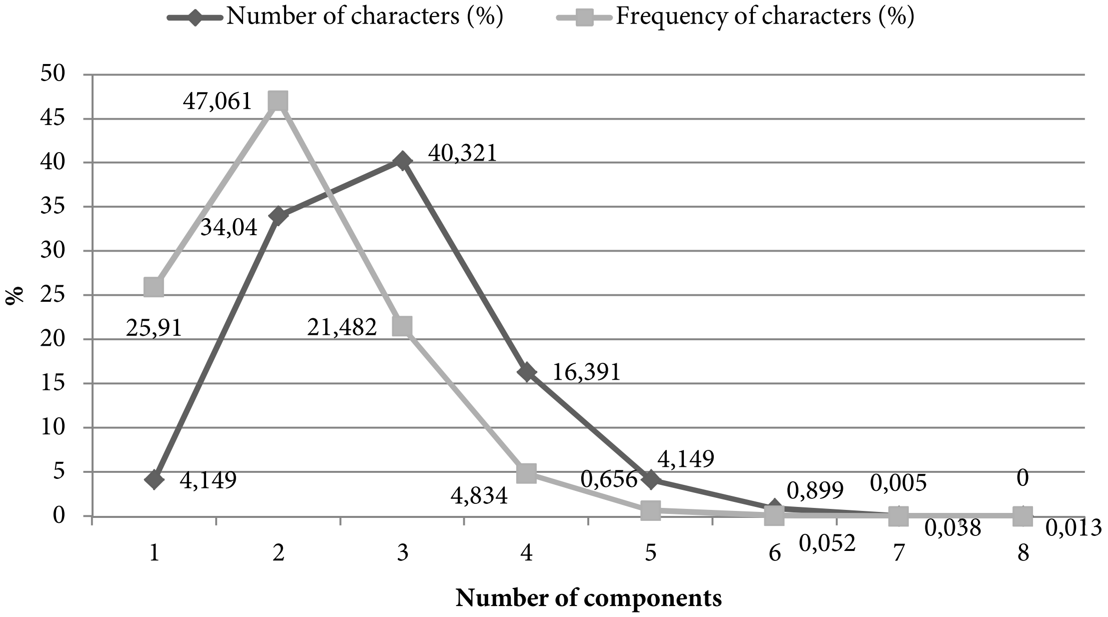


Fig. 6.3 Correlation of character complexity with their number and frequency

The data indicate that the *hànzi* consisting of 2 components are the most frequently appearing category, while *hànzi* composed of 3 components are the most numerous category in the character set.

The problem of complexity of characters in terms of the number of components will also be discussed in section 6.2.2. (Menzerath-Altmann hypothesis). Chapter 7 will provide detailed statistics on the subject in question from the graphotactical perspective.

6.1.4. Stroke statistics

Statistical analysis of Chinese script is also applicable to the atomic units of character structure – the strokes. Because of the practical aspect of the quantitative information on strokes, these types of studies are relatively common and offer diversified perspectives. Some of the most notable approaches are presented in this section.

6.1.4.1. Stroke number statistics

This type of study provides simple information on the total number of strokes by type in a given set of characters. This should not be confused with the stroke count which is a more common type of analysis and is discussed further on in this section. The exemplary data in Tab. 6.5 pertain to a set of 11,834 standard characters (正体字 *zhèngtǐzì*) containing both traditional and simplified characters out of a total number of strokes of 136,702 in the set. The statistical information regarding the number of the five basic stroke types is summarized in the Tab. 6.5:

Tab. 6.5[262] Stroke type statistics

Stroke type	Number of characters	Number of occurences	Frequency (%)
一	11,398	41,423	30.3023
丨	10,654	26,492	19.3792
丿	10,232	21,511	15.7261
丶	9,738	22,741	16.6351
乛	10, 890	24,535	17.9485

6.1.4.2. Stroke count

The most basic type of stroke number analysis is the calculation of the average number of strokes per character. The results depend heavily on the inventory type, with the most relevant features being the inventory size, degree of randomness (or the source of an inventory), and the traditional/simplified distinction. The calculations for a particular set of characters are very straightforward. For example, the average number of strokes per character for a frequency-based standardized set of 7,000 simplified characters (现代汉语通用字表 *xiàndài hànyǔ tōngyòngzì biǎo*) is 10.75.[263] The statistical analysis of stroke count may be much more sophisticated, especially in as a comparative aspect.

The simplification of Chinese script in the People’s Republic of China created two distinct sets of characters. One of the ways to measure the extent to which the sets differ, or what the measurable results of the simplification are, is the statistical analysis of strokes number in both traditional and simplified systems. This type of analysis facilitates language and character planning. It may also serve as a material basis for the psycholinguistic research on characters acquisition, recognition and speed of reading. Hard statistics of this type is often interpreted creatively by both sides of the debate on simplification. Political and non-academic discussions aside, the comparative study of stroke number offers interesting problems to explore and can deliver meaningful results. Also the fact that being the main target of script reform many high frequency characters were not simplified – a quick review of the top 20 lists in the previous sec-

[262] Su 2001: 71.

[263] Ibid., 67.

tion of the book reveals that on any of the four lists (the Japanese one may be considered irrelevant) there are at most only 5 simplified characters, which leaves a few interesting research possibilities. Below is a brief summary of the statistical and quantitative types of comparative stroke number analysis:

- one-to-one comparison of the simplified characters with their traditional equivalents;
- comparison of frequency graded characters sets;
- comparative analysis number of strokes in running traditional and simplified characters texts of the same size;
- comparative analysis of identical text corpuses in traditional and simplified characters.

It stands to reason that the first type of analysis reveals the largest difference in the average number of strokes. This is the case for any analysis confined to characters on the General List of Simplified Characters.

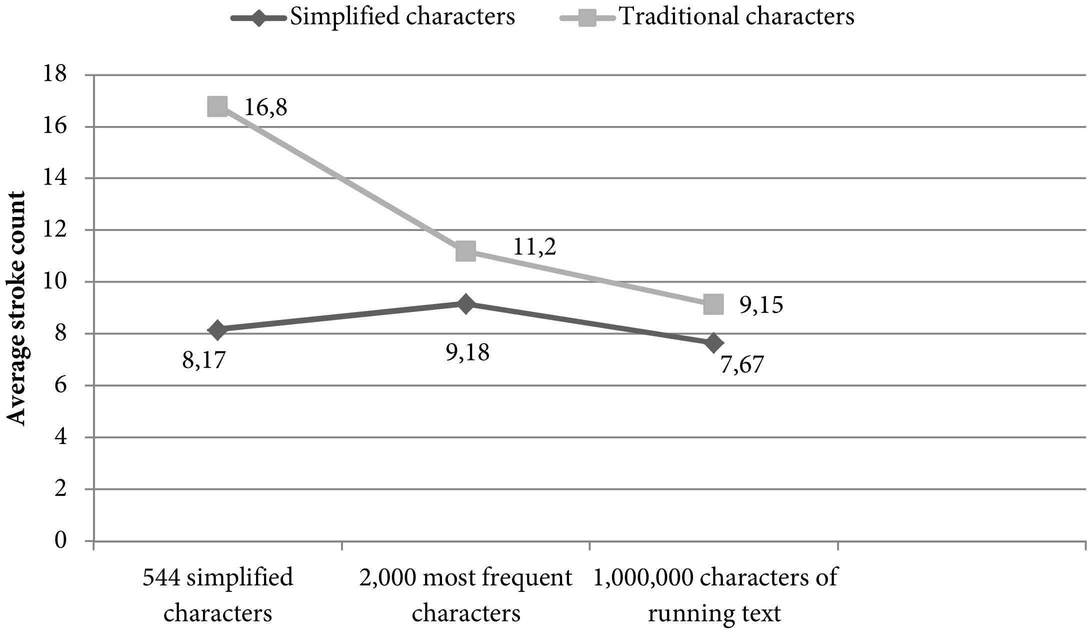


Fig. 6.4 Average stroke counts of traditional and simplified sets

Zhao & Baldauf (2008) cite interesting statistics regarding the quantitative outcome of the simplification reform that partially cover the above list.[264] The average number of strokes for 544 simplified characters is 8.17 compared to 16.08 for the traditional ones; this is a 50% reduction. The comparison of the frequency graded sets is more telling in terms of every day practice – in a set of 2,000 most frequently used characters the average number of strokes for traditional characters is 11.2 compared to 9.18 after

[264] Zhao &Baldauf 2008: 48.

simplification, which is a 12% reduction. The reduction is even less substantial in the case of corpus investigation – in the 1,000,000 characters of running texts the average number of strokes for traditional characters is 9.15 compared to 7.67 for the simplified ones, which is a 8.4% reduction (Fig. 6.4).

Guo (2009) performed a thorough analysis of differences in stroke numbers between traditional and simplified characters. Due to the need for cross-referencing the characters on the General List of Simplified Characters with the GB13000.1 standard (GB13000. 1 字符集:汉字字序(笔画序)规范 – 'Classifications of Chinese Radicals Character Set Specification') he compared 2,194 out of the 2,235 original characters (41 of the traditional equivalents are not found in GB13000.1). Guo's findings of the one-to-one comparison of the equivalent sets[265] are transposed into the diagram (Fig. 6.5).

The second diagram summarizes the results of Guo's analysis of the characters occurring in an actual corpus of texts. He found 1,128 characters from the General List of Simplified Characters that have traditional counterparts in the corpus. The statistical data on stroke count in both sets of 1,128 characters are summarized in Fig. 6.6.[266]

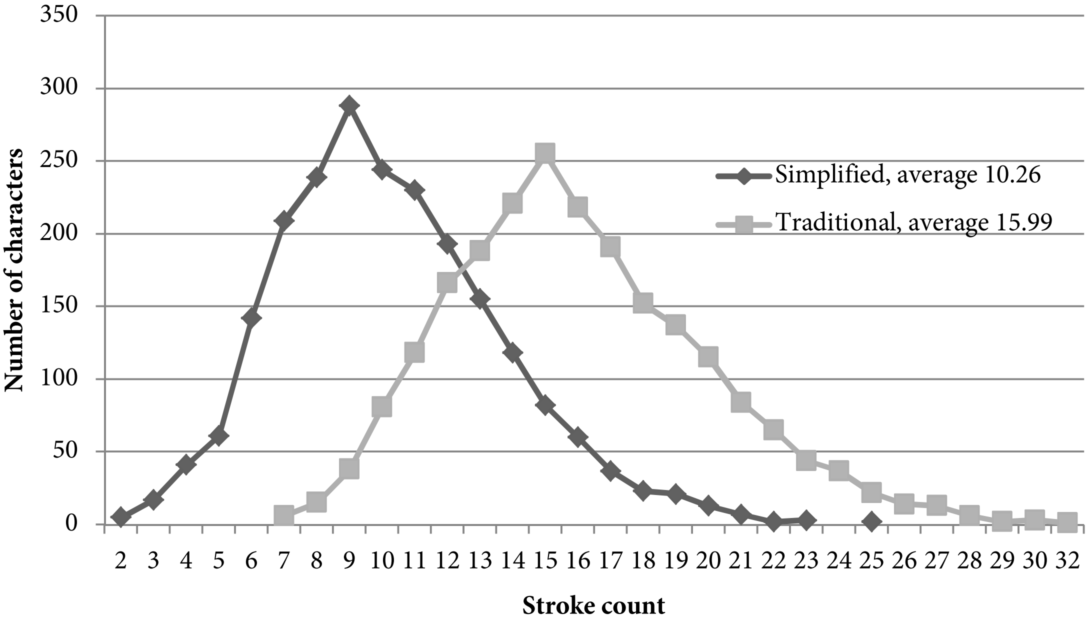


Fig. 6.5 Comparison of the characters on General List of Simplified Characters with traditional counterparts

Both charts indicate a similar shift in the number of characters belonging to each stroke number category as a result of the simplification. The most numerous categories of traditional characters are in the range of 12-18 strokes; the simplified ones in the

[265] Guo 2009: 52.

[266] Ibid.

range of 6-11 strokes. Before simplification, when one examines the 2,236 characters on the General List of Simplified Characters, there were only 141 characters consisting of less than 10 strokes; after the reform the increase was almost 10 fold to 1,236 characters, or up 6.3% to 56.4%.[267]

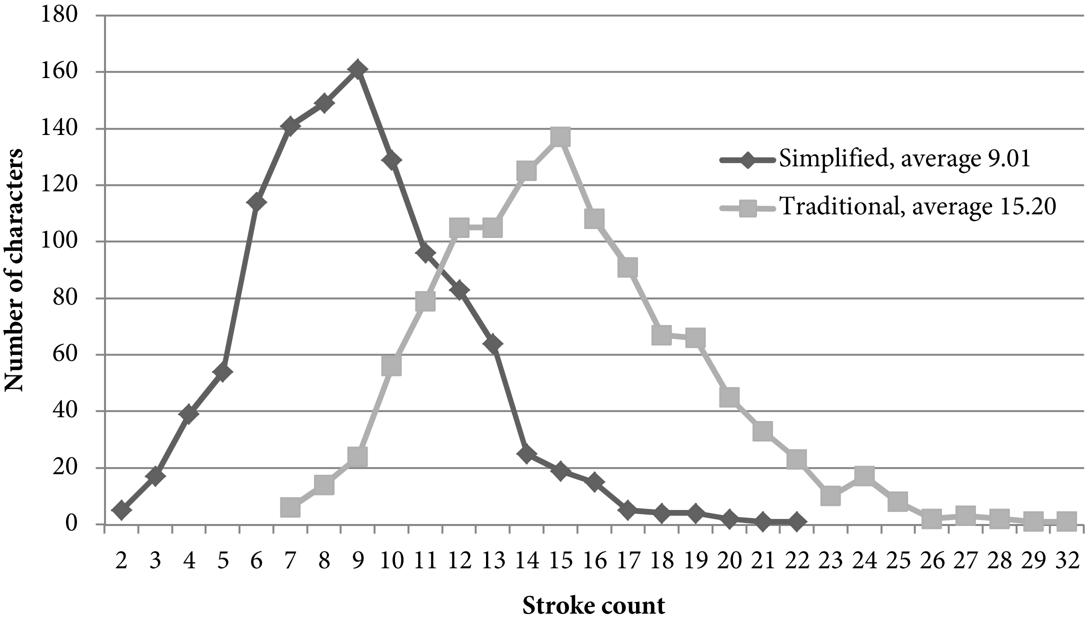


Fig. 6.6 Corpus-based comparison of simplified characters with traditional counterparts

Given the fact that the topics discussed in this section are not in the focus of this book, the remaining results of Guo's analysis will not be discussed here, but they are thorough and interesting enough to encourage the readers to become acquainted with the study.

The final part of this section presents a detailed results of stroke count investigation for different character sets filtered out from the Wenlin 4.1 database.[268] The sets analyzed in Table 6.6 differ in size and type of contained characters; all 5 were introduced in a more or less detailed way in Chapter 2.

The average values shown in Fig. 6.7 are not surprising, but without performing the actual calculations it can only be said with certainty that the traditional character sets (Big5) have a higher average stroke count. The relation of large sets containing both types of characters to the Big5 set in the discussed aspect is more difficult to estimate. In large heterogeneous sets the simplified *hànzi* are only a small fraction of the whole set and possibly can be balanced or outweighed by the numerous complex characters

[267] Zhao & Baldauf 2008: 48.

[268] An option available in the Wenlin software.

absent in the smaller sets of traditional characters. Closer examination of the results shows that the largest set by far it the one with the highest average stroke count, even though it contains a subset of simplified characters. The second largest set, also containing simplified characters, has a lower average than the traditional Big5 set. The 20,092 character set is not large enough to balance out the subset of simplified characters in comparison to the Big5 set, which is not much smaller. Unsurprisingly, the smallest set of simplified characters has the lowest average stroke count. The correlation of stroke count with the number of characters in each set are shown in the Fig. 6.8 and 6.9 – the sets were divided into two size categories for clearer comparison of the heterogeneous traditional set and the simplified set.

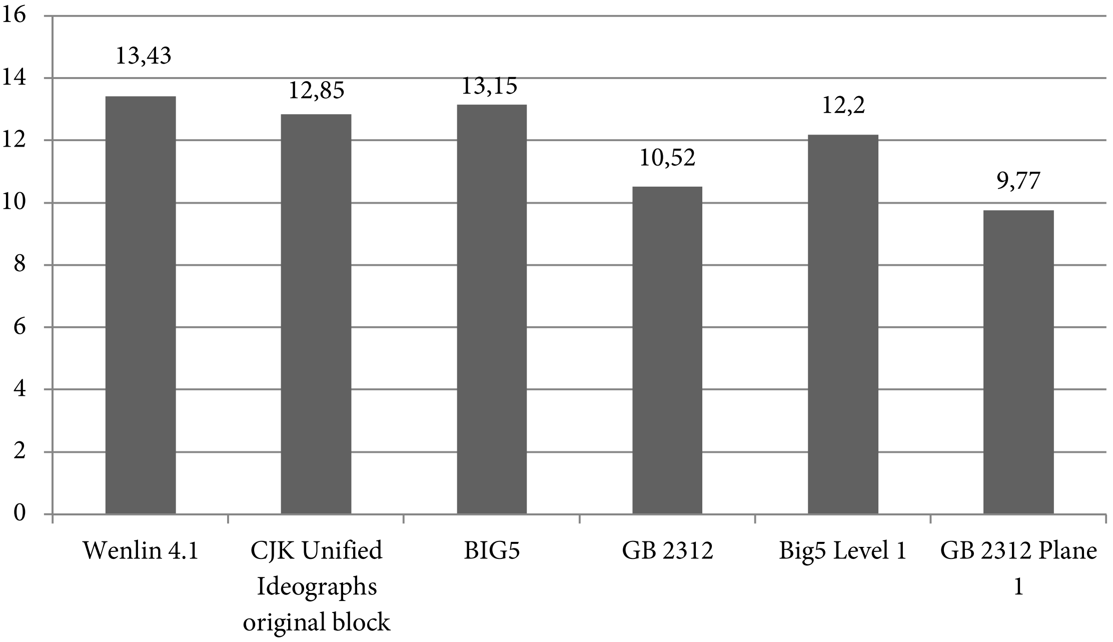


Fig. 6.7 Average stroke count for selected character sets

The plotted lines for large heterogeneous sets do not cross with each other and with Big5 set (Fig. 6.8), the lines for sets of simplified characters of different sizes also do not cross at any point. The only case when the graph lines cross is the case of simplified sets with traditional ones (Fig. 6.9).

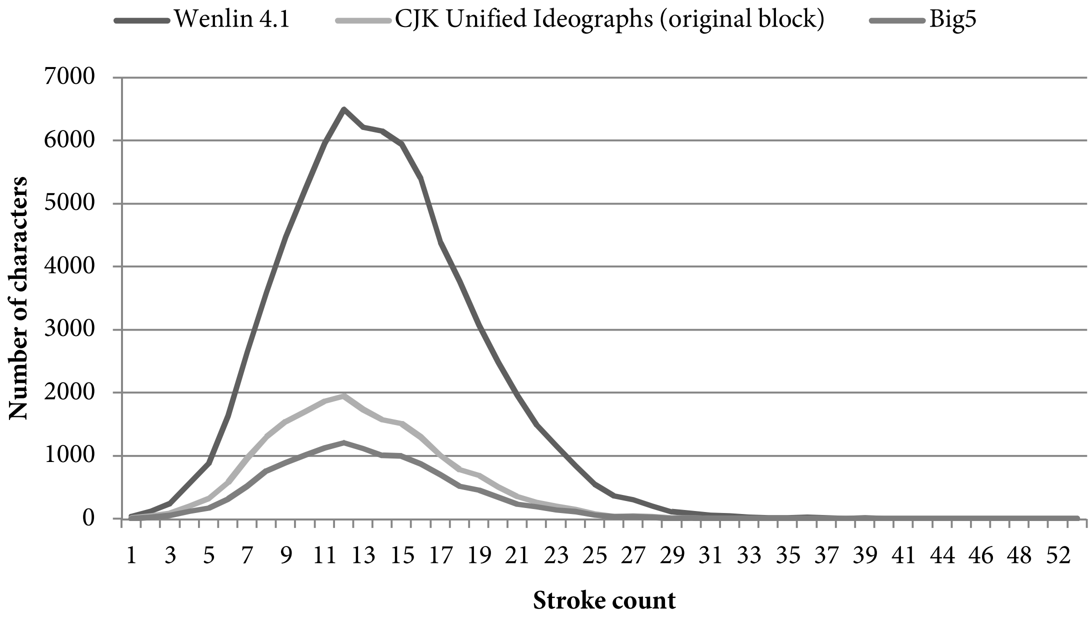


Fig. 6.8 Number of characters by stroke count categories in Wenlin 4.1, CJK Unified Ideographs (original block) and Big5.

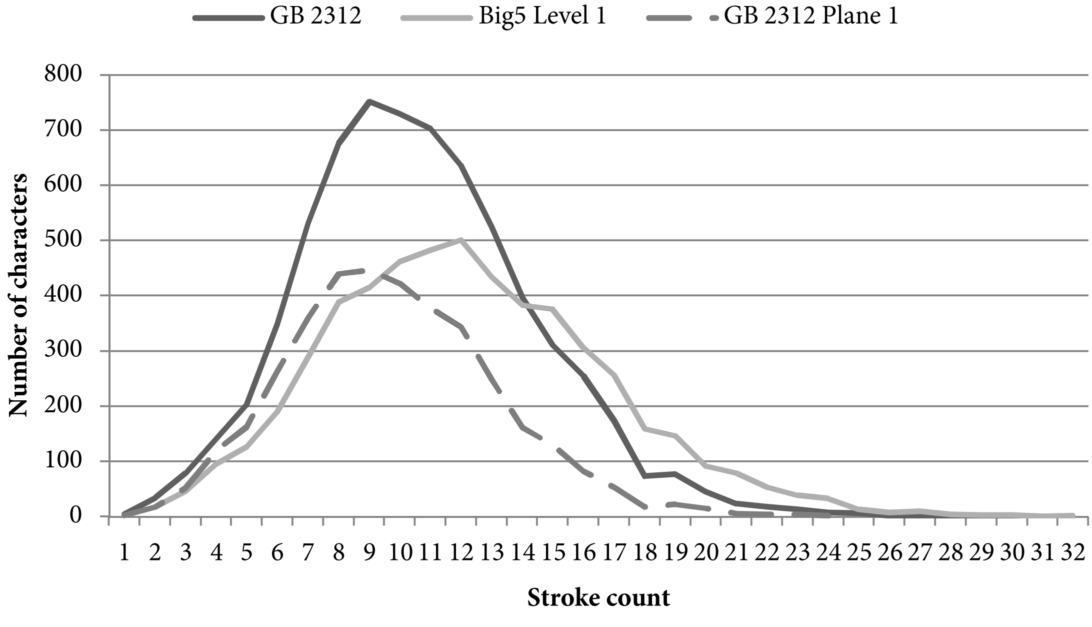


Fig. 6.9 Number of characters by stroke count categories in GB 2312, Big5 (Level 1), and GB 2312 (Plane 1)

Tab. 6.6 Stroke count statistics for selected character sets

Stroke count	Number of characters					
	Wenlin 4.1 (76,630)	CJK Unified Ideographs original block (20,092)	Big5 (13,061)[269]	GB 2312 (6,763)	Big5 Level 1 (5,411)	GB 2312 Plane 1 (3,755)
1	46	10	2	5	2	2
2	121	44	23	34	18	18
3	247	97	54	79	46	52
4	570	206	119	140	95	120
5	880	329	177	203	126	162
6	1,636	583	314	349	190	264
7	2,652	963	520	531	289	360
8	3,595	1,301	759	677	389	440
9	4,468	1,541	894	752	415	447
10	5,239	1,704	1,014	730	462	423
11	5,963	1,864	1,128	703	483	378
12	6,494	1,954	1,205	636	501	344
13	6,212	1,733	1,119	524	434	248
14	6,150	1,572	1,005	397	383	161
15	5,951	1,518	998	311	376	130
16	5,405	1,298	874	255	306	83
17	4,382	1,003	700	174	257	54
18	3,771	779	519	73	159	18
19	3,081	690	456	77	146	22
20	2,485	505	341	45	92	15
21	1,963	351	239	23	79	6
22	1,493	267	192	17	54	4
23	1,157	205	142	13	38	3
24	840	152	113	7	33	1
25	554	85	61	6	13	0
26	366	46	29	1	7	0
27	309	44	29	0	9	0
28	201	27	16	0	4	0
29	116	9	5	0	2	0
30	82	9	6	1	2	0
31	56	2	1	0	0	0
32	43	3	2	0	1	0
33	26	4	2	0	0	0

[269] The actual number of characters labeled as ‘Big5‘ in the Wenlin database.

Stroke count	Number of characters					
	Wenlin 4.1 (76,630)	CJK Unified Ideographs original block (20,092)	Big5 (13,061)[269]	GB 2312 (6,763)	Big5 Level 1 (5,411)	GB 2312 Plane 1 (3,755)
34	13	0	0	0	0	0
35	9	1	1	0	0	0
36	19	1	1	0	0	0
37	6	0	0	0	0	0
38	5	0	0	0	0	0
39	6	1	0	0	0	0
40	2	0	0	0	0	0
41	2	0	0	0	0	0
43	1	0	0	0	0	0
44	3	0	0	0	0	0
45	1	0	0	0	0	0
46	1	0	0	0	0	0
47	1	0	0	0	0	0
48	3	1	1	0	0	0
51	1	0	0	0	0	0
52	1	0	0	0	0	0
64	2	0	0	0	0	0
Average	13.90	12.85	13.16	10.62	12.21	9.77

6.1.4.3. Stroke count and character frequency

Another possible correlation of the discussed quantitative property of characters is the frequency of characters with a given stroke count. The results of this type of investigations were published by the Taiwan Ministry of Education (Tab. 6.7).[270]

Tab. 6.7. and Fig. 6.10 show that four of the stroke count based categories of traditional characters are most frequent: 8, 11, 6 and 9, meaning 8-stroke characters are most frequent, 11-stroke characters are ranked second, etc.

The types of quantitative analysis of Chinese script discussed in this section cannot be directly compared to any type of quantitative and statistical studies on alphabetical writing systems.

[270] http://www.edu.tw/files/site_content/M0001/86news/86rest6.html?open.

Tab. 6.7 Frequency of characters by stroke count categories

Rank	Number of Strokes	Number of Characters	Accumulated Characters	Frequency	Accumulated Frequency	Accumulated Frequency %
1	08	338	338	133,250	133,250	12.6760
2	11	438	776	93,539	226,789	21.5747
3	06	154	930	91,247	318,036	30.2553
4	09	345	1,275	88,149	406,185	38.6412
5	05	120	1,395	77,919	484,104	46.0539
6	12	411	1,806	68,954	553,058	52.6137
7	10	393	2,199	67,035	620,093	58.9909
8	13	384	2,583	66,998	687,091	65.3646
9	04	88	2,671	66,961	754,052	71.7348
10	07	254	2,925	57,274	811,326	77.1835
11	14	328	3,253	42,972	854,298	81.2716
12	03	36	3,289	35,887	890,185	84.6856
13	15	346	3,635	33,052	923,237	87.8299
14	02	17	3,652	26,695	949,932	90.3695
15	16	245	3,897	26,329	976,261	92.8743
16	17	230	4,127	20,209	996,470	94.7968
17	01	2	4,129	12,696	1,009,166	96.0046
18	19	119	4,248	9,682	1,018,848	96.9257
19	18	129	4,377	9,415	1,028,263	97.8214
20	20	82	4,459	6,773	1,035,036	98.4657
21	23	33	4,492	4,562	1,039,598	98.8997
22	21	64	4,556	4,144	1,043,742	99.2939
23	25	13	4,569	2,851	1,046,593	99.5651
24	22	52	4,621	2,740	1,049,333	99.8258
25	24	26	4,647	1,506	1,050,839	99.9691
26	27	9	4,656	116	1,050,955	99.9801
27	26	4	4,660	84	1,051,039	99.9881
28	29	1	4,661	44	1,051,083	99.9923
29	28	3	4,664	37	1,051,120	99.9958
30	32	1	4,665	36	1,051,156	99.9992
31	30	1	4,666	2	1,051,158	99.9994
32	31	1	4,667	1	1,051,159	99.0000

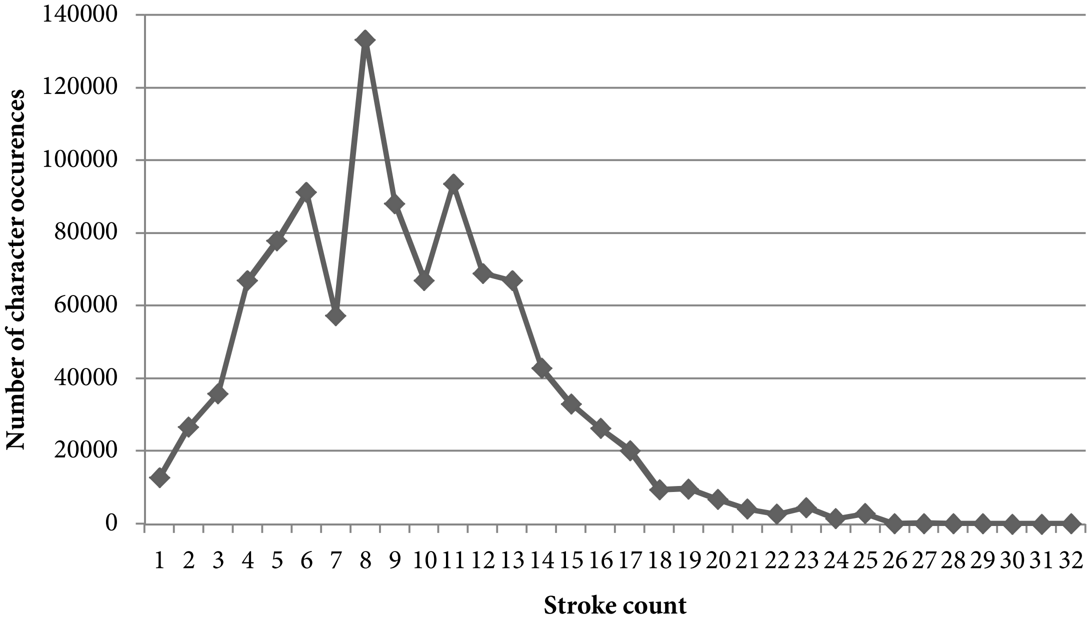


Fig. 6.10 Frequency of characters by stroke count categories

6.1.5. Quantitative properties of syllable-to-character mapping

The relations of characters to phonological units are also analyzed from the quantitative perspective. It is seemingly most natural to pursue the investigation of the correspondence of characters to syllables. It is a well-known fact that the restrictions on Chinese syllable structure result in a syllabary of just over 400 tonally undifferentiated syllables and over 1,300 including tonal distinctions.[271] From this perspective the characters may be investigated as a disambiguation device decreasing the homophone density. A rational signary should be established first to perform analysis of this kind. The size of the signary will immediately determine the average syllabic load of characters. In two articles Li (2011 and 2012) provided a detailed account of the syllable-to-character mapping based on a set of 9,212 characters in a modern dictionary, in which he identified a set of 1,280 tonally differentiated syllables. It is easy to calculate that in this particular set there is an average of 7.2 syllables per character. The number of syllables per character in the whole set is shown in Fig. 6.11. The 15 highest ranked syllables are shown in Tab. 6.8. The syllables are ordered according to the rankings in Li's analysis (second column); the third column contains the number of characters corresponding to a given syllable in the Wenlin 4.1 database (over 70,000 characters). It can be seen that the number of characters increases in each case (as expected), but the syllables change ranks. In both cases *yì* corresponds to the largest number of characters.

[271] Duanmu 2007: 95.

Tab. 6.8 The 15 syllables with highest character load

Syllable	Li (2012)	Wenlin 4.1
yì	83	506
xī	76	338
bì	58	332
yù	57	348
fú	52	238
zhì	50	347
jì	48	284
lì	47	358
yú	45	247
jī	43	229
qí	39	233
shì	39	202
jué	36	256
jí	34	249
huì	34	181

The syllables with the highest character load listed in Tab 6.8 are on one side of the scale, on the opposite side there are 203 syllables with the load equal 1. Statistics of this type are a useful for a discussion on the homophone load and homonymy in Mandarin Chinese, but this problem will not be pursued here any further.

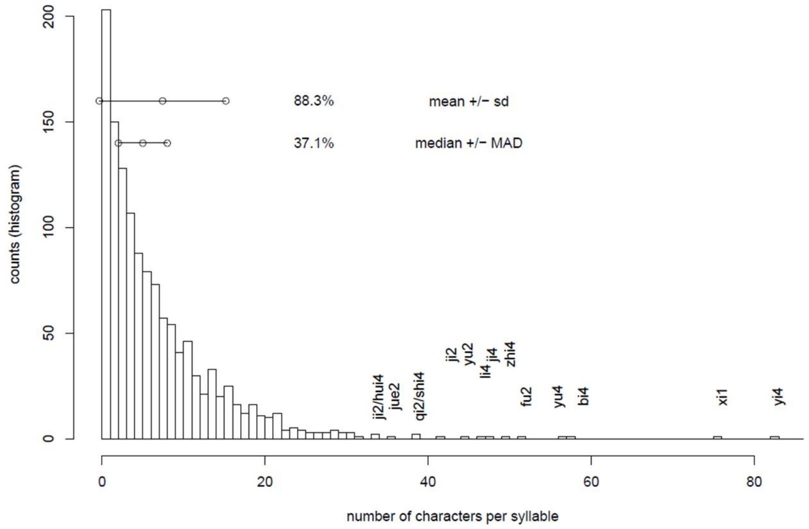


Fig. 6.11 Syllables per character[272]

[272] Li 2012: 4.

6.2. Quantitative linguistic laws

Chinese script was tested against the conformity to a number of laws proposed within statistical and quantitative linguistics. This section is intended as an exemplar presentation of the results of investigations pertaining to the Zipf's law and the Menzerath-Altmann law.

6.2.1. Zipf's Law

Zipf's law was originally formulated with regard to the distribution of word frequencies in a corpus of texts, stating that the frequency of a word is inversely proportional to its frequency rank where *C* is a constant:[273]

$$\text{frequency} = \frac{C}{rank}$$

In other words the dependency between frequency and rank is constant (*C* ≈ frequency x rank).[274]

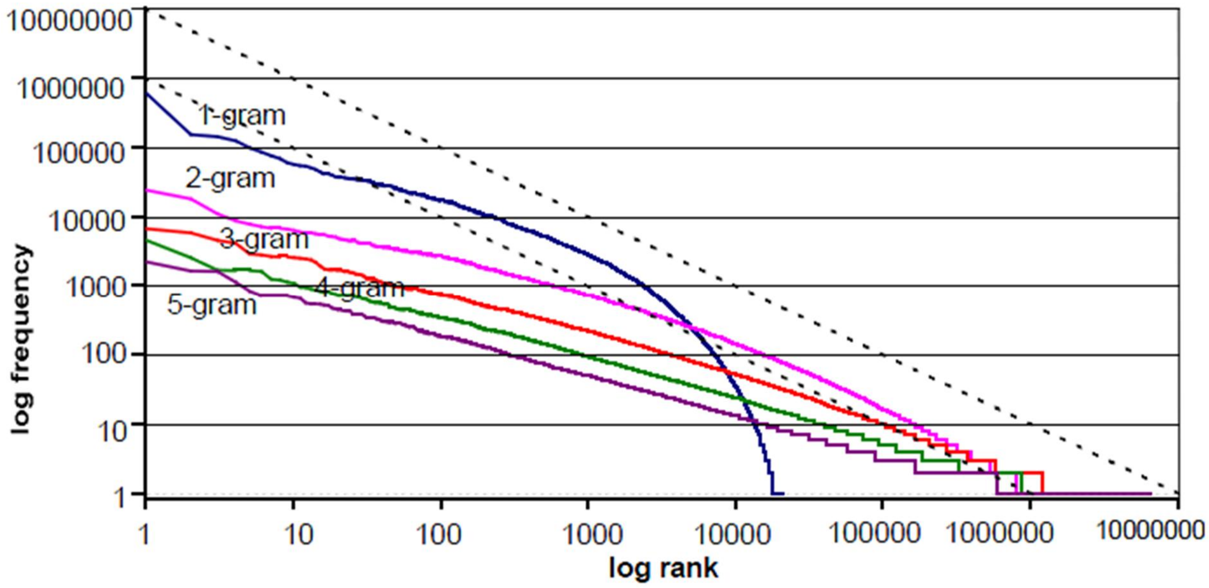


Fig. 6.12 Zipfian curves of N-grams for Chinese TREC corpus[275]

[273] Cantos Gomes 2013: 180.

[274] Ibid., 181.

[275] Ha et al. 2003: 87.

Most efforts testing the conformity of Zipf's law in Chinese texts pertain to the word level. One of the reasons for this is the fact that the character level of Chinese texts does not display conformity to Zipf's law. At a certain point the frequencies drop below frequencies predicted by Zipf's law. This finding was demonstrated by Ha et al. (2003), but also in a few other studies, including Clark et al. (1990) and Xiao (2008). Fig. 6.12 shows the results obtained by Ha for Chinese N-grams. Characters correspond to the 1-gram curve. The slope representing characters is less than 1 and drops rapidly around rank 1,000, indicating that *hànzi* do not satisfy the law in question. Xiao suggests that this property of Chinese characters stems from the fact that they form a closed set.[276]

6.2.2. Menzerath-Altmann Law

The quantitative approach to language offers not only the occurrence and frequency patterns of linguistic units that were introduced in previous sections, but also an insight into the inner patterns and nature of complex linguistic structures. One of those patterns manifests itself in a decrease in complexity of the component parts with an increase of complexity of the whole. In other words – the more complex the construction is, the simpler its parts. This regularity is captured by the Menzerath-Altmann Law. The law is formally expressed by the following equation (in complete form):

$$y = ax^{-b}e^{-cx}$$

The equation describes the relationship between the size (counted in parts) of a whole (x), and the mean size of its parts (y); a, b and c are the parameters. In recent years the Menzerath-Altmann law has been gaining increasing attention in quantitative linguistics.[277] The nature of the relationships described by this law makes it applicable to the graphic representation of language. Prün (1994) tested the validity of the law on the Chinese characters used in the Japanese *jōyō kanji* set of 1,945 frequently used characters. Prün did not conduct the decomposition by herself, but instead based her calculations on the inventory of 485 graphemes isolated by Stalph (1989), whose work is introduced with some detail in the section on the grammars of Chinese characters.

[276] Xiao 2008: 40.

[277] The Menzerath-Altmann law is also applicable in music, social groups and genomes research.

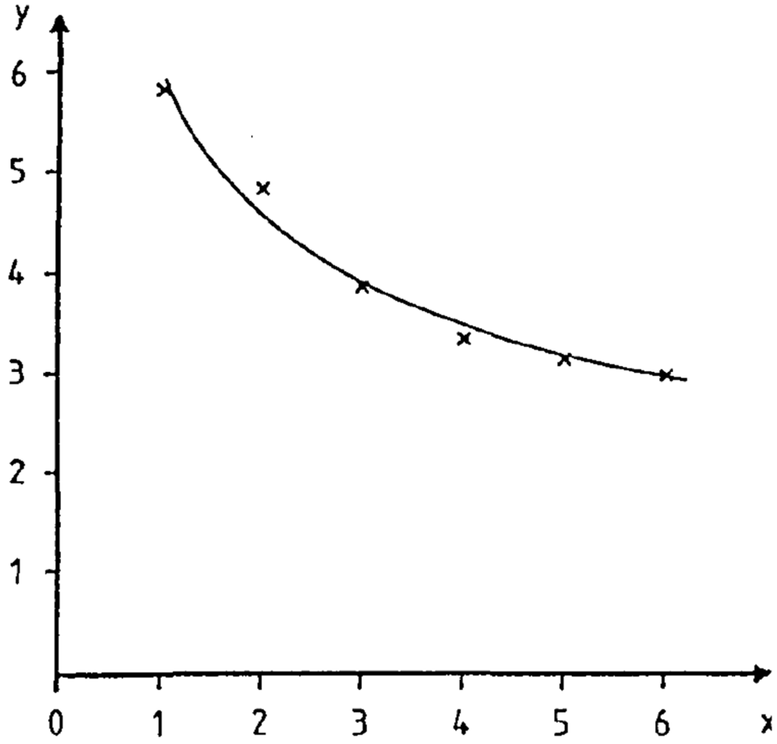


Fig. 6.13[278] Graphemes per *kanji* (x-axis) and average stroke count per grapheme (y-axis)

Fig. 6.13 shows that the curve is very close to the expected values which proves that the characters in the *jōyō kanji* set satisfy the Menzerath-Altmann hypothesis.

Bohn (2002) proved the validity of the hypothesis for GB 2312-80 character set on five different levels of Chinese texts, two of which pertain to the discussion in this section. He came up with two hypotheses:

> "Hypothese 1 (Komponentenebene): Je komplexer eine Komponente, gemessen in der Anzahl der Einzelstriche, desto einfacher die Striche.
> Hypothese 2 (Schriftzeichenebene): Je komplexer ein Schriftzeichen, gemessen in der Zahl seiner Komponenten, desto einfacher die Komponenten, gemessen in der Zahl ihrer Striche."[279]

Bohn proved both hypotheses to be correct. The analysis of the component level was also based on the Stalph's inventory, with necessary modifications. The results of the investigation of the complexity of components in terms of stroke count and the correlation of the stroke count of a component with the complexity of strokes is presented in Fig. 6.14.

[278] Prün 1994: 149.

[279] Bohn 2002: 128.

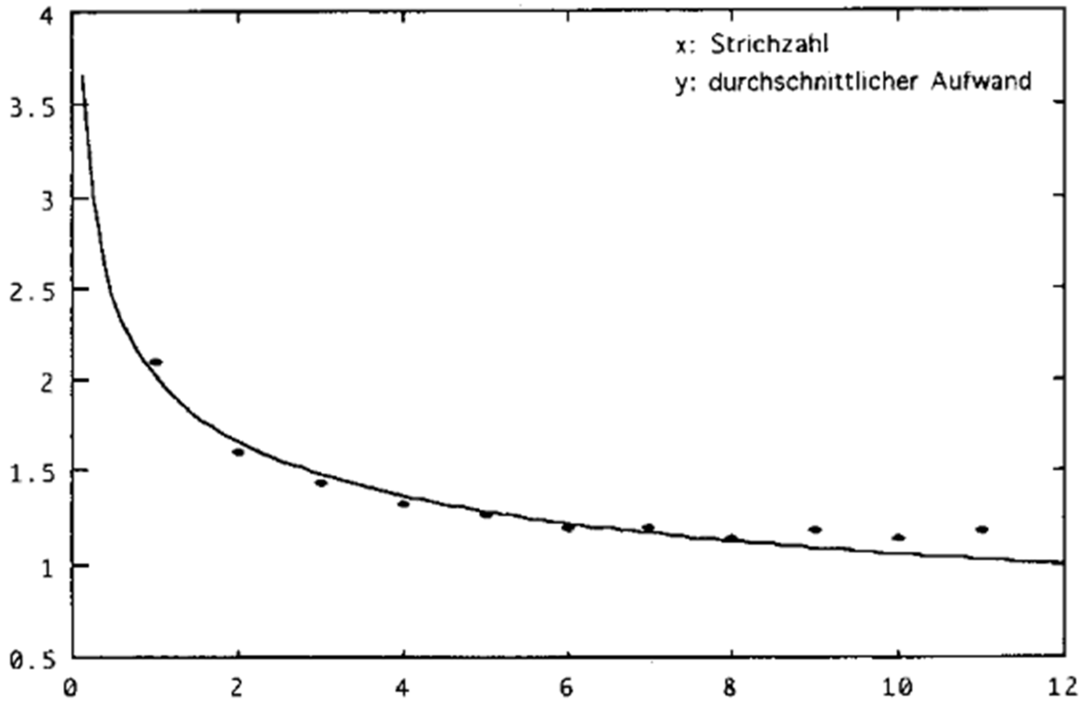


Fig. 6.14[280] Correlation of stroke count with stroke complexity

It is apparent that the results are close to the theoretical predictions, or, in more direct terms, they indicate that the higher the stroke count in a component, the simpler the strokes.[281]

Bohn's findings on the character level conform to the results obtained by Prün, and for that reason, they will not be presented here.

Menzerath-Altmann is presumed to be a quantitative diagnostic tool for testing the linguistic validity of a given level of analysis.[282] The validity of this assumption will not be discussed here, but in case it is legitimate, the results presented in this section confirm that the stroke-component and component-character levels are proper levels for the analysis of Chinese script.

6.3. Script complexity

One of the relatively unknown proposed approaches for script research within quantitative linguistics is an attempt to parameterize and formalize the graphical complexity of script. This method, with a focus on the applicability to the Chinese script,

[280] Ibid., 134.

[281] The stroke complexity is measured in a way very similar to the basic stroke count introduced in Section 4.5.

[282] Prün 1994: 150.

will be outlined in this section against a comparative background of already established methods of quantifying the degree of complexity of Chinese characters.

6.3.1. Compositional method[283]

The first of the quantitative methods of measuring the complexity of script analysis was proposed by Altmann (2004) – his main purpose was to devise a universal system of estimating the complexity of any type script, rather than a way of describing the individual signs. Any system of procedures which universally quantifies script complexity should:[284]

- be applicable to all scripts;
- be simple in use;
- be adaptable to the idiosyncrasies of individual scripts or styles.

In other words the measuring system should be able to capture the intuitively felt difference in complexity between the signs 'A', '𝐀', '𝒜', and '龍'.

Altmann's measuring criteria are twofold – one set of values pertains to the type of graphical elements constituting the signs, and the other to the type of interrelations between the elements. The graphical elements are divided into three categories (Tab.6.9). The elements in each category contribute different values to the computed graphical complexity. The interrelations concern the elements that are in contact within a given graphical sign. There are also three categories of contacts (Tab. 6.10); elements in each of these categories contribute different values to the total complexity of a sign. The system takes into account both the type of elements and their composition –these features are summarized in the tables below:

Tab. 6.9 Types of graphical elements and their values[285]

	Point of any size	Straight line of any size and direction	Arch of any size and direction
Value	1	2	3
Examples	•▪ ▶	–⁄ ∣ \!	⟍ ⟋ ∣) ⌉ ⌡ ⊃ ⊂ ∪ ∩

[283] The terms 'compositional' and 'intersectional' are borrowed from Peust (2006).

[284] Altmann 2004: 68.

[285] Ibid., 69.

Tab. 6.10 Types of contacts between elements and their values[286]

	Continuous contacts	Crisp contacts	Crossing
Value	1	2	3
Examples	O ~	┐┘ F┬┴⟨∠	× + ≠

Tab. 6.11 An example of applying Altmann's method to a few letters of Latin based script (computation of complexity in Altmann's method):[287]

	Types	Connections	Total
A	2 2 2	2 2 2	12
a	3 3 3	2 2	13
O	3 3	1 1	8
Ö	3 3 1 1	1 1	10

6.3.2. Intersectional method

The intersectional method was proposed by Peust (2006) as a different approach to the graphical complexity of script. The reasons for seeking an alternative were the heterogeneous criteria and arbitrariness of the values in Altmann's proposal.[288] As a result Peust proposed a method based on one criterium with one additional rule for compound signs:

Rule 1: The complexity of a sign is the maximal number of crossing points that can be achieved with a straight line. [289]

Rule 2: The complexity of a graphical cluster consisting of several disconnected components must not be computed with a single straight line. Instead, its complexity is defined as the sum of the complexities of its components. [290]

This proposal is less arbitrary in that it is not assigning values to different types of elements and is simpler in terms of the number of criteria and the ease of calculations.

[286] Ibid.

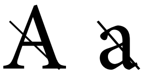
[287] Ibid., 70.

[288] Peust 2006: 11.

[289] Ibid.

[290] Ibid., 15.

The complexities of the letters ‘A’ and ‘a’ calculated in this method equal 3 and 4 respectively:



The direct reason for introducing the second rule is the fact that the simple rearrangement of components that intuitively should not influence the overall complexity, may in fact change the maximal number of intersection points. Peust uses Korean examples with rearranged *jamo* components:[291]

음 임

The intersectional complexity of the first sign is 5, while the second sign’s intersectional complexity equals 4, which is not justified by the actual difference in complexity. The second rule is also a solution of the resident problem of Chinese characters in this respect.

Peust’s demonstration of the advantage of the intersection method for different types of Latin fonts is quite convincing, but he limits the discussion of other types of script, including Chinese characters, to a minimum.[292] A further discussion of the two methods of quantifying the complexity of script here will be limited to Chinese script.

6.3.3. Complexity of Chinese script

Both proposed methods introduced above differ significantly from the established methods of quantifying the complexity of Chinese characters. In common practice the number of strokes is the natural determinant of the complexity of *hànzi*; in other words, the stroke count determines the categories of complexity. The traditional method of counting the strokes that is used for the purpose of classification and ordering of characters relies on the uninterrupted contact of the writing instrument to the writing surface. A more accurate method of estimating the complexity of characters should take into account the complexity of strokes, which is reflected in the distinction between basic and compound strokes.[293] It seems that the basic stroke count is a reliable method for quantifying the complexity of Chinese characters. Regardless of the stroke type this method is immune to changes in style, form and shape of characters that may be the result of the individual features of handwriting, differences between printing typefaces, computer fonts, etc. The two proposed methods in question are both very

[291] Ibid.

[292] Ibid., 14-15.

[293] For more details see Section 4.5.

sensitive to such changes. It does not mean they are inherently defective, but this sensitivity should be taken into account while formulating any conclusions about the results of analysis. For example, the character 電 *diàn* may be represented in the computer display by different glyphs, depending on the font chosen:

電(Simsun) 電(PMingLiU).

The distinctive part in the first character consists of four straight lines (compositional value 8); in the second character, the distinctive part consists of one of two arches, a straight line and a dot[294] (compositional value 9). In this particular example the intersection method renders identical results for both characters. Another example is the traditional character *guī* 'turtle', which in different fonts displays differences affecting both types of analyses:

龜(Simsun) 龜(PMingLiU)

The difference here is not only in the type of components, but also in their number and the number and type of connections. The compositional and intersectional complexity of both characters is calculated in Tab. 6.12.

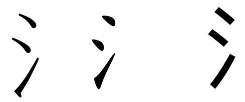
The instances of such relevant differences in the styles of computer fonts are not very numerous, but serve to prove the point. Handwriting styles, intuitively, display even more complexity-changing diversity. For this reason, the two methods in question can be used to quantify the complexity of individual styles of a given writing system, rather than writing systems, in general. On the other hand, those methods can be used freely across different writing systems. The stroke count based method is independent of stylistic differences in form and shape, but it is not universal. A comparative analysis of the results of calculations of graphical complexity using the two procedures with the well-established Chinese stroke count criteria should be a valid basis for assessing the adequacy of the two discussed methods. The issue of adequacy for measuring Chinese characters will be addressed by a brief examination of a small, but representational set of characters. Traditional characters were picked to represent different stroke counts and different structural features. Their simplified counterparts were also analyzed in an attempt to assess the accuracy of the discussed method by comparing changes in complexity with the changes in stroke count caused by simplification.

[294] This is subject to interpretation.

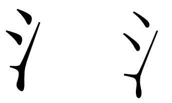
6.3.3.1. Compositional method and the Chinese script

Applying Altmann's method to the analysis of Chinese characters involves even more arbitrary commitments than there are in the method itself. The graphical aspects of the form of characters that are relevant for calculations – the shape, the contour and relative length of strokes, the distance between components resulting in a certain type of connection, or a lack of it, on one hand heavily depend on interpretation, while on the other hand depend on the individual handwriting styles (that includes the type of writing instrument) or font type. The interpretation of canonical shapes of strokes is not unambiguous. The reason for that is the fact that the shapes of strokes are not fixed or strictly defined along the lines of Altmann's concept. The shapes vary depending on individual styles, which in itself is not necessarily a drawback, as it was already assumed that this method is style sensitive. Nonetheless, in many cases the shapes in each style are subject to arbitrary interpretations. Another problem, not addressed by Altmann, is the randomness of connections. The rules of composition allow a certain indeterminacy with regard to the contacts between components and strokes in Chinese characters. This is particularly problematic in handwriting where the shapes of characters representing the same glyph written even by the same person may differ in significant ways. Some problems concerning the method of calculating the compositional complexity of Chinese characters are summarized below:

- the dot type stroke (點 *diǎn*) '丶' graphically reminds one not only of a dot, but even more of an arch – due to traditional analysis and relative simplicity compared to the elements treated as lines or arches, it is proposed to treat '丶' as a dot;[295]
- the rising type stroke (提 *tí*) displays a large array of possible interpretations, which is illustrated by the following examples (the stroke in question is the lowest in the top-bottom structure):

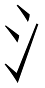


possibile interpretations: a straight line or an arch;



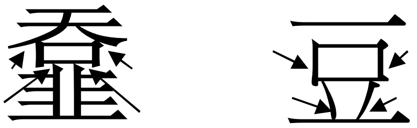
possible interpretations: two or three elements with a choice between different types (dots, straight lines and arches in different combinations);

[295] This is not an absolute rule. In some cases the 點 *diǎn* stroke is elongated enough to be treated as an arch.

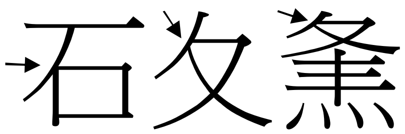


possibile interpretations: two elements – either two straight lines, or a straight line and a dot;

- in the compound SG stroke (vertical+hook 豎鉤 *shùgōu*) 亅 may be interpreted as an arch, or as two strokes (straight+straight, straight+dot, or straight+arch) with a connection point;
- handwriting or computer fonts imitate calligrafical shapes of strokes, such as in the straight type H stroke (橫 *héng*) 一 which may become arched 一一一;
- there are two possible types of problems with connections to be solved:
  - The randomness[296] of connections caused by the lack of explicit compositional rules for Chinese characters is often sufficient to raise uncertainty as to whether or not the contact or lack of contact is accidental, which intuitively can lead to the conclusion that the connections are not contributing to the complexity of a character. For inherently inconsistent handwriting this problem will remain unresolved. The different computer fonts should simply be treated as separate character sets, and therefore, their complexities should be calculated separately. The arrows in the example below indicate the examples of the random connections of elements:

  - The multiconnectivity of elements: it is possible that a few elements of a Chinese character can make multiple contact in one position of its structure. This issue was not addressed by Altmann. The following examples illustrate some typical multiple connections in one position of a character structure:

There are two possible methods of determining the number of connections in such cases. The simpler solution, rendering the lowest complexity, would be to treat multiconnections as one. The alternative is to calculate the number of connections using the simple formula x-1, in which x is the number of connecting elements. Closer examina-

[296] True randomness happens in the handwriting; in the computer fonts and printing typefaces the connections are a design feature.

tion of the examples reveals two types of multiple connections. In the case of 石 there is a connection of a left curving stroke 撇 *piě* (the element of 丆) with the two strokes connecting at the upper left corner of 口. There are three elements connecting at the indicated point. There is also a connection of three elements in 夆 but with a significant difference – here the the left curving stroke 撇 *piě* has a fixed connection with the horizontal stroke 橫 *héng*, while the right falling stroke 捺 *nà* is 'free', with respect to the point of connection with the horizontal stroke, is noticeable in the case of 夂. The enlarged font shows that in the standalone component the connection between 撇 *piě* and 橫 *héng* is in a different place than between 撇 *piě* and 捺 *nà*. In other words the indicated point is a connection of three independent elements, though the connection of all three is only an option. The situation is different in 石 – here the vertical (豎 *shù*) and the horizontal strokes are preconnected as an element of 口 which means that the connection with the left curving 撇 *piě* is inherently a multiple one. The tentative proposal here is:

(i) to treat the 'freely' (in the sense explained above) connecting elements as a multiconnection by calculating the number of connections using the formula x-1;

(ii) to treat the preconnected elements as one element while connecting with other elements.

Whereas the structure of Chinese characters is too complex in the context of the compositional method for a complete and thorough analysis here, some key problems are indicated. For simplicity's sake, in ambivalent cases the solutions proposed here tend to go along the lines of the traditional perspective on the structure of Chinese characters, rather than purely graphical criteria. Regardless of the solutions chosen, consistency in their application is of key importance.

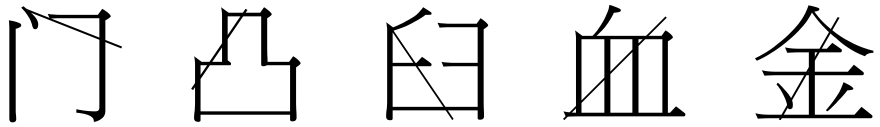
A tentative analysis of the sample set of characters is presented in Table 6.12. Given the nature of the compositional method it may only be claimed to pertain to the particular font type described.

6.3.3.2. Intersectional method and the Chinese script

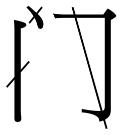
The alternative to Altmann's system is a relatively simple intersectional method that was briefly introduced earlier in this chapter. The idea is simple, but the answer to the question of why the number of intersecting points with a straight line is an indication of complexity is not understood intuitively. The more basic question is whether this method works with Chinese characters, since Peust (2006) only made a brief reference to them, pointing out their relatively high complexity while providing a few examples.[297] The two rules introduced earlier seem to work for Chinese script, but only to a certain extent. The method evidently works as intended with:

- structurally simple characters (e.g. 一，人，二，三，八，手);
- complex characters with a distinctively divided component structure (e.g. 众, 亿, 忤, 示, 䦎).

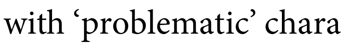
The category of ‘simple’ characters in the context of the intersectional method is quite different than the characters traditionally associated with this term. An examination of some ‘simple’ characters shows that a small number of strokes or single component composition does not guarantee that the intersectional method can be applied directly. It is difficult to pinpoint which relevant structural features make a character difficult to measure with the method in question. Obviously every simple character, that is, a character not having an apparent compenential structure, can be intersected with a single line. In some cases, however, a substantial part of the structure is left out, meaning the complexity of some non-compound characters cannot be properly represented by a number of points intersecting one straight line.



The characters above possess different types of structural features that result in the same problem – inappriopriateness of the single line treatment used in the intersectional method. The character 门, once the traditional concept of a compound character is abandoned, may simply be treated as an unproblematic complex character with three distinctive components:



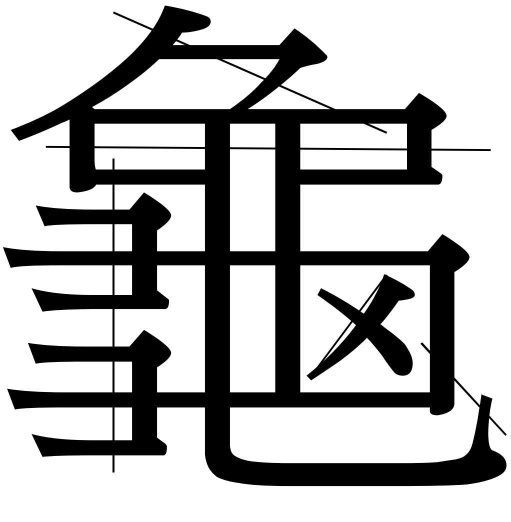
The remaining exemplary characters are not easily measured this way, but the componential idea, as a general rule, seems to be indispensable. Peust’s second rule in its original form cannot be applied to a significant number of complex characters with unclear component structures. The above examples should be treated uniformly with those problematic complex characters. Examples of problematic complex characters are given below:



It is not necessary to intersect the above characters with lines to illustrate the issues with ‘problematic’ characters.

[297] Peust 2006: 15.

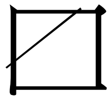
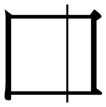
A uniform method of measuring the complexity of 'simple' and 'problematic' characters be devised by either modifying the second rule or formulating a new one. The existing formulation is too specific – it is designed for the structures with clearly delimited component parts. It is possible to reformulate it into a more general formula that also works with obscure composition or overlapping or intertwining components. The tentative approach proposed here assumes that the intersecting lines should not leave out any significant elements and that an element can only be intersected once. A more precise definition of a 'significant element' is desirable, but here only examples of such elements will be given. For instance, it is not clear how to calculate the complexity of the character below (one of the traditional variants for 'turtle'). On one hand it obviously is a very complex character, while on the other hand it is a non-decomposable radical[298] that should be measured with one line. Single component treatment cannot properly represent the complexity of 龜 with the intersectional method. Alternatively, the reformulated rule for handling the compound characters is applied in the following way (the enlargement helps to bring out the nuances of the structure):



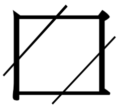
The complexity of the character measured with the proposed method is 20. The extension of the rules is intended to account for all the fragments and elements of the structure that contribute to the graphical complexity. The cost of the modification is a gain of more arbitrariness and a slight loss of some of the original simplicity. The introduction of the expanded rules also causes some simple characters measured by a single line to be classified as complex characters, which appears to be necessary in order to keep the homogeneity of the criteria. In other words, the criteria for both sin-

[298] This is the case, at least from the perspective of Chinese characterology.

gle line or multiple line treatment of characters should be method-specific – traditional characterology classification must be abandoned or at least play a secondary role. For example, the character 口 is simple enough to be measured by a single line:

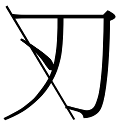
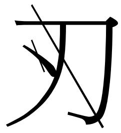
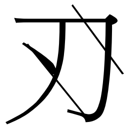
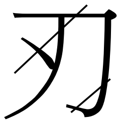
 or 

Despite the simplicity, it is evident that a substantial part of the structure is not measured by a single line. For that reason, in the system adopted here the multiple line measurement should be applied instead:



This treatment is supported by analysis of the characters 二 and 匚. In the one line treatment all 3 characters have the same complexity which evidently is not a desired result.

The same principle applies to all similar cases, regardless of their classification in Chinese characterology. This treatment of seemingly simple characters not displaying the traits of componential structure is a radical departure from Peust's original proposal, resulting in significantly different values of the complexities of characters.[299] Nonetheless, many problems still remain. There is no independent criteria that helps in choosing a solution. For example, the complexity of the character 刃 may be calculated on the basis of different paths of the straight lines:

(i)  (ii)  (iii)  (iv) 

There are more ways to draw the lines, but the four above are sufficient for the iilustration. Options (i) and (ii) render a complexity value 4, while (iii) and (iv) render the value 5. The difference is the treatment of the 亅 (豎鉤 *shùgōu*) stroke as either one element or as two elements. It is a practice that is purely arbitrary in its application within the context of the method in question. Even in Chinese characterology there are two perspectives (number of strokes vs. number of basic strokes). The choice between the options rendering the same results is of secondary importance; it is not clear whether any rules addressing similar problems are possible or necessary.

[299] There is at least one piece of evidence for the different intentions of the original proposal – the values of complexities of the Korean 음 and 임 calculated by Peust (2006: 14), which correspond to the single line measurement of the bottommost elements equivalent to the Chinese 口.

Despite the complicated picture emerging from the above discussion, the application of the modified method is quite intuitive. There seems little utility in pondering all manner of problems to be potentially solved by means of the intersectional method in the context of Chinese script, a task too immense to be adequately addressed here. The actual analysis of the sample set of characters is performed as a test of both discussed methods when they are applied. The results of the calculations are provided in Table 6.12.

6.3.3.3. Methods compared

Since Chinese characterology already offers methods for measuring the complexity of characters, it would be interesting to conduct a comparison between the established methods and the two discussed in this chapter. In order to accomplish that, the complexities of a substantial number of characters should be measured. The calculations for any standard set of characters[300] is beyond the scope of analysis here – both methods are too laborious in application for a set consisting of many thousands of characters. In this respect the compositional method is the more challenging of the two. Apart from the fact that it is the more complicated method, there seems to be no easier way than to calculate the complexity of every individual character separately. The intersectional method is more practical and, more importantly, it is possible to automatically conduct calculations, at least to some extent. Given that most Chinese characters have componential structure, complexity can be calculated by summing the complexity values of their components, though the precision of this procedure is yet to be tested. A necessary first step is assigning the values for each component in the system, but that is a relatively simple task even for a single researcher. This idea will not be pursued here, remaining a promising research perspective. Tab. 6.12 serves the following purposes:

- illustrates the application of the two discussed methods for the analysis of the Chinese script;
- provides data for the preliminary quantitive analysis of the complexity of Chinese characters using each of the methods;
- provides data for the preliminary comparative analysis of the methods;
- provides data for the preliminary comparative analysis of the discussed methods against the background of established stroke count analytical methods.

The comparative analysis involving stroke count methods has two goals:

- determine the quantitative relations between the results obtained with different methods;

[300] Any NCS or CCS – see Chapter 2.

- determine the correspondence of the increase of compositional and intersectional complexity with the increase of stroke count.

Due to the fact that the stroke count captures the intuitions about the complexity of characters quite well, the obtained results may also be a basis for an overall evaluation of the two methods as tools for analyzing Chinese script.

Tab. 6.12 consists of 17 stroke count categories with a total of 86 characters. Each category contains five characters with the exception of one stroke characters. Addtionally, there are four characters that were included for different reasons: two are Peust's examples that do not fit the 17 stroke count categories, and two illustrate complexity reduction. The sample characters representing each stroke count category were primarily chosen to represent diversified structure types, though random choice was also employed in assembling the sample characters. Three of the examples are Altmann's (悪, 握, 扱), and four are Peust's (瞧, 餐, 罐, 露).[301] The order of characters is based on traditional stroke count. Within each stroke count category the ordering is determined respectively by the basic stroke count,[302] intersectional complexity value, and compositional complexity value.

Tab. 6.12 Compositional complexity of Chinese script – sample set

Character	Component types	Connection types	Total compositional complexity value	Intersectional complexity value	Stroke count	Basic stroke count
一	1(2)	-	2	1	1	1
乙	1(2)+2(3)	1(2)	10	3	1	4
二	2(2)	-	4	2	2	2
十	2(2)	1(3)	7	2	2	2
人	2(3)	1(2)	8	2	2	2
丁	1(2)+1(3)	1(2)	7	3	2	3
九	2(2)+2(3)	2(2)+1(3)	17	4	2	5
工	3(2)	2(2)	10	3	3	3
上	3(2)	2(2)	10	3	3	3
才	1(2)+2(2)	1(2)+1(3)	11	3	3	4
刃	1(2)+3(3)	2(2)	15	4	3	5
及	3(2)+3(3)	4(2)+1(3)	26	6	3	6
升	2(2)+2(3)	1(2)+2(3)	18	4	4	4

[301] Altmann 2004: 70-71; Peust 2006: 15. In the case of the intersectional method values for Peust's examples, the results of the calculation with the modified rules are presented as the main values, while the original values are in parantheses.

[302] The method of counting the number of basic strokes is intuitive enough to restrain detailed explanation. It is in concordance with the stroke system presented in Su 2001, for example.

Character	Component types	Connection types	Total compositional complexity value	Intersectional complexity value	Stroke count	Basic stroke count
手	2(2)+2(3)	1(2)+2(3)	18	4	4	5
日	5(2)	6(2)	22	5	4	5
中	5(2)	4(2)+2(3)	24	5	4	5
月	3(2)+2(3)	6(2)	24	5	4	6
玉	1(1)+4(2)	2(2)+3(1)	16	5	5	5
正	5(2)	4(2)	18	5	5	5
石	5(2)+3(1)	5(2)	23	6	5	6
用	4(2)+2(3)	2(2)+2(3)	24	5	5	7
汀	2(1)+2(2)+2(3)	2(2)	16	6	5	7
耳	6(2)	7(2)+1(3)	29	6	6	6
式	1(1)+4(2)+1(3)	2(2)+1(3)	19	6	6	7
成	1(1)+2(2)+4(3)	3(2)+2(3)	29	7	6	9
臣	10(2)	8(2)	36	8	6	9
扱	5(2) +4(3)	5(2)+3(3)	41	10 (9)	6	10
夾	1(2)+2(1)+4(3)	3(2)+3(1)	25	7	7	7
形	3(2)+4(3)	2(2)+2(3)	28	7	7	7
豆	1(1)+8(2)	4(2)	25	8	7	8
孚	2(1)+3(2)+3(3)	2(2)+3(1)	24	9	7	9
囮	5(2)+3(3)	5(2)+3(1)	32	9	7	10
非	8(2)	6(2)	28	8	8	8
長	7(2)+3(2)	6(2)	32	9	8	9
伷	8(2)+3(1)	9(2)+4(3)	49	9	8	9
玥	7(2)+3(2)	8(2)+3(1)	39	9	8	10
𡇈	10(2)	11(2)+3(2)	48	10	8	10
挂	7(2)+2(3)	2(2)+4(3)	36	10	9	10
契	6(2)+4(3)	3(2)+4(3)	42	11	9	11
思	3(1)+6(2)+3(1)	8(2)+3(1)	37	11	9	12
耑	10(2)+3(1)	9(2)	41	12	9	12
飛	2(1)+5(2)+6(3)	7(2)+2(3)	50	13	9	13
𡶡	7(2)+4(3)	6(2)+3(1)	41	11	10	11
桌	9(2)+3(2)	10(2)+3(1)	47	12	10	11
這	2(1)+9(2)+1(3)	6(2)	35	13	10	13
圄	13(2)	12(2)+3(1)	53	13	10	13
赿	7(2)+4(3)	6(2)+3(3)	47	12	10	14
𤈦	4(1)+5(2)+3(3)	2(2)+4(3)	39	12	11	12
趆	8(2)+4(3)	8(2)+2(3)	50	13	11	13
域	1(1)+9(2)+2(3)	5(2)+3(3)	44	14	11	13
悪	3(1)+9(2)+1(3)	9(2)+4(3)	54 (55)	14	11	14
挏	1(1)+9(2)+3(3)	6(2)+2(3)	46	15	11	15
跌	10(2)+3(3)	10(2)+2(3)	55	13	12	13

Character	Component types	Connection types	Total compositional complexity value	Intersectional complexity value	Stroke count	Basic stroke count
葉	11(2)+2(3)	5(2)+6(3)	56	13	12	13
臯	1(1)+13(2)	13(2)+5(3)	68	14	12	14
握	1(1)+11(2)+2(3)	3(3) +8(2)	54	15	12	15
勤	11(2)+3(3)	6(2)+9(3)	70	16	12	16
琴	1(1)+11(2)+2(3)	7(2)+2(3)	49	14	13	14
鼓	1(1)+12(2)+2(3)	8(2)+3(3)	56	13	13	15
搖	13(2)+2(3)	7(2)+5(3)	61	17	13	16
[illegible]	6(2)+9(3)	10(2)+4(3)	71	17	13	17
庸	2(1)+11(2)+3(3)	14(2)+3(3)	70	18	13	18
熔	6(1)+5(2)+5(3)	10(2)	51	16	14	16
糖	2(1)+12(2)+2(3)	11(2)+5(3)	69	16	14	16
髣	2(1)+9(2)+5(3)	10(2)	55	16	14	17
壽	2(1)+13(2)+1(3)	8(2)+2(3)	53	17	14	17
齊	2(1)+10(2)+4(3)	10(2)	54	17	14	17
誰	2(1)+13(2)+1(3)	10(2)+2(3)	57	16	15	16
[illegible]	2(1)+9(2)+5(3)	13(2)+2(3)	67	17	15	17
[illegible]	5(1)+11(2)+1(3)	13(2)	56	17	15	18
[illegible]	15(2)+3(3)	15(2)+3(3)	78	18	15	18
[illegible]	1(2)+14(2)+3(3)	15(2)+7(3)	90	19	15	18
[illegible]	15(2)+2(3)	18(2)+1(3)	75	17	16	17
[illegible]	1(1)+14(2)+2(3)	14(2)+7(3)	84	19	16	17
[illegible]	4(1)+7(2)+6(3)	7(2)+2(3)	56	19	16	19
襃	4(1)+11(2)+4(3)	17(2)	72	19	16	19
餐	2(1)+11(2)+7(3)	16(2)+1(3)	76	19 (17)	16	20
瞧	5(1)+12(2)+1(3)	15(2)+2(3)	68	18 (16)	17	18
[illegible]	2(1)+10(2)+6(3)	11(2)+5(3)	77	18	17	18
[illegible]	1(1)+17(2)+2(3)	13(2)+2(3)	73	21	17	21
賢	16(2)+6(3)	21(2)+1(3)	95	23	17	23
[illegible]	1(1)+11(2)+9(3)	19(2)+3(3)	97	24	17	25
龜	1+17(2)+3(3)	18(2)+3(3)	89	20	17	23
龜	17(2)+4(3)	19(2)+5(3)	99	20	16	23
露	1(1)+21(2)+3(3)	17(2)+2(3)	92	25 (20)	21	25
罐	1(1)+23(2)+2(3)	18(2)+5(3)	104	26 (19)	23	26

The results of the calculations are plotted onto one diagram for a better illustration of quantitative relations (Fig. 6.15).

The complexity values for the intersectional method show a strong correlation with the basic stroke count method. The correlation of compositional values with the other two sets of values is more irregular. All three methods capture the general tendency of increasing complexity with an increase in stroke count.

Figure 6.16 shows the correlation of the two stroke count methods with the intersectional method in a more refined way.

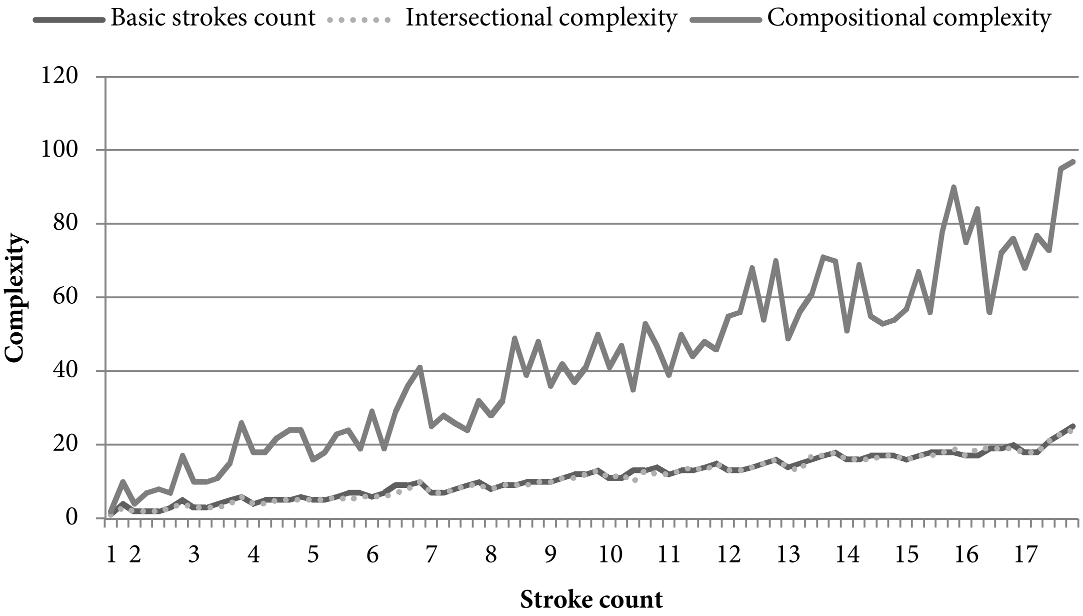


Fig. 6.15 Correlation of stroke count and complexity values

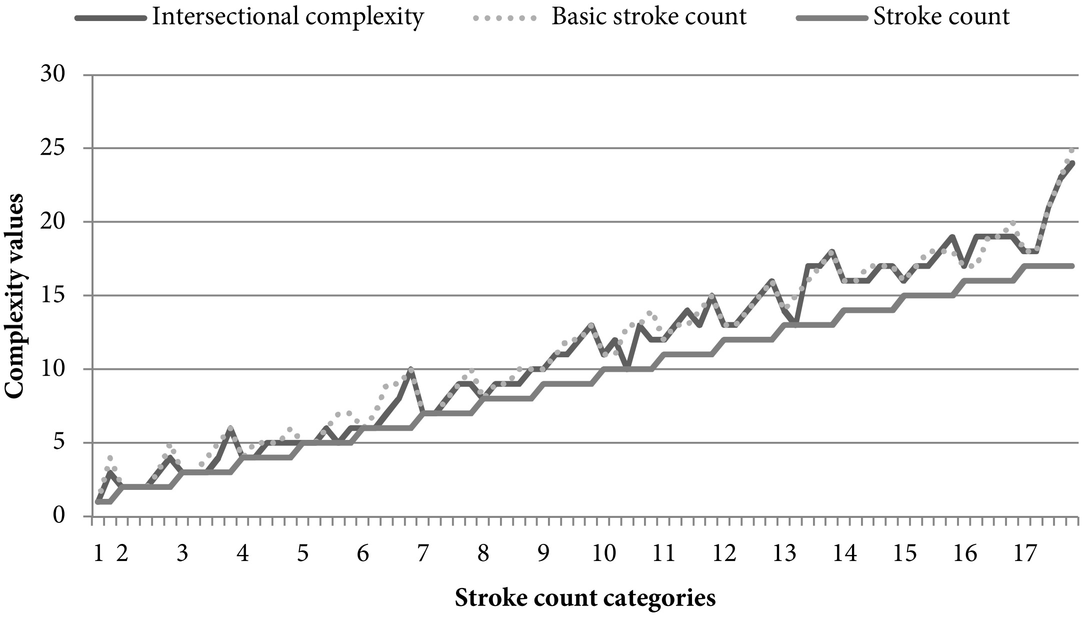


Fig. 6.16 Complexity values for stroke count, basic stroke count and intersectional methods

The differences are more noticeable, but the correlation of the intersectional method with basic stroke count is still evident.

Table 6.13 contains the average complexity values in 17 stroke count categories for all the discussed methods.

A pattern repeated in the values for all types of complexity measurement methods is a general increase in complexity correlated with an increase of the stroke count. A closer examination shows that the increase in intersectional complexity and basic stroke count is more regular, and that this correlation, at least in principle, is mirrored in the stroke count – the average increase in intersectional complexity per stroke count category is 1.13, and for basic stroke count it is 1.15. In both intersectional complexity and basic stroke count there is no decrease in the complexity value. Compositional complexity values display more fluctuations and bigger dispersions, and in three instances the values decrease. The average complexity values for the measures in question and their correlation are shown in Fig. 6. 17.

Tab. 6.13 Average complexity of characters

Stroke count	Avg. compositional complexity	Avg. intersectional complexity	Avg. basic stroke count
2	8.6	2.6	2.8
3	14.4	3.8	4.2
4	21.2	4.6	5.0
5	18.6	5.4	6.0
6	30.8	7.4	8.2
7	27.0	8.0	8.2
8	39.2	9.0	9.2
9	41.2	11.4	11.6
10	44.6	11.6	12.4
11	45.4	13.4	13.4
12	60.6	14.2	14.2
13	61.4	15.8	16.0
14	56.4	16.4	16.6
15	69.6	17.4	17.4
16	72.6	18.6	18.4
17	83.1	20.8	21.3

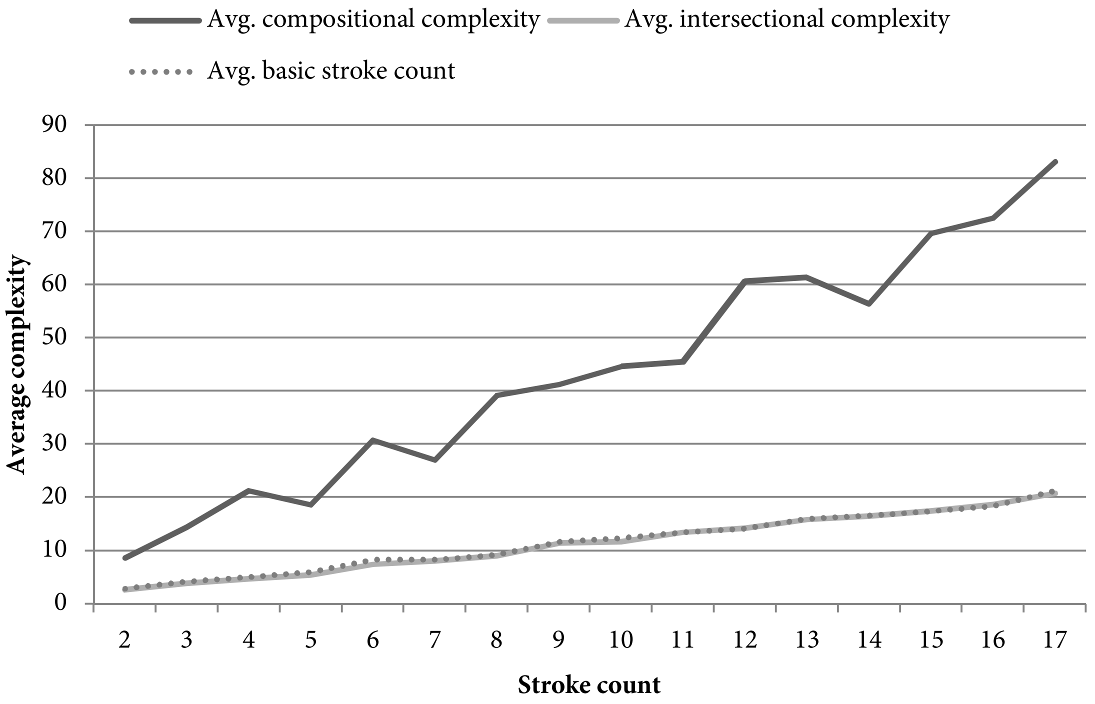


Fig. 6. 17 Correlation of average values of complexity

Unsurprisingly, the curves in Fig. 6.17 are similar to those in Fig. 6.15. The intersectional complexity and basic stroke count display close correlation while the compositional complexity is more irregular with respect to the two other measurement methods.

The data set was too small to draw any definitive conclusions, but it gives an insight into the nature and relations between different complexity measurement methods.

6.3.3.4. Complexity reduction

The effects of simplification of characters[303] may serve as another test for the viability of the two methods. It is assumed that the reduction in stroke count should be reflected in a similar reduction in complexity. For example, the number of strokes in 龟 is reduced by 58.82% compared to 龜, and by 56.25% compared to 龜. The simplifications in compositional complexity are 46.06% and 51.52% respectively. A small set of 18 traditional characters and their simplified counterparts was chosen to calculate the complexities and changes caused by simplification in terms of stroke count and complexity values.

[303] See Section 4.6.

Tab. 6.14 Complexity reduction

Traditional	Stroke count	Simplified	Stroke count	Stroke count reduction (%)	Compositional complexity reduction (%)	Intersectional complexity reduction (%)
壯	7	壮	6	14.2	40.0	25.0
協	8	协	6	25.0	52.0	27.2
後	9	后	6	33.3	39.1	41.6
這	10	这	7	30.0	25.7	23.0
習	11	习	3	72.7	79.1	64.2
嵐	12	岚	7	41.6	42.6	37.5
滅	13	灭	5	61.5	63.0	62.5
爾	14	尔	5	64.2	74.5	66.6
熱	15	热	10	33.3	26.0	22.2
龜	16	龟	7	56.2	55.5	50.0
總	17	总	9	47.0	47.6	47.8
繭	18	茧	9	50.0	42.3	52.1
蘋	19	苹	8	57.8	63.3	60.0
寶	20	宝	8	60.0	69.1	60.8
鐵	21	铁	10	52.3	39.2	56.5
體	22	体	7	68.1	76.2	75.8
韆	23	千	3	86.9	90.0	90.6
讓	24	让	5	79.1	80.9	76.6

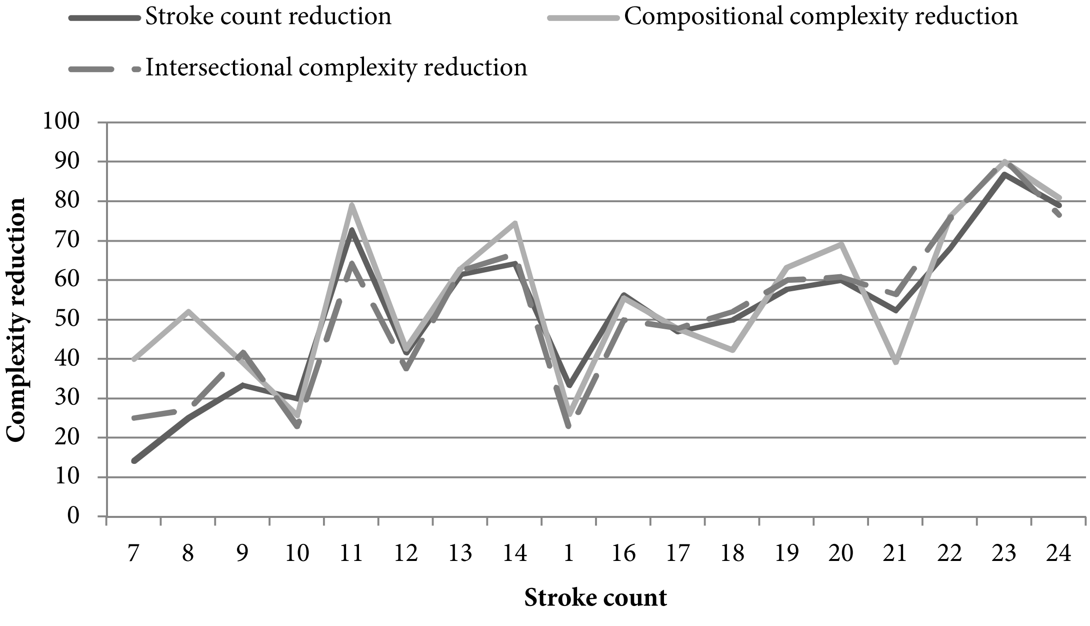


Fig. 6.18 Complexity reducton

Tab. 6.12, Tab. 6.14, Fig. 6.15, Fig. 6.17 and Fig. 6.18 show that the strongest correlation between the different types of complexity measures is observed between the basic stroke count and intersectional complexity. It would be interesting to further verify the degree of correspondence against the basic stroke count reduction. A relatively high degree of correspondence would be another confirmation of the validity of the intersectional method for the analysis of Chinese script. Tab. 6.15 contains the calculations of the basic stroke count reduction for a small sample of traditional characters along with their simplified counterparts. The results are then compared with the values obtained for intersectional complexity reduction (Tab. 6.14).

The curves in Fig. 6.19 display an expected similarity, though more extensive analysis of a larger sample set of characters is yet to be conducted. For the purpose of this study a tentative confirmation of the dependency in discussed systems measuring the graphical complexity of script with the recognized stroke count method study should be sufficient.

Tab. 6.15 Basic stroke count reduction

Traditional	Basic stroke count	Simplified	Basic stroke count	Basic stroke count reduction (%)
壯	8	壮	6	25.0
協	14	协	8	42.8
後	12	后	7	41.6
這	13	这	9	30.7
習	16	习	5	68.7
嵐	16	岚	10	37.5
滅	14	灭	5	64.2
爾	16	尔	7	56.2
熱	17	热	13	23.5
龜	23	龟	11	52.1
總	23	总	12	47.8
繭	23	茧	10	56.5
蘋	20	苹	8	60.0
寶	23	宝	9	60.8
鐵	23	铁	11	52.1
體	29	体	7	75.8
韆	33	千	3	90.9
讓	28	让	7	75.0

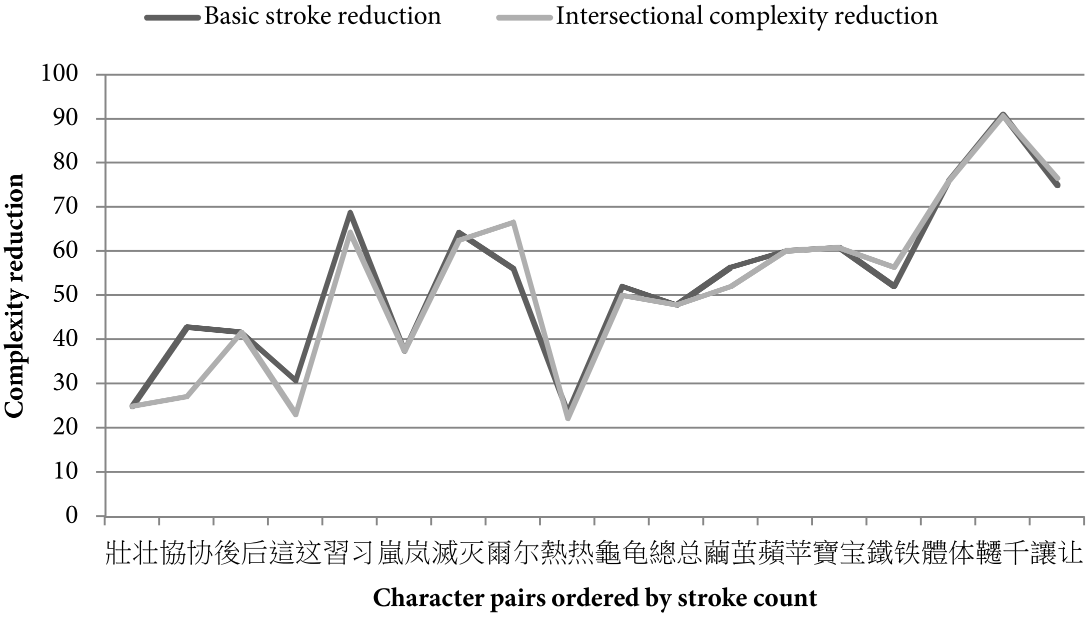


Fig. 6.19 Correlation of basic stroke reduction with intersectional complexity reduction

6.3.4. Summary

The system proposed in this section is far from being complete and controversy-free, but the tentative findings indicate that both methods are viable tools for measuring the complexity of Chinese script. The typological investigations that at this moment do not exist, should provide valuable evidence for further evaluation of the methods.

The compositional method uses heterogeneous criteria that almost completely dispenses with the traditional compositional structure of characters. The results rendered by this method are relatively divergent from the traditional stroke count method, but this fact is expected, given the different criteria of measurement. The results also display some independence from the number of strokes. The calculations in this method are more arduous and must be conducted for each character separately, which means that a calculation of complexity of a considerable set of characters would be a formidable enterprise, and will not likely be carried out by one person.

As the analysis has shown, the intersectional method is closely related to the basic stroke count. The preliminary results suggest that this method could be considered a universal equivalent of the Chinese stroke counting method.

Both discussed methods were proven to be correlated with the stroke count in reflecting the general increase or decrease in the number of strokes. The same dependency is also observed in the comparative analysis of the reduction of stroke count in individual characters – the resulting decrease in the stroke count is mirrored in decreasing complexity values in both methods. The measurement of script complexity yields most interesting results from the typological perspective. A discussion of the aforementioned methods must be restricted to the above analysis due to a lack of relevant research results, a limitation of space, and in no small measure, because of the main purpose of this book.

7. Graphotactic analysis of Chinese script

This chapter presents the detailed results of graphotactic analysis of Chinese script. The theoretical and practical prerequisites for conducting the analysis were outlined in chapters 2-5. Section 7.1. examines the graphotactic properties of Cangjie input method encoding, which is intended as an approximate equivalent of componental properties of *hànzi* to the orthotactic properties of words in alphabetical scripts. A proper graphotactic analysis of Chinese characters is provided in section 7.2.

7.1. Graphotactics of Cangjie input method

The term 'input methods' here, are actually devices for sorting and categorizing the characters that originally did not support computer technology. The most popular input methods are based on pronunciation. These pronunciation-based methods are primarily transliteration systems that are practical aids in information exchange, language teaching and learning, lexicographic ordering, the sorting of characters, etc. At this juncture, there is no need to even briefly introduce input methods other than Cangjie.

7.1.1. Introduction to the Cangjie input method (CIM)

The Cangjie input method (倉頡輸入法 *Cāng Jié shūrùfǎ*) was invented by Chu Bong-Foo in the 1970s and has been developed since. For reasons that will be explained in this section, the Cangjie method provides the means of conducting an alternative quantitative analysis within the graphotactic framework.

The inventor of the method was determined to enable Chinese script to be used in information processing, specifically to a level comparable to the alphabetic systems in terms of speed and practicality. The prerequisite to achieve that goal was to enable Chinese script input by means of a standard keyboard. The general idea was to base the input strategy on the structural features of characters and enable a convenient input for the defining features. For practicality's sake the descriptive components were classified into 24 categories denoted by Cangjie 'letters' (倉頡字碼 *Cāng Jié zìmǎ* or 倉頡碼 *Cāng Jié mǎ*). These components were chosen on the basis of frequency and structural features. The 24 Cangjie letters are mapped to the English alphabet (X and Z are the functional symbols), which allows their input using a standard keyboard. The project was successful – CIM was the first method allowing the input of more than 100 characters per minute.[304]

It would be logical to assume that each Cangjie component covers more than one component. In reality, Cangjie components do not represent more than one component. Quite often they represent the structural fragments of characters. The relationship of Cangjie letters to the component systems is complicated. The Cangjie system is a close relative of typical component/radical systems, but it is unique enough to require an independent introduction. An entirely new componential system was initially devised for the needs of CIM. This componential system evolved into a character encoding method, and inspired Chu Bong-Foo to introduce a fundamentally new perspective on Chinese characters that he called 漢字基因 *hànzijīyīn* 'the genes of Chinese characters'. The concept reflects some peculiar views of the author including substantial elements of ideology and mysticism. Additionally, 漢字基因 *hànzijīyīn* in fact advances the pictographic/ideographic stance on the nature of Chinese characters. Despite the controversies the 'genetic' theory of characters is worth looking into.[305] Due to space limitations the introduction here will be restricted to issues closely related to CIM.

One way of seeing the relationship between 漢字基因 *hànzijīyīn* and CIM is that the latter is a practical extension of the former (ignoring the fact that CIM was being developed first). Of the six components of the 'genes', four directly play practical roles in CIM:

- 字碼 *zìmǎ* – components,
- 字序 *zìxù* – order,
- 字形 *zìxíng* – form,
- 字辨 *zìbiàn* – identification.

All of these components must be correctly applied in order to render an appropriate encoding of a character. The 'genetic' theory is using 字義 *zìyì* ('meaning') to explain characters, while CIM is encoding characters with 字碼 *zìmǎ*. 漢字基因 *hànzijīyīn* distinguishes two basic types of theory-specific components:

- 字首 *zìshǒu* – initials, and
- 字身 *zìshēn* – bodies.[306]

字首 *zìshǒu* are similar to the notion of 'radicals' – their role is classificatory; 字身 *zìshēn* is what is left of a character after subtracting the initial part. The identification of the initials and the bodies of characters is a key issue in the 'genetic' theory. Chu provides a detailed explanation of his methodology that is based on a sophisticated

[304] The two most comprehensive sources of information on CIM are Chu (1990) and Chu & Shen (2006). Most of the general introduction in this section is based on these two sources.

[305] http://cbflabs.com/book/dnahtml/dnabase/dnabase01.htm.

[306] https://zh.wikipedia.org/zh/%E5%AD%97%E9%A6%96.

semantic classification of components.[307] Chu's semantic classification is reflected in CIM with such labels as 'philosophical', 'strokes', 'human body' and 'character structure' – the categories into which the 24 Cangjie symbols are divided.

In principle 字義 *zìyì* and 字碼 *zìmǎ* are similar – CIM rules of decomposition (or generation) are grounded in the 'genetic' rules, but for reasons of practicality CIM sometimes uses a different order of generating the characters. To express it differently, CIM strictly follows the rule stating that the initial is always on the left, on the top, or is the most external element of structure, depending on the form of a character. In these cases the 'genetics' follow the etymological interpretations. This results in discrepancies.[308] For example, the 字義 *zìyì* ('genetic') analysis of the character 頭 *tóu* 'head' designates the '頁' as an initial (in accordance with etymology and the radical systems), and '豆' as the body of the character. The 字碼 *zìmǎ* (CIM) rules state otherwise – the order is reversed. In some cases the two ways of decomposition render completely different component structures, with the character '條' *tiáo* being an example:

- 字義 – initial: '木', body: '攸' (in accordance with etymology and radical systems),
- 字碼 – initial: '亻', body: '丨攵木'.[309]

Another instance of discrepancy is the treatment of the simple/complex structure of characters, as in the example of the character '兆' *zhào*:

- 字義 – simple character,
- 字碼 – initial: '中', body: '一山人'.[310]

The system recognizes 594 initials and 9,897 bodies, which according to Chu Bong-Foo, are the most primitive and basic forms of characters, and therefore, as a set of components with formation rules have the capability to constitute any other character. This set of components is arranged in a number of letters corresponding to a prescribed number of letters in the English alphabet. The maximum length of CIM encoding strings is five symbols. The reason for this is not the number of components in the most complex Chinese characters, but purely arithmetical calculation that the system of one to five place strings out of 24 symbols can encode more than 10,000 components.

The CIM representation of a character may consist of English alphabet letters or the Cangjie symbols (Cangjie letters). The computer input is based on the standard QWERTY keyboard, and the Cangjie descriptive elements are assigned to corresponding letters on the keyboard. The symbols/letters (primary shapes) also cover a set of

[307] http://open-lit.com/bookindex.php?gbid=311.

[308] http://zh.wikipedia.org/zh-hant/%E5%AD%97%E9%A6%96.

[309] https://zh.wikipedia.org/zh/%E5%AD%97%E9%A6%96.

[310] Ibid.

76[311] auxiliary shapes (輔助字形 *zhuǎnzhù zìxíng*) that complete the system, thus forming the set of Cangjie roots or simply the Cangjie components (轉助字根 *zhuǎnzhù zìgēn*). The CIM component system is introduced in Tab 7.1.

The primary purpose of the method was different than a pure description of the form and structure of characters.[312] Therefore, Cangjie encoding is not a character description language. Its popularity is not a result of its adequacy as a CDL, but in large measure because of its practicality and speed as a computer input method. On the other hand, the Cangjie method is a description language, despite Cangje encoding not being considered such. It was already mentioned that the inventor consciously integrated CIM into the larger 漢字基因 *hànzijīyīn* project. CIM's popularity, the fact it can be learned, and that it enables encoding of characters with a high degree of adequacy and relatively low ambiguity, all constitute a strong indication that CIM provides valuable information on the structure of characters in a consistent and analyzable manner. Cangjie componential decomposition – the sequences of components as representations of characters – is either unorthodox or controversial in at least four ways:

- the inventory of components is extremely reduced;
- some elements in more complex characters are omitted;
- the number of component parts is limited to five;
- and the coding has some ambiguity.

The inventory of components must be reduced, simply because there are only 24 Cangjie 'letters' to represent a few hundred elements and thousands of characters. The number of components is reduced by distribution of all possible constituents between the 24 categories denoted by the 'Chinese letters' with a few general guidelines in mind:

- encoding redundancy and ambiguity must be minimalized;
- elements classified into one CIM category must display similarity in form and shape to a corresponding 'letter';
- concordance is established with the habitual use of a given component.[313]

The first rule is to guarantee system economy; the last two rules address the learnability and the speed of input.

[311] The number of auxiliary shapes and, consequently, the number of roots differ according to different sources and different generations of CIM.

[312] The best source of detail regarding the invention of the method and the ideas behind it are in Chu 1990 and Chu 2002.

[313] http://cbflabs.com/book/gif_cg/gif_cg/index.html.

Tab 7.1[314] The Cangjie system of symbols and letters

Category	Cangjie Symbol	Corresponding Letter	Pithy Name	Auxiliary Shapes
Philosophical	日	A	日	rotations of 日, 曰
	月	B	月	first four strokes of 目, 冂, ⺥, 冖, 夕 , 夕 , first four strokes of 骨
	金	C	金	丷, 八, 儿, 儿
	木	D	木	寸, first two strokes of 也 and 皮
	水	E	水	氵,氺, 又
	火	F	火	小, 灬
	土	G	土	
Strokes	竹	H	斜	⺮, 丿 and the short slant ノ
	戈	I	點	丶, 广, 厶
	十	J	交	crossing strokes shape, 宀
	大	K	叉	乂, ナ, 亠, X shaped elements
	中	L	縱	丨，衤， 肀
	一	M	橫	㇀,厂, 工
	弓	N	鉤	亅
Human Body	人	O	人	㇏,𠂆, 𠂉, last two strokes of 兆
	心	P	心	忄, ⺗ , second stroke in 心, 匕, 七, 勹, last two strokes in 代
	手	Q	手	
	口	R	口	
Character Structre	尸	S	側	first two strokes of 己, ㇆, ㇁, ⺕
	廿	T	並	卄, 廾, 艹 (also in broken form)
	山	U	仰	an enclosing structure with an open top
	女	V	紐	right hook, V shaped elements, 以
	田	W	方	口, enclosed shapes also with elements inside
	卜	Y	卜	辶, 辶, ⺀
Disambiguation Symbol		X		
Special Symbol		Z		

[314] The contents of the table are based on various sources, among which the following were used the most by the author: http://cbflabs.com; http://www.chinesecj.com/newlearncj/cj5/cj3.php; http://en.wikipedia.org /wiki/ Cangjie_input_method; and http://www.hkpe.net/cj/cjtable.htm.

The examples in Tab 7.2 illustrate the strategies of Cangjie encoding, its advantages and deficiencies. One potential source of problems for CIM encoding are the similarities in form and shape between characters.

Tab 7.2 Cangjie descriptions of selected characters

Character	Cangjie Description	Character	Cangjie Description	Character	Cangjie Description
士	十一	土	土		
工	一中一	干	一十		
匕	山竹	七	十山		
找	手戈	我	竹手戈		
未	十木	末	木十		
天	一大	夭	竹大		
石	一口	古	十口	右,夻	大口
己	尸山	已	尸山	巳	口山
唱	日日日	晶	日日日		
暃	日中一卜	晱	日中一卜		
易	日心竹竹	昒	日心竹竹		
昺	日一人月	昞	日一人月		
晚	日弓日山	冕	日弓日山		
閃	日弓人	𨳇	日弓人		
曄	日廿一十	曅	日廿一十		
分	金尸竹	釖	金尸竹		
东	大木	杀	大木	奈	大木
痳	大木木	森	大木木		
釦	金口	㕣	金口		
昍	日山山	昢	日山山		
照	日口火	煦	日口火		
釣	金心戈	鈎	金心戈		
鼈	中人中尸一	鼇	中人中尸一		
酥	一田竹木	粟	一田竹木		
酟	一田卜口	䣧	一田卜口		
角	弓月土	墮	弓月土	堕	弓月土
鷊	一月竹日火	鸝	一月竹日火	鶮, 鶮	一月竹日火
勿	心竹竹	匆	心大大		
搞	手卜口月	篙	竹卜口月		
因	田大	困	田木		
間	日弓日	閒	日弓月		
購	月金廿廿月	構	木廿廿月		

On the most abstract level the problems with ambiguous coding in CIM may result from:

- perceptional similarities,
- structural similarities,
- similar or identical components.

The interrelation of the three sources is complicated and will be discussed below in concert with the given CIM encoding examples.

The first seven entries were chosen to illustrate the coding efficiency and flexibility of CIM – they contain structurally similar characters with unambiguous CIM sequences of symbols assigned. The similarity in question actually manifests itself in several types of singular differences that are often difficult to perceive to an untrained eye:

- length of a single stroke: 匕七 and 士土,
- relative position of strokes: 工干 and 未末,
- type of strokes: 天夭,
- number of strokes: 找我,
- combinations of the above with possible different types (for example, different angles of strokes, or rotations of elements): 石古右.[315]

CIM unambiguity in the above cases is not a result of the application of some rules directly addressing the lengths of strokes, angles, rotations, relative positions, etc., but rather it is a consequence of the CIM system's primary components and auxiliary shapes, which were able in those cases to capture and encode the differences, despite the similarities.

The next 20 entries are also examples of similar characters, but in these cases CIM fails to provide unambiguous codes. It should be stressed that the *ad hoc* classification here is a result of an analysis from the perspective of the traditional treatment of shapes, forms, structures, components and strokes, and the same can be said of the judgments regarding the affinity of elements (related/unrelated). The classification of the ambiguous cases is more complicated:

- length of a single stroke: 己已 (巳 is assigned a distinctive code);
- relative position of a component: 唱晶, 𣈅暃, 易昒, 昺昞, 晚冕, 閃𨳇, 曄曅;
- CIM specific structural equivalence of unrelated components: 分釖, 东杀奈, 痳森, 釦台, 呂咄;
- perceptual similarity of unrelated components: 照煦, 釣鈎, 鳖鳖, 酥粟, 酟鉆, 角墮堕, 鷊鸍鴯鳾.

[315] The only purpose of the classification of the types of similarities between characters is an exemplification of the CIM encoding strategies and their effectiveness, presentation of cases when it works, and when CIM displays deficiencies. Some of the categories require more elaboration to serve more general purposes. For example, the relationship of distinctive single strokes to the whole character.

The remaining entries are the examples of the most typical cases of partially related characters with one distinctive component that CIM handles unambiguously. One of the regular sources for ambiguity of CIM codes is the inability to represent the spatial layout of components. The codes are linear in that they follow the temporal order of writing, but ignore the spatial arrangement.

The CIM structural affinities of elements are also the reason for treating the traditionally distinctive elements as belonging to the same symbol category. The same can be said of perceptually similar elements, since they are different, despite their visual resemblance. Some examples illustrate the difficulty of classification. For example, the five characters: 丘, 仝, 仜, 亼, 𠓛 share the same Cangjie code – OM (人一) and the types of similarities between these characters are diverse. The examples from the above table are chosen to highlight both actual and possible encoding problems. Leaving the subject without further comment, however, might create a distorted view of CIM. The evaluation of CIM as an input method is not a concern in this book, but perhaps it should be pointed out as an input method the system is distinctively efficient compared to other methods. From the perspective of an input method encoding ambiguity is a minor, if not negligible, problem in CIM. In an overwhelming majority of cases at least one character in an ambiguously encoded pair (or in more numerous sets) is extremely rarely used and encountering that character in regular input is unlikely. Another source of ambiguity that has not been mentioned so far is the traditional/simplified distinction. Again, from the input method point of view it is not a real problem, since Chinese input is usually predetermined by one of the character sets. The simplified character 堕 and the traditional character 墮 share the same code. This is a generalization rather than a rule. The two sets of characters are in principle treated as unrelated and are coded by the same sequence only when their CIM composition is identical; in other cases the coding is distinctive, as in the case of the following: 門: 日弓 – 门 (and 冂): 中尸，爱: 月月大水 – 愛: 月月心水，脏: 月戈土 – 髒: 月月廿一廿.

Unorthodox from the traditional perspective, but coherent in a system like CIM, and as farfetched as it may seem, the CIM dcodes allow us to glimpse into the Chinese script from the perspective of system layout similar to alphabetic ones. This is the main purpose of the graphotactic analysis of CIM structural descriptions.

7.1.2. Graphotactic analysis of Cangjie codes

A quantitative analysis of CIM encoding has been performed on the corpus of 27,607 codes representing 30,301 characters.[316] The difference is the result of ambiguity in coding. The calculations revealed 14,152 CIM tactographemes, which makes possible a determation of the average graphotactemicity. The units corresponding to graphemes in the analysis of CIM codes are CIM letters. The basic quantitative data pertaining to the investigated corpus of Cangjie codes of Chinese characters is summed up below:

- number of CIM graphotactemes: 27,607;
- number of CIM tactographemes: 14,152;
- number of encoded characters: 30,301;
- average tactographemic efficiency: 1.95;
- average graphotactemic efficiency: 0.53;
- average length of CIM code sequence: 4.23;
- average tactographemic dispersion: 2,367.96;
- average graphotactemic dispersion: 4,285.6.

The total number of CIM graphotactemes (CIM codes of Chinese characters) is significantly smaller than in the case of IDS based corpuses investigated in the next sections, but is representative enough, even though the set contains both simplified and traditional characters. Due to the encoding ambiguity the total number of encoded characters is higher than the number of CIM graphotactemes; the data indicates an encoding ambiguity of 8.89%. Any commentary on this number requires a proper perspective. From the point of view of input method efficiency, the result is better than any other structural method,[317] not to mention the notoriously ambiguous phonetic methods. From the perspective of an encoding system similar in function to character description languages (CDLs), any ambiguity is undesirable, especially since there is no practical aspect of CDLs that accounts for the inaccuracies and ambiguities. To put it simply, the ambiguity ratio of coding is a straightforward indication of the number of

[316] The corpus is freely available at: http://en.wiktionary.org/wiki/Wiktionary:Chinese_Cangjie_index. The number of unique codes is very close to the CJK Unified Ideographs including Extension A. That database contains 27,484 characters. No additional information on the relationship of the set to the official standards is provided, but it is reasonable to assume that the Cangjie codes database is closely related to the Unicode/Unihan database.

[317] This refers to the methods designed to encode the same type of character set. The 五筆 *wŭbĭ* method is probably less ambiguous in this respect, but is limited when applied to the traditional characters. A comparison of CIM with other methods can be found at: http://zh.wikipedia.org/zh-hant/%E5%80%89%E9%A0%A1%E8%BC%B8%E5%85%A5%E6%B3%95#.E8.88.87.E5.85.B6.E4.BB.96.E5.BD.A2.E7.A2.BC.E8.BC.B8.E5.85.A5.E6.B3.95.E7.9A.84.E6.AF.94.E8.BC.83.

character descriptions that are not captured by the system. It was shown in Section 5.1 that the CDLs are not free of a certain degree of ambiguity, but the ratio is much smaller than in CIM.[318]

The average tactographemic efficiency is much higher than calculated for Polish letters (1.36) and Chinese *pīnyīn* transliteration (1.11).[319] It is also expected to be higher compared to the efficiency of the actual Chinese tactographemes, regardless of the analyzed subset of *hànzi*.[320]

The average length of CIM code sequences is a measure of the average complexity of CIM graphotactemesin terms of CIM graphemes per character.

The average dispersion numbers pertain to the distribution of CIM graphemes in tactographemes and graphotactemes. A thorough account of dispersional properties of CIM graphemes will be provided in a further part of this section.

The more detailed results of the graphotactic analysis of Chinese characters in terms of CIM graphemes are introduced in the remaining part of this section. Part of the results pertaining to the family of categories of graphemicity is summarized in Tab. 7.3.

Table 7.3 Quantitative properties of CIM tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	24	78	3.25
2	291	1,205	4.14
3	1,936	6,729	3.47
4	6,720	13,510	2.01
5	5,181	6,085	1.17
Total:	**14,152**	**27,607**	

The quantitative properties of CIM graphemes are best presented in a graphical way – the data in Tab. 7.3 will be introduced in separate diagrams with necessary comments.

7.1.2.1. CIM tactographemicity and t-graphotactemicity

In Fig. 7.1 the tactographons are plotted on the x-axis according to their graphemicity, while tactographemicity is plotted on the y-axis.

[318] There is no concrete numerical data to support this claim, but the intuition seems to be strongly justified.

[319] Bańczerowski 2009: 15-19.

[320] See Section 7.2.

Fig. 7.1. shows that most CIM tactographemes consist of 4 and 5 CIM graphemes. Since there are 24 CIM graphemes, there may only be 24 1-grapheme CIM tactographemes ('X' by itself cannot generate any characters). There are only five categories of graphemicity, so the shape of the curve can only be approximated to the expected Gaussian one.

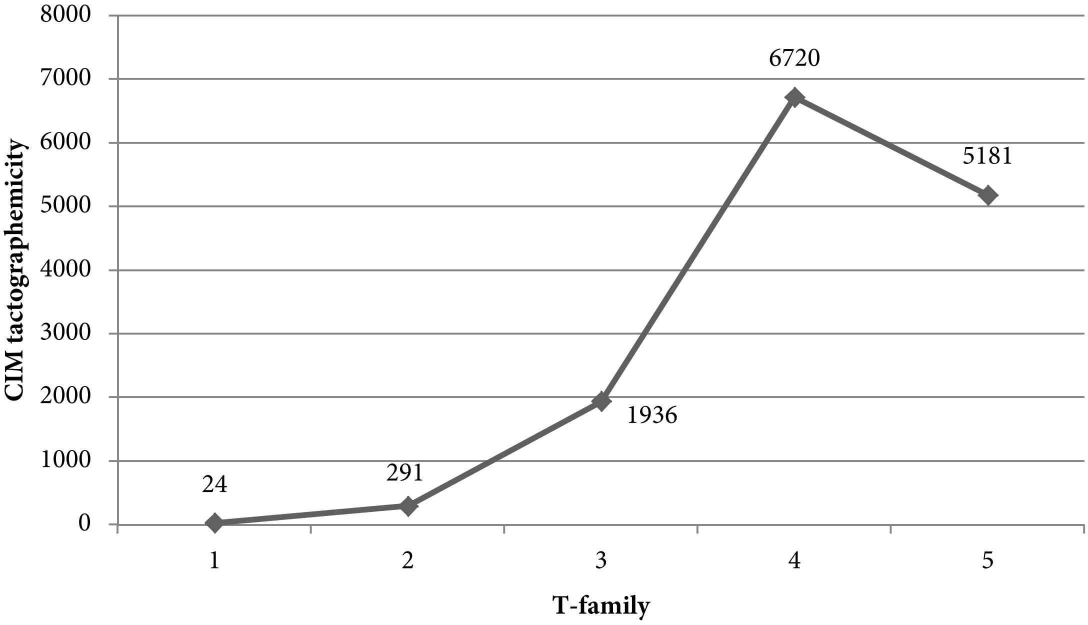


Fig. 7.1 CIM tactographemicityby tactographons(t-families)

In Fig. 7.2 the tactographons are plotted on the x-axis according to their graphemicity, while t-graphotactemicity is plotted on the y-axis.

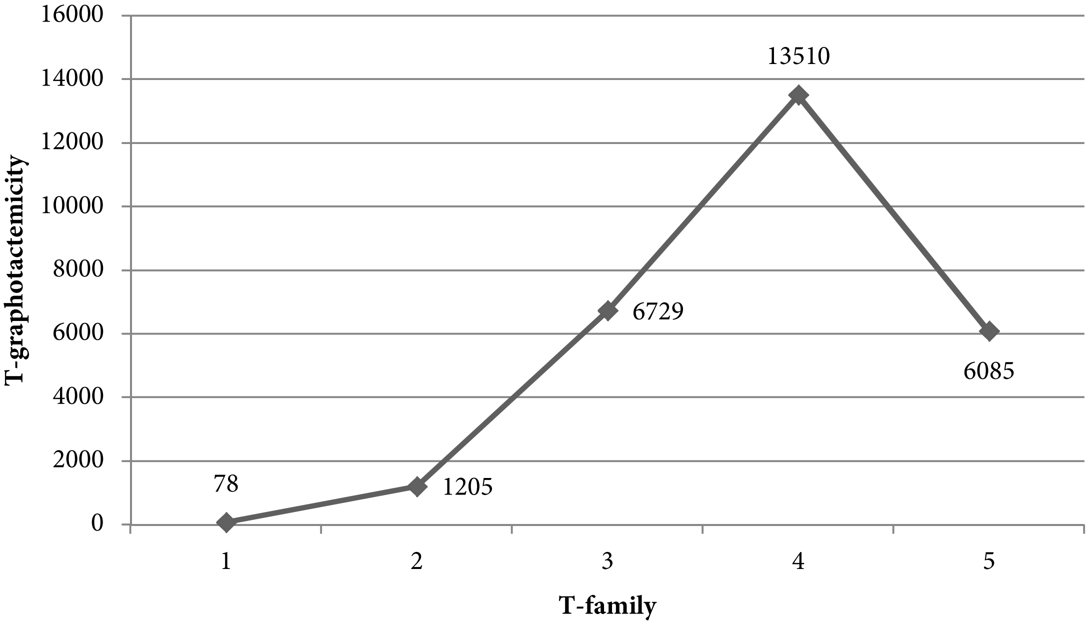


Fig. 7.2 CIM t-graphotactemicity by tactographons (t-families)

Most numerous are the characters generated by the CIM tactographons with graphemicity values 4, 3 and 5, respectively. In other words the tactographemes consisting of 4, 3 and 5 CIM graphemes generate most graphotactemes (characters). Also in this case the shape of the curve may be approximated to a Gaussian one. The data on t-graphotactemicity is also pertinent to the quantification of the complexity of characters in terms of CIM structure. The correlation is not direct, because tactographons and tactographemes do not account for graphemes that occur multiple times in a particular grapotacteme.

CIM tactographemicity and t-graphotactemicity can be compared in one diagram to better visualize the quantitative relations between the number of CIM tactographemes in each t-family and CIM graphotactemes generated out of each t-family. This is shown in Fig. 7.3. The curves of CIM tactographemicity and t-graphotactemicity are very similar.

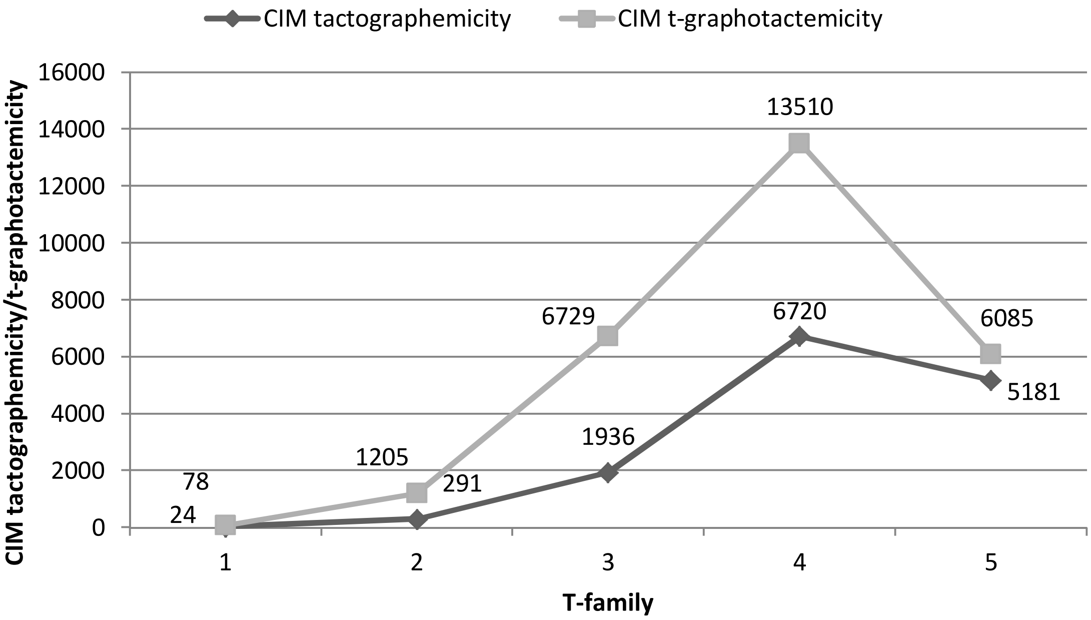


Fig. 7.3 CIM tactoraphemicity and t-graphotactemicity compared

7.1.2.2. T-efficiency

T-efficiency pertains to the graphotactemic efficiencies of individual tactographons (members of t-family)

Fig. 7.4 shows that the tactographemes consisting of a smaller number of CIM graphemes (graphemicity 1 to 3) on average produce most CIM graphotactemes.

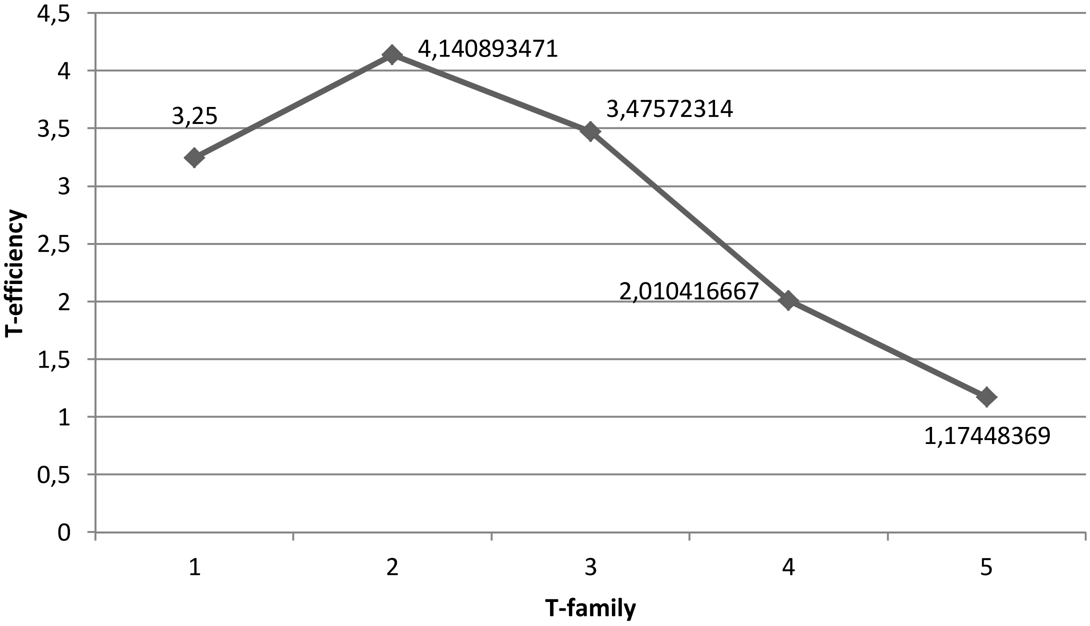


Fig. 7.4 CIM t-efficiency

7.1.2.3. Tactographemic t-efficiency

The next quantitative property of the CIM tactographemes is the tactographemic t-efficiency that can be expressed by the following formula:

tactographemic t-efficiency = $\frac{g}{tx}$

Where g is CIM t-graphotactemicity (3rd column in Tab. 7.3), t is CIM tactographemicity (2nd column in Tab. 7.3) and x is t-graphemicity (graphemicity of a tactographon (1st column in Tab. 7.3). Tactographemic t-efficiency provides information on the average graphotactemic efficiency of tactographemes belonging to a given tactographon. Tactographemic t-efficiency differs from t-efficiency in that it reflects the correlation of the number of generated graphotactemes with the number of generating graphemes – t-efficiency only indicates the number of generated graphoactemes. This property is shown in Fig. 7.5.

Tactographemic t-efficiency drops rapidly with the increasing graphemicity along a curve similar to a logarithmic one.

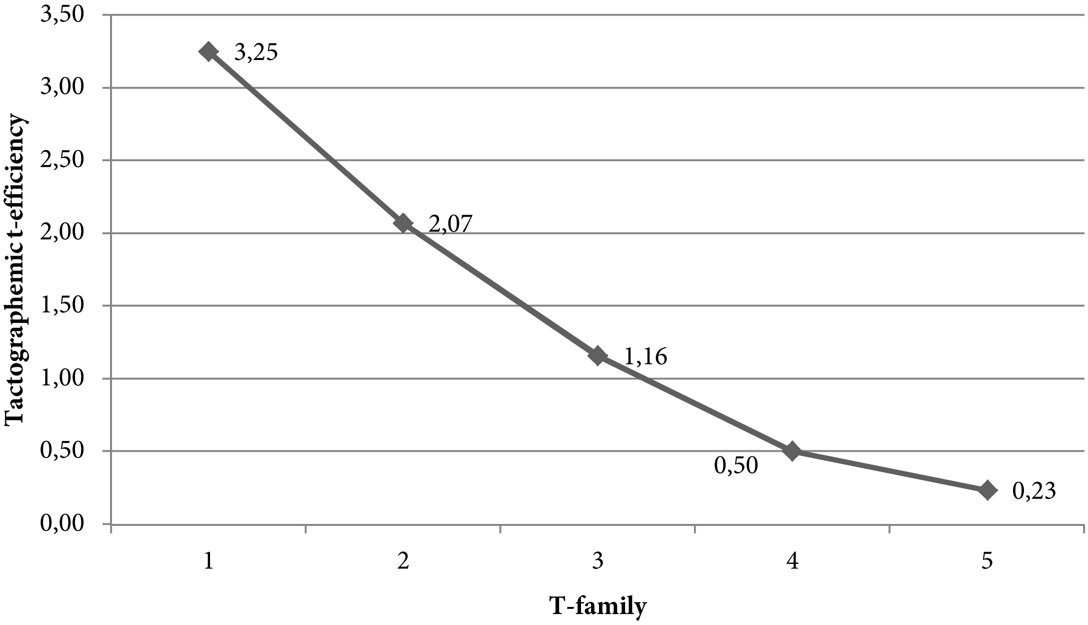


Fig. 7.5 CIM graphotactemic t-efficiency by the categories of graphemicity

7.1.2.4. CIM categorial graphotactemic efficiency

Another type of graphotactic data that can be extracted from the corpus is the cardinality of individual CIM tactographons.

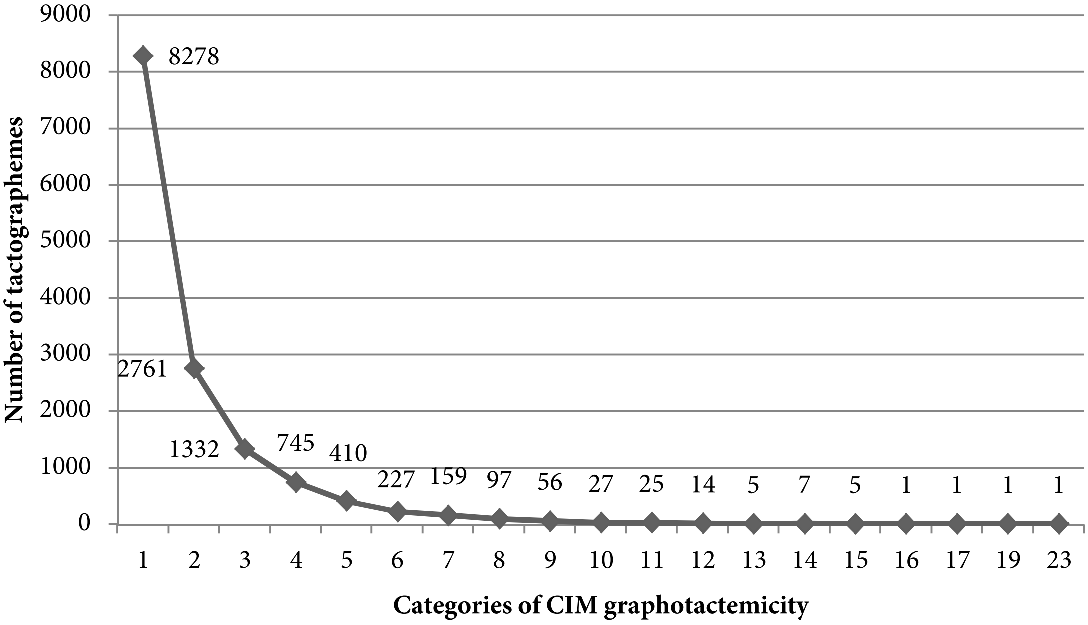


Fig. 7.6 The number of CIM tactographemes generating a given number of graphotactemes

Fig. 7.6 illustrates the numerosity of categories of graphotactemicity revealed by an investigation of the graphotactemic loads of all CIM tactographemes. A category of graphotactemicity is understood as a set of all tactographemes characterized by the same graphotactemic load (graphotactemicity), i.e. generating the same number of graphotactemes. The graphotactemic loads of CIM tactographemes range from 1 to 23 establishing 19 categories of graphemicity. The curve assumes a logarithmic shape indicating a large number of CIM tactographemes with low graphotactemicity, a medium number with medium graphotactemicity, and a small number of CIM tactographemes with high graphotactemicity.

7.1.2.5. CIM graphemes dispersion

This section introduces the results of the investigation of CIM tactographemic and graphotactemic dispersions of the CIM graphemes. The analysis concentrates on the dispersion numbers and statistics for both types of dispersion.

7.1.2.5.1. CIM tactotactemic dispersion

The distribution of the CIM graphemes between the CIM tactographemes can be presented in two forms – using CIM letters or CIM graphemes, both of which are sampled below:

in CIM letters:

A: {A, AB, ABC, ABCD, ABCDP, ABCE, ABCF, ABCFH, ABCFI, ABCFM, ABCG, ABCH, ABCHK, ABCHM, ABCHP, ABCHX, ABCIL, ABCIM, ABCJK, ABCJM, ABCJN, ABCJS, ABCKN, ABCKQ, ABCL, ABCM, ABCMO, ABCMV, ABCMW, ABCN, ABCNU, ABCOW, ABCP, ABCQ, ABCR, ABCRY, ABCSV, ABCT, ABCU, ABCUW, ABCV, ABCW, ABCY, ABDFH, ABDH, ABDHJ, ABDHL, ABDHN, ABDI, ABDJ, ABDJY, ABDN, ABDQT, ABDT, ABE, ABEFH, ABEH, ABEHW, ...}.

The same sample in the corresponding CIM symbols:

日: {日, 日月, 日月金, 日月金木, 日月金木心, 日月金水, 日月金火, 日月金火竹, 日月金火戈, 日月金火一, 日月金土, 日月金竹, 日月金竹大, 日月金竹一, 日月金竹心, 日月金竹 X, 日月金戈中, 日月金戈一, 日月金十大, 日月金十一, 日月金十弓, 日月金十尸, 日月金大弓, 日月金大手, 日月金中, 日月金一, 日月金一人, 日月金一女, 日月金一田, 日月金弓, 日月金弓山, 日月金人田, 日月金心, 日月金手, 日月金

口, 日月金口卜, 日月金尸女, 日月金廿, 日月金山, 日月金山田, 日月金女, 日月金田, 日月金卜, 日月木火竹, 日月木竹, 日月木竹十, 日月木竹中, 日月木竹弓, 日月木戈, 日月木十, 日月木十卜, 日月木弓, 日月木手廿, 日月木廿, 日月水, 日月水火竹, 日月水竹, 日月水竹田, …} (1804 CIM tactographemes).

Due to the size of the tactographemic dispersion for each CIM grapheme (a total of 59,199 occurences) the small samples must suffice at this point and the data in question will be introduced below based on the dispersion numbers. In the case of the tactographemic dispersion CIM graphemes are distributed between 14,152 CIM tactographemes. The dispersion numbers for each CIM grapheme are listed in descending order in the table below:

Tab. 7.4 Tactographemic dispersion of CIM graphemes

Tactographemic Dispersion Numbers					
1.	M 一	3,824	14.	K 大	2,185
2.	H 竹	3,701	15.	V 女	2,157
3.	B 月	2,907	16.	D 木	2,150
4.	I 戈	2,851	17.	S 尸	2,144
5.	N 弓	2,768	18.	U 山	2,130
6.	O 人	2,742	19.	G 土	1,895
7.	F 火	2,618	20.	P 心	1,894
8.	R 口	2,564	21.	E 水	1,890
9.	Y 卜	2,556	22.	A 日	1,804
10.	T 廿	2,543	23.	W 田	1,715
11.	L 中	2,369	24.	Q 手	1,641
12.	J 十	2,339	25.	X	1,586
13.	C 金	2,226			
Average: 2,367.96 Median: 2,226					

CIM tactographemic dispersion of graphemes can also be characterized by the standard statistical measures of dispersion and distribution:

- standard deviation: 566.50;
- average deviation: 431.64;
- median absolute deviation: 423.64.

The data in Tab. 7.4 is presented below in the form of a diagram (Fig. 7.7):

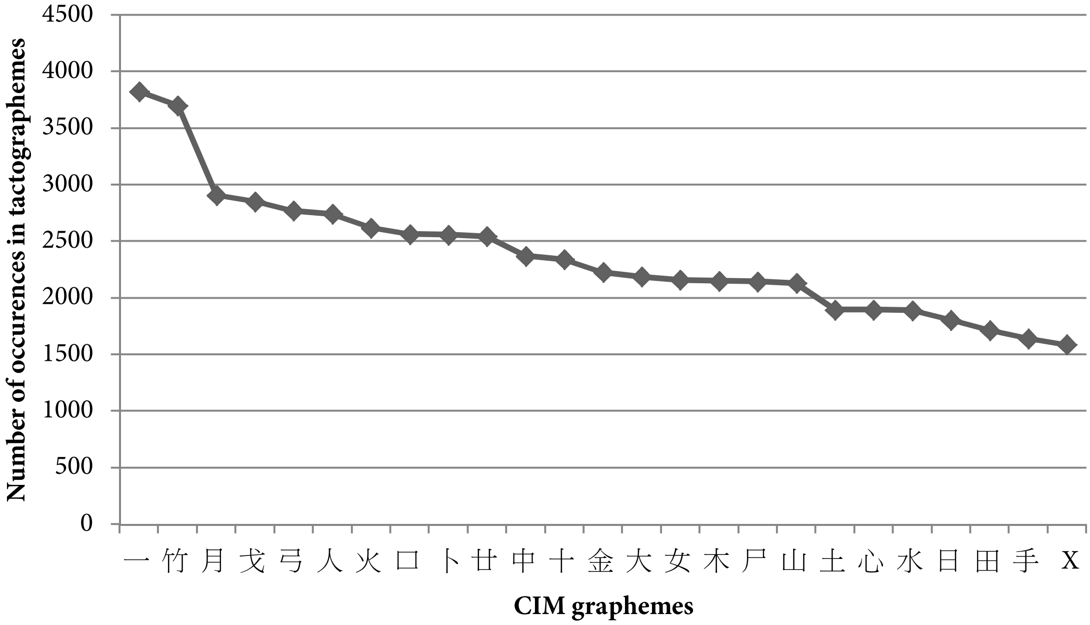


Fig. 7.7 Tactographemic dispersion numbers of CIM graphemes

The tactographemic dispersion numbers of CIM graphemes are not spread over a large range of values. In the case of tactographemes this feature is not a part of the design, simply because the notion of a tactographeme is specific to the graphotactic framework.

7.1.2.5.2. CIM graphotactemic dispersion

The same procedure as in the previous section can be applied to the graphotactemic dispersion of the CIM graphemes. A CIM graphotacteme belongs to the dispersion of a CIM grapheme if the CIM grapheme is a part of it. In other words the dispersion of a CIM grapheme is a set of all CIM graphotactemes (characters) that have this grapheme as their constituent. The samples of graphotactemic dispersion of CIM graphemes are presented below:

A: {A, AA, AAA, AAAH, AAAM, AAAV, AABT, AABUU, AAF, AAHAF, AAHM, AAHML, AAJV, AAM, AAMH, AAMJ, AAMU, AAPH, AAPV, AATE, AAVF, AAYF, AB, ABAC, ABBE, ABBT, ABBUU, ABF, ABGR, ABHA, ABHAF, ABHF, ABIK, ABJCM, ABJJ, ABKF, ABKQ, ABME, ABMGI, ABMR, ABMS, ABOF, ABOU, ABT, ABU, ABUG, ABUU, ABWI, ACI, ACIM, ACMBC, ACNH, ACR, ACSH, ACWA, AD, ADD, ADHAF, ...}, The following is the same sample in the corresponding CIM symbols:

日: {日, 日日, 日日日, 日日日竹, 日日日一, 日日日女, 日日月廿, 日日月山山, 日日火, 日日竹日火, 日日竹一, 日日竹一中, 日日十女, 日日一, 日日一竹, 日日一十, 日日一山, 日日心竹, 日日心女, 日日廿水, 日日女火, 日日卜火, 日月, 日月日金, 日月月水, 日月月廿, 日月月山山, 日月火, 日月土口, 日月竹日, 日月竹日火, 日月竹火, 日月戈大, 日月十金一, 日月十十, 日月大火, 日月大手, 日月一水, 日月一土戈, 日月一口, 日月一尸, 日月人火, 日月人山, 日月廿, 日月山, 日月山土, 日月山山, 日月田戈, 日金戈, 日金戈一, 日金一月金, 日金弓竹, 日金口, 日金尸竹, 日金田日, 日木, 日木木, 日木竹日火, ...}.

In the case of the graphotactemic dispersion the CIM graphemes are distributed between 27,607 CIM graphotactemes. The dispersion numbers for each CIM grapheme are listed in descending order in Tab. 7.5.

Tab. 7.5 Dispersion of CIM graphemes

Graphotactemic Dispersion Numbers					
1.	M 一	8,004	14.	K 大	3,942
2.	H 竹	7,133	15.	S 尸	3,735
3.	B 月	5,790	16.	C 金	3,708
4.	O 人	5,464	17.	V 女	3,683
5.	I 戈	5,313	18.	U 山	3,665
6.	R 口	5,255	19.	G 土	3,389
7.	N 弓	5,087	20.	P 心	3,376
8.	T 廿	4,606	21.	E 水	3,343
9.	Y 卜	4,429	22.	A 日	2,913
10.	F 火	4,427	23.	W 田	2,675
11.	L 中	4,298	24.	Q 手	2,647
12.	D 木	4,057	25.	X	2,187
13.	J 十	4,014			
Average: 4,285.6 Median: 4,014					

CIM graphotactemic dispersion of graphemes can also be characterized by the standard statistical measures of dispersion:

- standard deviation: 1,355.283181;
- average deviation: 1,013.152;
- median absolute deviation: 987.2.

Analogously to the previous section, the data in Tab. 7.5 is presented below in the form of a diagram (Fig. 7.8).

As Fig. 7.8 shows the distribution of CIM graphemes between CIM graphotactemes is relatively even. In this case it is a result of the presupposed criteria of component (CIM grapheme) selection for character structure representation in CIM. Since CIM graphotactemes correspond to individual characters, it is possible to directly tie the even dispersion of CIM graphemes to the component selection criteria formulated by the CIM inventor. From this point of view one might expect an even more balanced graphotactemic dispersion. It must be noted, however, that the selection of the 24 graphemes was originally done for a much smaller number of characters than the corpus investigated here.

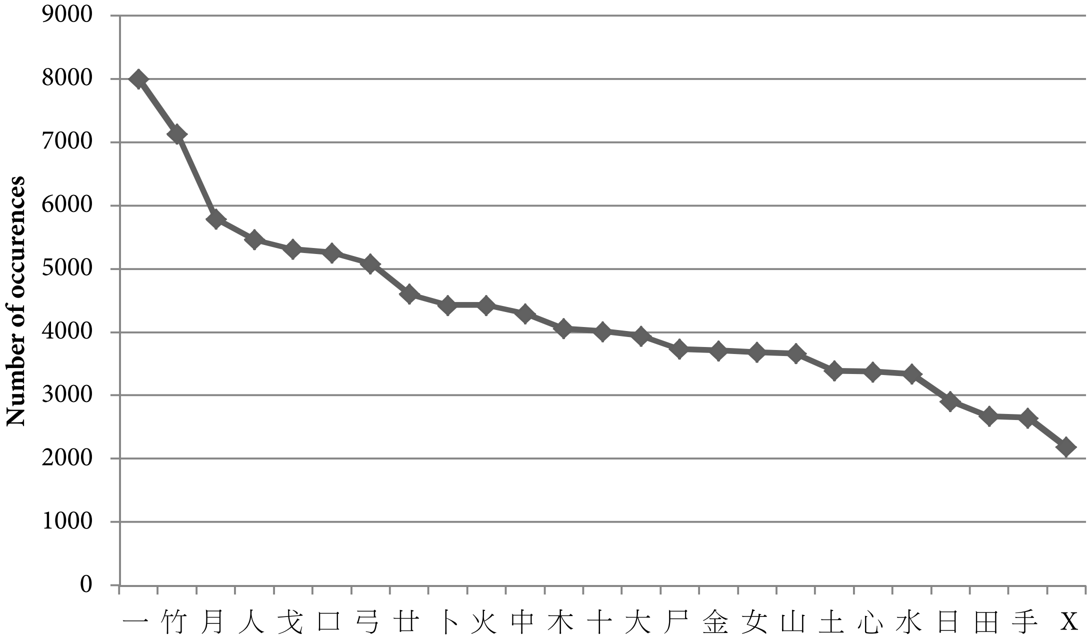


Fig. 7.8 Graphotactemic dispersion numbers of CIM graphemes

As it was already mentioned, at this point the data are insufficient to speculate on the relationship between tactographemic and graphotactemic dispersion.

7.1.2.5.3. Tactographic and graphotactic dispersions compared

A visual comparison of tactographemic and graphotactemic CIM dispersion curves is shown in Fig. 7.9.

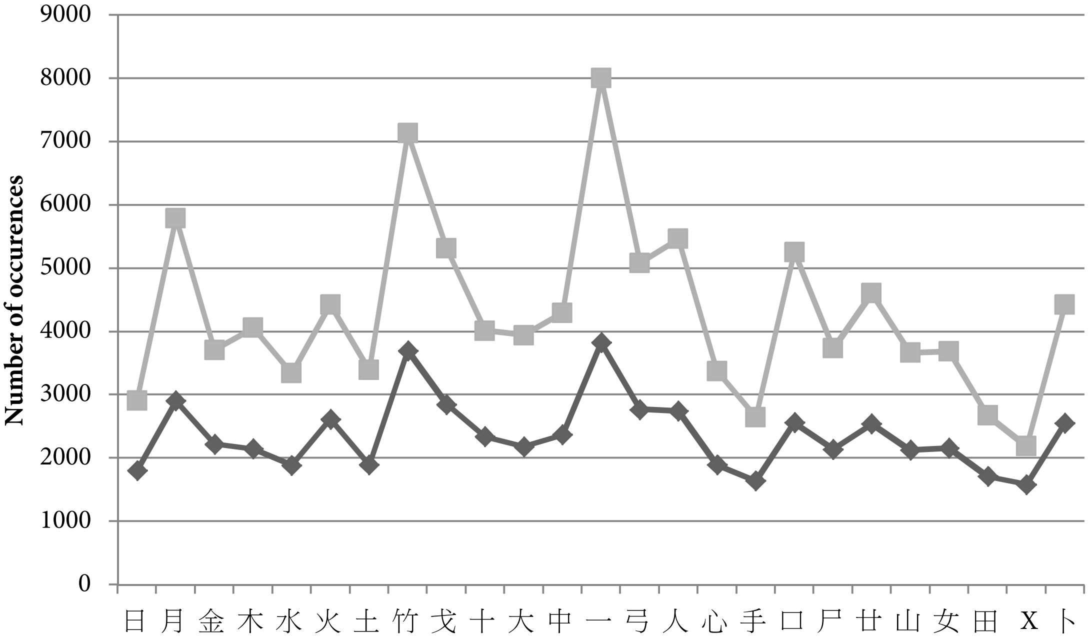


Fig. 7.9 Comparison of tactographemic and graphotactemic dispersions

The above diagram reveals a striking similarity between both types of dispersion. The relatively even distribution of the CIM graphemes seems to be another indication of the consistent and efficient design of the Cangjie input method.

In summary, based on the results, it can be stated that the analytic model outlined in Bańczerowski (2009) proved to be a legitimate research tool. More complete interpretation of the results depends substantially on comparative analysis against a background of wider range of graphotactic investigations. The results provided in Section 7.2. may be a good start.

7.2. Graphotactics of Chinese script

This section presents the results of graphotactic analysis of selected sets of Chinese characters. In order to capture the diversity hidden behind the term ‘Chinese characters’ it was necessary to diversify the investigation accordingly. Because of the lack of comparable research it seems natural to provide results that can be a basis for at least preliminary conclusions and generalizations. *Hànzi* offer a wide choice of investigative range – adoption of a combination of different criteria leads to a number of significantly different character sets. The criteria may be geographic (different locales), structural (traditional vs. simplified), pragmatic (official vs. variant forms), historical (currently used vs. abandoned), and practical (for the purposes of education, language planning, etc.). Different set of characters will be used for the investigation of the general properties of Chinese script, which are different from the character sets used for the purposes of primary school education.

This study investigates the general graphotactical properties of Chinese script. It is for that reason that the Unihan database, the largest available set, is analyzed. The details and pitfalls of investigating a database of this size are explained in section 7.2.5.

The results would not be complete without a graphotactic inquiry into the largest possible set of homogeneous characters. That leaves a choice between traditional, simplified and Japanese *kanji* sets. The size criterium leaves the traditional sets as the sole option. Among the traditional sets, Big5 (CNS 11643:2007 Plane 1 and 2) presents the most attractive alternative. The details are provided in section 7.2.3.

To complete the investigation the largest set of simplified characters is also included in the analysis. In order to obtain more graphotactic data the simplified set is contrasted with a traditional set comparable in size.

Ordering of the sections in this chapter reflects the prominence of the sets and interpretability of the obtained results – Big5 being the most important and offering the richest options for interpretation, and Unihan being the most difficult to interpret, though still very important. The detailed results are provided in sections concerning these two sets. Not all of the diagrams illustrating graphotactic properties of characters in the investigated set are interpreted and explained – in most cases, for the sake of the reader’s convenience, only the first figure pertaining to a certain type of properties and/or certain type of elements is explained in the simpliest possible terms. The explanations are provided independently of Section 7.1.2. The conclusions will be discussed in the summary sections.

7.2.1. Levels of analysis

Following the distinctions made in Sections 4.2. and 4.4., the graphotactic analysis conducted in this section recognizes two levels of decomposition of characters, or, in other words, two types of components (graphemes) – immediate and basic. In practice it means that two separate computations have to be performed for each investigated set of characters. The issue of establishing the respective sets of immediate and basic components is discussed in Section 7.2.2.

7.2.2. Database

The process of adopting the most appropriate corpus of Chinese characters for graphotactic analysis should be a conscious one. At the current stage of digitalization of Chinese script, including the structural data, many limitations are lifted. The size and type of corpus depends on the specific purpose of the analysis. The most general purpose of graphotactic analysis is the quantification of combinational properties of components, or, in other words, the properties of characters in terms of their constituent parts – to meet a requirements of this type of inquiry the largest possible database should be adopted. The large size criterion is best met by the CHISE and KDP databases that correspond to the Unihan database (CJK Unified Ideographs set). The necessary details regarding the two databases, as well as the reasons for selecting the KDP, were presented in Section 5.1.1.

Because of the flexibility it offers, a database like KDP would be necessary, regardless of the scale of investigation. Having a large database at one's disposal, it is possible to extract any of its subsets for analysis. This procedure was adopted in this study. The first step was to establish a list of all unique immediate components in KDP. This is a very straightforward procedure, since the components contained in the IDS descriptions are considered to be immediate components. The process of establishing a list of basic components requires more effort, but it is also automatic. KDP lists only immediate components of characters, most of which are further decomposable into more basic constituents. The decomposition data for such components are to be found elsewhere in the database – the data can be found by the search algorithm and assigned to the analyzed component. The basic components are searched recursively – in the event that the assigned components are still decomposable into even more basic constituents, the procedure is repeated. This operation is carried out for every single character in KDP and a list of unique basic components is compiled based on the results. The actual recursive search operation and the assignment of basic components to a character can be demonstrated with a relatively simple example ⿱竹梢:

⿱竹梢 → 竹梢[immediate components: 竹梢] →
竹 → 亇亇 →
亇 → 𠂉亅 →
𠂉 → 𠂉[non-decomposable basic component: 𠂉] →
亅 → 亅[non-decomposable basic component: 亅] →
梢 → 木肖 →
木 → 木 [non-decomposable basic component: 木] →
肖 → ⺌月 →
⺌ → ⺌[non-decomposable basic component: ⺌] →
月 → 冂冫 →
冂 → 冂[non-decomposable basic component: 冂] →
冫 → 冫[non-decomposable basic component: 冫]

As a result the character ⿱竹梢 is assigned two immediate components {竹, 梢} and six basic ones {𠂉, 亅, 木, ⺌, 冂, 冫}. Both assignments pertain to separate levels of analysis, and both sets of components are tactographemes – {竹, 梢} on the immediate level, and {𠂉, 亅, 木, ⺌, 冂, 冫} on the basic level. 竹 and 梢 are added to the inventory of immediate components; 𠂉, 亅, 木, ⺌, 冂 and 冫 are added to the inventory of basic components.

The extraction of componential data for any subset of KDP is relatively simple, but the inventory of basic components must be extracted from the entire KDP, not just the investigated subset. The reason for this is the fact that some basic components are extracted from the description of characters that are not in a given subset.

It should stressed that this procedure results in a significantly different inventory in comparison to the basic component sets mentioned in different parts of this book (CDP, GF3001-1997, Stalph, etc.). The approach adopted here and described above leads to a purely graphical level of representation and to a substantial reduction in the number of basic elements. This may be exemplified in contrast to the approach advanced by Chuang and Teng (2009). Although they declare a graphical approach to the decomposition of characters, in some cases the depth of purely graphically motivated decomposition is limited. They formulate this limitation explicitly: "*basic components are the smallest units of graphical identity of characters, components with this function should not be decomposed into smaller basic components. For example, the component '貝' in '寶' should not be further decomposed into '目' and '八', because '貝' cannot be identified by the components {目, 八} and '寶' cannot be identified by {宀, 王, 缶, 目,*

㇒㇏}."[321] A quick review of KDP shows that '貝' is further decomposed. It should be kept in mind to what type of constituents the analytic results for basic components pertain.

KDP in its raw form is not immediately suitable for analytic purposes. There are two reasons for this:

- it contains additional information on sources of decomposition;
- it contains non-character entries.

The latter concerns only the analysis of KDP itself, and for that reason it will be addressed in Section 7.2.5. The former issue needs to be resolved for the entire KDP prior to extracting the subsets,[322] otherwise the results would be substantially distorted. This problem will be shortly discussed below. The additional information provided in KDP pertains to a different decomposition of some characters in different sources. Examples were given in Section 5.1.1.2., but for the sake of clarity some additional will be supplied below:

U+4EA0	亠	⿱丶一[GTK]	⿱丨一[J]
U+4EB6	亶	⿱亩旦[GTJ]	⿱亩且[K]
U+4F14	伔	⿰亻冗[GT]	⿰亻⿱冖儿[K]
U+7E90	纐	⿰糸頝[J]	⿰糹頝[G]
U+4391	䎑	⿰羽录[G]	⿰羽彔[T]
U+4331	䌱	⿰糹𣎆[T]	⿰糸𣎆[K]

The examples illustrate the alternative decompositions with an indication of their source. KDP contains the following number of entries with individual alternative sources:

[G] China	2,423	entries
[J] Japan	1,454	entries
[K] Korea	1,890	entries
[T] Taiwan	2,479	entries
[V] Vietnam	437	entries
[X] unidentified[323]	113	entries

[321] Chuang & Teng 2009: 80-81.

[322] In the case of this study the subsets include: Big5, PRC's 通用规范汉字表 *tōngyòng guīfànhànzì biǎo* 'Common Standard Chinese Characters Table', and the joint Taiwanese list 常用國字標準字體表 *chángyòng guózì biāo zhǔn zìtǐ biǎo* 'List of Standard Forms of Frequently Used Characters', and 次常用國字標準字體表 *cìchángyòng guózì biāozhǔn zìtǐ biǎo* 'List of Standard Forms of Less Frequently Used Characters'.

[323] The author was unable to identify the reference of 'X'.

Some entries contain alternative decompositions without any indications of the source. The examples show all kinds of configurations of sources. For the sake of consistency the [T]aiwan source was selected as a primary one. In cases where there is no [T] source, [G] is selected instead, which guarantees the primacy of Chinese sources. In the remaining cases, including instances in which there is no indication of the source of decomposition, the first alternative is selected.

The Ideographic Description Characters are ignored in this study, but it seems probable that IDCs might be essential for equigraphic and disgraphic study of Chinese script.[324] The algorithms for character decomposition with IDCs can be found in Lu et al. (2002).

The final format of the input data for analysis is a result of long and multiple testing, but it is probable, that despite the adjustments the database is not clear of hidden problems. At this point it can be said that the results do not raise suspicions concerning the input data or the data handling.

To the author's knowledge there is no font covering all characters in CJK Unified Ideographs, not to mention the fonts representing character components. For that reason presentation of some of the the qualitative aspects of analysis, especially component lists, was affected by the inability to properly display and print certain contents. In cases of problems insolvable or too time consuming respective glosses are provided.

7.2.3. Big5 (CNS 11643 Plane 1 and 2)

The selection of a set of traditional characters was limited to the Big5 set (or CNS 11643:2007 Plane 1 and 2) for technical reasons. Big5 is not the largest set of traditional characters – all the characters it contains are a subset of CNS 11643:2007 standard.[325] The componential descriptions that are extracted from the IDS sequences in the KDP database are encoded in Unicode. CNS 11643 uses its own encoding, and as a result, a large portion of characters in Plane 3 and beyond are not decoded properly. There are, however, independent reasons for selecting Big5. It is a well established set in the Chinese-writing community in Taiwan and Hong Kong, meaning that any references to it as a basis for analysis are easily recognized. Finally, but importantly, it is the only set known to the author that was partially analyzed in a manner similar to the graphotactic framework. Chuang and Teng (2009) offer some details of their investigation of the Big5 set in terms of CDP components that are directly pertinent to graphotactics.[326]

[324] Bańczerowski 2009: 20-21, Bańczerowski 2013 and Section 3.2.1.1.

[325] See Section 2.2.1.2.

[326] Chuang & Teng 2009: 79-81.

In terms of CDP components there are 301 characters that are not composed of unique sets of components (tactographemes). These characters (graphotactemes) can be classified into two types:

- 234 characters consisting of non-recurring components:

加 (另, 叻), 旭 (旯, 旮), 架 (枷, 柺), 翌 (翊, 翋), 可 (叮), 召 (叨), 只 (叭), 妃 (改), 吝 (呅), 否 (吥), 呆 (杏), 含 (吟), 岑 (岒), 岌 (岋), 旱 (旰), 李 (杍), 防 (邡), 阮 (邧), 阬 (邟), 坌 (坋), 周 (坰), 坪 (幸), 委 (姀), 帕 (帛), 招 (拐), 易 (昒), 杳 (杲), 盲 (盳), 邸 (阺), 旼 (旻), 炅 (炚), 邴 (陃), 保 (咻), 垢 (垕), 某 (柑), 查 (柦), 毗 (毘), 泵 (砅), 珀 (皇), 省 (眇), 耑 (峏), 衍 (洐), 郁 (陏), 咡 (咠), 呰 (呲), 咰 (昫), 峗 (峞), 峉 (峈), 峇 (峆), 昶 (昹), 郲 (陎), 哲 (哳), 員 (唄), 宴 (晏), 晌 (晑), 案 (桉), 烘 (粪), 杦 (染), 郟 (陜), 啦 (掊), 唯 (售), 垩 (垭), 娶 (娵), 脣 (脤), 部 (陪), 都 (陼), 陴 (郫), 陬 (郰), 唼 (婄), 婇 (桵), 勛 (賀), 啻 (啼), 景 (晾), 晷 (喒), 棘 (棗), 椎 (集), 渺 (渻), 郵 (陲), 嵒 (喦), 棷 (棸), 棐 (棑), 棼 (棥), 渮 (菏), 鄃 (隃), 暉 (暈), 蜃 (蜄), 訾 (訿), 賅 (賌), 隕 (鄖), 楙 (楺), 輁 (輂), 塵 (塷), 墓 (塻), 幕 (幙), 摧 (嶊), 窪 (漥), 蜥 (蜤), 障 (鄣), 嗺 (嶉), 蜒 (蜑), 暮 (暯), 槸 (槷), 槻 (槼), 篎 (箵), 諆 (諅), 澫 (蕅), 鴪 (鴥), 麋 (麊), 鏖 (鏕), 櫑 (櫐), 櫙 (藲), 譪 (藹), 礧 (礨), 鷍 (鸐), 蠬 (蠪)

- 67 characters consisting of recurring components:

比, 爻, 圭 (垚), 多, 朋, 林, 炎 (焱, 燚), 玨, 品, 奻 (姦), 哥, 弱, 玆, 昍 (晶), 棘, 棗, 兢, 赫, 砳 (磊), 聶, 蟲, 競, 轟, 矗, 劦, 孖, 屾, 幵, 艸, 戔, 沝 (淼), 牪, 芔 (茻), 甡, 秝, 兟, 惢, 掱, 毳, 猋, 皕, 聑, 臸, 覞, 豩, 賏, 畾, 皛, 虤, 雔 (雥), 贔, 嚞, 鑫, 鱻, 羴, 驫, 麤, 龘.

A comparison with the graphotactic analysis will require some refinements. The discussed 301 characters cannot be counted directly among the tactographemes. The 234 characters are generated out of 115 unique sets, which means there are 115 tactographemes in this set. The 67[327] characters formed by a multiplication of one component should also be recounted with regard to the number of tactographemes. An easy calculation reveals 58 tactographemes of this type. The total number of CDP tactographemes generating more than one graphotacteme (character) is then 115+58=173. The total number of CDP tactographemes equals the number of tactographemes generating 1 graphotacteme + the number of tactographemes generating more than 1 graphotacteme: 12,817[328]+173=12,990. Hence, the average CDP tactographemic efficiency in the BIG5 set is 1.0047.

The remainng part of this section is devoted to a graphotactic analysis of the Big5 character set.

[327] From a graphotactic perspective the number is larger, since most of the recurring components also form simple characters.

[328] Chuang & Teng 2009: 79.

7.2.3.1. Big5 – immediate components

7.2.3.1.1. General properties

The analysis of IDS descriptions extracted for the Big5 set reveals 2,420 immediate components. It must be repeated that this is a functional category, different from the CDP compound components, and for that reason the numbers and lists cannot be compared directly. Other general results of the analysis are provided below and in Tab. 7.6:

- number of graphotactemes: 13,051;
- number of tactographemes: 12,939;
- number of graphemes (immediate components): 2,420;
- average tactographemic efficiency: 1.01;
- average graphemicity: 2.01.

Table 7.6 General quantitative properties of Big5 immediate tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	187	203	1.09
2	12,443	12,539	1.01
3	268	268	1.00
4	33	33	1.00
5	7	7	1.00
6	1	1	1.00
Total:	**12,939**	**13,051**	

The data in Tab. 7.6 will be depicted in diagrams in successive subsections. For the sake of the reader, the following is a quick and convenient reminder: tactographons (jointly called t-family) are sets of tactographemes with the same graphemicity; '1' in the 'Tactographon (T-family)' column indicates tactographons that consist of tactographemes containing 1 grapheme (graphemicity equal to 1); tactographemicity indicates the number of tactographemes in a given tactographon; t-graphotactemicity indicates the number of graphotactemes (characters) generated by tactographemes of a given tactographon; and t-efficiency pertains to average graphotactemic efficiency of a given tactographon.

7.2.3.1.2. Big5 immediate tactographemicity

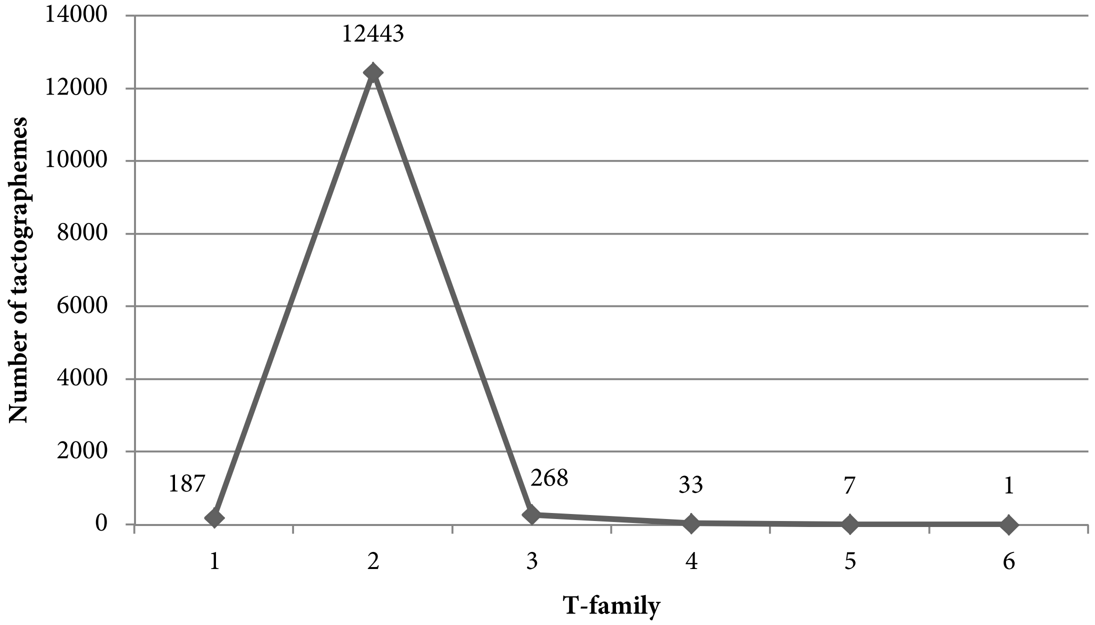


Fig. 7.10 Big5 immediate tactographemicity

Figure 7.10 shows the number of tactographemes in each tactographon, or in other words, how many tactographemes consist of a given number of graphemes. The tactographemic curve spikes sharply for the tactographon with a graphemicity of 2. More than 95% of the immediate Big5 tactographemes belong to this category.

7.2.3.1.3. Big5 immediate t-graphotactemicity

Figure 7.11 shows the number of graphotactemes generated by each tactographon, or in other words, how many characters are generated by tactographemes consisiting of a given number of graphemes. The t-graphotactemic curve spikes for the number of tactographemes consisting of 2 graphemes. Tactographemes in this category of graphemicity generate 96% of all graphotactemes.

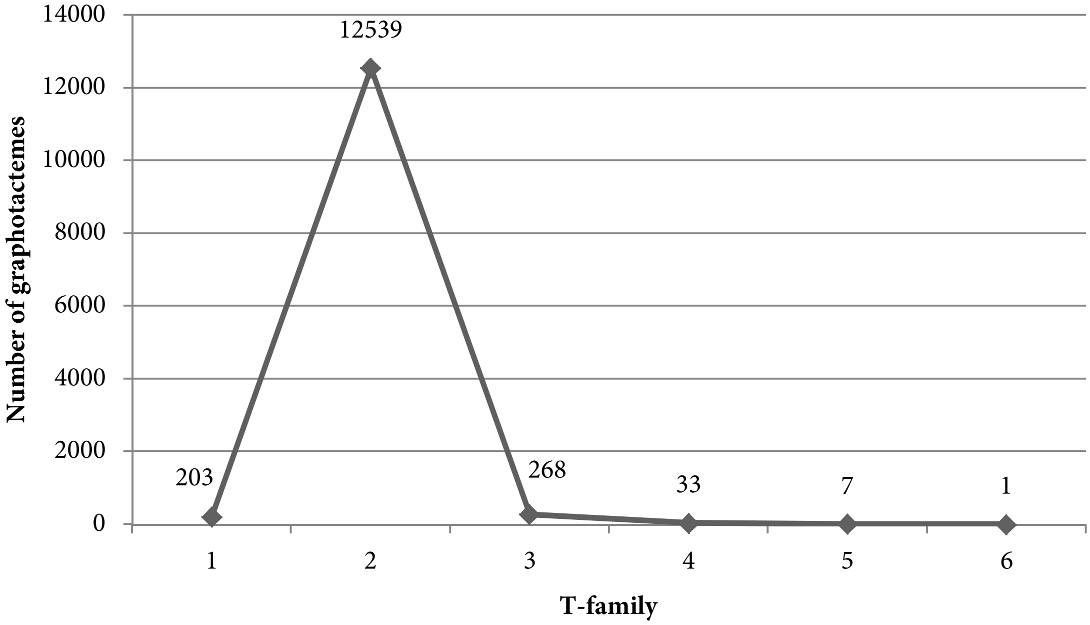


Fig. 7.11 Big5 immediate t-graphotactemicity

7.2.3.1.4. Big5 immediate tactographemicity and t-graphotactemicity compared

The comparison of quantitative properties of tactographemicity and t-tactographemicity should reveal the regularities, or the lack of them, between the number of equigraphemic tactographemes and the number of graphotactemes generated by them. This comparison will be repeated for every investigated set of characters and level of analysis.

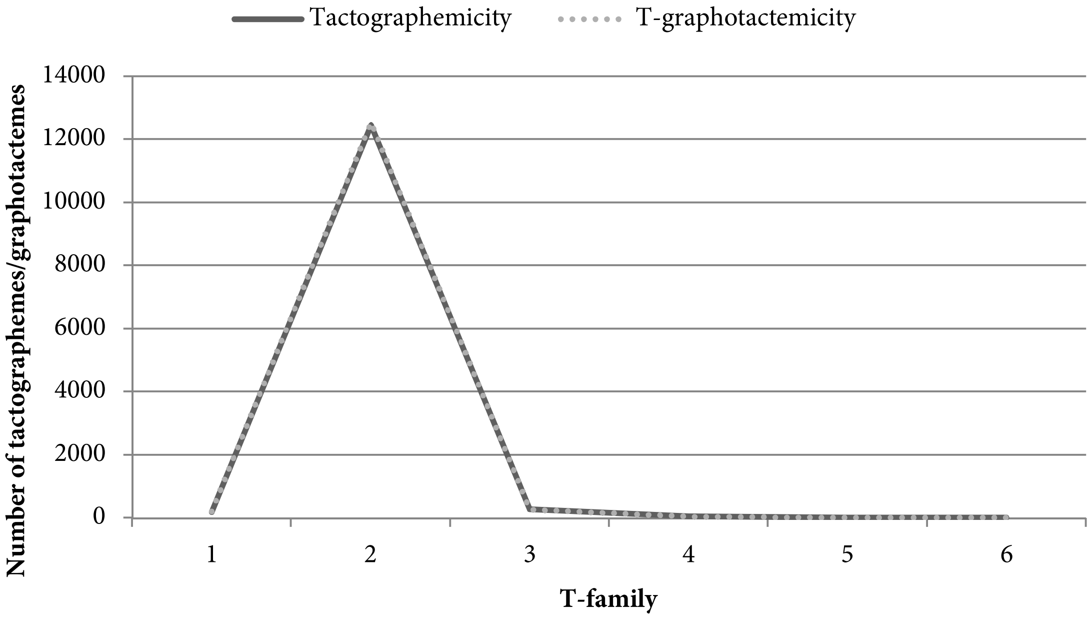


Fig. 7.12 Big5 immediate tactographemicity and t-graphotactemicity

The similarity of both curves is striking, and if further analysis shows the same or simlar degree of correlation, it will give an empirical basis for theoretical claims.

7.2.3.1.5. Big5 immediate t-efficiency

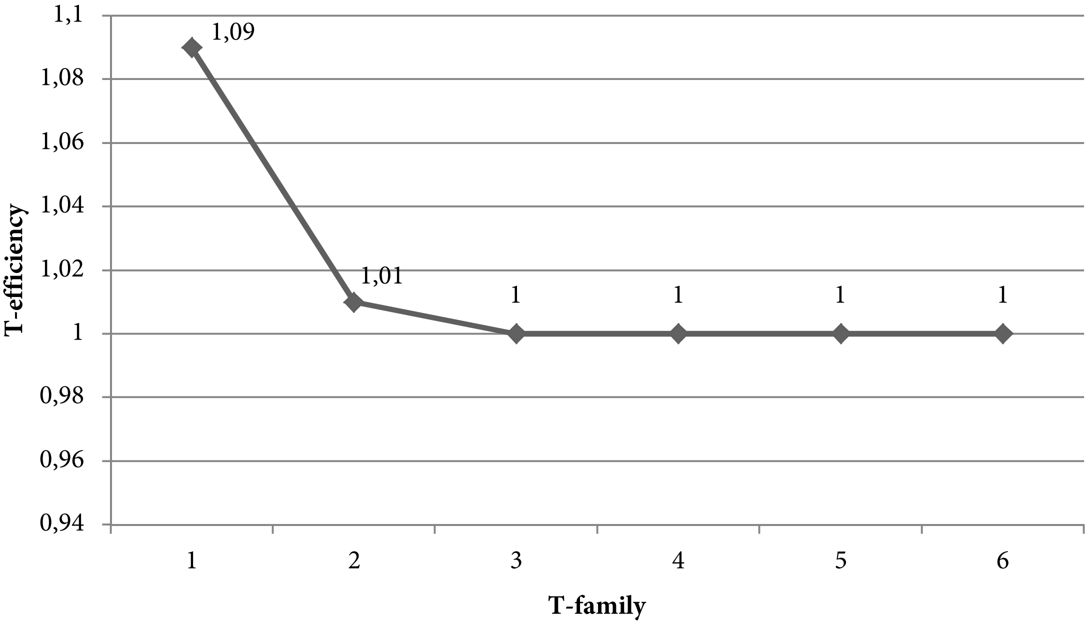


Fig. 7.13 Big5 immediate t-efficiency

T-efficiency accounts for the graphotactemic efficiency of each member of the t-family (a ratio of the number of graphotactemes to the number of tactographemes restricted to individual tactographons). As the diagram shows, only two immediate Big5 tactographons generate more graphotactemes than the number of tactographemes of which they consist. The t-efficiency value cannot drop below 1.

7.2.3.1.6. Big5 immediate tactographemic t-efficiency

Fig. 7.14 indicates the average graphotactemic efficiency of individual tactographemes belonging to a given tactographon which is a correlation of the number of generated graphotactemes with the number of generating graphemes. In other words, it shows the average graphotactemic efficiency value of tactographemes in a given tactographon. The value of tactographemic t-efficiency indicates the average number of graphotactemes generated by single graphemes in a given t-family. For example, graphemes belonging to tactographemes with a graphemicity equal to 2 will, on average, generate 0.5 graphotactemes. A steady decrease of efficiency along a logarithmic type of curve can be observed.

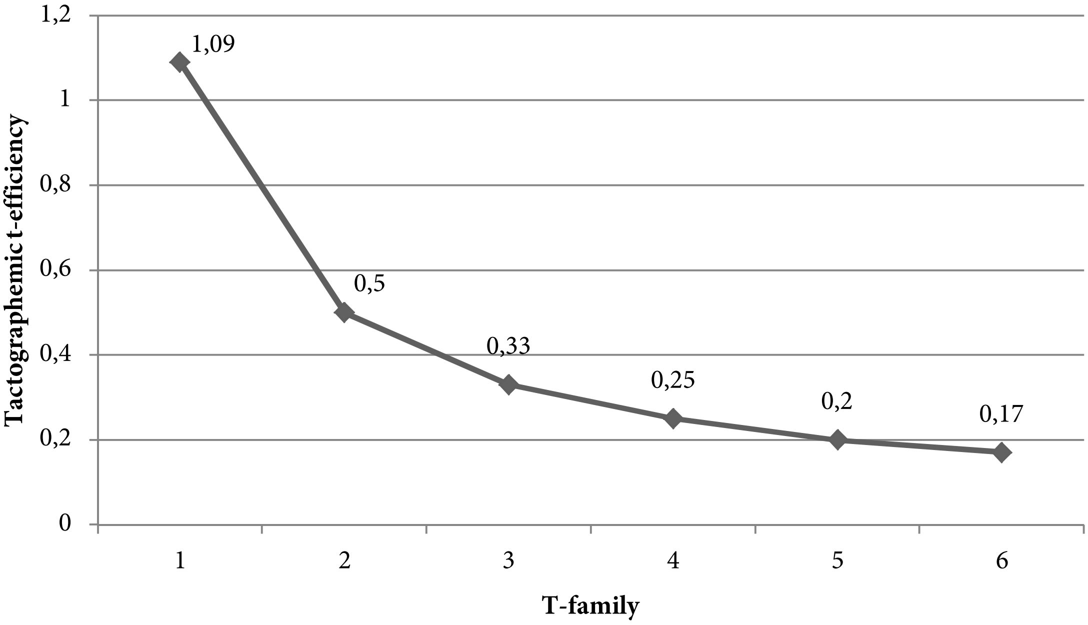


Fig 7.14 Tactographemic t-efficiency of Big5 immediate tactographons

7.2.3.1.7. Big5 immediate categorial graphotactemic efficiency

Fig. 7.15 shows how many Big5 tactographemes belong to a given category of graphotactemic efficiency, or in other words, how many tactographemes generate a given number of graphotactemes.

The results shown in Fig. 7.15 indicate that, as expected, the largest number of Big5 immediate tactographemes generate one graphotacteme; that is, the data shows that 99% of unique component sets generate just 1 character.

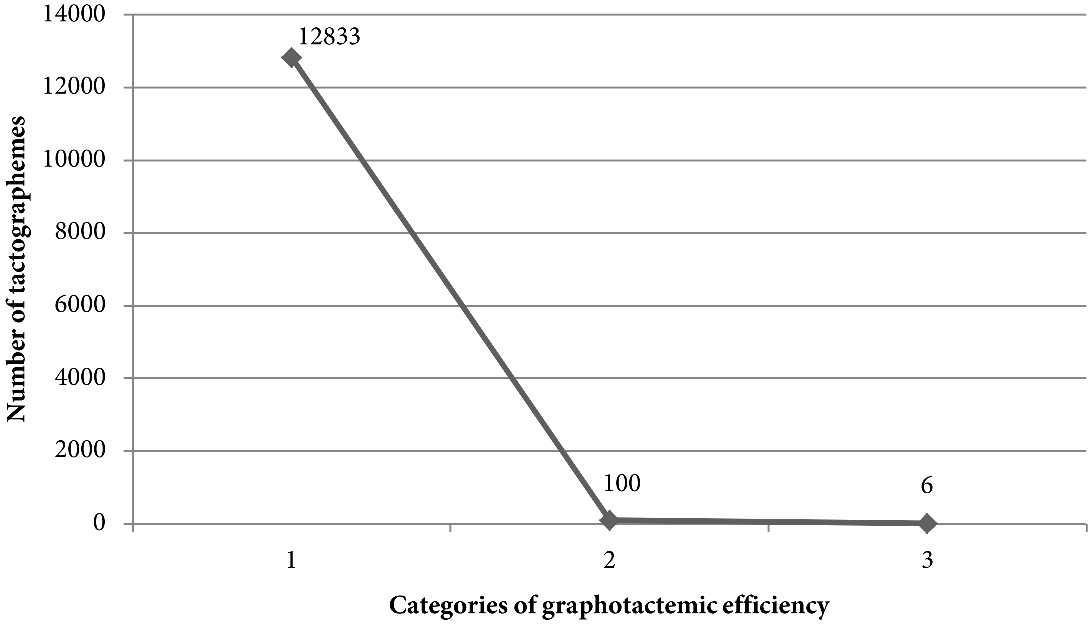


Fig. 7.15 The number of tactographemes generating given number of graphotactemes

7.2.3.2. Big5 basic components

7.2.3.2.1. General properties

Graphotactic analysis of the Big5 set revealed 304 basic components. This study concentrates on the quantitative properties of an investigated domain, but, given the existence of the CDP basic component system, listing the 304 graphemes for contrastive purposes is justified:[329]

⺍乏个⺁𦥑畢八卝甘歹里⺈⺁𠂢𠂉𠣜乡馬尔𣅀兒𣌭氏𣎆爻鳥皿𠂤𦉼𠃑𦥮艮与冂㐅重刂華𣏟⺈氏冖戶巳小鳥
匚亘𠂤朋业𢦏巨冂㇒く⺄冂⺊丷牛㐅肙㔾吕一丈丏丐丑且世丘丨丩丱丶丷丹丿乀乁
久乍乎乑乙乚乛乜九也乡亅事于井人亻以先兆入八冂冉冖冘冫几凵凸凹刂勹匚
匸十卌卍卜丬卩厂厶又及口史囗夂夊央女子孑孓宀寸小尢尸尺山川州工巨己已
巳巴巾干年广廴廿弋弓彐彑彡彳心忄戊戌我手扌才承攵斗旡日曰曲曳月木朩未
末本朱朿東東柬欠止毋毌母比毛氏民氵永灬為熏爪爿片牙牛牜犬犭瓜瓦甘生田
由甲申疋疌疒白皮皿目示礻禹禺米缶罒耂耳肅肉臣自臼舟艮虫衤襾角言谷豆豸
身車辶酉重金阝隶隹非革頁飛黽龜⻊マ龵龶盟戸𠂆ナ𠂉ク歹生乚厶㇆㇆亅ㄣ肀
尸㇄乃𠆢几小く朩少⺪ 臣𧘇豕 彐丨𦘒 五

[329] For the technical reasons, mentioned in Section 7.2.2., only 298 basic components can be presented. A list of CDP basic components was provided in Section 5.1.2.

The examination of the Big5 set in terms of basic components reveals sharp quantitative differences in comparison to the immediate components analysis. General results of the analysis are provided below and in Tab. 7.7:

- number of graphotactemes: 13,051;
- number of tactographemes: 12,141;
- number of graphemes (basic components): 304;
- average tactographemic efficiency: 1.07;
- average graphemicity: 4.8.

Table 7.7 General quantitative properties of Big5 basic tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	158	189	1.20
2	1,148	1,260	1.10
3	2,163	2,403	1.11
4	2,824	3,076	1.09
5	2,596	2,763	1.06
6	1,836	1,916	1.04
7	945	968	1.02
8	340	345	1.01
9	89	89	1.00
10	29	29	1.00
11	11	11	1.00
12	1	1	1.00
13	1	1	1.00
Total:	**12,141**	**13,051**	

7.2.3.2.2. Big5 basic tactographemicity

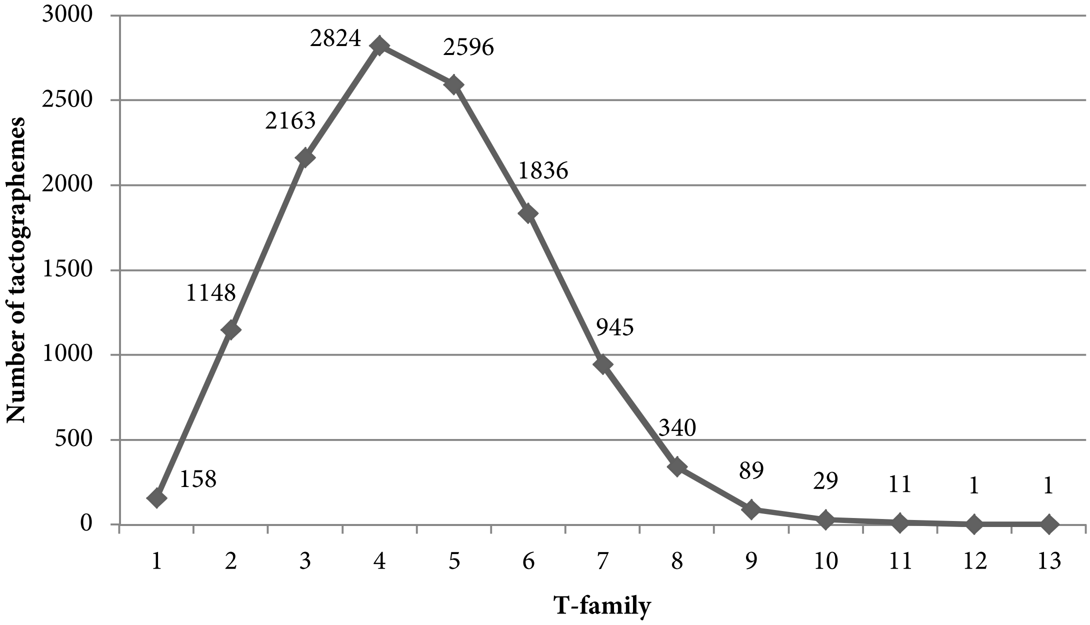


Fig. 7.16 Big5 basic tactographemicity

7.2.3.2.3. Big5 basic t-graphotactemicity

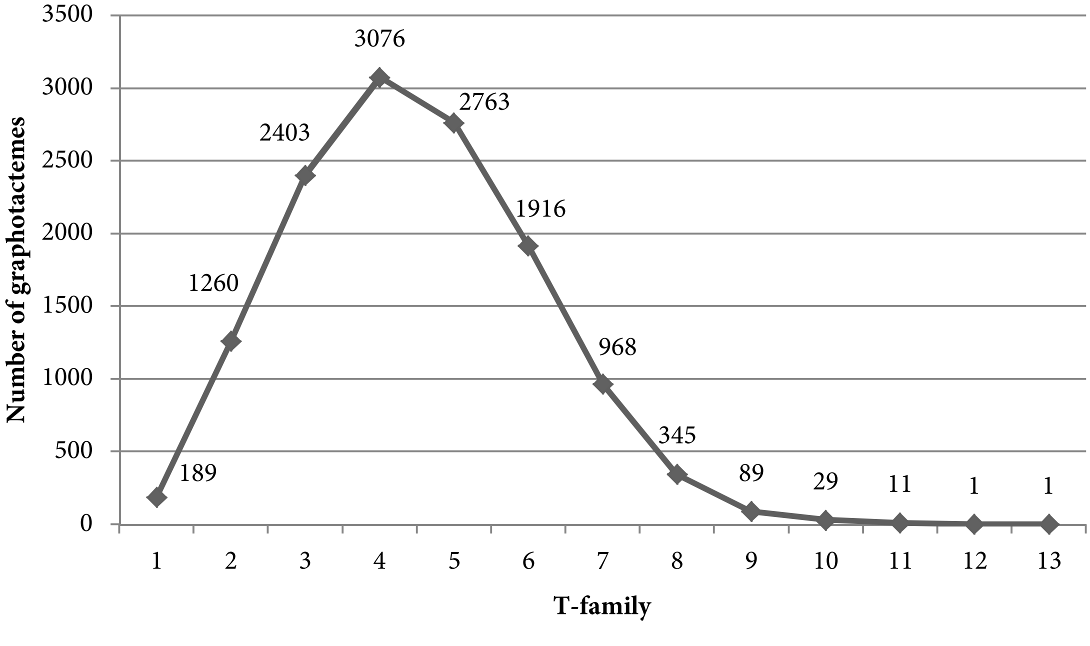


Fig. 7.17 Big5 basic t-graphotactemicity

7.2.3.2.4. Big5 basic tactographemicity and t-graphotactemicity compared

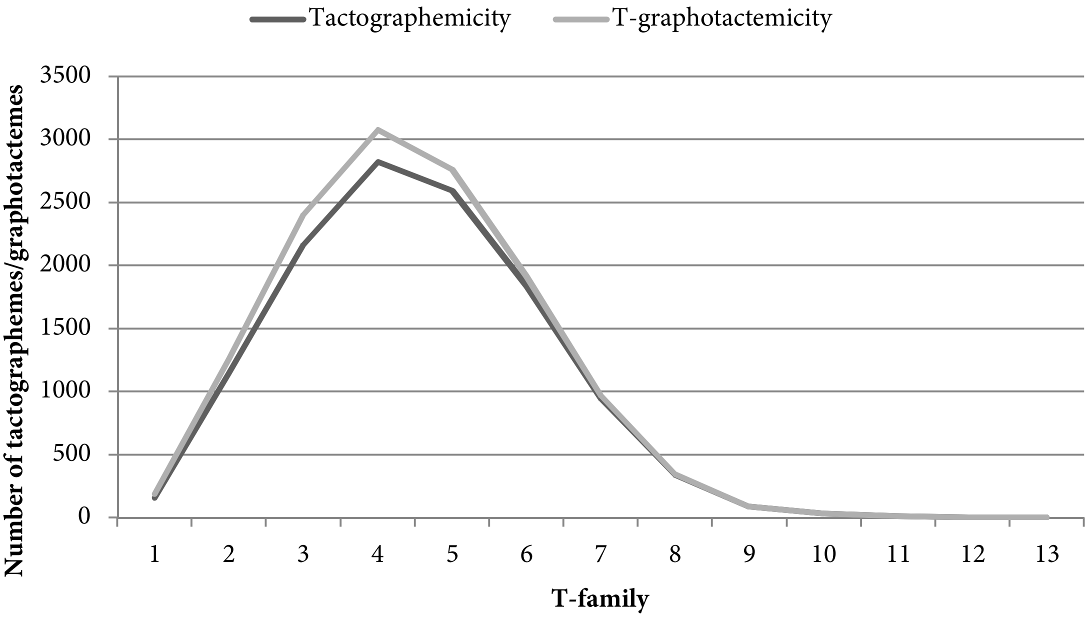


Fig. 7.18 Comparison of Big5 basic tactographemicity and t-graphotactemicity

It is evident that there is a strong correlation also in the case of basic tatographemicity and t-graphemicity.

7.2.3.2.5 Big5 basic t-efficiency

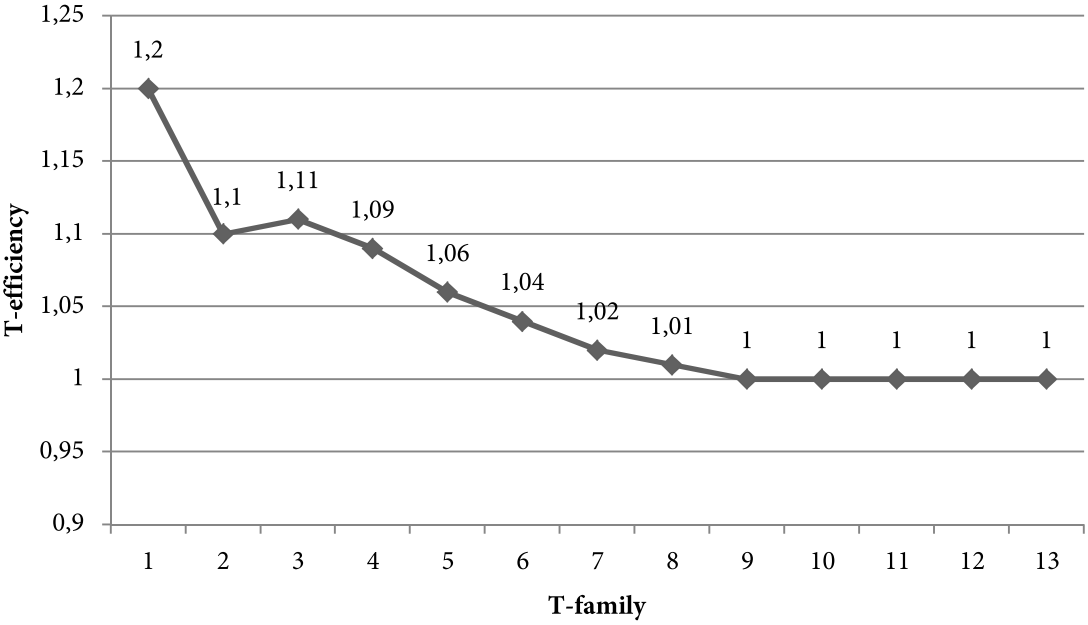


Fig 7.19 T-efficiency of Big5 basic tactographons

7.2.3.2.6. Big5 basic tactographemic t-efficiency

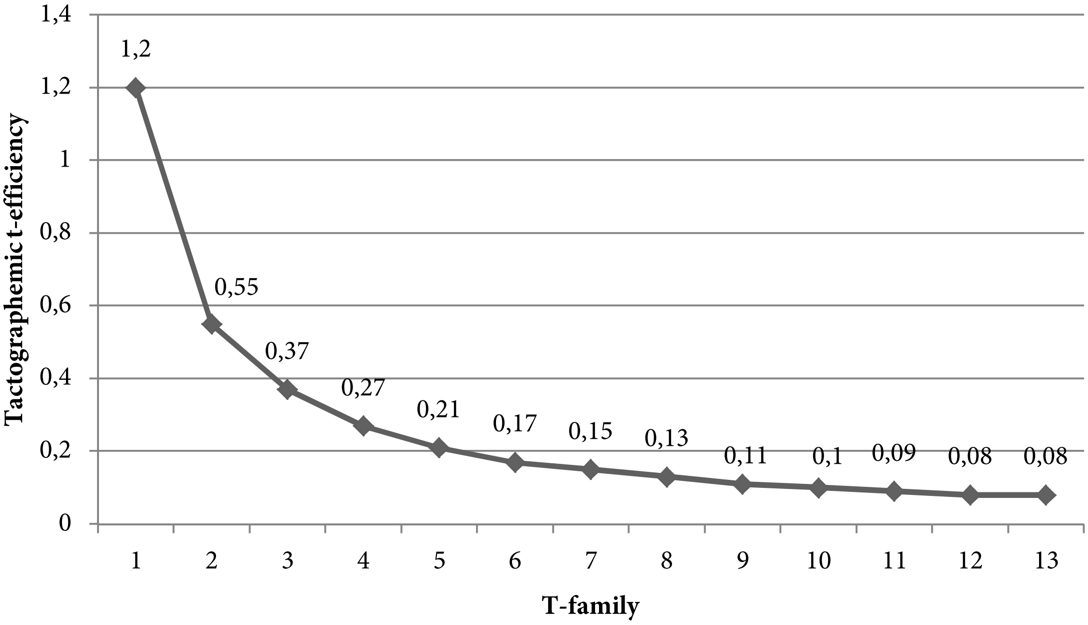


Fig. 7.20 Tactographemic t-efficiency of Big5 basic tactographons

7.2.3.2.7. Big5 basic categorial graphotactemic efficiency

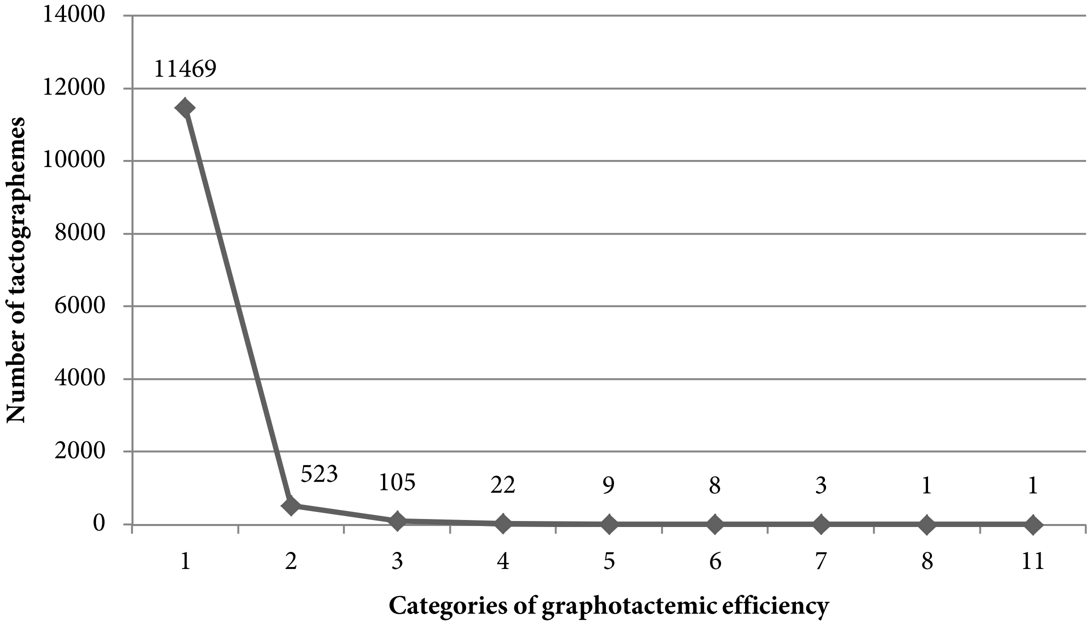


Fig. 7.21 The number of tactographemes generating given number of graphotactemes

7.2.4. Comparative analysis

An investigation of the Big5 set provides a good insight into the graphotactic properties of traditional Chinese script. This study is aimed at the totality of Chinese characters and such approach would not be complete without at least a glimpse at simplified *hànzi*. For reasons explained in section 4.6., simplified character sets are comparatively smaller in size than the traditional and open sets. For that reason, Big5 is too large to be directly confronted with any homogeneous set of simplified characters. As it was mentioned in section 2.1.1.1., the largest official list not containing traditional and variant forms is 通用规范汉字表 *tōngyòng guīfànhànzì biǎo* (TYGFZB) 'Common Standard Chinese Characters Table', which was published in 2009 and contains 8,300 characters. A set serving as a comparative background is not easy to find, since Taiwanese sets of traditional characters are either much smaller or considerably larger. Section 2.1.1.2. lists two sets of interest: one containing 4,808 characters (常用國字標準字體表 *chángyòng guózì biāozhǔn zìtǐ biǎo* 'List of Standard Forms of Frequently Used Characters'), and another containing 6,341 characters (次常用國字標準字體表 *cìchángyòng guózì biāozhǔn zìtǐ biǎo* 'List of Standard Forms of Less Frequently Used Characters'). The former is too small to compare with TYGFZB, and the latter is not comparable in content. A solution adopted here, albeit not a perfect one, is to treat both Taiwanese sets jointly as a comparative background for TYGFZB. This imperfect solution results in a set that is still considerably larger than TYGFZB. The joint sets will be abbreviated here as TCYZB (台灣常用字表 *Táiwān chángyòng zìbiǎo* 'Taiwanese List of Frequently Used Characters').

7.2.4.1 Immediate components

7.2.4.1.1. General properties

The results of the analysis of both sets in terms of immediate components are provided in Tab. 7.8, Tab. 7.9., and 7.10.

Tab. 7.8 General graphotactic properties of TYGFZB and TCYZB (immediate components)

	TYGFZB	TCYZB
Number of graphotactemes:	8,300	11,146[330]
Number of tactographemes:	8,236	11,064
Average tactographemic efficiency:	1.01	1.01
Number of graphemes (immediate components):	1,897	2,266
Average graphemicity:	2.01	2.02

Tab. 7.9 General quantitative properties of TYGFZB immediate tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	175	187	1.07
2	7,808	7,860	1.01
3	236	236	1.00
4	17	17	1.00
Total:	**8,236**	**8,300**	

[330] The number should be 11,149, but for some reason only 11,146 were recognized by the computer system.

Tab. 7.10 General quantitative properties of TCYZB immediate tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	173	184	1.06
2	10,595	10,666	1.01
3	256	256	1.00
4	33	33	1.00
5	6	6	1.00
6	1	1	1.00
Total:	**11,064**	**11,146**	

7.2.4.1.2. Immediate tactographemicity

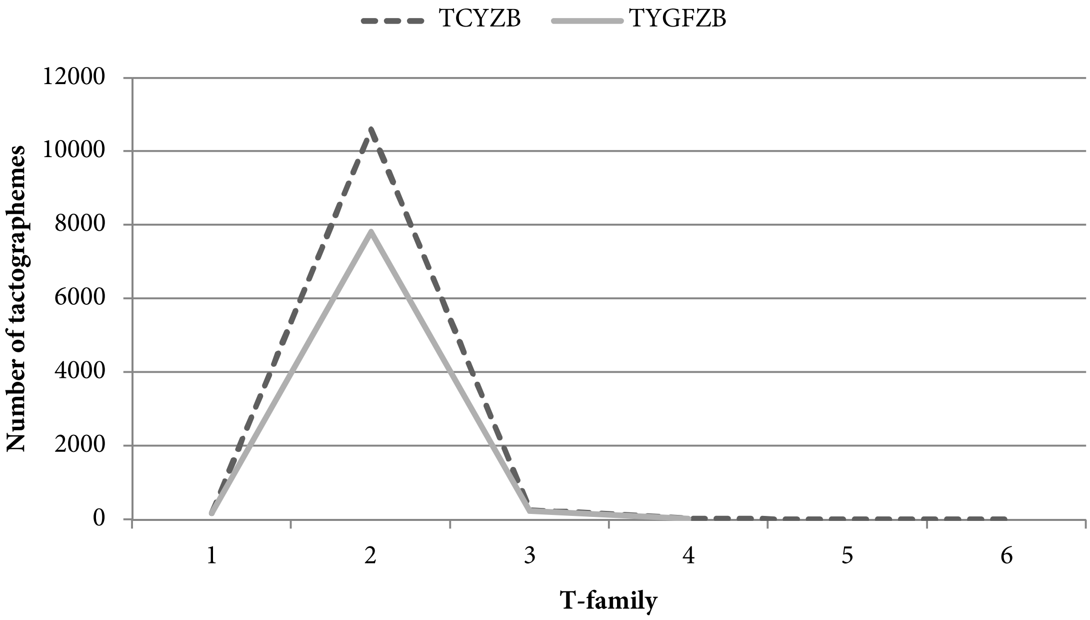


Fig. 7.22 TYGFZB and TCYZB immediate tactographemicity

7.2.4.1.3. Immediate t-graphotactemicity

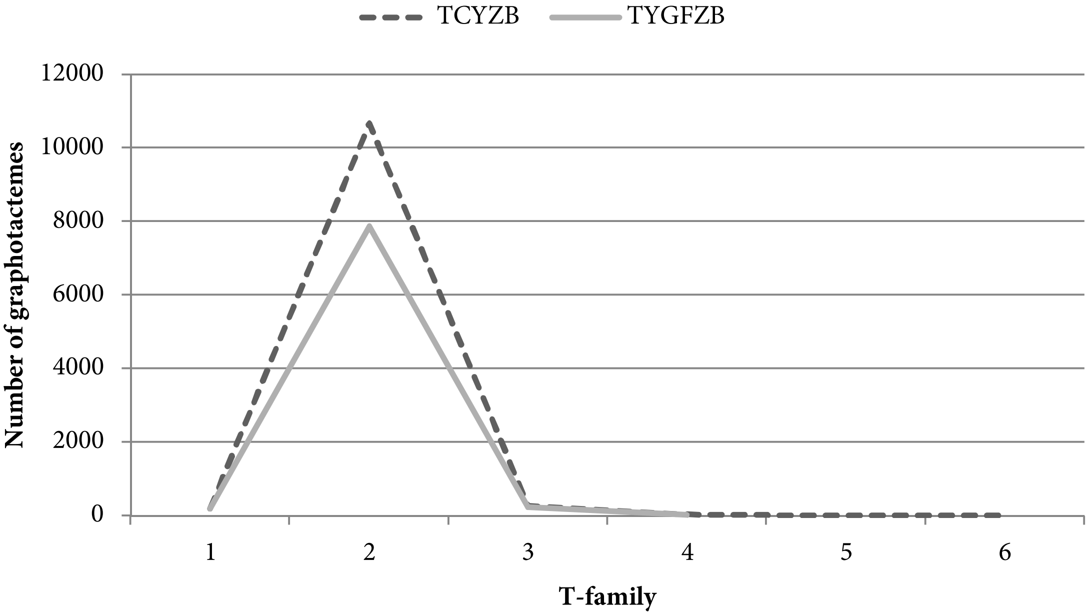


Fig. 7.23 TYGFZB and TCYZB immediate t-graphotactemicity

7.2.4.1.4. Immediate tactogaphemicity and t-graphotactemicity compared

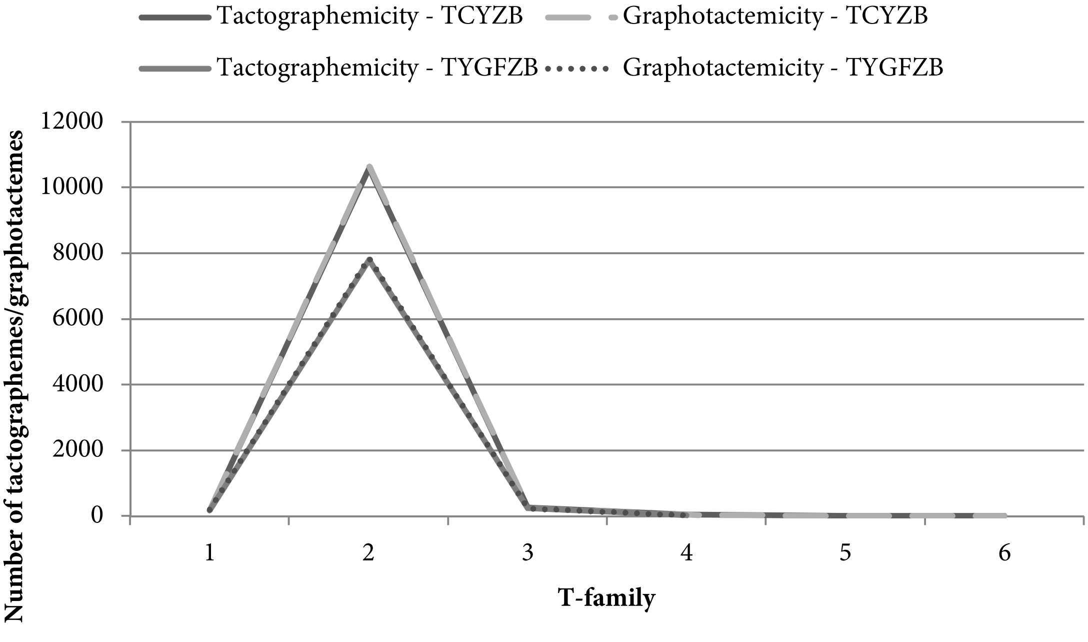


Fig 7.24 TYGFZB and TCYZB immediate tactographemicity and t-graphotactemicity

7.2.4.1.5. Immediate t-efficiency

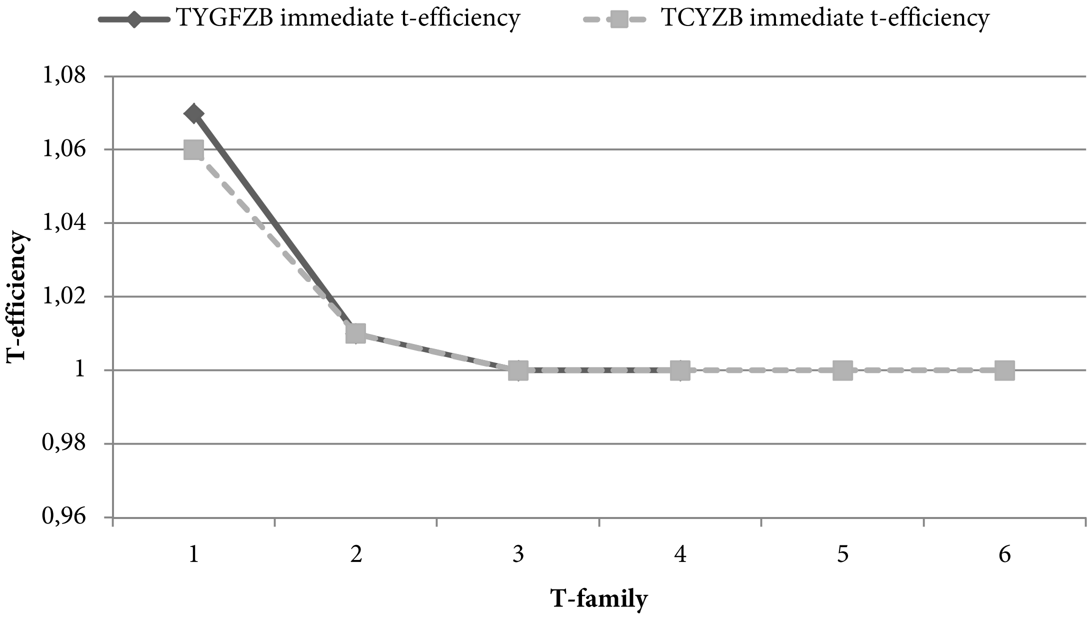


Fig. 7.25 T-efficiency of TYGFZB and TCYZB immediate tactographons

7.2.4.1.6. Immediate tactographemic t-efficiency

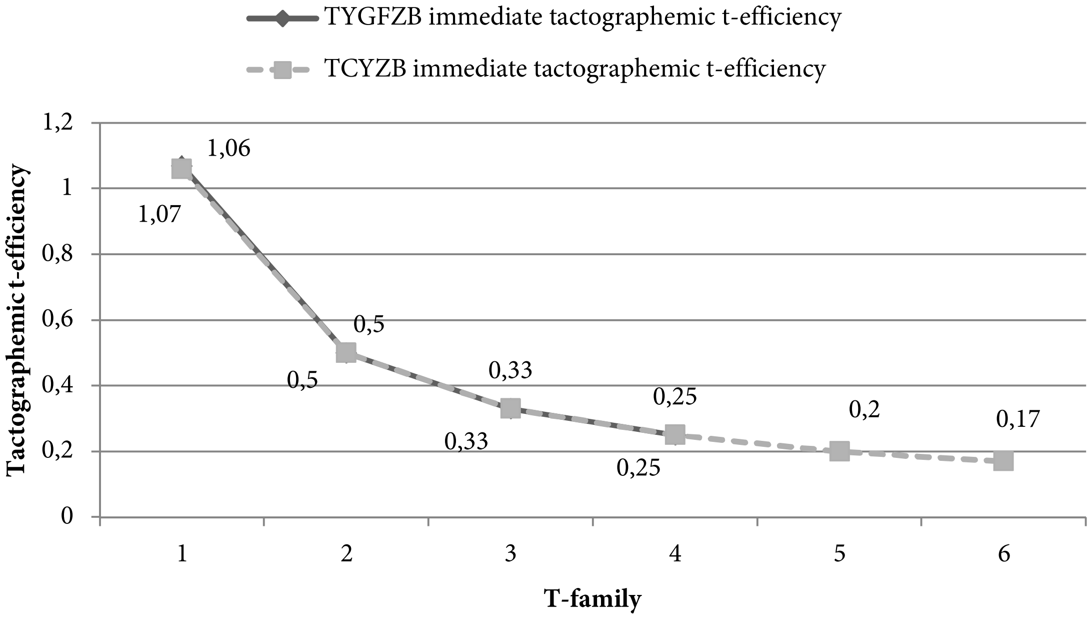


Fig. 7.26 Tactographemic t-efficiency of TYGFZB and TCYZB immediate tactographons

7.2.4.1.7. Immediate categorial graphotactemicity

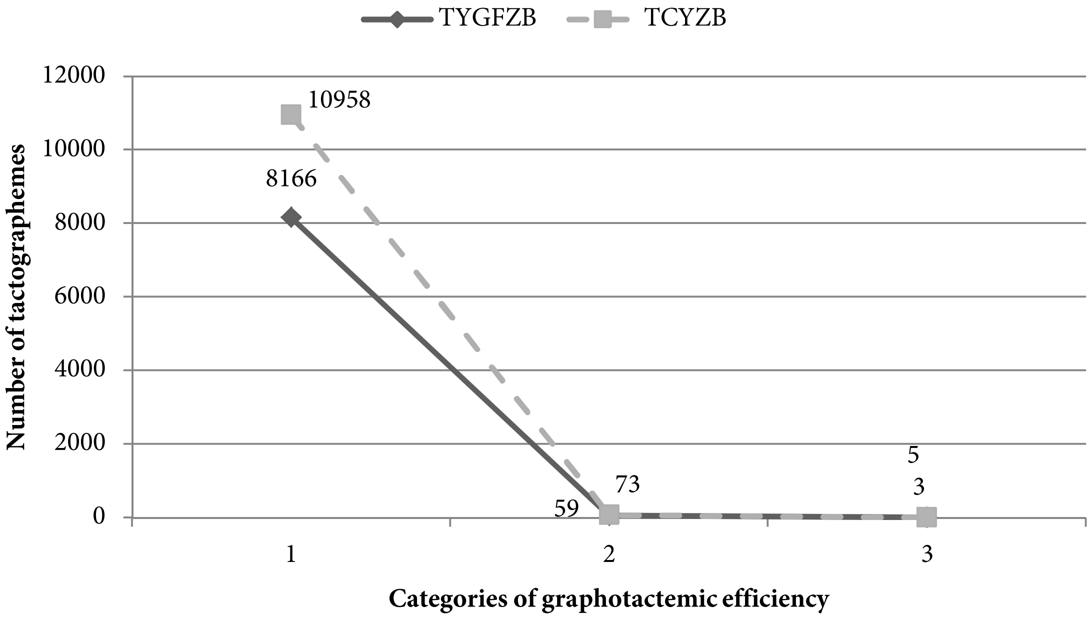


Fig 7.27 Number of tactographemes generating given number of graphotactemes

7.2.4.2. Basic components

7.2.4.2.1. General properties

Results of the graphotactic analysis of TYGFZB and TCYZB in term of basic components are provided in Tab. 7.11, 7.12, and 7.13.

Tab. 7.11 General graphotactic properties of TYGFZB and TCYZB (basic components)

	TYGFZB	TCYZB
Number of graphotactemes:	8,300	11,146
Number of tactographemes:	7,798	10,482
Average tactographemic efficiency:	1.06	1.06
Number of graphemes (basic components)	307	303
Average graphemicity:	4.04	4.50

Tab. 7.12 General quantitative properties of TYGFZB basic tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	161	183	1.14
2	1,034	1,114	1.08
3	1,744	1,911	1.10
4	2,016	2,143	1.06
5	1,591	1,664	1.05
6	814	841	1.03
7	328	333	1.02
8	90	91	1.01
9	16	16	1.00
10	4	4	1.00
Total:	**7,798**	**8,300**	

Tab. 7.13 General quantitative properties of TCYZB basic tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	154	179	1.16
2	1,016	1,109	1.09
3	1,857	2,038	1.10
4	2,417	2,596	1.07
5	2,249	2,359	1.05
6	1,555	1,615	1.04
7	829	843	1.02
8	292	294	1.01
9	76	76	1.00
10	25	25	1.00
11	10	10	1.00
12	1	1	1.00
13	1	1	1.00
Total:	**10,482**	**11,146**	

7.2.4.2.2. Basic tactographemicity

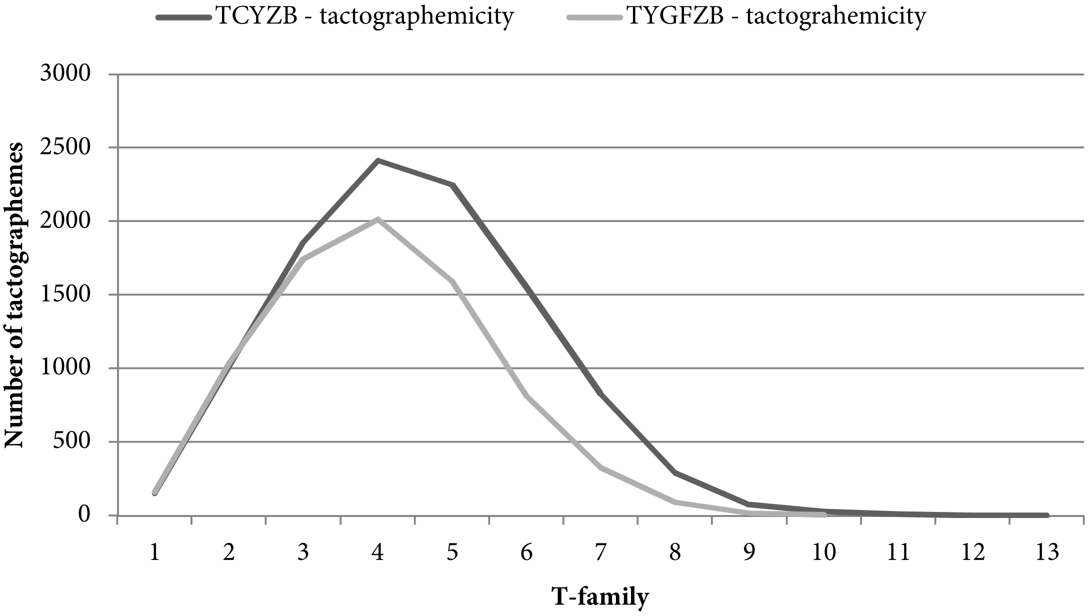


Fig. 7.28 TYGFZB and TCYZB basic tactographemicity

7.2.4.2.3. Basic t-graphotactemicity

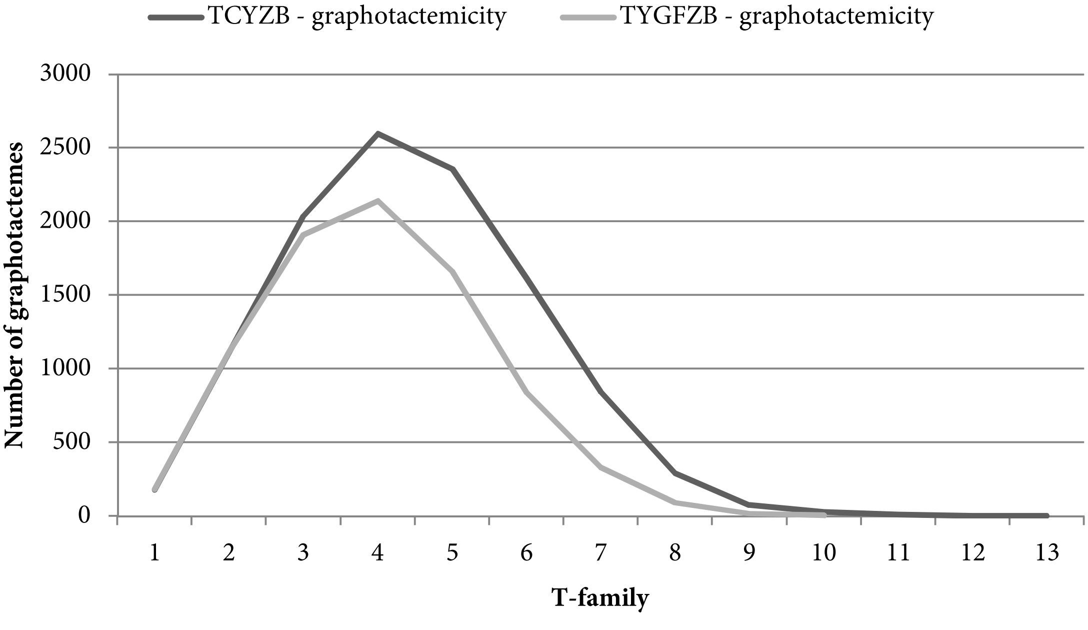


Fig. 7.29 TYGFZB and TCYZB basic t-graphotactemicity

7.2.4.2.4. Basic tactographemicity and t-graphotactemicity compared

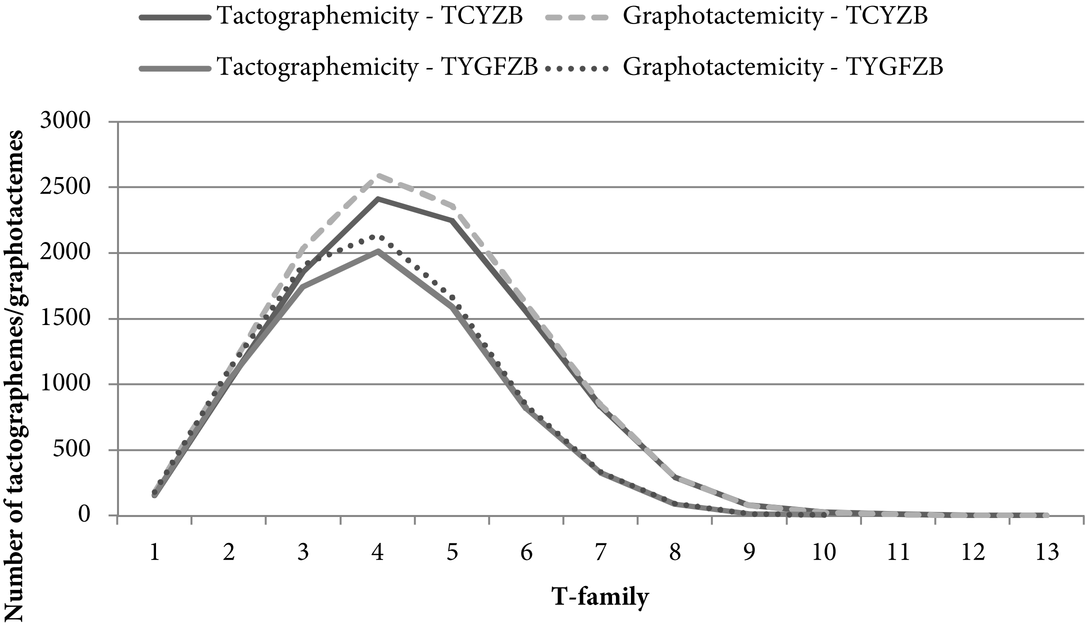


Fig. 7.30 TYGFZB and TCYZB basic tactographemicity and t-graphotactemicity

7.2.4.2.5. Basic t-efficiency

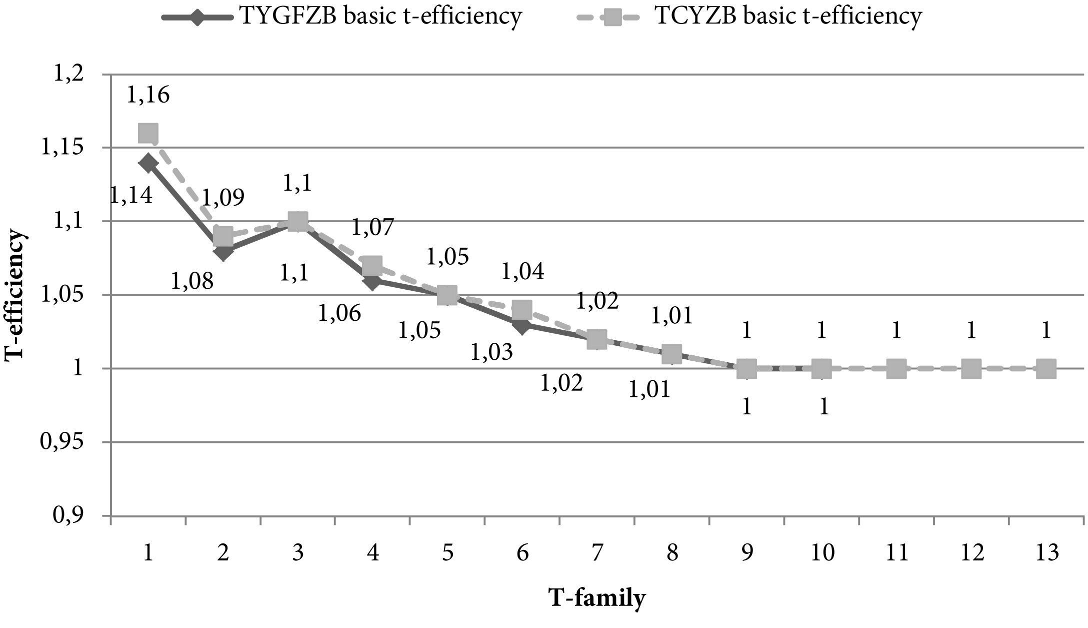


Fig. 7.31 T-efficiency of TYGFZB and TCYZB basic tactographons

7.2.4.2.6. Basic tactographemic t-efficiency

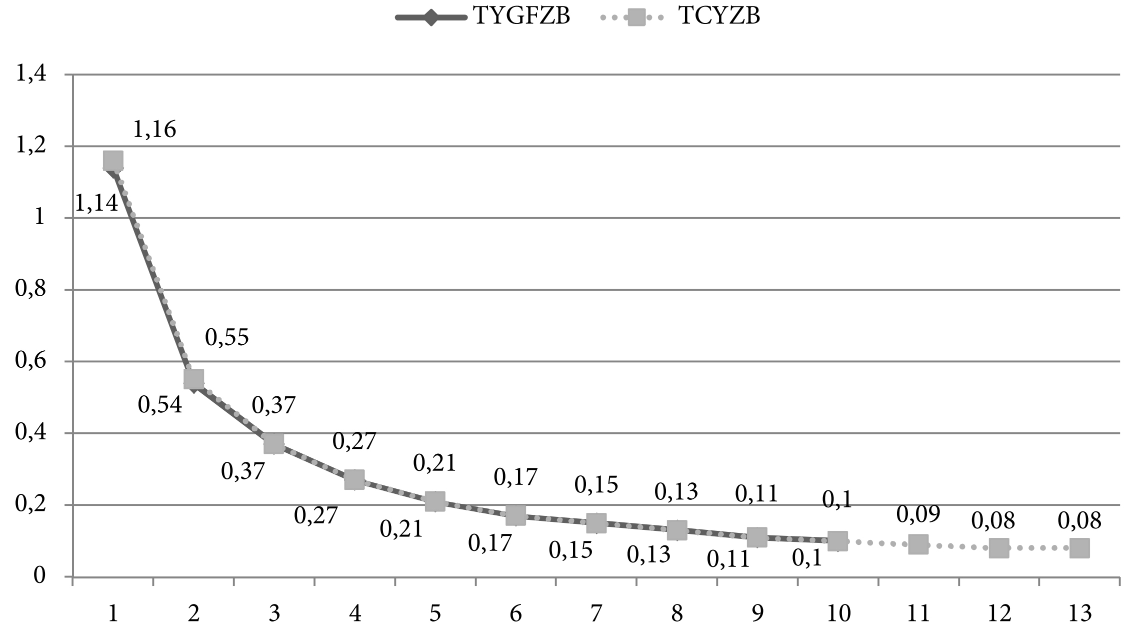


Fig. 7.32 Tactographemic t-efficiency of TYGFZB and TCYZB basic tactographons

7.2.4.2.7. Basic categorial graphotactemic efficiency

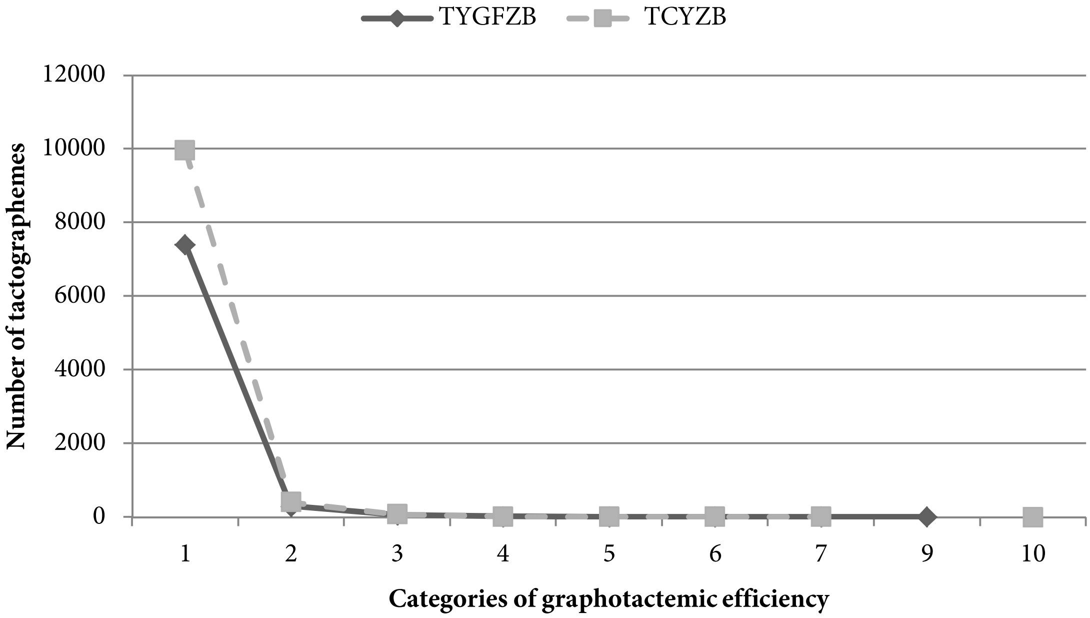


Fig. 7.33 The number of tactographemes generating a given number of graphotactemes

7.2.5. Unihan

The Unihan database[331] is the largest of the character sets investigated in this study. The simple reason for this is that it is also the largest available set with corresponding componential descriptions of contained characters. The size and contents of Unihan are a source of a few serious analytical problems. The most basic of them was already mentioned before – the contained characters are very heterogeneous, i.e. they come from diverse sources (China and Taiwan, simplified and traditional Chinese, Japan, Korea, Vietnam) and have a very diverse status (frequently and rarely used basic forms; frequently and rarely used variant forms; abandoned, obsolete and historical forms[332]). This problem can be quite easily sorted by extracting the homogenous subsets of CJK Unified Characters, based on numerous local NCSes and CCSes[333] – which is exactly what was done in the previous sections of this chapter.

It could be argued that investigating a corpus so diverse is comparable to analyzing all words, including all known historical forms. It might be analogous to doing the same with Romance languages and trying to draw viable conclusions. This argument would not be completely without sense, but due to the inaccuracies in the analogy, this kind of argument is for the most part easily refuted. Structural and compositional properties of characters (as described in Chapter 4) are the same, regardless of the source of the characters. It is absolutely viable to investigate the entire Unihan – from a statistical perspective, neither source of origin, nor the pragmatic status of the characters have a significant influence on their structural and compositional properties. Also historical factors play no role at all – Unihan contains only characters in modern regular script. The fact that a character's history dates back nearly 2,000 years does not correlate with structural differences in archaic and modern forms. The graphotactic analysis of the CJK Unified Ideographs set renders viable results, but their interpretation should be deliberate and careful.

The raw format of KDP requires some adjustments; those concerning the investigation of all subsets of KDP were discussed at the beginning of this chapter. The adjustments concerning only the analysis of KDP will be briefly reviewed here.[334] From

[331] It should be remembered that the analysis in this chapter is conducted not directly on the CJK Unified Ideographs set, but on the IDS descriptions in the Kanji Database Project maintained by T. Kawabata (see Section 5.1.1.2.).

[332] The distinction between 'abandoned', 'obsolete' and 'historical' is not formal and serves only as an illustration of the diversification of character status. These descriptions are used in the literature in this context, typically without attention to clarifying the exact meaning. Some discussion may be found in Sections 4.4.1.1. and 4.4.1.2.

[333] See Chapter 2.

[334] See Section 5.1.1.2.

a technical point of view the entries in KDP reflect the structure of CJK Unified Ideographs. There are, however, entries which facilitate decomposition, and should not be analyzed as graphotactemes. This pertains to 695 CDP components that were eliminated from the graphotactic analysis of KDP.[335] Also certain entries in CJK Unified Ideographs were eliminated – this includes compatibility ideographs and supplement radicals. The modified KDP contains 74,810 graphotactemic entries (characters).

7.2.5.1. Immediate components

7.2.5.1.1. General properties

General results of the analysis are provided below and in Tab. 7.14:

- number of graphotactemes: 74,810;
- number of tactographemes: 71,588;
- number of graphemes (immediate components): 8,673;
- average tactographemic efficiency: 1.05;
- average graphemicity: 2.16.

Tab. 7.14 General quantitative properties of Unihan immediate tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	621	800	1.29
2	61,717	64,708	1.05
3	6,812	6,857	1.01
4	1,909	1,915	1.00
5	435	436	1.00
6	79	79	1.00
7	15	15	1.00
Total:	**71,588**	**74,810**	

[335] CDP components were included in compiling the inventory of immediate components and the recursive extraction of the inventory of basic components, but were not analyzed as graphotactemes (characters).

7.2.5.1.2. Unihan immediate tactographemicity

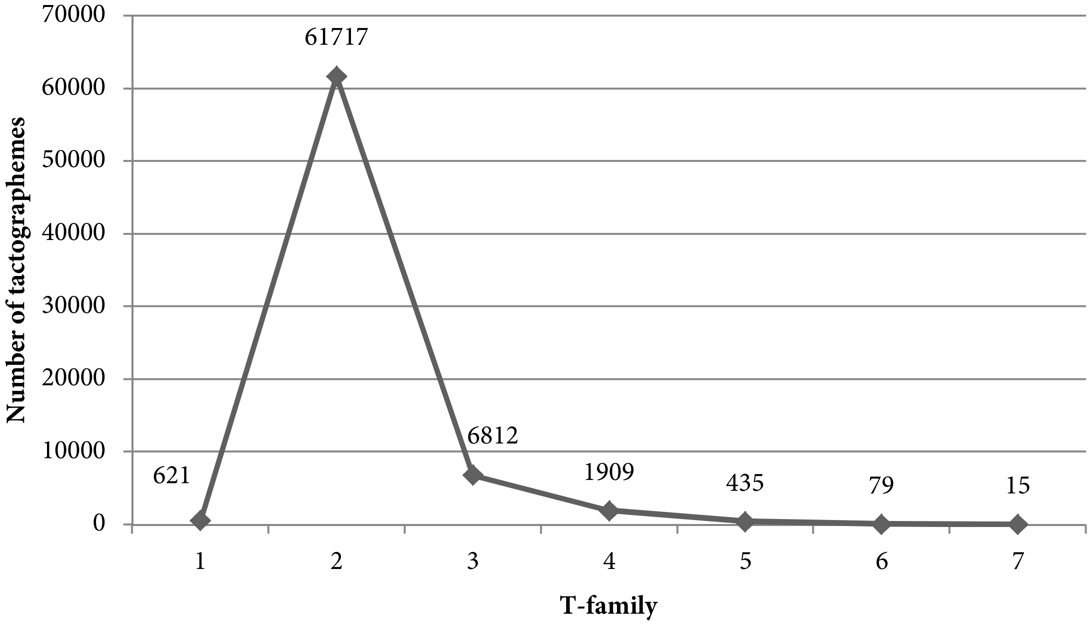


Fig. 7.34 Unihan immediate tactographemicity

7.1.5.1.3. Unihan immediate t-graphotactemicity

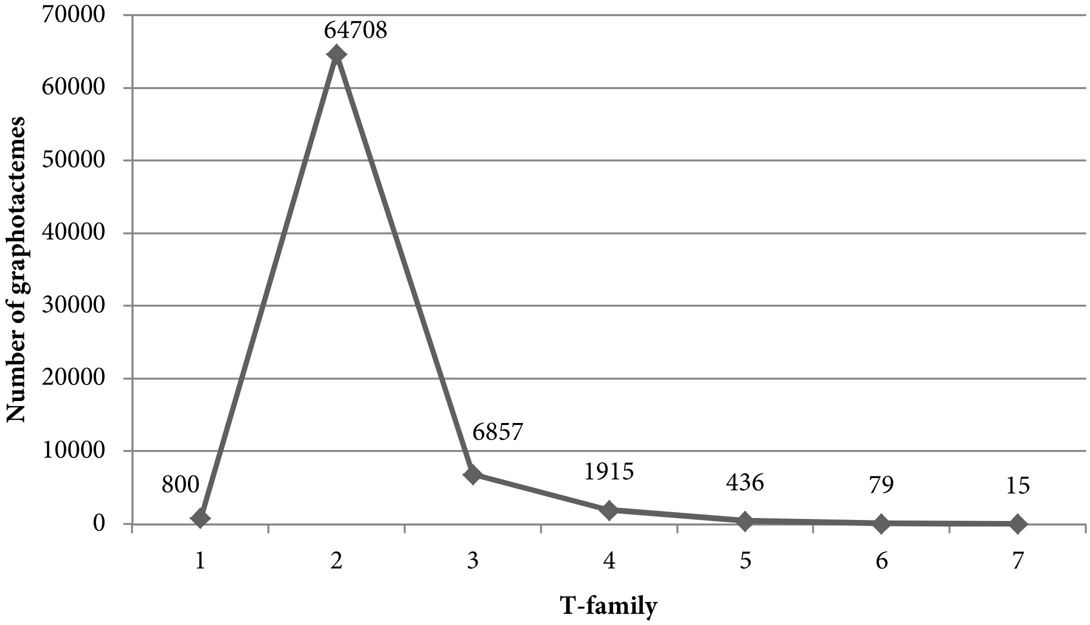


Fig. 7.35 Unihan immediate t-graphotactemicity

7.1.5.1.4. Unihan immediate tactographemicity and t-graphotactemicity compared

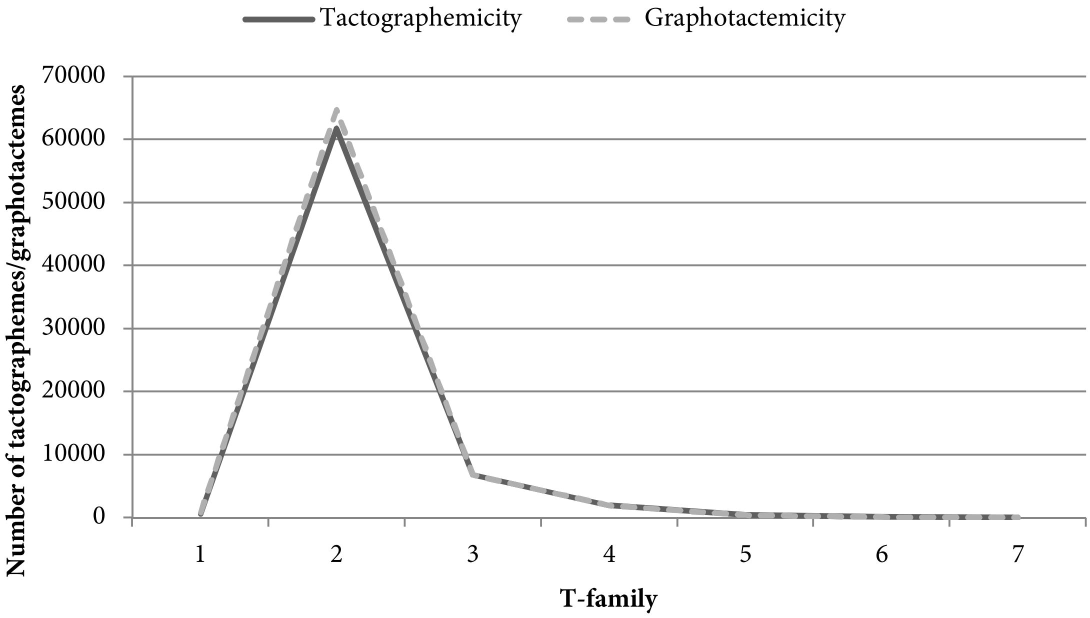


Fig. 7.36 Unihan immediate tactographemicity and t-graphotactemicity

7.1.5.1.5. Unihan immediate t-efficiency

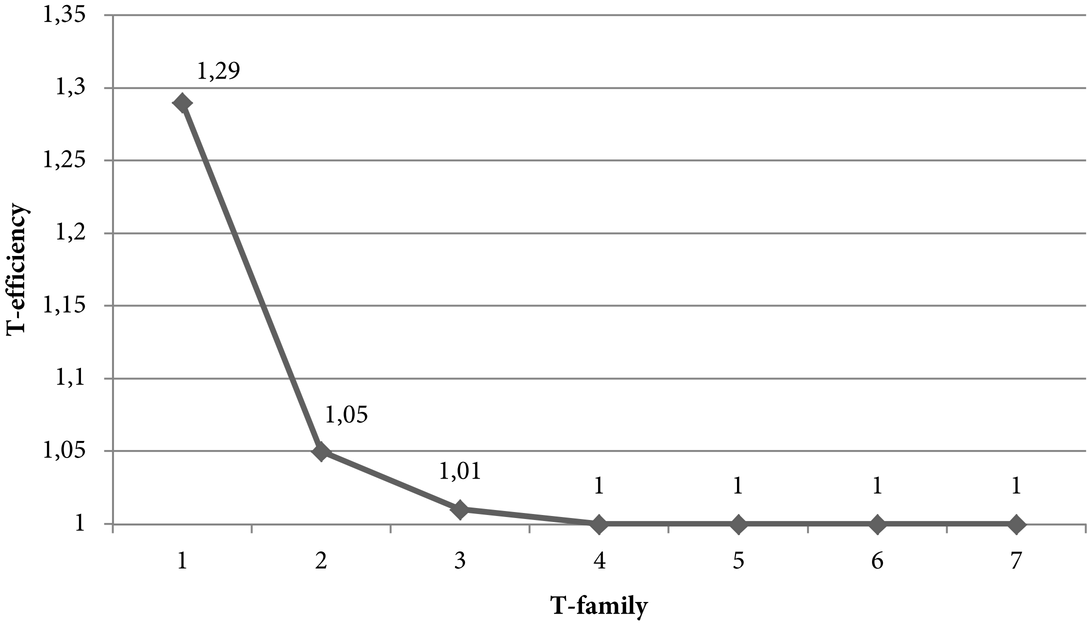


Fig. 7.37 T-efficiency of Unihan immediate tactographons

7.1.5.1.6. Unihan immediate tactographemic t-efficiency

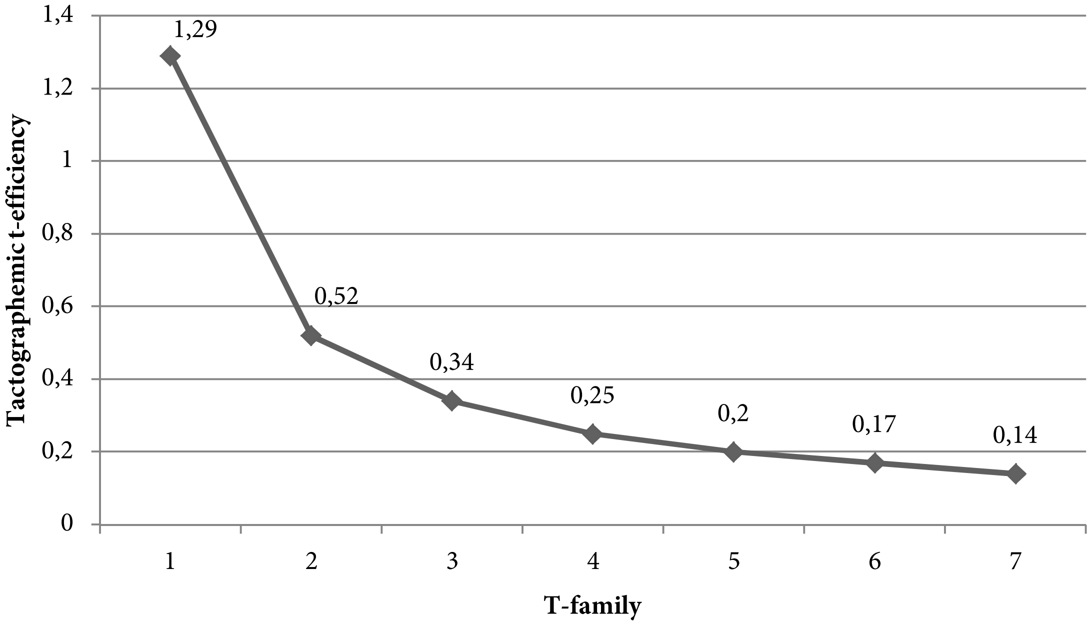


Fig. 7.38 Tactographemic t-efficiency of Unihan immediate tactographons

7.1.5.1.7. Unihan immediate categorial graphotactemicity

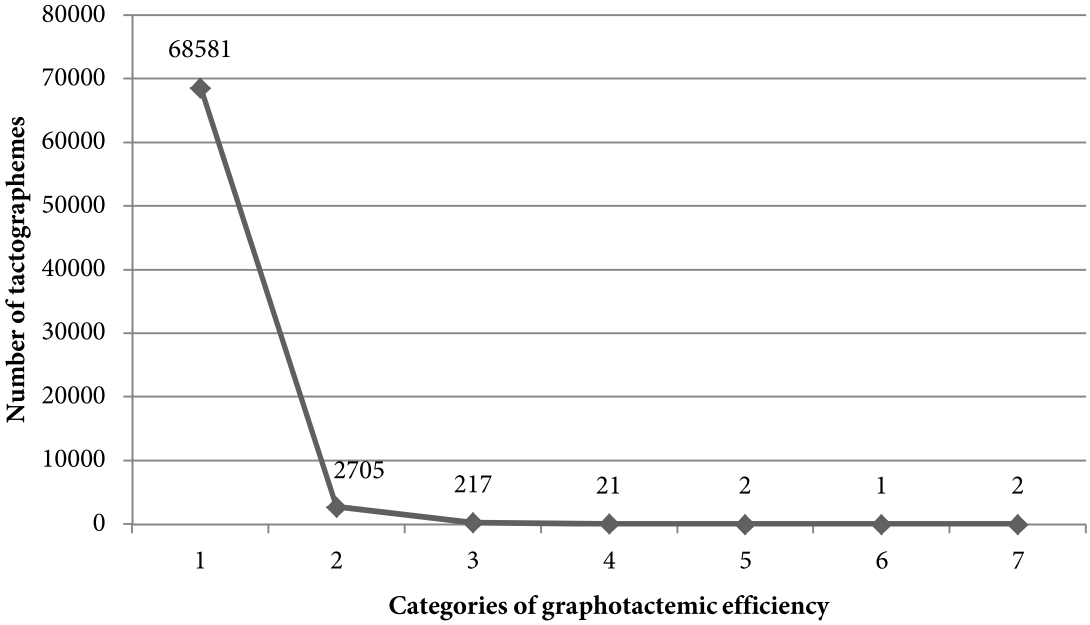


Fig. 7.39 The number of tactographemes generating a given number of graphotactemes

7.2.5.2. Unihan basic components

7.2.5.2.1. General properties

General results of the analysis are provided below and in Tab. 7.15:

- number of graphotactemes: 74,810;
- number of tactographemes: 63,600;
- number of graphemes (basic components): 593;
- average tactographemic efficiency: 1.18;
- average graphemicity: 4.82.

Tab. 7.15 General quantitative properties of Unihan basic tactographons

Tactographon (T-family)	Tactographemicity	T-graphotactemicity	T-efficiency
1	394	622	1.58
2	4,332	5,556	1.28
3	9908	12,323	1.24
4	13,846	16,819	1.21
5	14,178	16,452	1.16
6	10,728	11,987	1.12
7	6,085	6,640	1.09
8	2,706	2,899	1.07
9	1,005	1,070	1.06
10	297	312	1.05
11	90	95	1.06
12	23	25	1.09
13	5	7	1.40
14	2	2	1.00
16	1	1	1.00
Total:	**63,600**	**74,810**	

7.2.5.2.2. Unihan basic tactographemicity

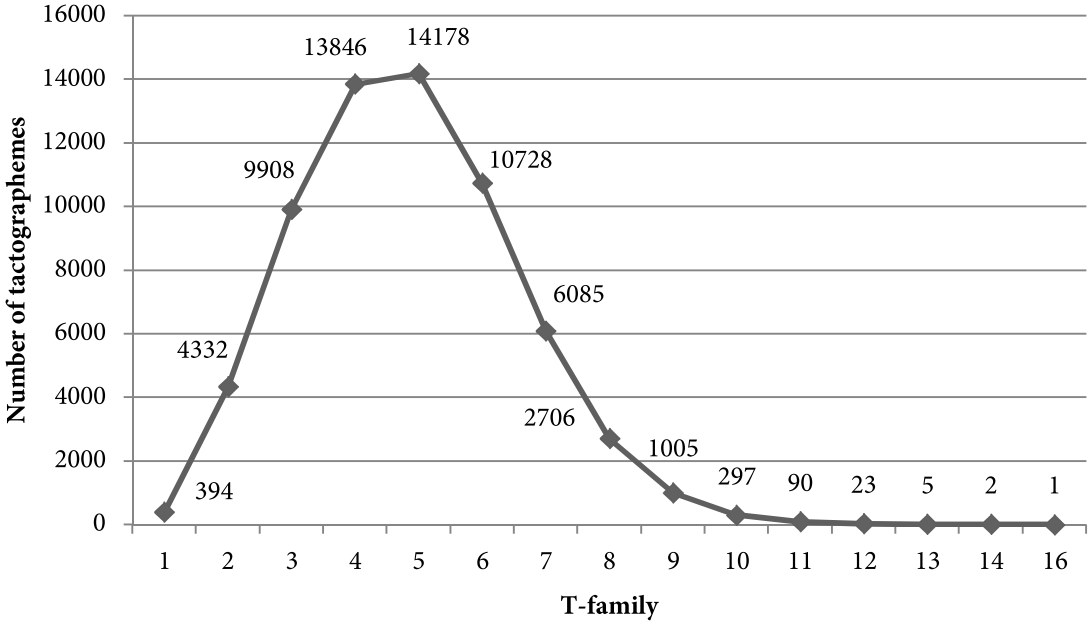


Fig. 7.40 Unihan basic tactographemicity

7.2.5.2.3. Unihan basic t-graphotactemicity

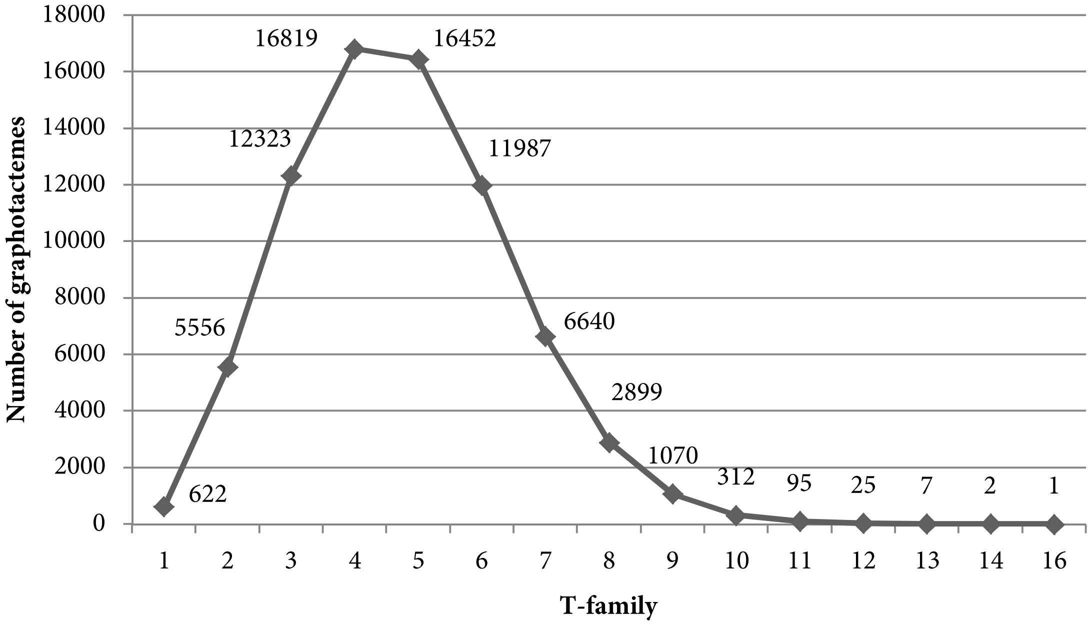


Fig. 7.41 Unihan basic t-graphotactemicity

7.2.5.2.4. Unihan basic tactographemicity and t-graphotactemicity compared

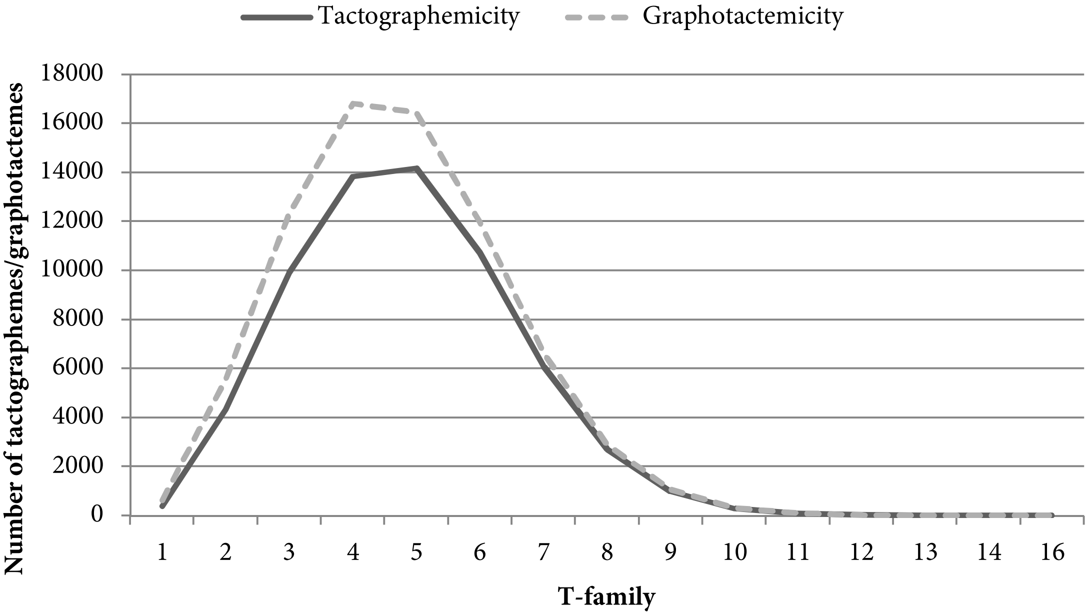


Fig. 7.42 Unihan basic tactographemicity and t-graphotactemicity

7.2.5.2.5. Unihan basic t-efficiency

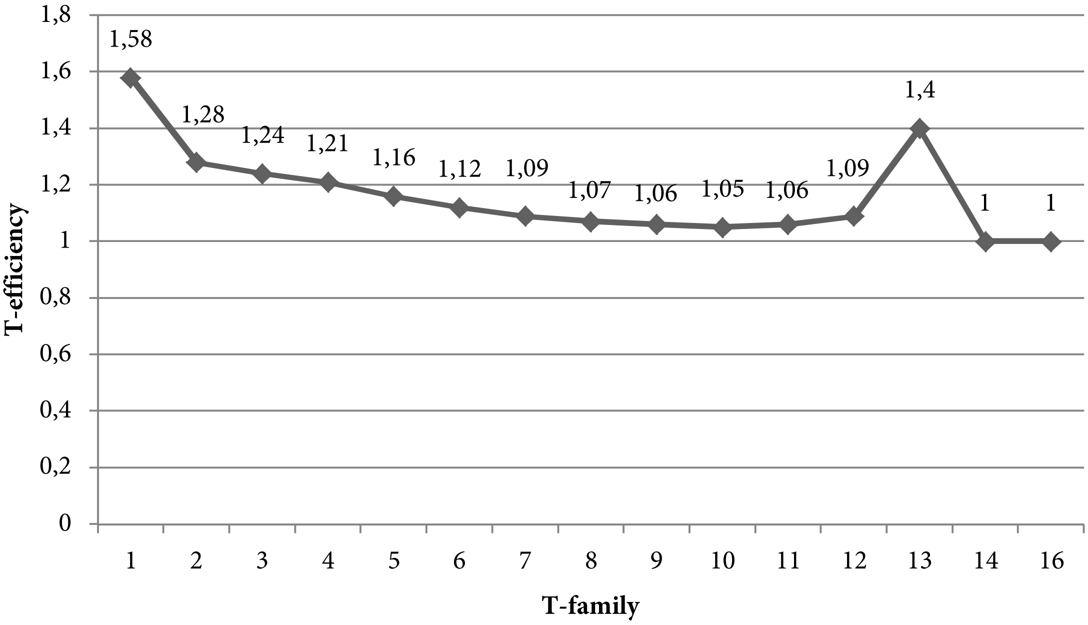


Fig. 7.43 T-efficiency of Unihan basic tactographons

7.2.5.2.6. Unihan basic tactographemic t-efficiency

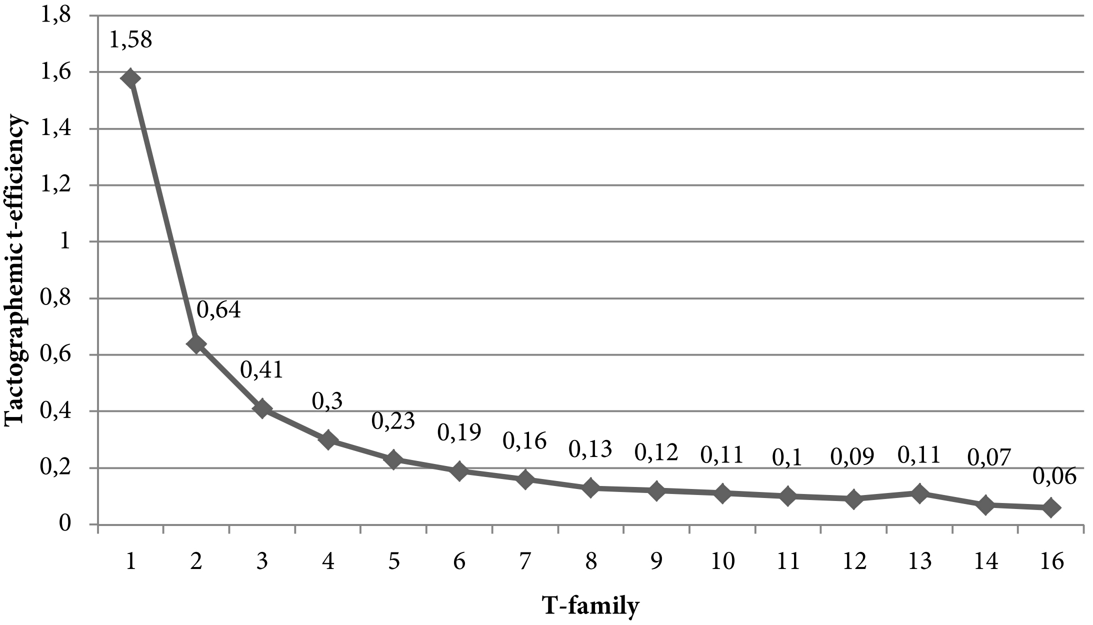


Fig. 7.44 Tactographemic t-efficiency of Unihan basic tactographons

7.2.5.2.7. Unihan basic categorial graphotactemic efficiency

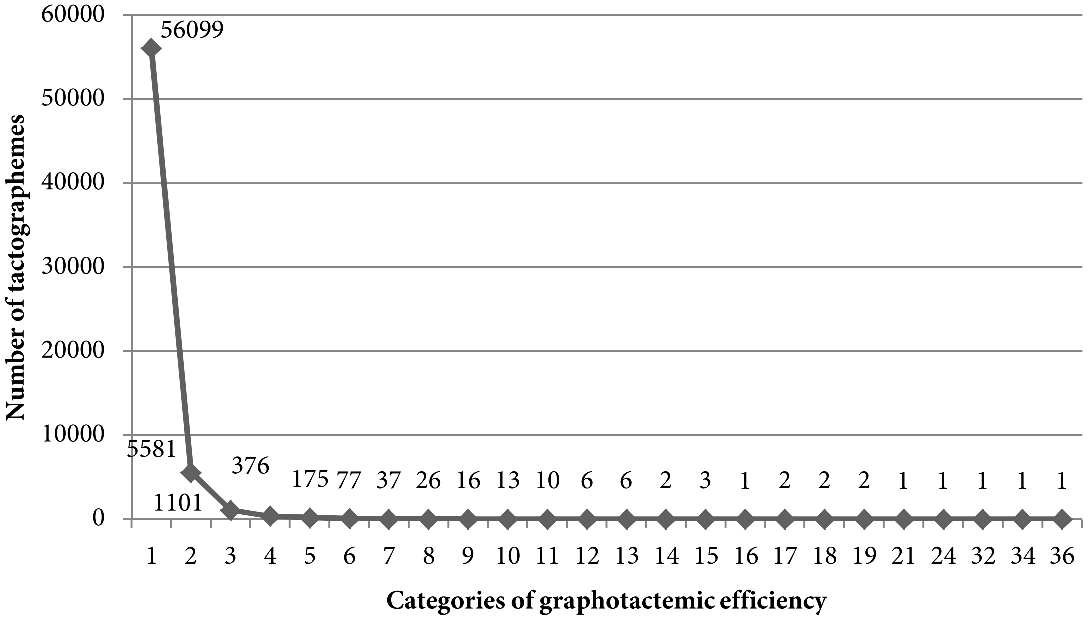


Fig. 7.45 The number of tactographemes generating a given number of graphotactemes

7.2.6. Summary

The formulation of complete and final conclusions about the portion of graphotactic data presented in the above sections is difficult and probably impossible without further investigations. Some of the results present a rather straightforward picture, but in other cases there is no simple explanation. The following summary concerns the results of graphotactic investigation of the four sets presented so far.

7.2.6.1. Tactographemicity and t-graphotactemicity

The gathered data pertaining to tactographemicity and t-graphotactemicity of tactographemes in an investigated set allows a few conclusions to be drawn. This part of the analysis is comparable with the results obtained by Bańczerowski and Wierzchoń for Polish and Chinese orthographic systems[336] that were presented in Bańczerowski (2009). The results are shown in Fig. 7.46 and Fig. 7.47.[337]

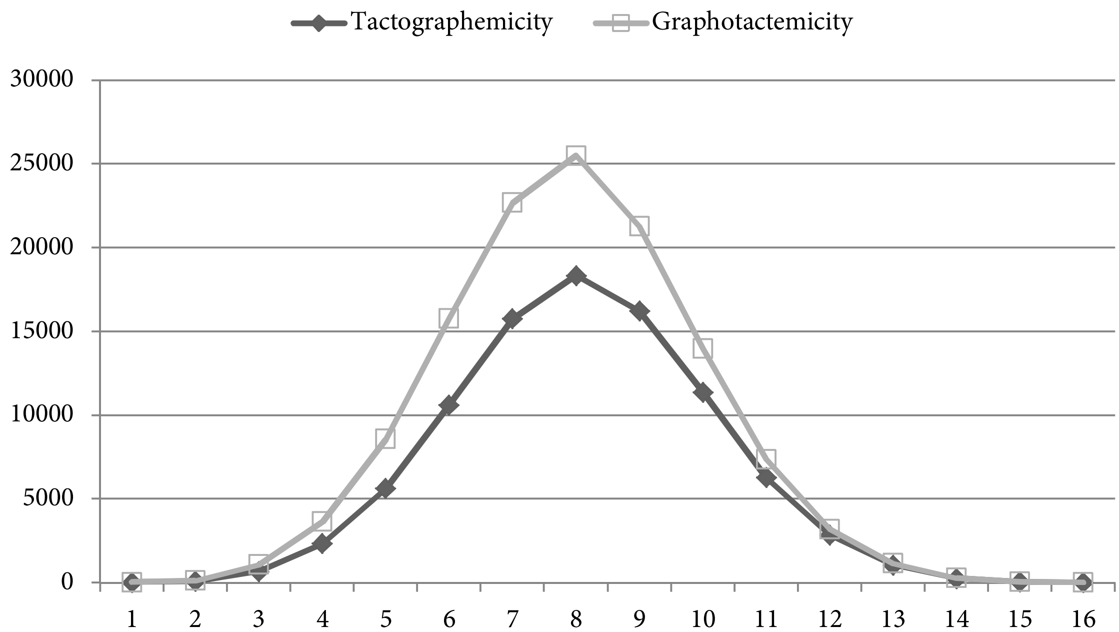


Fig. 7.46 Tactoorthemicity and t-orthotactemicity of Polish tactoorthonomes[338]

[336] Bańczerowski 2009, also see section 3.1.1.

[337] To avoid terminological confusion the terms for the units involved were changed to those relating to the orthotactical level.

[338] Bańczerowski 2009: 18.

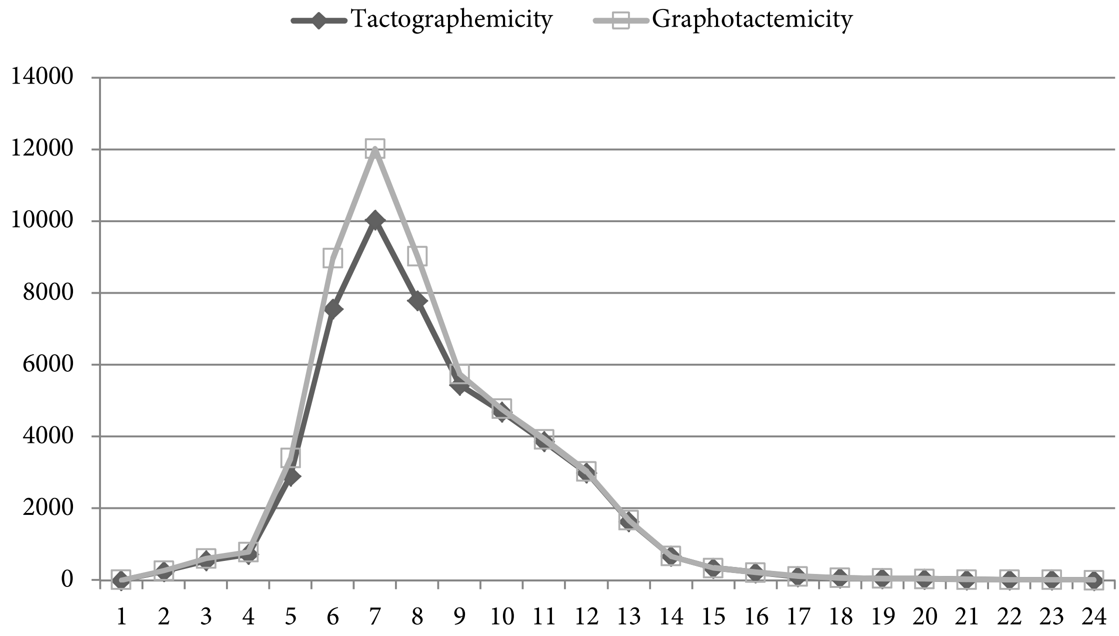


Fig. 7.47 Tactoorthemicity and t-orthotactemicity of Chinese *pīnyīn* tactoorthonomes[339]

Both diagrams show a close correlation between the properties.[340]

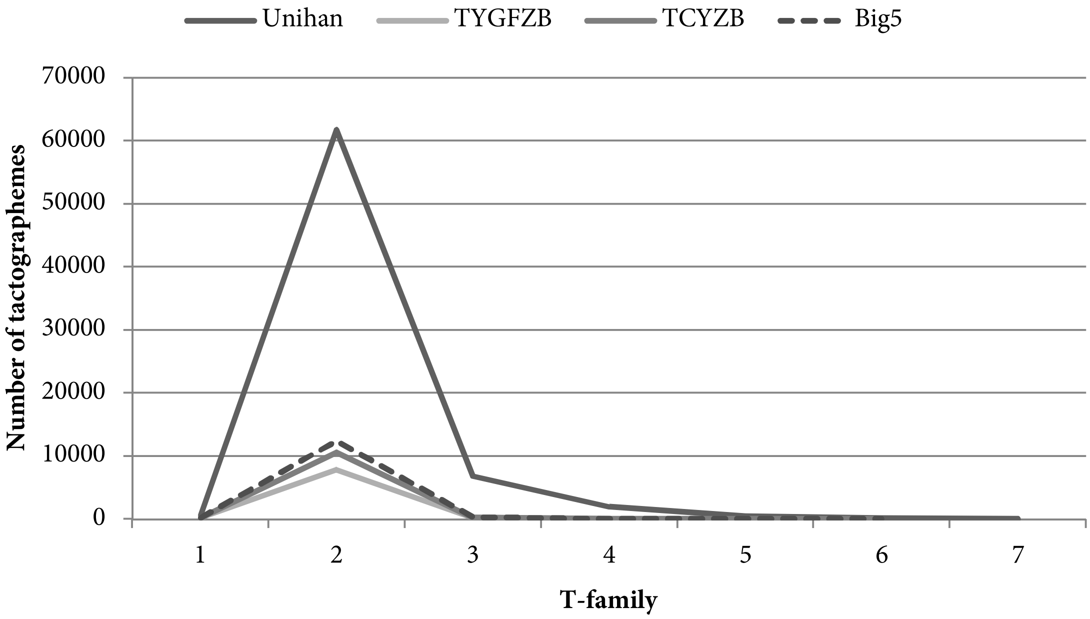


Fig. 7.48 Tactographemicity of immediate tactographons

[339] Ibid., 19.

[340] Tactoorphmicity and t-orthotactemicity correspond to tactogaphemicity and t-graphotactemicity in graphotactics.

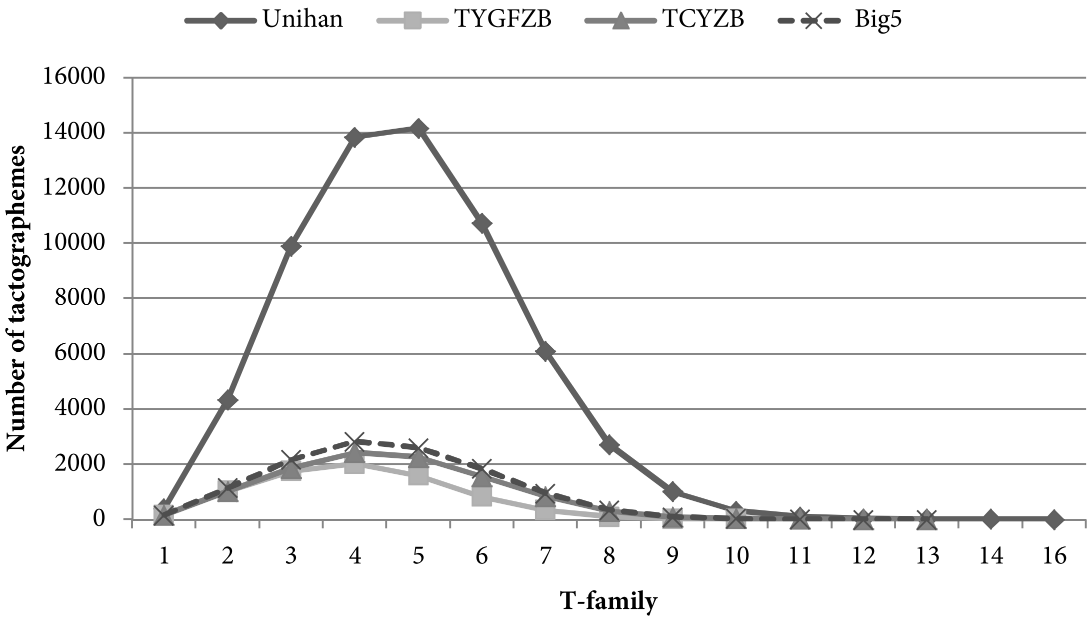


Fig. 7.49 Tactograhemicity of basic tactograpons

The situation with Chinese graphemes is more complicated, since there are two types of them, rendering two different levels of analysis. Fig. 7.48 and Fig. 7.49 collate the representations of immediate and basic tactographemic properties for all four sets.

It can be observed that in both cases the shape of the tactographemic curves is very similar. The differences can be explained by the different sizes of the sets.

Fig. 7.50 and Fig. 7.51 collate the curves which represent the immediate and basic t-graphotactemic properties for all four sets.

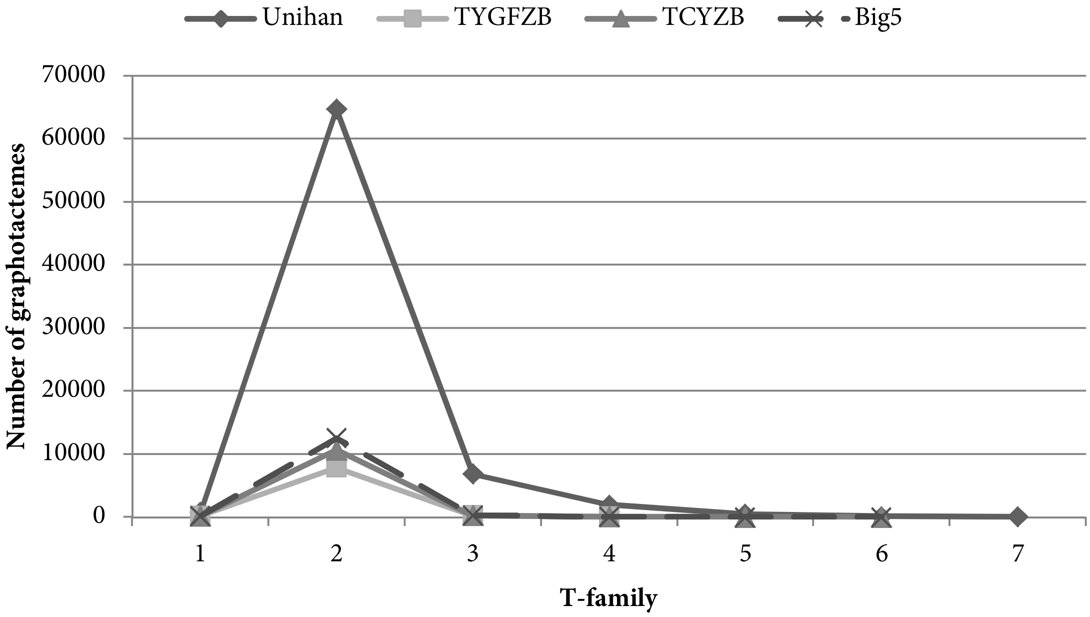


Fig. 7.50 Immediate t-graphotactemicity

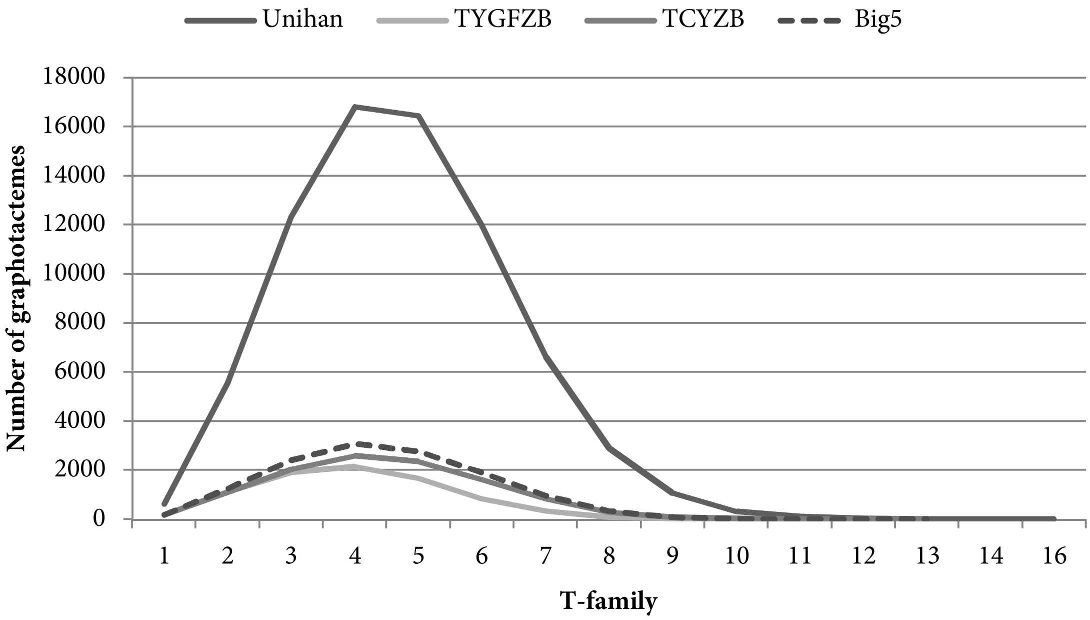


Fig. 7.51 Basic t-graphotactemicity

At this point it may be observed that the curves reflecting the same type of components and the same type of graphotactic properties are very similar. Further investigation reveals that the quantitative similarities with regard to the discussed graphotactic properties go even further. Collation of graphical representations of tactographemic and t-graphotactemic properties in all sets performed separately for the immediate and basic levels of analysis, shows that the shape of the curves is independent of the type of counted graphotactic units (tactographemes or graphotactemes). This is illustrated in Fig. 7.52 and Fig. 7.53.

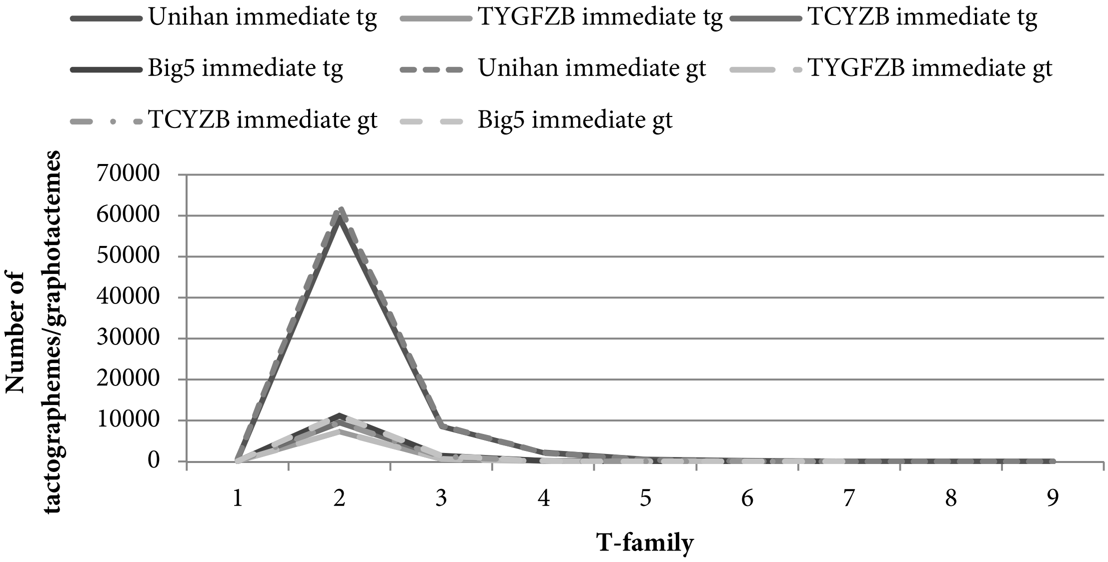


Fig. 7.52 Immediate tactographemicity and t-graphotactemicity

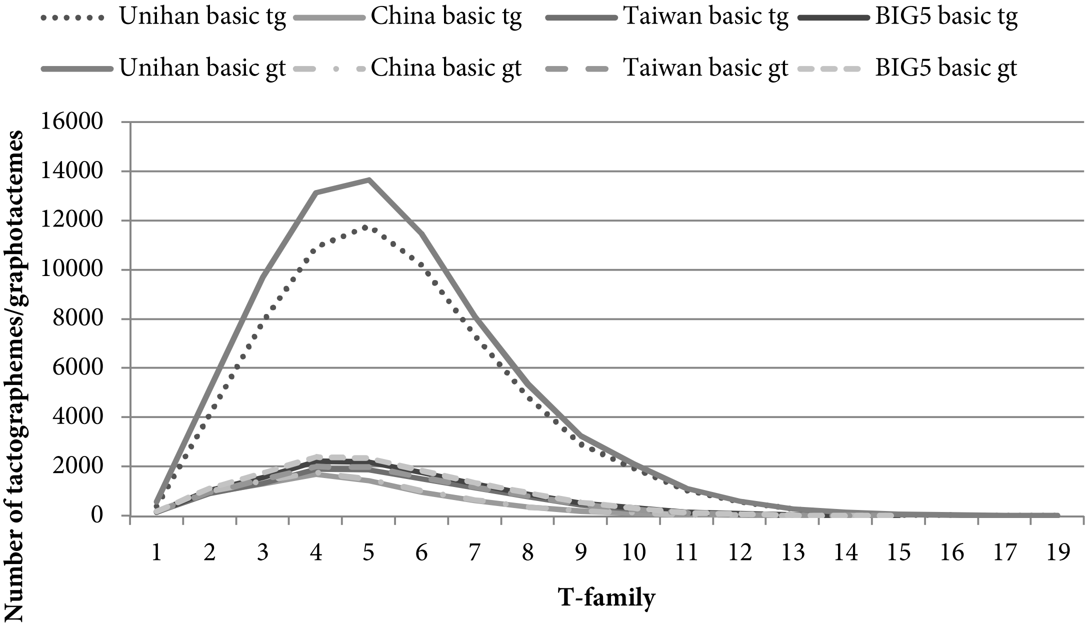


Fig. 7.53 Basic tactographemicity and t-graphotactemicity

The shape of the curves representing the quantitative relations between the investigated units, i.e. tactographemes and graphotactemes, correlated with the family of tactographons, and are found to be independent of the type of units (tactographemes or graphotactemes), as well as independent of the size and type of sets. Nonetheless, the shape of the curves are sensitive to the level of analysis (type of graphemes).

The graphotactic evidence is supported by the results of Bańczerowski's (2009) analysis. His pioneering investigation of Polish orthography and Chinese *pīnyīn* transliteration mentioned on several occasions in this work revealed exactly the same correlation between the tactoorthemicity and t-orthotactemicity, as Fig. 7.54 shows.[341]

Given that the investigation of Cangjie codes showed the same type of relationship, it can be hypothesized that the correlation between the number of tactographemes and graphotactemes in the tactographonic family (t-family) is a universal feature.

The same procedure of collating the graphical representations of quantitative properties for all sets separately for the two levels of analysis will be applied to the remaining results pertaining to t-efficiency, tactographemic t-efficiency and categorial graphotactemic efficiency.

[341] Based on Bańczerowski 2009: 18-19.

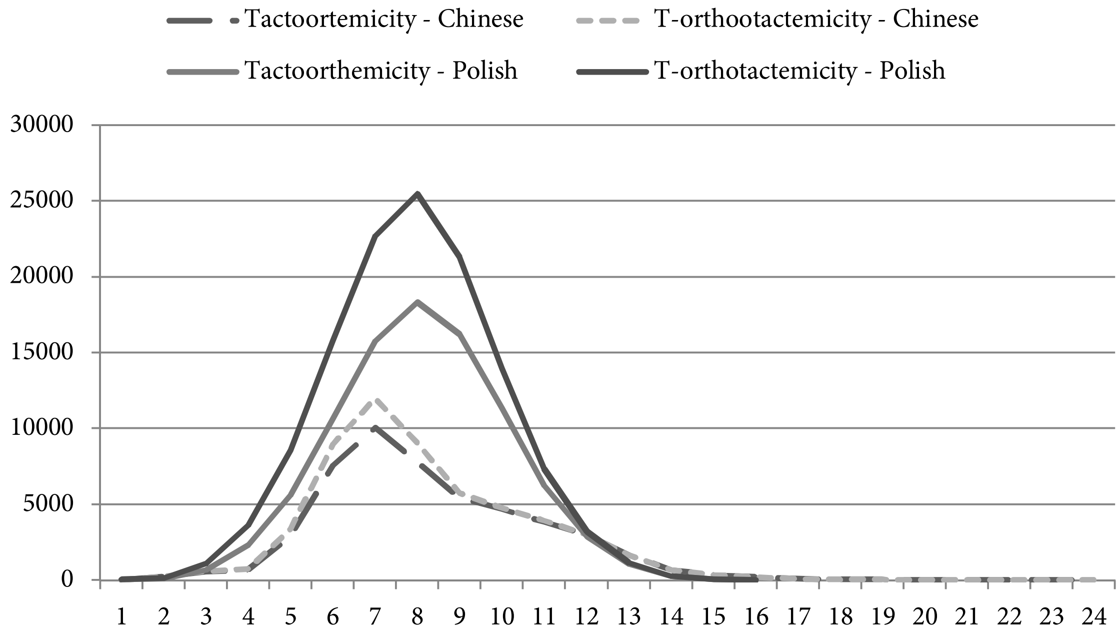


Fig. 7.54 Polish and Chinese tactoorthemicity and t-orthotactemicity

7.2.6.2. T-efficiency

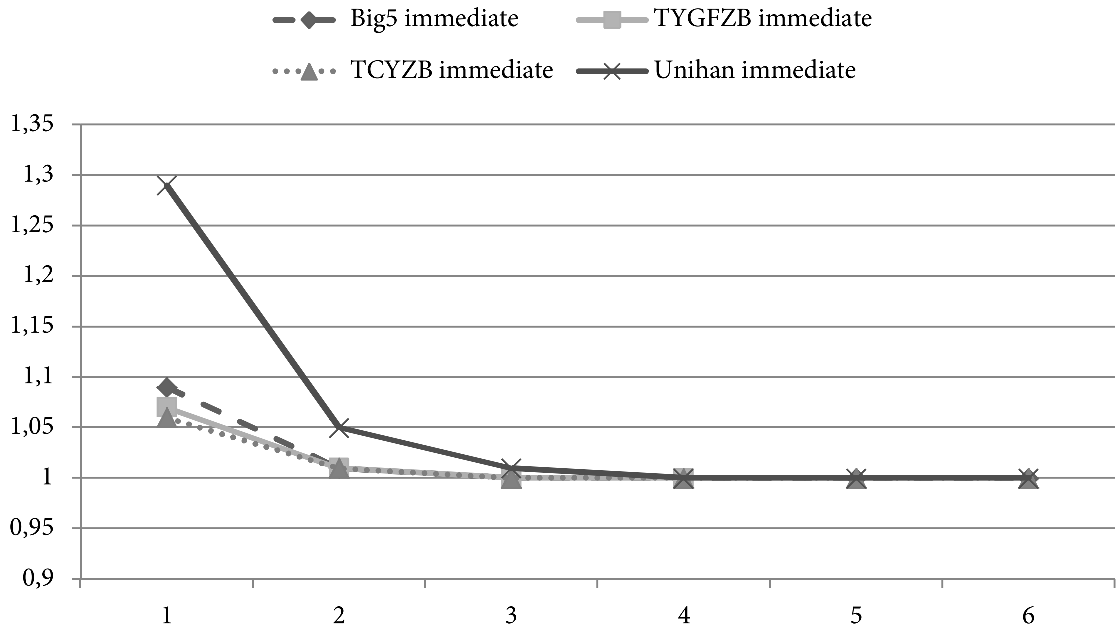


Fig. 7.55 Immediate t-efficiency

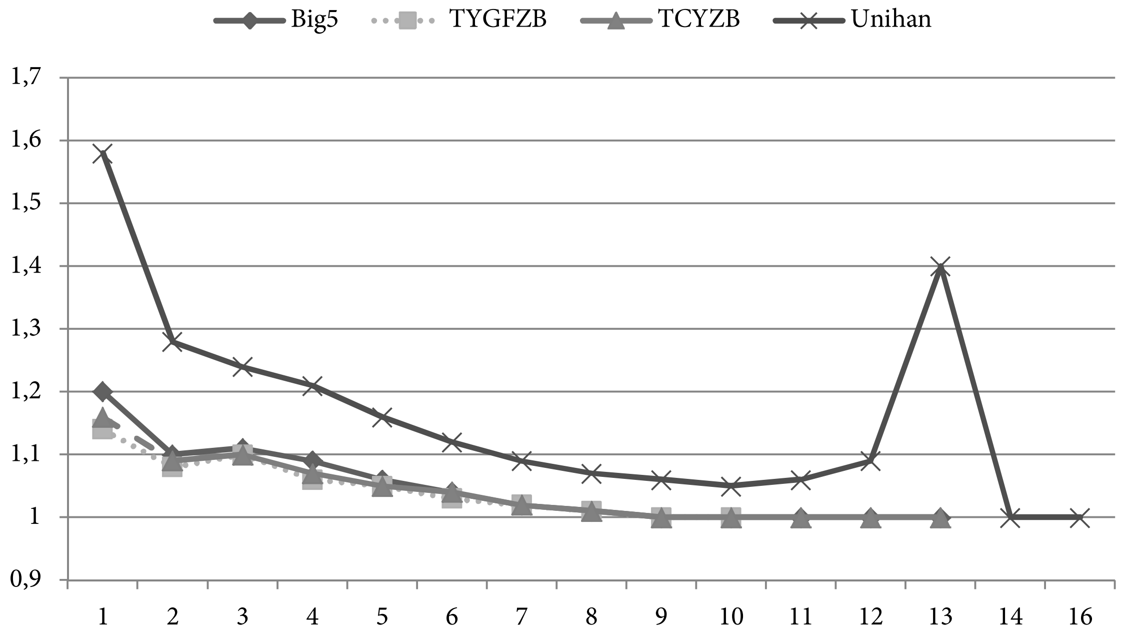


Fig. 7.56 Basic t-efficiency

Except for one irregularity on the basic level of Unihan the curves are very similar both in Fig. 7.55 and Fig. 7.56.

7.2.6.3. Tactographemic t-efficiency

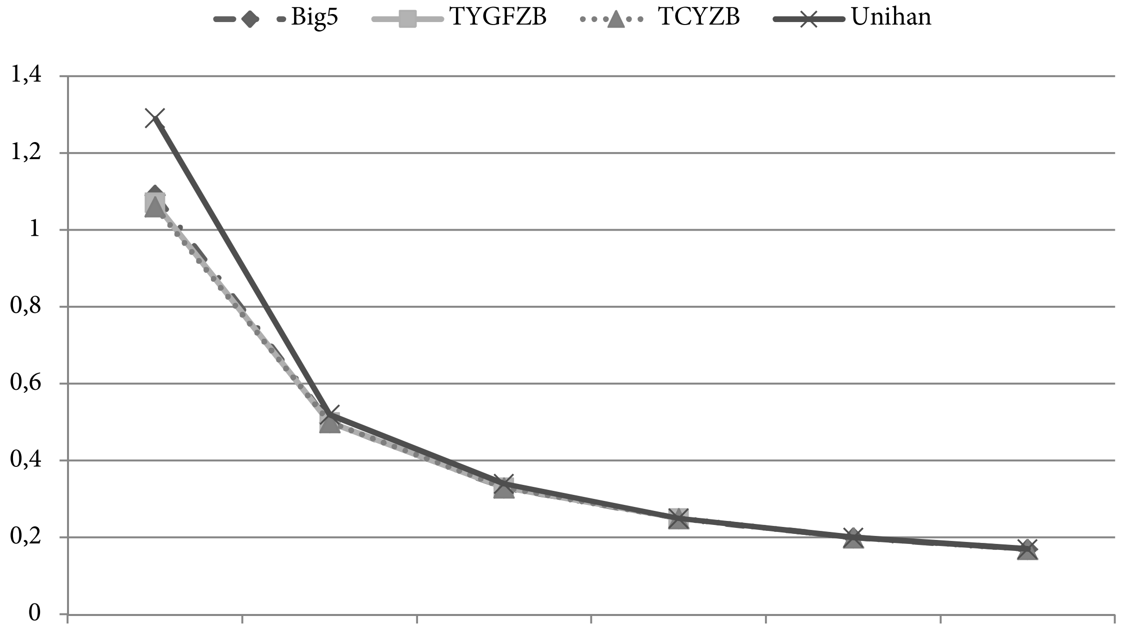


Fig. 7.57 Immediate tactographemic t-efficiency

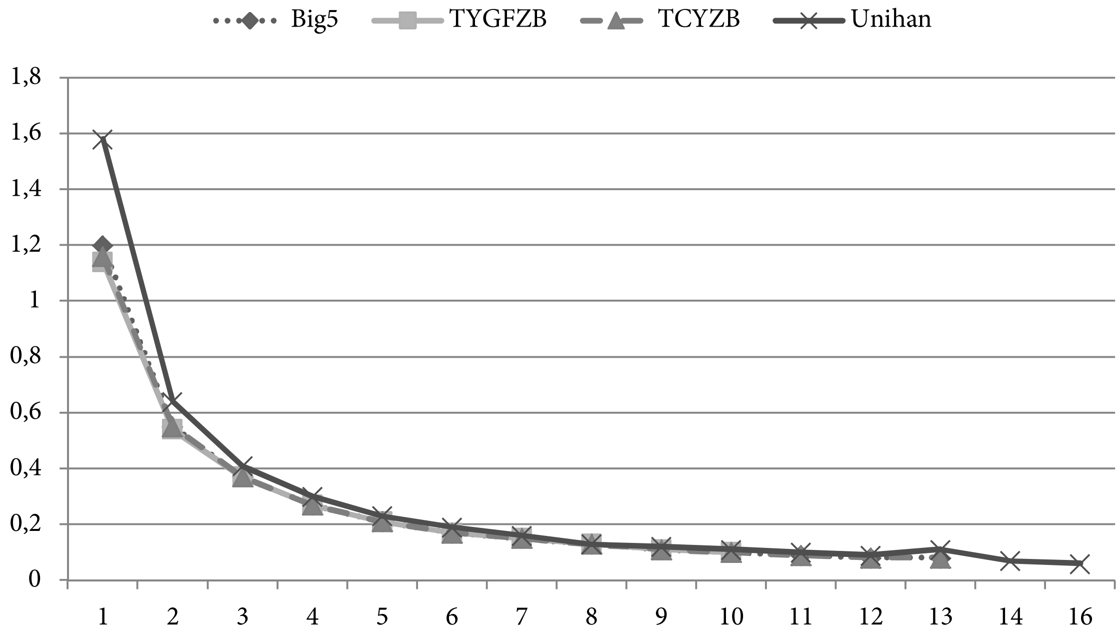


Fig. 7.58 Basic tactographemic t-efficiency

The curves in Fig. 7.57 and Fig. 7.58 are almost identical.

7.2.6.4. Categorial graphotactemic efficiency

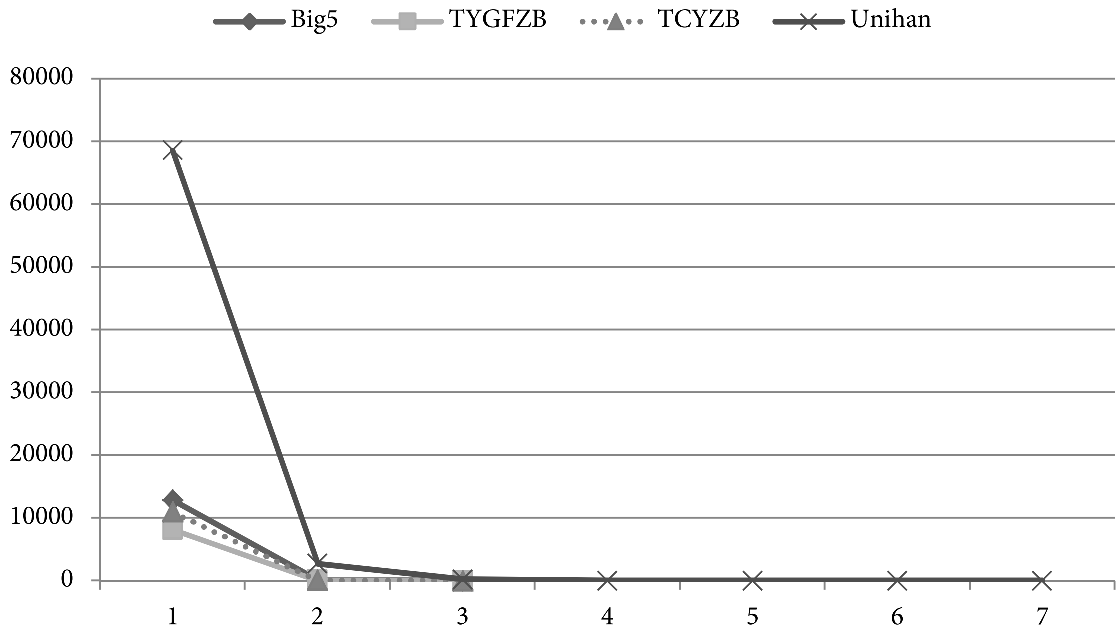


Fig. 7.59 Immediate categorial graphotactemic efficiency

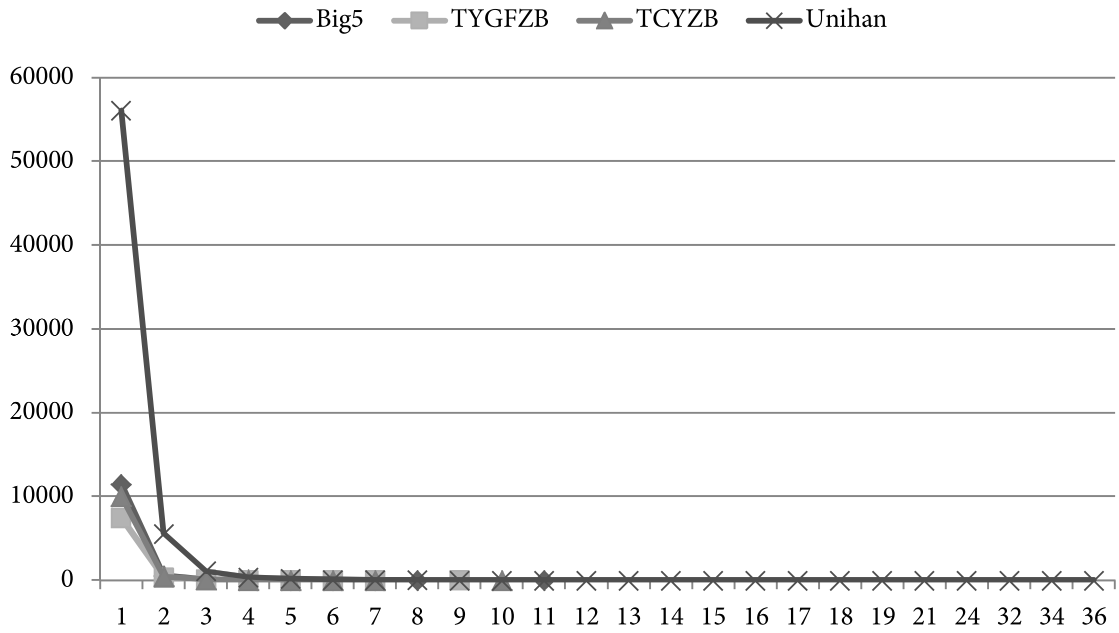


Fig. 7.60 Basic categorial graphotactemic efficiency

Besides the spike in the Unihan curve that can be explained by the size of the corpus, the shapes are almost identical.

The findings so far point to one inescapable conclusion – for a given writing system the size of the investigated corpus is irrelevant for the general quantitative relations, or in other words, the quantitative relations in the subsets reflect the quantitative relations in the superset. At this point it can be stated that this hypothesis is valid for the subsets with frequency motivated contents. It is plausible to predict that it does not hold for random sets/corpuses.

Investigating most types of script does not require, nor does it allow, differentiation between two levels of analysis. In the case of Chinese script the quantitative relations of the investigated types hold on both levels – it may be an independent graphotactic evidence that both immediate and basic components are legitimate levels of analysis of Chinese script.

7.2.7. Graphemic dispersion

The discussion so far has not touched upon the subject of dispersion of graphemes. This section provides detailed results of dispersion-related analysis of all four sets. Due to the size of the investigated corpuses the dispersion data for individual graphemes are difficult to present. Instead, only the graphs of graphemic dispersion will be shown.

7.2.7.1. Immediate components

7.2.7.1.1. Tactographemic dispersion

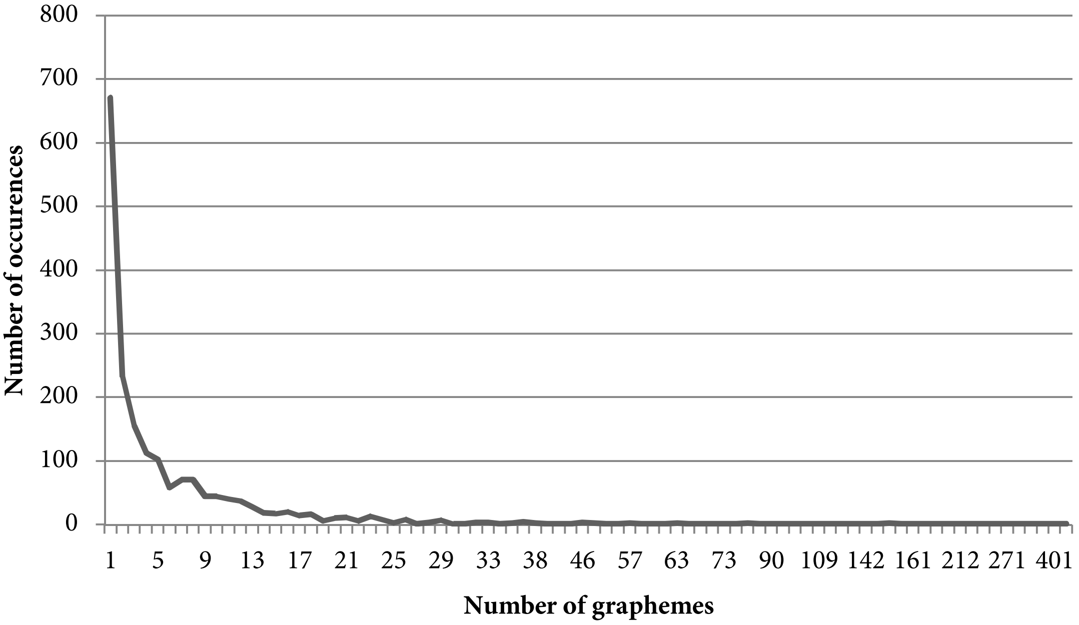


Fig. 7.61 Tactographemic dispersion numbers of TYGFZB graphemes

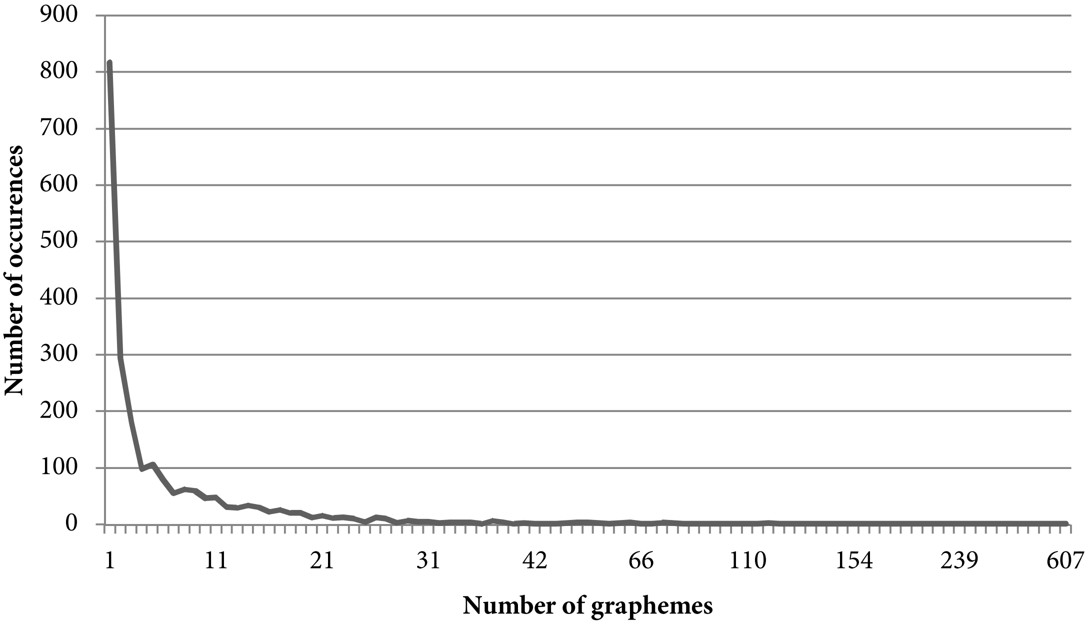


Fig. 7.62 Tactographemic dispersion numbers of TCYZB graphemes

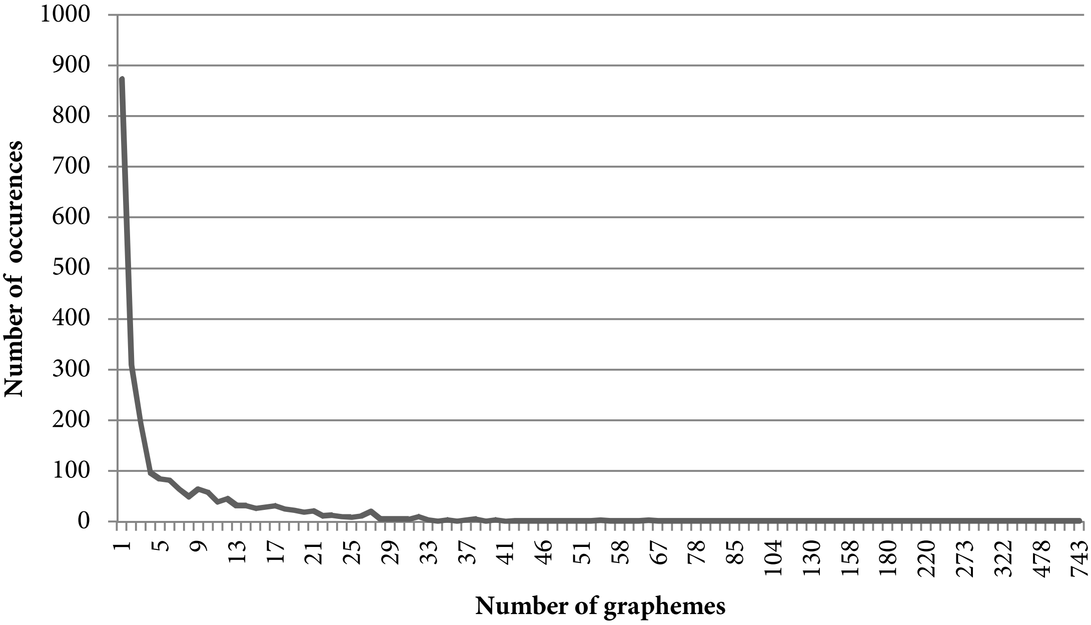


Fig. 7.63 Tactographemic dispersion numbers of Big5 graphemes

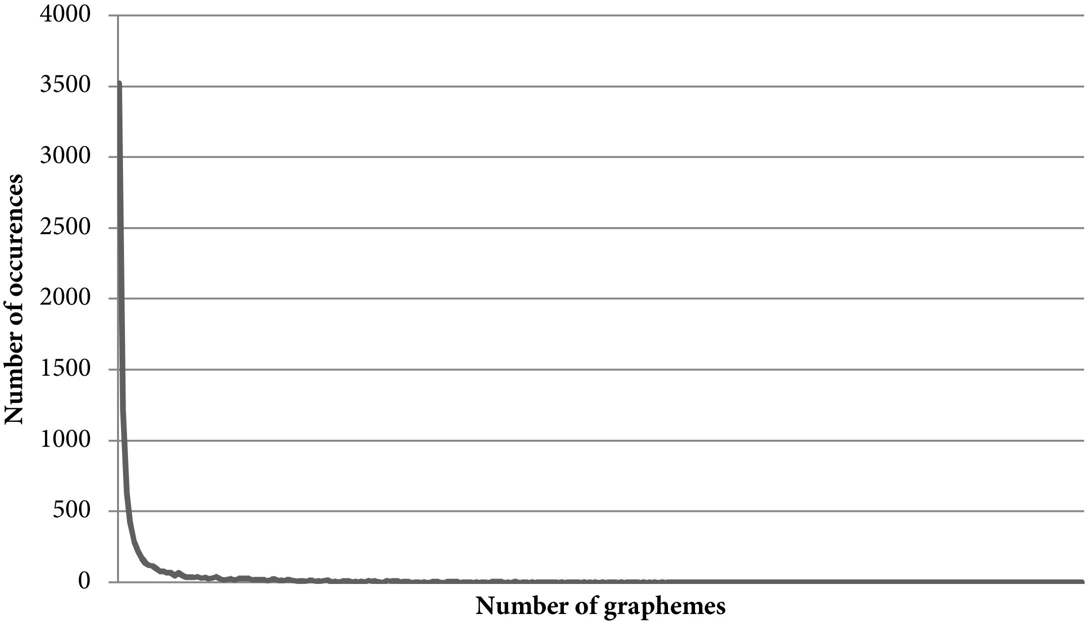


Fig. 7.64 Tactographemic dispersion numbers of Unihan graphemes

Fig. 7.61 to Fig. 7.64 all show similar dispersional properties of graphemes – a small number of graphemes occurring in a large number of tactographemes (high dispersion numbers) and a large number of graphemes occurring in small number of tactographemes (low dispersion numbers). This is represented by a logarithmic curve.

7.2.7.1.2. Graphotactemic dispersion

Fig. 7.65 to Fig. 7.67 indicate similar dispersional properties of graphemes as is the case with tactographemic dispersion – a small number of graphemes occurring in a large number of graphotactemes and a large number of graphemes occurring in small number of graphotactemes.

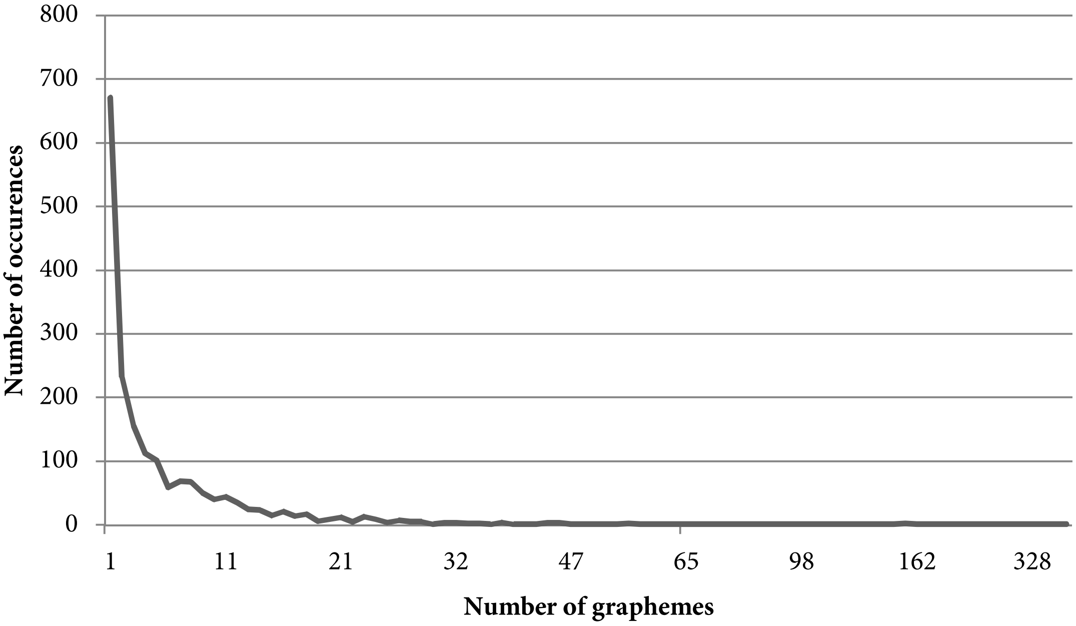


Fig. 7.65 Graphotactemic dispersion numbers of TYGFZB graphemes

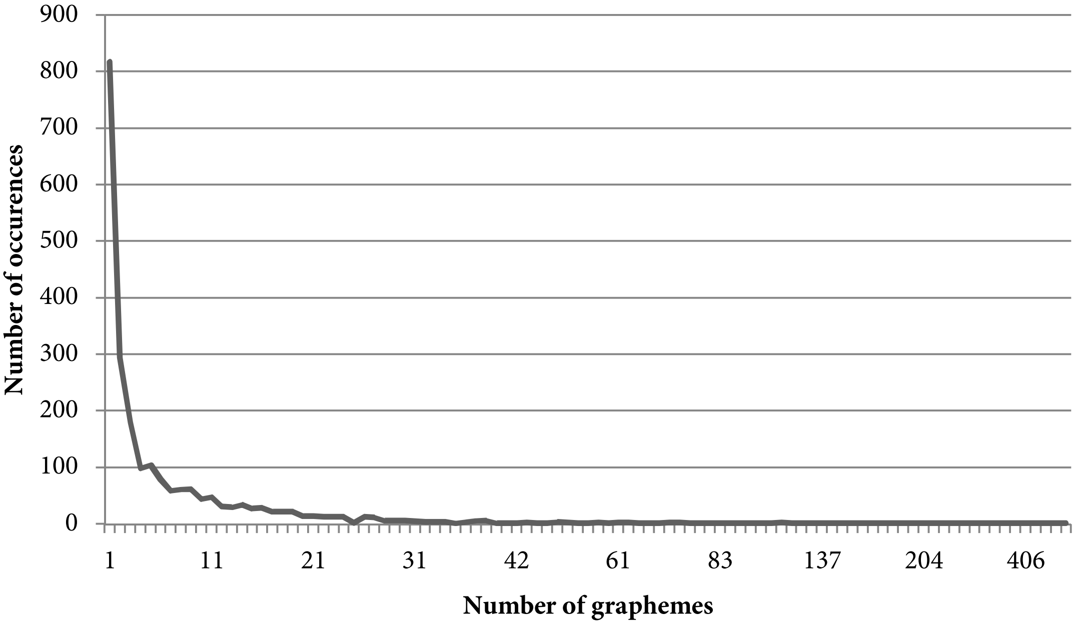


Fig. 7.66 Graphotactemic dispersion numbers of TCYZB graphemes

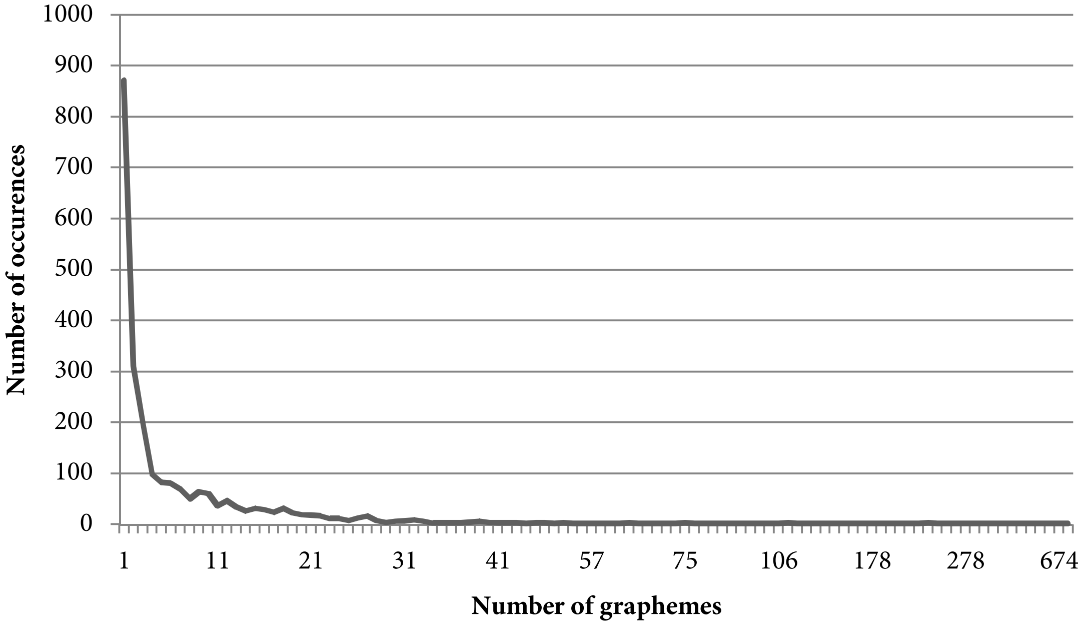


Fig. 7.67 Graphotactemic dispersion numbers of Big5 graphemes

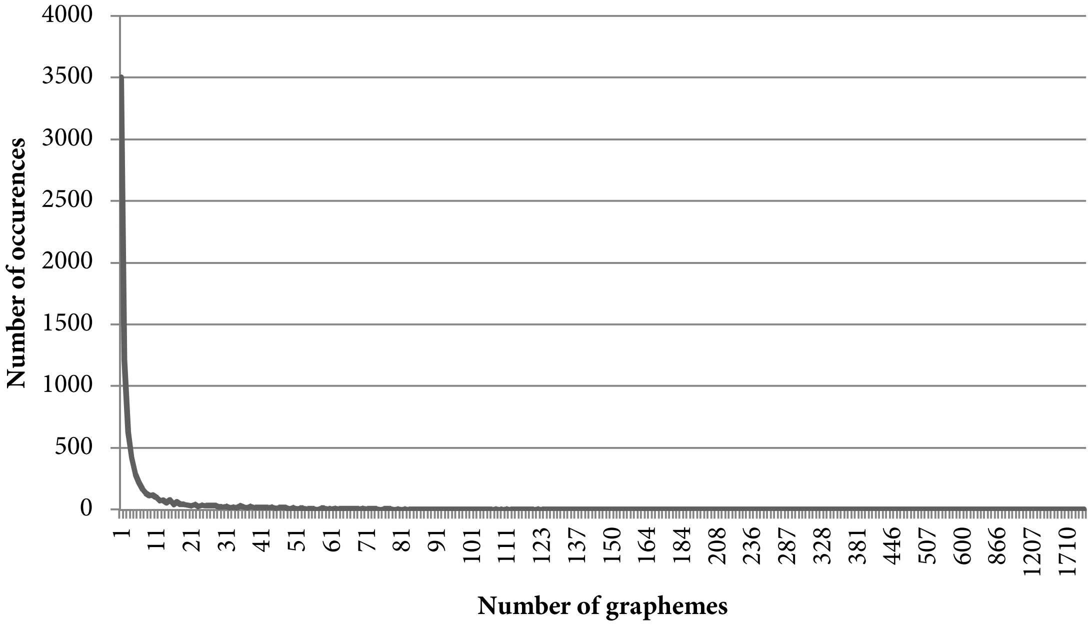


Fig. 7.68 Graphotactemic dispersion numbers of Unihan graphemes

7.2.7.2. Basic components

In the case of the basic component dispersion, an irregular distribution of graphemes can be observed. Only the dispersion curves for the largest set display a resemblance to the immediate dispersion of components illustrated in the previous section.

At this point, providing reasons for the irregularities would be pure speculation. This issue needs to be studied further.

Fig. 7.69 to Fig. 7.71 show a small number of graphemes occurring in a large number of tactographemes (high dispersion numbers), and a large number of graphemes occurring in a small number of tactographemes (low dispersion numbers); the middle values are very irregular. The same applies to Fig. 7.73 to Fig. 7.75. The curves for Unihan dispersion (Fig. 7.72 and Fig. 7.76) show less irregularities for the middle values.

7.2.7.2.1. Tactographemic dispersion

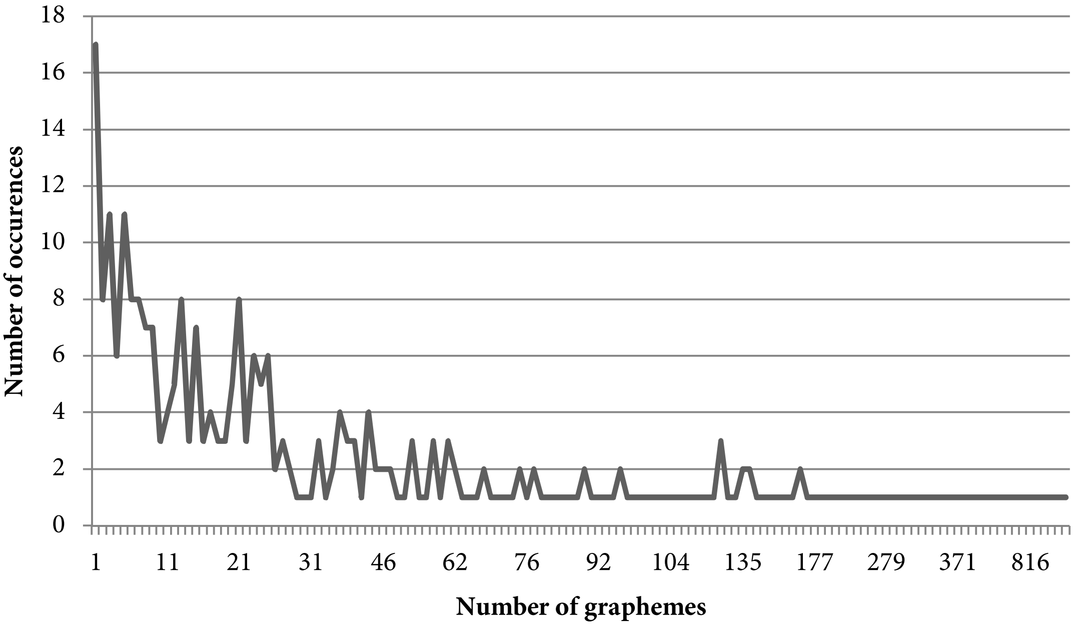


Fig. 7.69 Tactographemic dispersion numbers of TYGFZB graphemes

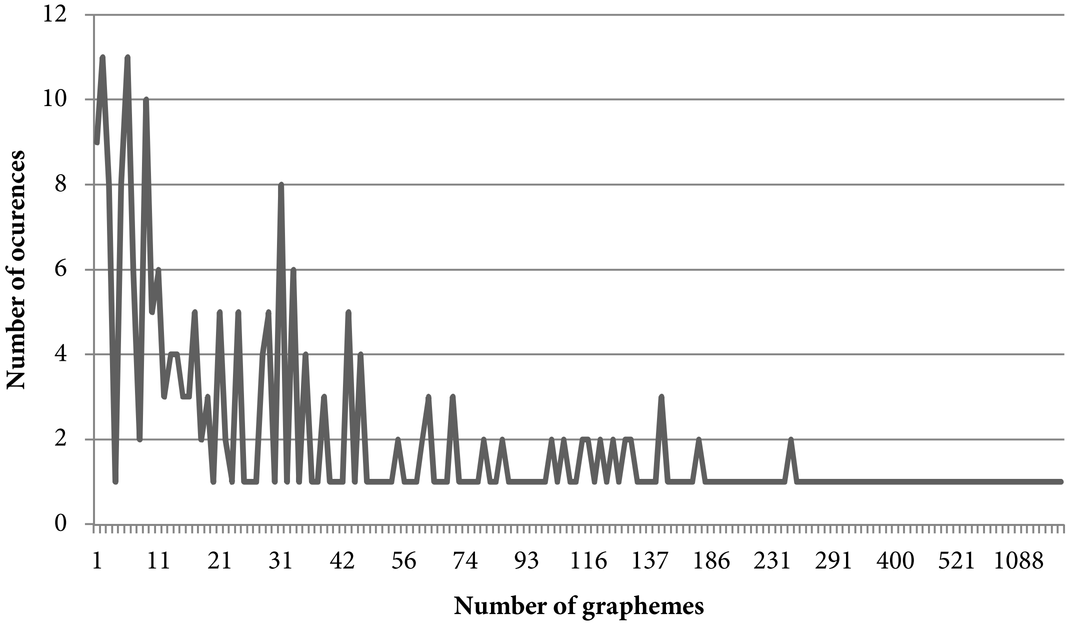


Fig. 7.70 Tactographemic dispersion numbers of TCYZB graphemes

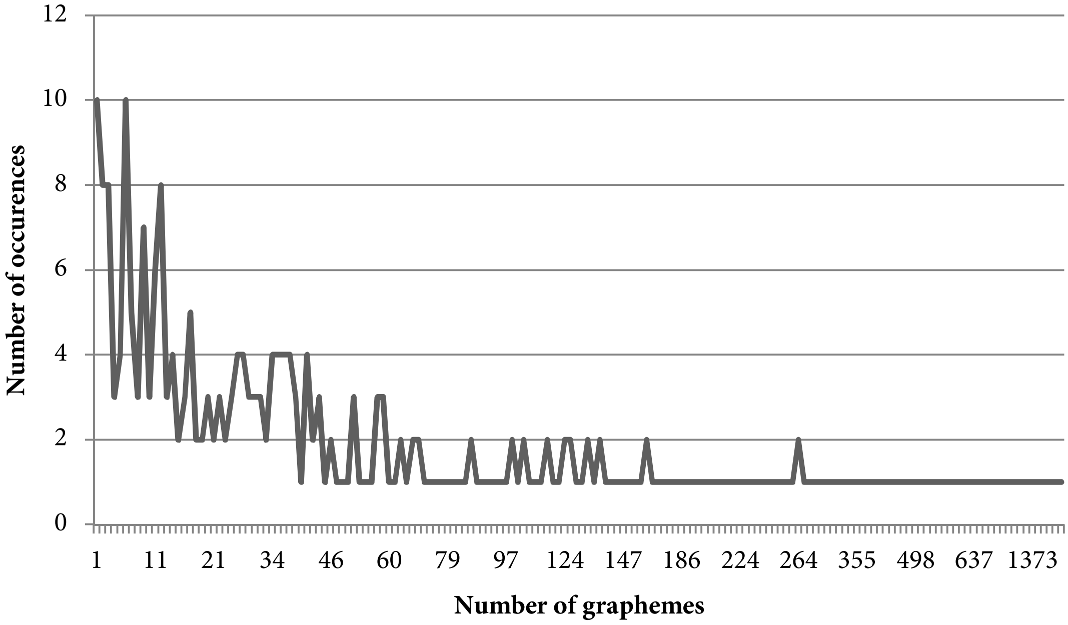


Fig. 7.71 Tactographemic dispersion numbers of Big5 graphemes

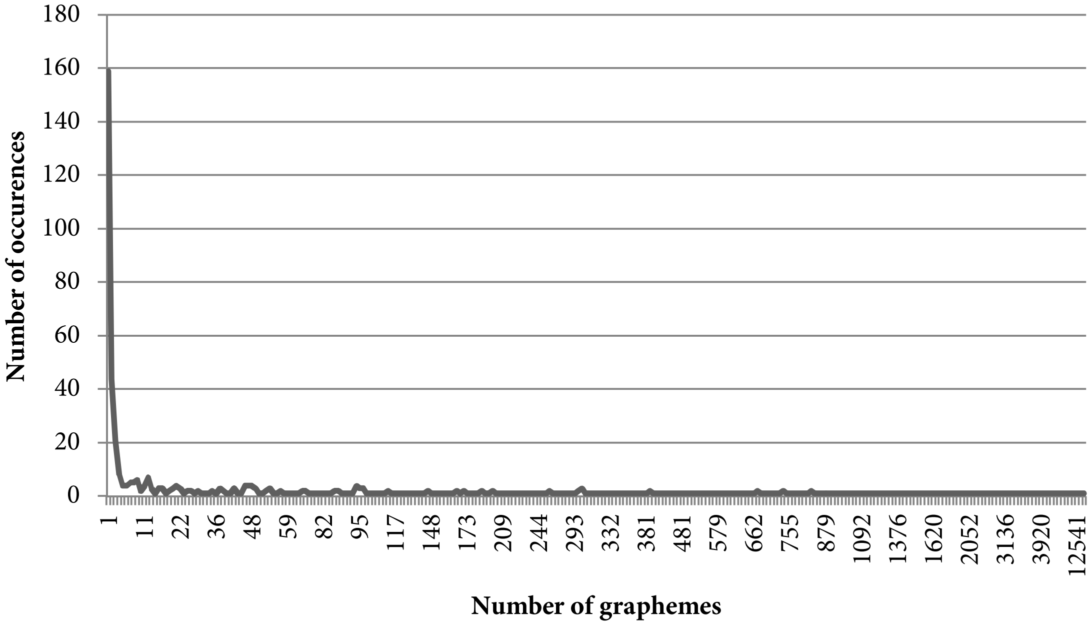


Fig. 7.72 Tactographemic dispersion numbers of Unihan graphemes

7.2.7.2.2. Graphotactemic dispersion

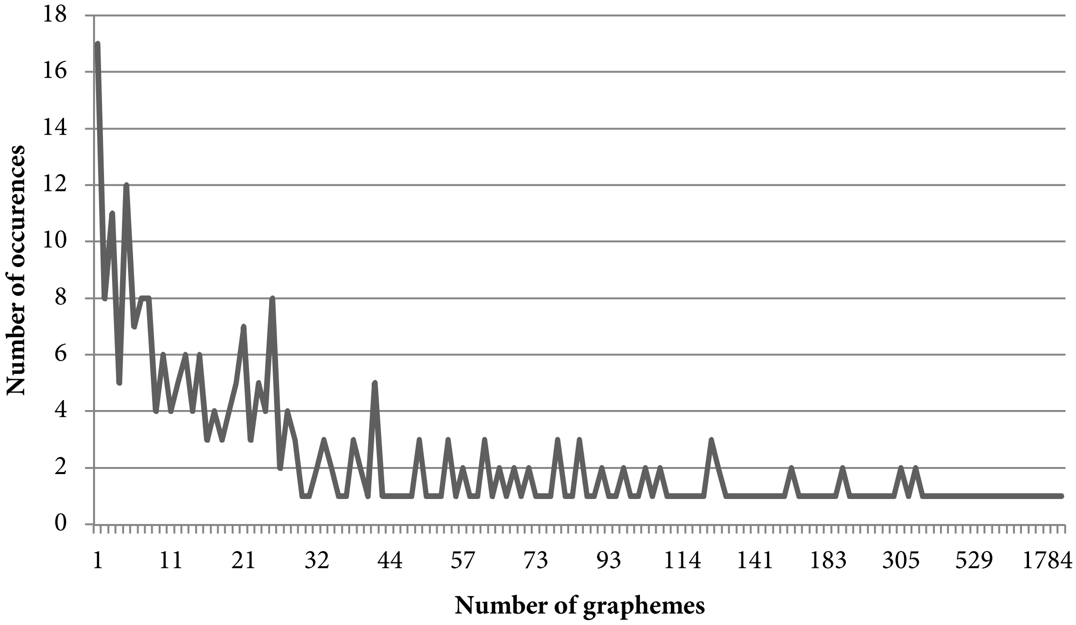


Fig. 7.73 Graphotactemic dispersion numbers of TYGFZB graphemes

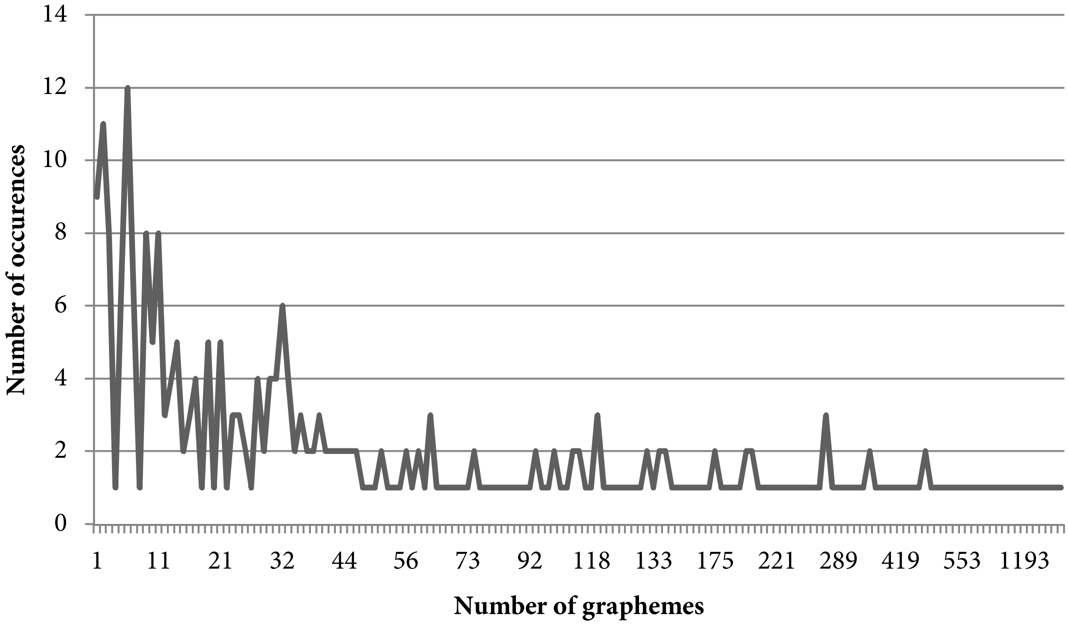


Fig. 7.74 Graphotactemic dispersion numbers of TCYZB graphemes

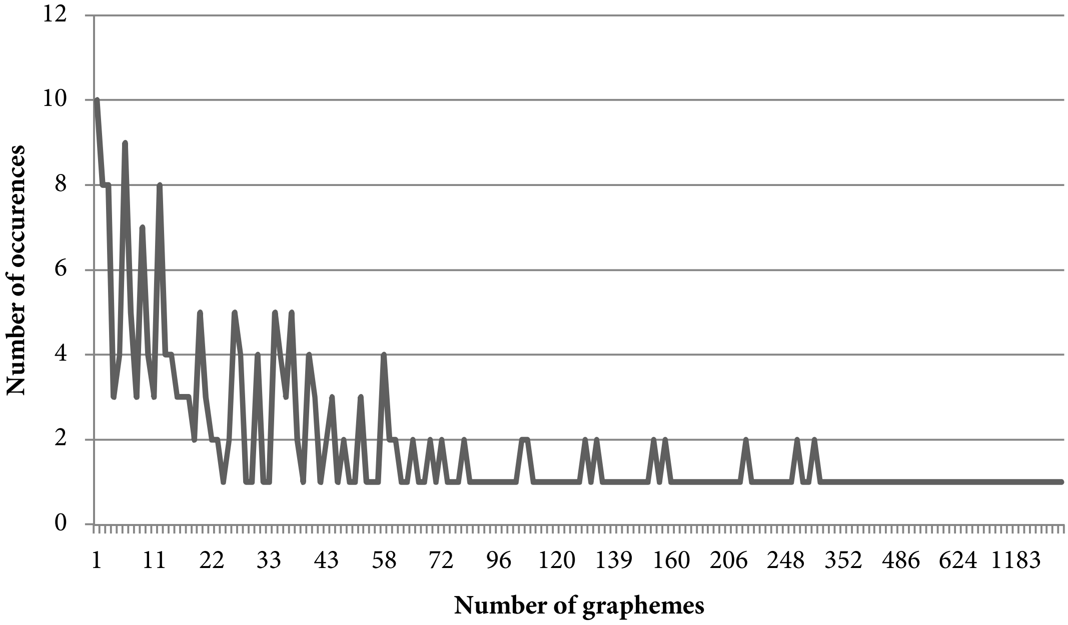


Fig. 7.75 Graphotactemic dispersion numbers of Big5 graphemes

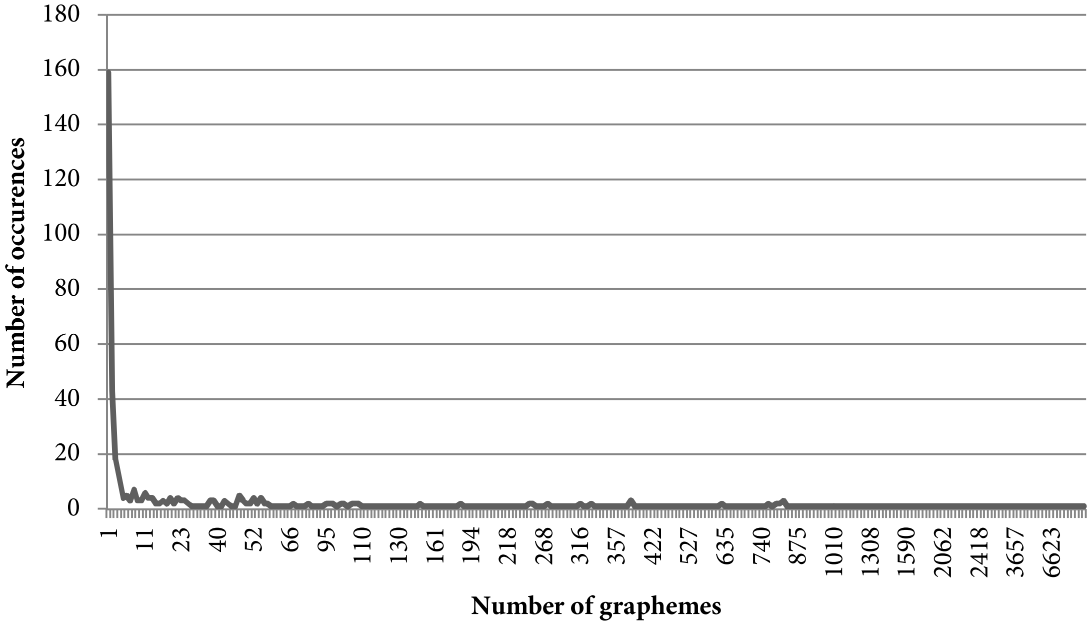


Fig. 7.76 Graphotactemic dispersion numbers of Unihan graphemes

7.2.7.3. Summary

The properties of graphemes in different sets pertaining to tactographemic and graphotactemic dispersion that were presented in this section can be summarized in a few points:

- dispersion of immediate components (graphemes) is regular, and can be represented by a logarythmic curve – there is a small number of graphemes with high dispersion numbers and a large number of graphemes with low dispersion numbers;
- dispersion of basic components (graphemes) is irregular, and only in terms of a very general tendency can it be stated that there is a small number of graphemes with high dispersion numbers and a large number of graphemes with low dispersion numbers;
- only the Unihan dispersion curve for basic components is similar to the curve for immediate components;
- graphemes display similar dispersional properties within the same grapheme types (immediate and basic).

At this point it is difficult to speculate on what the results would have been in an analysis conducted with different sets of basic components (CDP, GF3001-1997). It is certain, however, that such an analysis would provide valuable contrastive data that would help to interpret the results obtained here.

In all the examined sets the distribution of graphemes expressed in dispersion numbers vary over a broad range of values. Most graphemes behave idiosyncratically in this respect. It means that the distributional properties of graphemes, regardless of the their type and the size of the set, are not well represented by standard measures indicating a central tendency and dispersion. For that reason only average dispersion values are provided it Tab. 7.16. The results confirm predictions based on common sense – the average values increase with the size of character sets. The size of a set is directly reflected in the number of graphotactemes which in turn must be directly reflected in graphotactemic dispersion numbers. As it was shown in Sections 7.2.3. to 7.2.5., the number of graphotactemes (graphotactemicity) is correlated with the number of tactographemes. It can therefore be assumed that the size of the character set is also reflected in tactographemic dispersion numbers. Because there are always more graphotactemes than tactographemes the average tactographemic and graphotactemic dispersion values are expected to vary respectively. These predictions are confirmed by the analytic results – larger sets always have higher corresponding average values, and average graphotactemic dispersion is always higher for corresponding sets and grapheme types than in average tactographemic dispersion.

Tab. 7.16 Average tactographemic and graphotactemic dispersion of graphemes

	Avg. tactographemic dispersion		Avg. graphotactemic dispersion	
	Immediate	Basic	Immediate	Basic
TYGFZB	8.72	102.69	8.78	108.43
TCYZB	9.84	155.75	9.90	163.98
Big5	10.76	180.26	10.85	191.77
Unihan	17.84	517.30	18.57	596.56

7.2.8. Complexity of graphotactemes in terms of graphemes

The graphotactic analysis provides data that do not directly pertain to the subject matter of graphotactology, whereas the data are significant from a more general perspective. Statistics on complexity of characters in terms of the number of components were presented in previous chapters.[342] The same type of information will be shown in this section, based on the data obtained by graphotactic investigations. This understanding of the complexity of characters is different than the graphemicity of tac-

[342] See Section 6.1.3. and Tab. 5.2.

tographemes – tactographemes do not contain recurring graphemes while recurring components in a structure of characters contribute to the total component count.

7.2.8.1. Immediate components

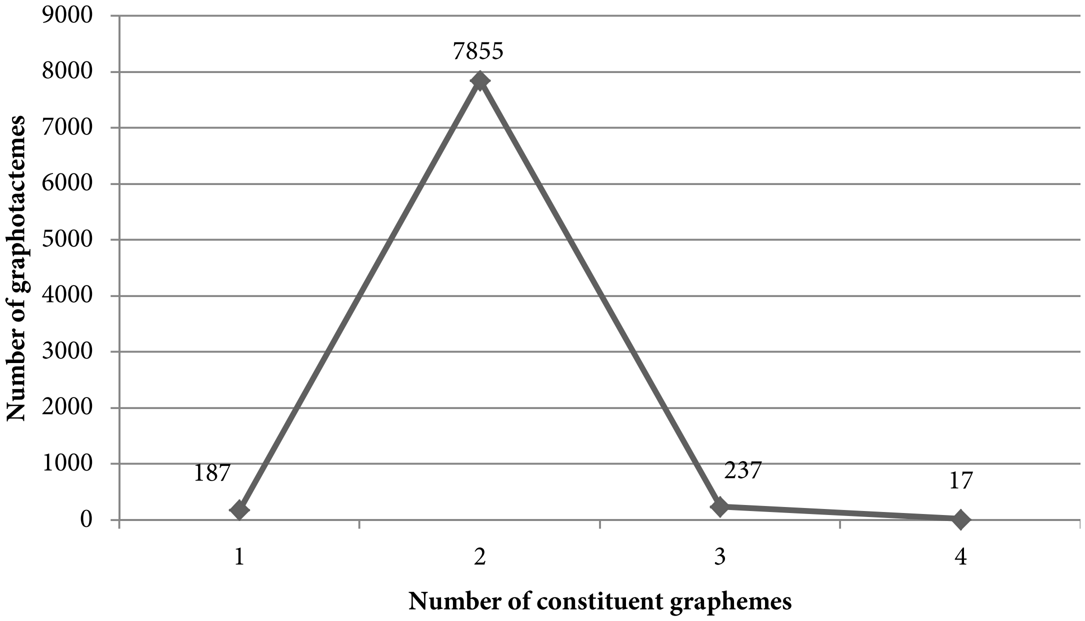


Fig. 7.77 Complexity of TYGFZB graphotactemes

Fig. 7.77 indicates that 7,855 graphotactemes (95% of all characters) consist of 2 graphemes (immediate components), 237 graphotactemes consist of 3 graphemes (3%), 187 consist of 1 grapheme (2%), and 17 consist of 4 graphemes. As expected, the curve takes a Gaussian shape.

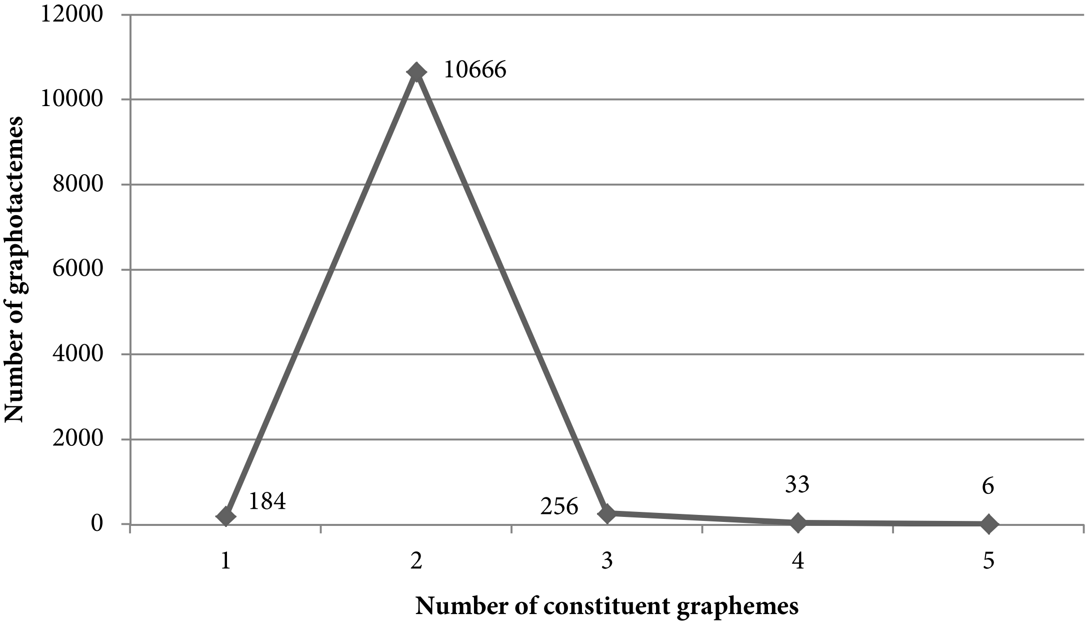


Fig. 7.78 Complexity of TCYZB graphotactemes

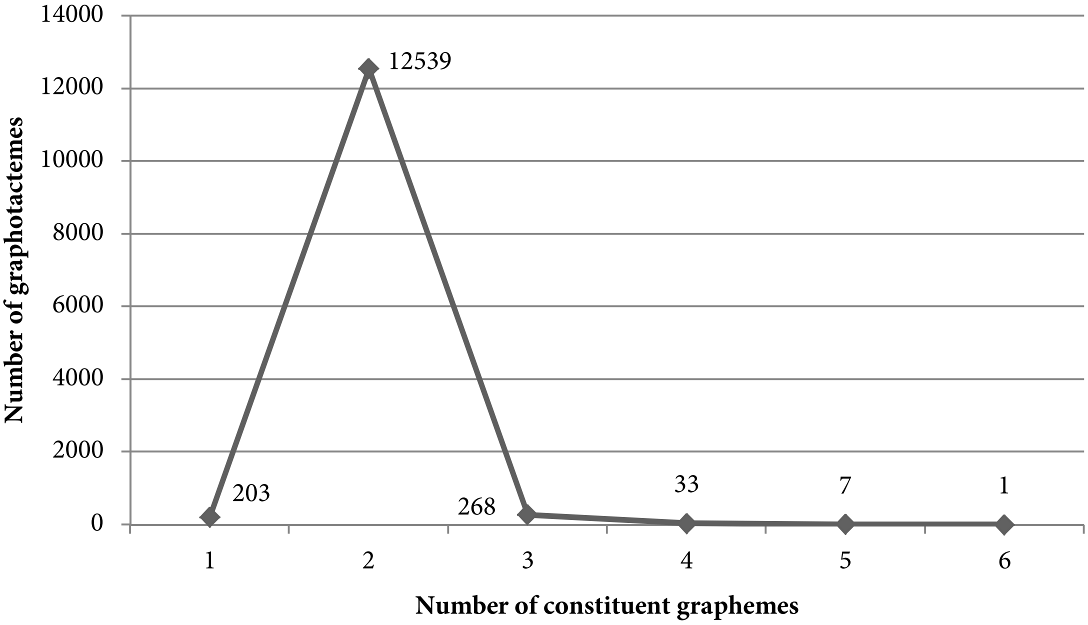


Fig. 7.79 Complexity of Big5 graphotactemes

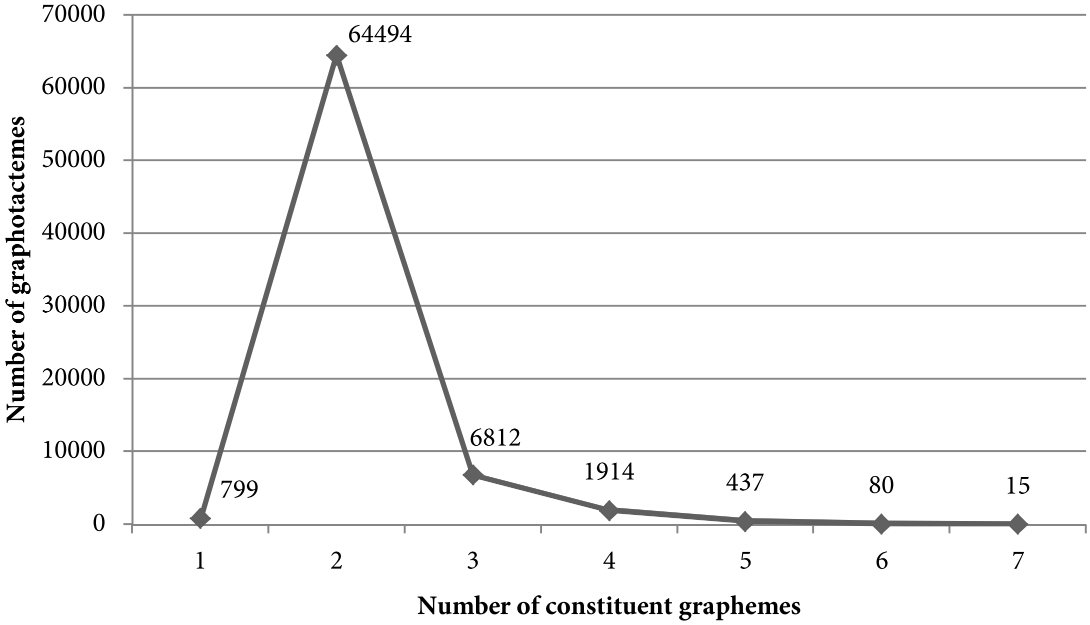


Fig. 7.80 Complexity of Unihan graphotactemes

7.2.8.2. Basic components

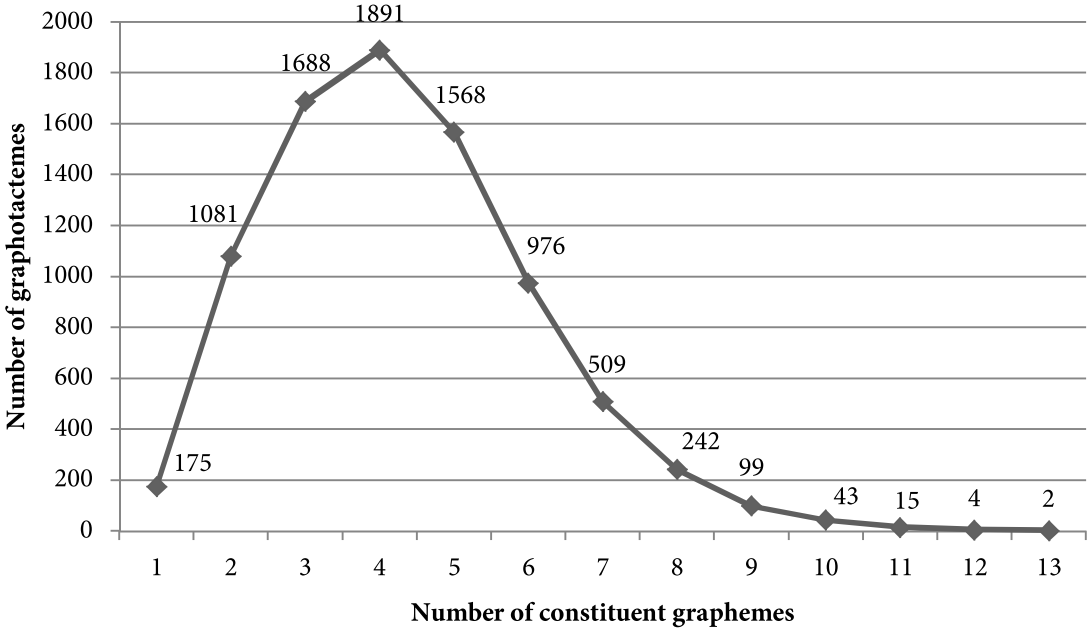


Fig. 7.81 Complexity of TYGFZB graphotactemes

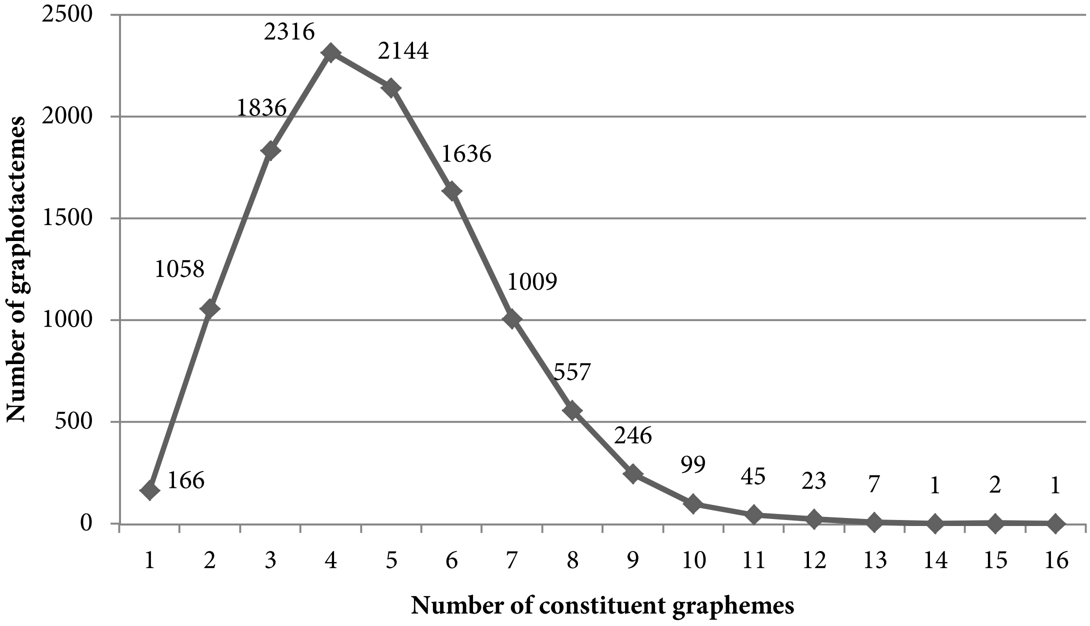


Fig. 7.82 Complexity of TCYZB graphotactemes

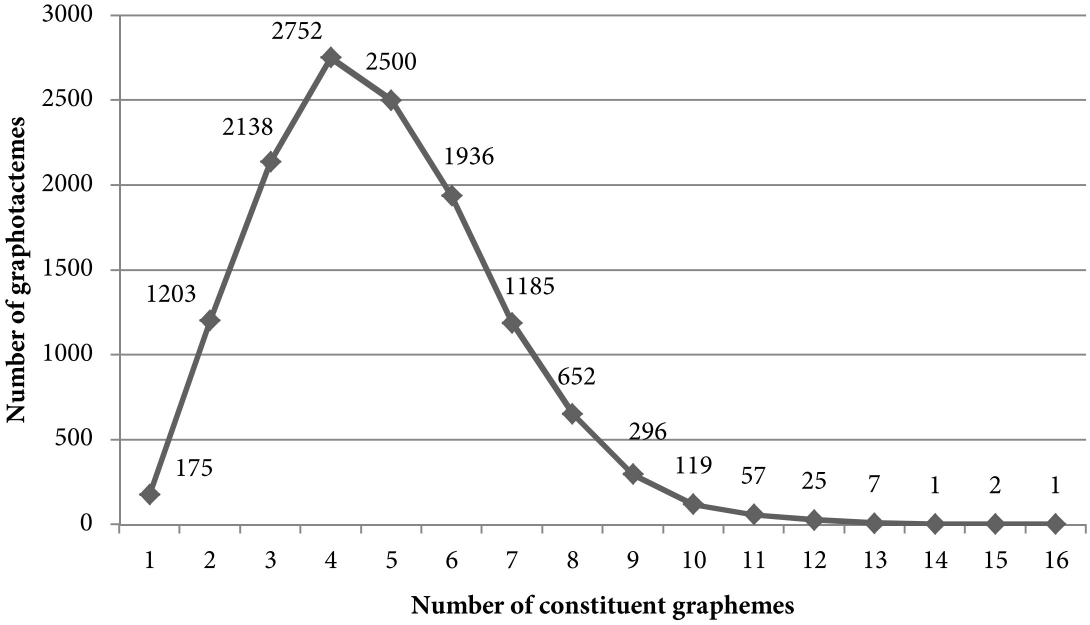


Fig. 7.83 Complexity of Big5 graphotactemes

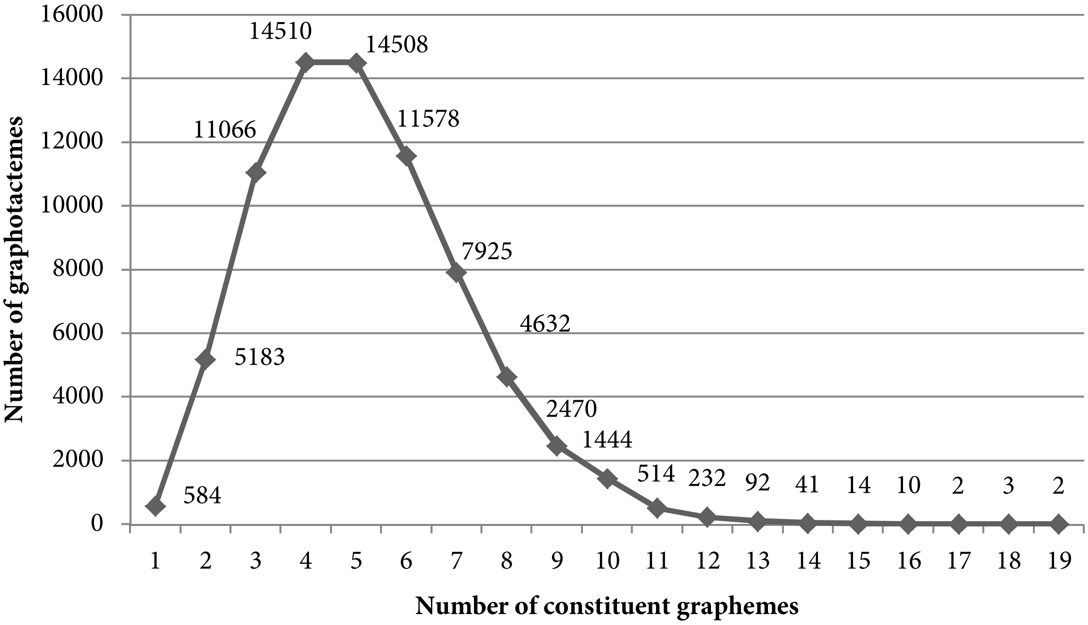


Fig. 7.84 Complexity of Unihan graphotactemes

7.2.8.3. Summary

In conclusion it can be safely stated that the statistical properties of character sets with respect to the componential complexity of characters are very regular:

- distribution scores for each type of components are represented by a Gaussian curve;
- the number of characters consisting of two immediate components exceeds 95% of all characters in every set, except in the CJK Unified Ideographs where the number exceeds 86%;
- the most numerous categories are similar for all sets within given types of components, with one expected exception being the basic component level of the TYGFZB set, in which simplification of characters must be reflected in the categorial shift;
- the other exception, caused by the size of the corpus, is the immediate components level of the CJK Unified Ideographs set where the category of characters consisting of four components is more numerous than the category of characters consisting of one component.

The discussed conclusions are summarized in Tab. 7.17.

Tab. 7.17 Most numerous categories of componential complexity

Character set	Immediate components	Basic components
TYGFZB	2, 3, 1	4, 3, 5, 2
TCYZB	2, 3, 1	4, 5, 3, 6
Big5	2, 3, 1	4, 5, 3, 6
Unihan	2, 3, 4	4, 5, 6, 3

Finally, in Tab. 7.18 the average values of complexity in terms of number of components are provided for all sets and for both types of components.

Tab. 7.18 Average complexity of graphotactemes

Character set	Immediate graphemes	Basic graphemes
TYGFZB	2.01	4.31
TCYZB	2.01	4.78
Big5	2.01	4.80
Unihan	2.15	5.13

7.2.9. Summary and concluding remarks

The aim of this chapter, as well as the main purpose of this study, was to provide extensive quantitative data reflecting the graphotactic properties of Chinese script. The analysis was designed to provide results as complete as possible, but within the confines of modern script. In case of *hànzi* it means covering the widest possible range of modern characters, while at the same time capturing the diversity of script from the structural and pragmatic perspectives.

The theoretical grid adopted from Bańczerowski's ideas[343] turned out to be a flexible tool, capable of providing the desired results. Their interpretation is a completely different problem. Segmentotactology can hardly be called an established discipline, and the results obtained here do not have a natural research environment in which the results can be discussed, analyzed and compared. From this perspective this study should be regarded as a contribution to the formation and maturation of a new discipline. Nevertheless, the results presented in this chapter demonstrated interesting regularities that were supported by the results of Bańczerowski's investigation of Polish and Chi-

[343] See Chapter 3.

nese orthography.[344] It is Bańczerowski's research that opened a potential space for hypothetical approaches about the general segmentotactic nature of script.

The quantitative data rendered by the graphotactic analysis of diverse character sets is a valuable result by itself. Beyond the data, however, there are two general hypotheses derived from the analysis:

Hypothesis I: The number of tactographemes and number of graphotactemes are closely correlated in each equigraphemic category (tactographon).

Hypothesis II: The quantitative segmentotactic relations that hold for a given system of script also hold for any frequency motivated segmental subset that is viable for statistical analysis.

Hypothesis I is a generalization of the findings on the relations between tactographemicity and t-graphotactemicity. Hypothesis II is a generalization of the findings pertaining to other graphotactic properties in all investigated sets. 'Segmental' pertains to different types of investigated units – in this study 'segment' pertains to characters, whereas in alphabetic scripts it pertains to orthographic words, etc. The conditions connected with the type of subset are to exclude random sets and the ones that are too small for quantitative analysis. Needless to say, since both hypotheses were established for Chinese script with the support of Canjie encodings and the results of Bańczerowski's investigation, further testing and verification is required. At this point it can be claimed that the investigation of graphotactic properties of diverse graphical lingual systems rendered meaningful results that display some significant regularities. The graphotactic analysis captured the distinctions between alphabetic scripts (Bańczerowski 2009), quasi-alphabetic Cangjie encoding, and the Chinese characters, but it also revealed similarities that allow a better understanding of the nature of graphical encoding of linguistic information.

It is hoped that the presented data and results will initiate a qualitative discussion of the investigated subject. It is beyond the scope of this book to reflect on the reasons the two hypotheses should hold, whether a system working on different principles is possible, or if these general claims are mutually valid, completely unanalyzed, domains of segmentotactology.

Other direct outcomes of the analysis performed in this chapter include a purely graphical compilation, an IDS based inventory of immediate, and more importantly, basic components, and lastly, the providing of evidence for the legitimacy of both immediate and basic levels of analysis of Chinese script.

[344] Bańczerowski 2009.

7.2.10. Perspectives

The present book has contributed empirical evidence for further study and for typological investigations. Many aspects of Chinese graphotactics remain unexplored, partially due to space limitations, but also due to the foundational character of this study. Given the complexity of the investigated subject, it is safe to assume that there are research areas yet to be realized. A glimpse into the unexplored segmentotactical territory is offered in the latest theoretical work by prof. Bańczerowski on the subject.[345] Given the typological diversity of the lingual systems that are subjects of interest in segmentotactology and the gamut of possible approaches, it is hoped that segmentotactology, and graphotactology in particular, will become an endeavor for a more substantial number of researchers.

Refering to the question of unexplored research territories – one very natural type of analysis still remains – the graphotactic analysis of Chinese characters in terms of strokes. Possibly two different levels of analysis need further investigation – immediate and basic strokes, analogously to the componential analysis. It is yet to be determined how arduous this task would be. Hopefully, it would be enough to assign stroke representations to the preexisting basic components and the rest of the process could be done automatically.

Some of the results obtained from the analysis of characters in terms of basic components revealed some irregularities[346] that were absent at the immediate level. It would be interesting and, in fact, necessary to perform the same analysis based on a different set of basic components.[347] The results could show whether the irregularities are inherent to the basic level, or just to the types of graphemes used in this study. Of course, results obtained with a different set of basic components pertaining to other aspects of Chinese graphotactics would also be interesting to examine.

Chinese graphotactics may be viewed as an auxiliary discipline which provides data for graphotactically non-releted research. For example, the decomposition procedure has produced inventories of basic and immediate components that can be used for retesting the conformity of Chinese script to statistical laws, and to the Menzerath-Altmann hypothesis, in particular.

One of the main purposes of this study was to demonstrate, through examples of graphotactics applied to Chinese characters, that segmentotactology is a flexible analytical tool rendering meaningful results, and in so doing, interest in the discipline would be ignited. It is hoped that future developments will prove this study to be at least partially successful in this respect.

[345] Bańczerowski 2013.

[346] Here, irregularities refers to the dispersion of graphemes.

[347] The CDP inventory seems like the most natural candidate.

References

Altmann, G. 1980. Prolegomena to Menzerath's Law. *Glottometrika* 2: 1-10.

___ 2004. Script Complexity. *Glottometrics* 8: 68-74.

___ 2008. Towards a Theory of Script. In: Altmann, G., Fan, F. (eds.), 149-164.

Altmann, G., Fan, F. (eds.). 2008. *Analyses of Script. Properties of Characters and Writing Systems.* Berlin – New York: Mouton de Gruyter.

Augst, G. (ed.) 1986a. *New Trends in Graphemics and Orthography.* Berlin & New York: Walter de Gruyter.

___ 1986b. Descriptively and Explanatorily Adequate Models of Orthography. In: Augst, G. (ed.), 25-42.

Bańczerowski, J. 2009. Aspects of Chinese Phonotactics Against a Comparative Background of Polish. *Scripta Neophilologica Posnaniensia* X: 7-22. (http://keko.amu.edu.pl/sites/default/files/Scripta%20Neophilologica%20Posnaniensia%20X.pdf#page=7, accessed 19.10.2012)

___ 2013. Izofoniczny aspekt fonotaktyki, In: Migdał, J., Piotrowska-Wojaczyk, A. (eds.): *Cum reverentia, gratia, amicitia... Księga jubileuszowa dedykowana Profesorowi Bogdanowi Walczakowi*, vol. I, 127-139. Poznań: Wydawnictwo Rys.

Bańczerowski, J., Tư, Lê Đình. 2012. Phonetic and Morphological Coding of Minimal Syntactic Units in Isolating Languages. *Rocznik Orientalistyczny* LXV.1: 7-23.

Baron, N. S. 1981. *Speech, Writing, and Sign. A Functional View of Linguistic Representation.* Bloomington: Indiana University Press.

Best, K.-H., Altmann, G. 2008. Script Ornamentality. In: Altmann, G., Fan, F. (eds.), 91-104.

Bi, Y., Xu, Y., Caramazza, A. 2009. Orthographic and Phonological Effects in the Picture–Word Interference Paradigm: Evidence from a Logographic Language. *Applied Psycholinguistics* 30: 637–658. (http://www.wjh.harvard.edu/~caram/PDFs/2009_Bi_Xu_Caramazza.pdf, accessed 18.06.2011)

Bishop, T., Cook., R. 2003. A Specification for CDL Character Description Language. (http://www.wenlin.com/cdl/cdl_spec_2003_10_31.pdf, accessed 25.12.2010)

___ 2007. A Character Description Language for CJK. *Multilingual* 18.7: 62-68. (http://www.wenlin.com/cdl/MLC-CDL.pdf, accessed 25.12.2010)

Bohn, H. 2002. Untersuchungen zur chinesischen Sprache und Schrift. In: Köhler, R. (ed.), 127-177.

Boltz, W. G. 1994. *The Origin and Early Development of the Chinese Writing System.* New Haven: American Oriental Society.

___ 2006. Pictographic Myths. *Bochumer Jahrbuch zur Ostasienforschung* 30: 39-54.

Brice, W. C. 1976. The Principles of Non-phonetic Writing. In: Haas, W. (ed.), 29-44.

Butler, Ch. S. 1985. *Statistics in Linguistics.* Oxford: Basil Blackwell.

Cantos Gomez, P. 2013. *Statistical Methods in Language and Linguistic Research*. Sheffield: Equinox Publishing Ltd.

Caramazza, A., Miceli, G. 1990. The Structure of Graphemic Representations. *Cognition* 37.3: 243-297.
(http://www.wjh.harvard.edu/~caram/PDFs/1990_Caramazza_Miceli_COG.pdf, accessed 28.03.2011)

Catach, N. 1986. The Grapheme: Its Position and its Degree of Autonomy with Respect to the System of the Language. In: Augst, G. (ed.), 1-10.

Chang, H. 1996. Semiographemics: A Peircean Trichotomy of Classical Chinese Script. *Semiotica* 108.1-2: 31-43.

陳學志 (Chen Hsueh-Chih), 張瓅勻 (Chang Li-Yun), 邱郁秀 (Chiou Yu-Shiou), 宋曜廷 (Sung Yao-Ting), 張國恩 (Chang Kuo-En). 2011. 中文部件組字與形構資料庫之建立及其在識字教學的應用 [Chinese Orthography Database and Its Application in Teaching Chinese Characters]. *教育心理學報43卷閱讀專刊* [*Bulletin of Educational Psychology (Special Issue on Reading)*] 43: 269-290.

Chen, T., Dell, G. S., Chen, J. 2007. A Cross-linguistic Study of Phonological Units: Syllables Emerge from the Statistics of Mandarin Chinese, but not from the Statistics of English. *Chinese Journal of Psychology* 49: 137-144.
(http://www.cogsci.northwestern.edu/cogsci2004/papers/paper263.pdf, accessed 08.10.2011)

陳奕全 (Chen Yi-Chuan), 葉素玲 (Yeh Su-Ling). 2009. 辨識理論模型中的部件表徵 [Component Representation in Models of Chinese Speech Recognition]. *應用心理研究* [*Research in Applied Psychology*] 43: 177-205.
(http://www.appliedpsyj.org/paper/43/07.pdf, accessed 15.10.2011)

陈原（Chen Yuan）主编（ed.）1989. *现代汉语定量分析* [*Quantitative Analysis of Modern Chinese*]. 上海教育出版社 (Shanghai: Jiaoyu Chubanshe).

Chen, P. 1994. Four Projected Functions of the New Writing System of Chinese. *Anthropological Linguistics* 36.3: 366-380.

Chen, P. 2004. *Modern Chinese. History and Sociolinguistics.* Cambridge University Press.

Chikamatsu, N., Yokoyama, S., Nozaki, H., Long, E., Fukuda, S. 2000. A Japanese Logographic Character Frequency List for Cognitive Science Research. *Behavior Research Methods, Instruments, & Computers* 32.3: 482-500.

朱邦復 (Chu Bong-Foo). 1990. 倉頡輸入法與中文字形產生器 [The Cangjie Input Method and the Chinese Character Forms Generator].
(http://cbflabs.com/book/gif_cg/gif_cg/index.html, accessed 25.06.2010)

朱邦復 (Chu Bong-Foo), 沈紅蓮 (Shen Hong-Lien). 2006. 第五代倉頡輸入法手冊 [Manual for the Fifth Generation of Cangjie Input Method]. 博碩文化出版 (Boshi Wenhua Chuban) (http://cbflabs.com/book/ocj5/ocj5/01.htm, accessed 25.06.2010)

莊德明 (Chuang Der-ming)，鄧賢瑛 (Teng Hsian-ying). 2009. 漢字構形資料庫的研發與應用 [Research and Development of Chinese Characters Information Database and Its Application].
(http://cdp.sinica.edu.tw/service/documents/T090904.pdf, accessed 15.03.2011)

Clark, J. L., Lua, K. T., McCallum, J. 1990. Conformance of Chinese Text to Zipf's Law. *PARBASE-90 International Conference on Databases, Parallel Architectures and Their Applications*, 533.

Cook, R. S. 2001. The Extreme of Typographic Complexity: Character Set Issues Relating to Computerization of The Eastern Han Chinese Lexicon "說文解字" *Shuowenjiezi. 18th International Unicode Conference.*
(http://linguistics.berkeley.edu/~rscook/pdf/IUC18-SWJZZ.wp3-600.pdf, accessed 22.09.2010)

___ 2003. 說文解字 – 電子版 *Shuo Wen Jie Zi - Dianzi Ban: Digital Recension of the Eastern Han Chinese Grammaticon*. UC Berkeley, Dept. of Linguistics, Ph.D diss.

Coulmas, F. 1989. *The Writing Systems of the World*. London: Basil Blackwell.

___ 2003. *Writing systems. An introduction to their linguistic analysis.* Cambridge University Press.

Cramer, I. M., 2005. Das Menzerathsche Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), 659-687.

Culicover, P. W. 2013. *Grammar & Complexity. Language at the Interface of Competence and Performance*. New York: Oxford University Press.

Da, J. 2004. A Corpus-based Study of Character and Bigram Frequencies in Chinese E-texts and its Implications for Chinese Language Instruction. In: Zhang, P., Xie, T., Xu, J. (eds.): *The Studies on the Theory and Methodology of the Digitalized Chinese Teaching to Foreigners: Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese*, 501-511. Beijing: Qinghua University Press.
(http://lingua.mtsu.edu/academic/dajun-4thtech.pdf, accessed 03.04.2011)

Daniels, P. T., Bright, W. (eds.) 1996. *The World's Writing Systems*. New York: Oxford University Press.

DeFrancis, J.. 1984a. Digraphia. *Word* 35(1): 59–66.

___ 1984b. *The Chinese Language: Fact and Fantasy.* Honolulu: University of Hawaii Press.

___ 1989. *Visible Speech: The Diverse Oneness of Writing Systems.* Honolulu: University of Hawaii Press.

Ding, G., Peng, D., Taft, M. 2004. The Nature of the Mental Representation of Radicals in Chinese: A Priming Study. *Journal of Experimental Psychology-Learning Memory and Cognition* 30.2: 530-539. (http://www2.psy.unsw.edu.au/users/mtaft/dingpengandtaft2004.pdf, accessed 03.04.2011)

Doerman, D., Jaeger, S. 2008. *Arabic and Chinese Handwriting Recognition.* Berlin – Heidelberg: Springer.

Doležel, L., Prucha, J. 1966. A Statistical Law of Grapheme Combinations. *Prague Studies in Mathematical Linguistics* 1: 33-43.

Dowdy, S., Weardon, S., Chilko, D. 2004. *Statistics for Research*. Hoboken, New Jersey: Wiley-Interscience.

Duanmu, S. 2007. *The Phonology of Standard Chinese.* Oxford University Press.

___ 2008. *Syllable Structure. The Limits of Variation.* New York: Oxford University Press.

范可育 (Fan Keyu). 1990 (2002). 汉字部件分解的原则 [Principles of Decomposition of Chinese Characters into Components]. In: Su Peicheng 2002a (ed.), 102-110.

费锦昌 (Fei Jinchang). 1997 (2002). 现代汉字笔画规范刍议 [Proposal on Standarization of Strokes of Chinese Characters]. In: Su Peicheng 2002a (ed.), 90-101.

Feldman, L. B., Siok, W. T. 1997. The Role of Component Function in Visual Recognition of Chinese Characters. *Journal of Experimental Psychology: Learnng, Memory and Cognition* 23.3: 776-781. (http://web.haskins.yale.edu/Reprints/HL1042.pdf, accessed 13.06.2011)

___ 1999. Semantic Radicals in Phonetic Compounds: Implications for Visual Character Recognition in Chinese. In: Wang, J., Inhoff, A. W., Chen, H.-Ch. (ed.), 19-36.

Figge, U. L. 1990. Design Features of Human Language. In: Koch, W.A. (ed.), *Evolution of Culture. Proceedings of the International and Interdisciplinary Symposium, Sept. 19-23, 1988, Loveno di Menaggio, Italy = Evolution der Kultur. Bochum: Brockmeyer, 1990. (Bochum Publications in Evolutionary Cultural Semiotics; 22)*, 94-99.

French, M. A. 1976. Observations on the Chinese Script and the Classification of Writing Systems. In: Haas, W. (ed.), 101-130.

Frost, R., Katz, L. 1992. *Orthography, Phonology, Morphology and Meaning.* Amsterdam: North-Holland Elsevier Science Publishers.

Fujimura, O., Kagaya, R. 1969. Structural Patterns of Chinese Characters. *Annual Bulletin of the Research Institute of Logopedics & Phoniatrics, Faculty of Medicine University of Tokyo* 3: 131-148.

高更生 (Gao Gengsheng). 1999. *汉字研究* [*The Chinese Characters Study*] 济南: 山东教育出版社 (Jinan: Shandong Jiaoyu Chubanshe).

高家莺 (Gao Jiaying), 范可育 (Fan Keyu). 1985 (2002). 建立现代汉字学刍议 [*Proposal on Establishing Modern Chinese Characterology*]. In: Su Peicheng 2002a (ed.), 34-42.

Gelb, I. J. 1963. *A Study of Writing*. Chicago: University of Chicago Press.

Gordon, M. K. 2006. *Syllable Weight: Phonetics, Phonology, Typology*. New York & London: Routledge.

顾小凤 (Gu Xiaofeng) 2000 (2002). 汉字在计算机中如何表示 [Representation of Chinese Characters in Computers]. In: Su Peicheng 2002a (ed.), 253-259.

郭曙纶 (Guo Shulun), 方有林 (Fang Youlin). 2005. 网络汉字的大规模统计与分析 [Large Scale Statistics and Analysis of Chinese Characters on the Internet]. *汉字研究* [*Chinese Characters Research*] 1: 12-18.

郭曙纶 (Guo Shulun) 2009. 简化字与繁体字笔画数的动态统计与比较 [Dynamic Counting and Contrast between the Stroke Numbers of Simplified and Traditional Chinese Characters]. *北华大学学报(社会科学版)* [*Journal of Beihua University (Social Sciences)*] 10.2: 50-56.

Ha, L., Sicilia-Garcia, E. I., Ming, J., Smith, F. J. 2003. Extension of Zipf's Law to Word and Character *N*-grams for English and Chinese. *Computational Linguistics and Chinese Language Processing* 8.1: 77-102.

Haas, W. (ed.). 1976a. *Writing without Letters*. Manchester University Press.

___ 1976b. Writing: The Basic Options. In: Haas, W. (ed.), 131-208.

___ 1983. Determining the Level of a Script. In: Coulmas, F., Ehlich., K. (eds.), *Writing in Focus* (Trends in Linguistics, Studies and Monographs 24), 15-27. Berlin, New York and Amsterdam: Mouton.

Halpern, J. 2002. Lexicon-based Orthographic Disambiguation in CJK Intelligent Information Retrieval. *COLING '02 Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization - Volume 12*, 17-23. (http://www.cl.cs.titech.ac.jp/alr/WS/3rd/ALR02.pdf, accessed 17.06.2012)

韩布新 (Han Buxin). 1994. 汉字部件信息数据库的建立 – 部件和部件组合频率的统计分[Development of Database of Chinese Constituents Information – Statistical Analysis of the Frequency of the Constituents and Their Combination]. *心理学报* [*Acta Psychologica Sinica*] 26.2: 147–52.

___ 1995. 部件组合 – 潜在的汉字结构层次 *[Combination of Chinese Character Constituents – A Latent Structural Unit]. 中文信息学报* [*Journal of Chinese Information Processing*] 9.3: 27-32.

___ 2002. Frequency and Position Effects of Component Combination in Chinese Character Recognition. In: Kao, H. S. R., Leong, C., Gao, D. (eds.), *Cognitive Neuroscience Studies of the Chinese* Language, 207-223. Hong Kong: Hong Kong University Press.

Han, B., Luo, L. 2001. Use of Modern Chinese Language. In: Karwowski, W. (ed.), *International Encyclopedia of Ergonomics and Human Factors*, 746-749. London & New York: Taylor & Francis.

Han, J. 2012. *Chinese Characters*. New York: Cambridge University Press.

Handel, Z. 2009. Towards a Comparative Study of Sinographic Writing Strategies in Korean, Japanese, and Vietnamese. *Proceedings of the SCRIPTA 2008. Hunminjeongeum and Alphabetic Writing Systems*, 105-135. (Also in *Scripta* 1: 89-126).

Hannas, W. C. 1997. *Asia's Orthographic Dilemma.* Honolulu: University of Hawaii Press.

___ 2003. *The Writing on the Wall. How Asian Orthography Curbs Creativity.* University of Pennsylvania Press.

___ 2005. Script and Cognition. The Impact of Orthography on Western and Eastern Patterns of Thought. *Occidental Quarterly* 5.3: 53-63.

Haralambous, Y. 2011. Seeking Meaning in a Space Made out of Strokes, Radicals, Characters and Compounds. *Computation and Language* (http://arxiv.org/pdf/1104.4321v1, accessed 18.08.2012)

Hegenbarth-Reichardt, I., Altmann, G. 2008. On the Decrease of Complexity from Hieroglyphs to Hieratic Symbols. In: Altmann, G., Fan, F. (eds.), 105-114.

Hill, Archibald A. 1967. The Typology of Writing Systems. In: William M. Austin (ed.), *Papers in Linguistics: in Honor of Leon Dostert*, 92–99. The Hague: Mouton.

Householder, F. W. 1971. *Linguistic Speculations*. Camridge: Cambridge University Press.

Hsieh, C. C., Chang, C. T., Huang, J. K. T. 1990. On the Formalization of Glyph in the Chinese Language. *Journal of Library & Information Science* 16.1: 1-26.

黃沛榮 (Huang Pei-Rong). 2003. *漢字教學的理論與實踐* (*Theory and Practice of Teaching Chinese Characters*). 台北: 樂學 (Taipei: Le Xue).

Huang, K. T., Huang, T. D. 1989. *An Introduction to Chinese, Japanese and Korean Computing*. Singapore: World Scientific.

黄普书 (Huang Pushu). 2006. *汉字. 字源篇.* [*Chinese Characters. Etymological Study*] 上海: 学林出版社 (Shanghai: Xue Lin Chubanshe).

Hyman, M. 2006. Of Glyphs and Glottography. *Language and Communication* 26: 231-249.

Ingulsrud, J. E., Allen, K. 2003. First Steps to Literacy in Chinese Classrooms. *Current Issues in Comparative Education* 5.2: 103-116.

Iwanowski. M. 2004. *Segmentacja grafów Tōyō-Kanji drukowanego pisma japońskiego.* Warszawa: Dialog.

Jin, P., Carroll, J., Wu, Y., McCarthy, D. 2012. Distributional Similarities for Chinese: Exploiting Characters and Radicals. *Mathematical Problems in Engineering,* volume 2012.

Johnson, K. 2008. *Quantitative Methods in Linguistics.* Malden – Oxford – Carlton: Blackwell Publishing.

Juang, D., Wang, J., Lai, C., Hsieh, C., Chien, L., Ho, J. 2005. Resolving the Unencoded Character Problem for Chinese Digital Libraries. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries JCDL '05*, 311-319. (http://www.iis.sinica.edu.tw/papers/hoho/2151-F.pdf, accessed 12.11.2012)

康加深 (Kang Jiashen). 1993 (2002). 现代汉语形声字形符研究 [Analysis of the Pictographic Elements in Picto-phonetic Characters]. In: Su Peicheng 2002a (ed.), 126-140.

Koda, K., Zehler, A. M. (eds.) 2008. *Learning to Read Across Languages. Cross-Linguistic Relationships in First- and Second-Language Literacy Development.* New York & London: Routlege.

Köhler, R. (ed.). 2002. *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik.* Universität Trier, Linguistische Datenverarbeitung.

___ 2005. Gegenstand und Arbeitsweise der Quantitativen Linguistik. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), 1-15.

___ 2008. Quantitative Analysis of Writing Systems: An Introduction. In: Altmann, G., Fan, F. (eds.), 3-9.

Köhler, R., Altmann, G., Piotrowski, R. G. (eds.). 2005. *Quantitative Linguistik - Quantitative Linguistics – Ein Internationales Handbuch – An International Handbook.* Berlin – New York: Walter de Gruyter.

Kordek, N. 2012. Segmentotactics of Mandarin Chinese. *Rocznik Orientalistyczny* LXV.1: 107-119.

Kułacka, A. 2011. *Statystyczne prawa językowe. Na przykładzie prawa Menzeratha-Altmanna w składni języków polskiego i angielskiego.* Wrocław: Oficyna Wydawnicza ATUT – Wrocławskie Wydawnictwo Oświatowe.

Künstler, M. J. 1970. *Pismo chińskie.* Warszawa: PIW.

Lai, P. K., Yeung, D. Y., Pong, M. C. 1996. A Heuristic Search Approach to Chinese Glyph Generation Using Hierarchical Character Composition. *Computer Processing of Oriental Languages* 10.3: 307-323.

Landsberg, M. E. 1980. The Icon in Semiotic Theory. *Current Anthropology* 21.1: 93-95.

黎纪 (Li Ji). 1995. 汉字结构分析法在汉语教学中地位和作用[Status and Function of the Analysis of Chinese Characters Structure in Chinese Language Teaching]. *武当学刊 (哲学社会学版)* [*Journal of Wudang (Philosophy and Social Sciences)*]15.1: 82-86.
(http://cd2.wudisk.com/filestores/010%E5%AD%A6%E6%9C%AF%E7%A0%94%E7%A9%B6/pdf41/%E6%B1%89%E5%AD%97%E7%BB%93%E6%9E%84%E5%88%86%E6%9E%90%E6%B3%95%E5%9C%A8%E6%B1%89%E5%AD%97%E6%95

%99%E5%AD%A6%E4%B8%AD%E7%9A%84%E5%9C%B0%E4%BD%8D%E5%92%8C%E4%BD%9C%E7%94%A8.pdf, accessed 29.05.2012)

Li, W.. 2011. Fitting Chinese Syllable-to-Character Mapping Spectrum by the Beta Rank Function. *Physica A* 391: 1515-1518. (http://www.nslij-genetics.org/wli/pub/physicaa11-aam.pdf, accessed 26.09.2012)

___ 2012. Characterizing Ranked Chinese Syllable-to-Character Mapping Spectrum: A Bridge Between the Spoken and Written Chinese Language. (http://arxiv.org/pdf/1205.1564.pdf, accessed 26.09.2012)

李燕 (Li Yan), 康加深 (Kang Jiashen). 1993. 现代汉语形声字形符研究 [Analysis of the Pictographic Elements in Picto-phonetic Characters]. In: Su Peicheng 2002a (ed.), 141-154.

李翊綺 (Li Yi-Chi). 2009. *從筆尖現象探討中文字的字形表徵* [*The Orthographic Representations of Chinese Characters as Revealed by the Tip-of-the-Pen Phenomenon*]. National Cheng Kung University, Ph.D. diss. (http://etdncku.lib.ncku.edu.tw/ETD-db/ETD-search-c/getfile?URN=etd-0724109-225344&filename=etd-0724109-225344.pdf, accessed 15.02.2011)

李兆麟 (Li Zhaolin). 1988 (2002). 三种字频统计资料的比较 [Comparison of Three Types of Statistical Studies on Chinese Characters]. In: Su Peicheng 2002a (ed.), 43-51.

Lin, J., Lin. F, 2012. Unicode Han Character Lookup Service Based on Similar Radicals. *International Journal of Smart Home* 6.3: 99-106. (http://www.sersc.org/journals/IJSH/vol6_no3_2012/13.pdf, accessed 18.12.2012)

林联和 (Lin Lianhe). 1980 (2002). 关于汉字统计特征的几个问题 [On Some Aspects of Statistical Characteristics of Chinese Characters]. In: Su Peicheng 2002a (ed.), 227-241.

林民 (Lin Min), 宋柔 (Song Rou). 2007. 汉字字形形式话描述方法研究 [Formal Description of Chinese Character Glyph]. *计算机科学* [*Computer Science*] 34.11: 185-188. (http://file.lw23.com/9/94/946/9467b781-5b8a-4a9f-ab59-09c9c2a4f404.pdf, accessed 16.11.2011)

Linell, P. 2005. *The Written Language Bias in Linguistics. Its nature, origins and transformations.* New York & London: Routledge.

Liu, C. 2008. Handwritten Chinese Character Recognition: Effects of Shape Normalization and Feature Extraction. In: Doerman, D., Jaeger, S. (eds.), 104-128.

Liu, C., Lin, J. 2008. Using Structural Information for Identifying Similar Chinese Characters. *Proceedings of ACL-08: HLT, Short Papers (Companion Volume),* 93–96. (http://aclweb.org/anthology/P/P08/P08-2024.pdf, accessed 09.04.2011)

Liu, C., Lai, M., Chuang, Y., Lee, C. 2010. Visually and Phonologically Similar Characters in Incorrect Simplified Chinese Words. *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*: 739-747. (http://dl.acm.org/citation.cfm?id=1944651, accessed 09.04.2011)

Lo, M., Hue, C., Tsai, F. 2007. Chinese Readers' Knowledge of How Chinese Orthography Represents Phonology. *Chinese Journal of Psychology* 49: 315-334.

Lu Q., Chan S. T., Li Y., Li N. L. 2002. Decomposition for ISO/IEC 10646 Ideographic Characters. *COLING '02 Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization - Volume 12*, 1-7. (http://dl.acm.org/citation.cfm?doid=1118759.1118768, accessed 09.04.2011)

Lunde, K. 2008. *CJKV Information Processing.* Sebastopol: O'Reilly.

Lupker, S. J. (1982). The Role of Phonetic and Orthographic Similarity in Picture-word Interference. *Canadian Journal of Psychology* 36: 349-367.

吕洁 (Lü Jie). 2006. *汉字学十讲* [*Ten Lectures on Chinese Characterology*]. 上海: 学林出版社 (Shanghai: Xue Lin Chubanshe).

吕永进 (Lü Youngjin). 2004. *汉字规范. 形音义* [Standarization of Chinese Characters. Form, Sound and Meaning]. 上海辞书出版社 (Shanghai: Shanghai Cishu Chubanshe).

Manning, C. D., Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge – London: The MIT Press.

Matsunaga, S. 1994. *The Linguistic and Psycholinguistic Nature of Kanji: Do Kanji Represent and Trigger Only Meanings?* University of Hawaii, Ph.D. diss.

Mattingly, I. G. 1992. Linguistic Awareness and Orthographic Form. In: Frost, R., Katz, L. (eds.), 11-26.

Menzel, C. 2002. Das synergetische Basismodell der Lexik und die chinesische Schrift. In: Köhler, R. (ed.), 179-207.

Morioka, T. 2008. CHISE: Character Processing Based on Character Ontology. In: Takenobu Tokunaga, T., Ortega, A. (eds.): *Large-Scale Knowledge Resources. Construction and Application*. Lecture Notes in Computer Science 4938: 148-162, Berlin – Heidelberg: Springer.

Ney, H. 2005. The Statistical Approach to Natural Language Processing. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), 808-820.

Nisbett, R. 2003. *The Geography of Thought*. New York: The Free Press.

Nöth, W. 1995. *Handbook of Semiotics.* Bloomington and Indianapolis: Indiana University Press.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Packard, J. L. 2000. *The Morphology of Chinese. A Linguistic and Cognitive Approach.* Cambridge: Cambridge University Press.

Peebles, D. G. 2007. *SCML: A Structural Representation for Chinese Characters*. Dartmouth College Technical Report TR2007-592.
(http://www.cs.dartmouth.edu/reports/TR2007-592.pdf, accessed 07.04.2011)

Perfetti, C. A., Liu, Y. 2006. Reading Chinese Characters: Orthography, Phonology, Meaning, and the Lexical Constituency Model. In: Li, P., Tan, L., Bates, H. E. & O. J. L. Tzeng (eds.), *Handbook of East Asian Psycholinguistics*, 225-236. New York: Cambridge University Press.

Peust, C. 2006. Script Complexity Revisited. *Glottometrics* 12: 11-15.

Partee, B. H., ter Meulen, A., Wall, R. E. 1990. *Mathematical Methods in Linguistics.* Studies in Linguistics and Philosophy 30, Dordrecht – Boston – London: Kluwer.

Prün, C. 1994. Validity of Menzerath-Altmann's Law: Graphic Representation of Language, Information Processing and Synergetic Linguistics. *Journal of Quantitative Linguistics* 1(2): 148-155.

Pulgram, E. 1976. The Typologies of Writing Systems. In: Haas, W. (ed.), 1-28.

钱乃荣 (Qian Nairong) 1990. 字符的分类和归类 [Classification and Categorization of Characters]. In: Su Peicheng 2002a (ed.), 118-125.

裘锡圭 (Qiu Xigui) 1988. *文字学概要* [*Outline of Characters Study*]. 北京商业印书馆 (Beijing: Shangye Yinshuguan).

Rankin, B. K. 1965. *A Linguistic Study of the Formation of Chinese Characters*. University of Pennsylvania, Ph.D. diss.

Rankin, B. K., Siegel, S., McClelland, A., Tan, J. L. 1966. *A Grammar for Component Combination in Chinese Characters*. Washington, D.C.: United States Department of Commerce, Technical Note 292.

Rankin, B. K., Tan, J. L. 1970. *Component Combination and Frame-Embedding in Chinese Character Grammars.* Washington, D.C.: United States Department of Commerce, Technical Note 492.

Robinson, A. 2009. *Writing and Script. A Very Short Introduction.* New York: Oxford University Press.

Rogers, H. 1995. Optimal Orthographies. In: Taylor, I., Olson, D. R. (eds.), 31-43.

___ 2005. *Writing Systems: A Linguistic Approach*. Oxford: Blackwell.

Sampson, G. 1985. *Writing Systems. A Linguistic Introduction*. London: Hutchinson.

___ 1994. Chinese Script and the Diversity of Writing Systems. *Linguistics* 32, 117–32.

沙宗元 (Sha Zongyuan). 2004. 百年来文字学通论性著作关于汉字结构研究的综述 [Summary of Hundred Years of Theoretical Graphemic Studies on Chinese Characters Structure]. *安徽大学学报(哲学社会科学版)* [*Journal of Anhui University (Philosophy and Social Sciences)*] 28.2: 126-129.

佘延 (She Yan). 1997. 20世纪汉字结构的理论研究 [Research on the Theory of Chinese Character Structure in the 20th Century]. *汉字文化* [*Chinese Characters and Culture*] 3: 17-22.

沈克成, 沈迦 (Shen Kelong, Shen Jia). 1998. *汉字部件学* [*Theory of Chinese Characters Composition*]. 北京: 机械工业出版社 (Beijing: Jixie Gongye Chubanshe).

宋业瑾，贾娇燕 (Song Yejin, Jia Jiaoyan). 2003. *实用汉字* [*Functional Study of Chinese Characters*]. 合肥: 安徽教育出版社 (Hefei: Anhui Jiaoyu Chubanshe).

Stallings, W. 1975. The Morphology of Chinese Characters. A Survey of Models and Applications. *Computers and the Humanities* 1: 13-24.

Stalph, J. 1989. *Grundlagen einer Grammatik der sinojapanischen Schrift*. Wiesbaden: Harrasowitz Verlag.

苏培成 (Su Peicheng). 2001. *现代汉字学纲要* [*The Outline of Modern Chinese Characteorology*]. 北京大学出版社 (Beijing: Beijing Daxue Chubanshe).

___ 2002a (ed.). *现代汉字学纲要. 参考资料* [*The Outline of the Study of Modern Chinese Characters. Reference Materials*]. 北京大学出版社 (Beijing: Beijing Daxue Chubanshe).

___ 2002b. 汉字的性质 [On the Nature of Chinese Characters]. In: Su Peicheng 2002a (ed.), 7-23.

Sun, C. 2006. *Chinese: A Linguistic Introduction*. Cambridge – New York – Melbourne – Madrid – Cape Town – Singapore – São Paulo: Cambridge University Press.

Taft, M., Zhu, X., Peng, D. 1999. Positional Specificity of Radicals in Chinese Character Recognition. *Journal of Memory and Language* 40.4: 498-519.

Tan, L. H., Laird, A. R., Li, K., Fox, P. T. 2005a. Neuroanatomical Correlates of Phonological Processing of Chinese Characters and Alphabetic Words: A Meta-Analysis. *Human Brain Mapping* 25: 83–91.

Tan, L.H., Spinks, J. A., Eden, G. F., Perfetti, Ch. A., Siok, W. 2005b. Reading Depends on Writing, in Chinese. *Proceedings of the National Academy of Sciences of the United States of America.* 102(24): 8781–8785.
(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1150863/, accessed 17.07.2010)

Taylor, I., Olson, D. R. (eds.). 1995. *Scripts and Literacy. Reading and Learning to Read Alphabets, Syllabaries and Characters.* Dordrecht: Kluwer.

Taylor, I., Taylor, M. 1995. *Writing and Literacy in Chinese, Korean and Japanese*. Amsterdam: John Benjamins.

Taylor, J. K., Cihon, C. 2004. *Statistical Techniques for Data Analysis*. Boca Raton – London – New York – Washington D.C.: Chapman & Hall/CRC.

Taft, M., Zhu, X. 1997. Submorphemic Processing in Reading Chinese. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 23(3): 761-775.

Thompson, R. M. 1980. *A Descriptive Procedure for Coding and Decoding Chinese Ideographs.* Indiana University, Ph.D diss.

Trager, G. L. 1974. Writing and Writing Systems. In: Thomas Sebeok (ed.), *Current Trends in Linguistics,* volume XII: *Linguistics and Adjacent Arts and Sciences*, 373–496. The Hague: Mouton Publishers.

Unger, J. M. 1990. The Very Idea: The Notion of Ideogram in China and Japan. *Monumenta Nipponica* 45: 391–411.

___ 1991. Minimum Specifications for Japanese and Chinese Alphanumeric Workstations. In: V. H. Mair and Y.Q. Liu (eds.) *Characters and Computers.* Amsterdam: IOS Press, 131-140.

___ 2004. *Ideogram: Chinese Characters and the Myth of Disembodied Meaning*. Hawaii: University of Hawaii Press.

Vanderschot, L. 2011. *Word Processing in Languages Using Non-Alphabetic Scripts. The Cases of Japanese and Chinese.* Netherlands Graduate School of Linguistics / Landelijke – LOT.

王凤阳 (Wang Fengyang). 1989. 频率与汉字的简化 [Frequency and the Simplification of Chinese Characters]. In: Su Peicheng 2002a (ed.), 52-68.

王惠 (Wang Hui). 2006. 新加坡华语教材用字的频率与分布 [A Quantitative Analysis on the Chinese Characters in Singapore Textbook Corpus]. *Journal of Chinese Language and Computing* 16.4: 239-252.

Wang, J. Ch. 1983. *Toward a Generative Grammar of Chinese Character Structure and Stroke Order.* University of Wisconsin-Madison, Ph.D diss.

Wang, J., Inhoff, A. W., Chen, H.-Ch. (ed.) 1999. *Reading Chinese Script. A Cognitive Analysis.* Mahwah – London: Lawrence Erlbaum Associates.

Wang, M., Yang, C. 2008. Learning to Read Chinese. Cognitive Consequences of Cross-language and Writing System Differences. In: Koda, K., Zehler, A. M. (eds.), *Learning to Read Across Languages Cross-Linguistic Relationships in First- and Second-Language Literacy Development*, 125-153. New York and London: Routledge.

王宁 (Wang Ning), 王立军 (Wang Lijun), 齐元涛 (Qi Yuantao), 宋继华 (Song Jihua), 陈淑梅 (Chen Shumei) 2001. *汉字学概要* [*Outline of Chinese Characterology*]. 北京: 北京师范大学出版社 (Beijing: Beijing Shifan Daxue Chubanshe).

王宁 (Wang Ning). 2002. *汉字构形学讲座* [*Lectures on the Theory of Chinese Characters Formation*]. 上海教育出版社 (Shanghai: Jiaoyu Chubanshe).

王作新 (Wang Zuoxin). 1999a. *汉字结构系统与传统思维方式* [*System of Chinese Character Structures and the Traditional Chinese Thought*]. 武汉: 武汉出版社 [Wuhan: Wuhan Chubanshe].

___ 1999b. 汉字结构的成分位序 [Ordering of Components in the Structure of Chinese Characters]. *湖北三峡学院学报* [*Journal of Hubei Three Gorges University*] 21.1: 65-69.

Wierzchoń, P. 2004a. *Gramatyka diakrytologiczna. Studium ortograficzno-kwantytatywne.* Poznań: Wydawnictwo Naukowe UAM.

___ 2004b, Z zagadnień diakryzy w języku koreańskim – perspektywa morfosyntaktyczna. The Cross-Cultural Conflicts of the Eastern & Balkan Europe after 10 Years of Democratization, 155–178. Seoul: HUFS Press.

___ 2005. Podstawy gramatyki diakrytologicznej. *Speech and Language Technology* 8: 29–63.

Wu, S., Zheng, S. 2009. A Structure Character Modeling for Chinese Character Glyph Description. *2009 International Conference on Electronic Computer Technology,* 245-248.

晓东 (Xiao Dong). 1994. 现代汉字独体与合体的再认识 [Towards a Better Understanding of Distinctions between the Simple and Compound Characters]. In: Su Peicheng 2002a (ed.), 111-117.

Xiao, H. 2008. On the Applicability of Zipf's Law in Chinese Word Frequency Distribution. *Journal of Chinese Language and Computing* 18.1: 33-46.

Xing, H., Shu, H., Li, P. 2004. The Acquisition of Chinese Characters: Corpus Analyses and Connectionist Simulations. *Journal of Cognitive Science* 5: 1-49.

Xue Y., Gu Y. 2012. A Survey on Digital Description of Chinese Character Glyph. *International Journal of Knowledge and Language Processing* 3.2: 39-42.

尹斌庸 (Yin Binyong). 1984."多余度"与文字优劣 ["Redundancy" and the Strengths and Weaknesses of Script]. In: Su Peicheng 2002a (ed.), 242-252.

Yin, B., Rohsenow, J. S. 1994. *Modern Chinese Characters.* Beijing: Sinolingua.

Yin, W. 1991. *On Reading Chinese Characters – A Neuropsychological and Experimental Study*. University College London, Ph.D diss.
(http://discovery.ucl.ac.uk/1348979/1/337404.pdf, accessed 23.09.2010)

Yiu, C. L. K., Wong, W. 2003. Chinese Character Synthesis Using METAPOST. *TUGboat 24.1 Proceedings of the 2003 Annual Meeting*: 85-93.
(http://www.tug.org/TUGboat/Articles/tb24-1/yiu.pdf, accessed 15.08.2011)

Zajdler, E. 2008. Pismo chińskie i klucze semantyczne jako nośnik znaczenia leksykalnego. In: Pawlak, N. (ed.), *Języki Azji i Afryki w komunikacji międzykulturowej*, 113-124. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

___ 2012. Graficzna reprezentacja świata w piśmie chińskim. In: Kardela, H., Muszyński Z., Rajewski, M. (eds.), *Empatia, obrazowanie, kontekst jako kategorie kognitywistyczne*, 223-237. Lublin: Wydawnictwo UMCS.

张普 (Zhang Pu). 1992. 步入信息社会的汉语和汉字 [Chinese Language and Chinese Characters at the Beginning of the Information Society]. In: Su Peicheng 2002a (ed.), 260-286.

张轴材 (Zhang Zhoucai) (ed.) 2008. *CJK 汉字构件集. CJK Component Set. 文件汇编* [*Collection of Documents*]. 北京书同文数字化技术有限公司 [Beijing Shu Tongwen Shuzihua Jishu Youxian Gongsi].
(http://hanzi.unihan.com.cn/CoolHanzi/data/download.asp?file=INFO%5FFOR%5FCJKDecomposed20080425V312.pdf, accessed 05.06.2011)

张玉金 (Zhang Yujin). 2000. 论汉字的部件拆分和字符拆分 [On the Decomposition of Chinese Characters into Components and Symbols]. *辽宁师范大学学报 (社會科學版)* [*Journal of Liaoning Normal University (Social Sciences)*] 4: 67-69.

Zhao, S., Zhang, D. 2007. The Totality of Chinese Characters – A Digital Perspective. *Journal of Chinese Language and Computing* 17.2: 107-125. (http://www.colips.org/journal/volume17/JCLC_2007_V17_N2_04.pdf, accessed 04.02.2011)

Zhao, S., Baldauf, R. B. 2008. *Planning Chinese Characters. Reaction, Evolution or Revolution?*. Dordrecht: Springer.

周殿生 (Zhou Diansheng). 2008. 汉字结构中的信息和对外汉字教学 [Information Embedded in Chinese Characters Structure and Chinese Teaching to International Students]. *新疆大学学报 (哲学人文社会科学版)* [*Journal of Xinjiang University (Philosophy, Humanities and Social Sciences)*] 36.3: 136-138. (http://file.lw23.com/3/33/332/332860ff-7776-49b5-8e5d-bb729f997cbe.pdf, accessed 09.10.2011)

周有光 (Zhou Youguang). 1980 (2002). 现代汉字学发凡 [*Introduction to Modern Chinese Characterology*]. In: Su Peicheng 2002a (ed.), 24-33.

___ 1984 (2002). 现代汉语用字的定量问题 [Problem of Quantity of Characters in Modern Chinese]. In: Su Peicheng 2002a (ed.), 69-80.

朱德熙 (Zhu Dexi). 1988. 在“汉字问题学术讨论会”开幕式上的发言 [Speech at the Opening Ceremony of “Symposium on Chinese Characters”]. In: Su Peicheng 2002a (ed.), 1-6.

WWW references

http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg34/IRGN1646Confirmed.doc (accessed 17.12.2012)

http://cbflabs.com (accessed 17.07.2010)

http://cbflabs.com/book/dnahtml/dnabase/dnabase01.htm (accessed 17.07.2010)

http://cbflabs.com/book/gif_cg/gif_cg/index.html (accessed 17.07.2010)

http://ccl.pku.edu.cn:8080/ccl_corpus/ (accessed 08.03.2011)

http://ccl.pku.edu.cn:8080/ccl_corpus/CCL_CC_Sta_Xiandai.pdf (accessed 08.03.2011)

http://cdp.sinica.edu.tw/cdphanzi/ (accessed 02.02.2011)

http://cdp.sinica.edu.tw/cdphanzi/documents/history1010417.pdf (accessed 08.03.2011)

http://code.google.com/p/cjklib/ (accessed 10.05.2010)

http://code.google.com/p/cjklib/wiki/Features (accessed 10.05.2010)

http://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition (accessed 10.04.2012)

http://corpus.leeds.ac.uk/frqc/i-zh-char.num (accessed 12.11.2011) http://ctext.org/dictionary.pl?if=en&char=%F0%A0%88%8C (accessed 21.08.2011)

http://dict.variants.moe.edu.tw/ex.htm (accessed 10.05.2011)

http://en.wikipedia.org/wiki/ Cangjie_input_method (accessed 30.04.2010)

http://en.wikipedia.org/wiki/GB_18030 (accessed 18.01.2011)

http://en.wikipedia.org/wiki/GB_2312 (accessed 18.01.2011)

http://en.wiktionary.org/wiki/Wiktionary:Chinese_Cangjie_index (accessed 28.07.2011)

http://glyphwiki.org (accessed 11.04.2010)

http://kanji-database.sourceforge.net/ (accessed 15.03.2011)

http://kanji-database.sourceforge.net/ids/ids-analysis.html?lang=en (accessed 15.03.2011)

http://lingua.mtsu.edu/chinese-computing/statistics/char/list.php?Which=MO (acessed 09.12.2011)

http://open-lit.com/bookindex.php?gbid=311 (accessed 15.03.2012)

http://proj1.sinica.edu.tw/~cdp/service/documents/T960419.pdf (accessed 02.02.2011)

http://zh.wikipedia.org/zh-hant/%E5%AD%97%E9%A6%96 (accessed 12.04.2011)

http://zh.wikipedia.org/zh-hant/%E5%80%89%E9%A0%A1%E8%BC%B8%E5%85%A5%E6%B3%95#.E8.88.87.E5.85.B6.E4.BB.96.E5.BD.A2.E7.A2.BC.E8.BC.B8.E5.85.A5.E6. B3.95.E7.9A.84.E6.AF.94.E8.BC.83 (accessed 12.04.2011)

http://zh.wikipedia.org/zh/%E7%AC%94%E7%94%BB (accessed 12.04.2011)

http://www.accu.or.jp/litdbase/policy/chn/index.htm (accessed 12.04.2011)

http://www.baike.com/wiki/%E4%BB%8A%E6%98%94%E6%96%87%E5%AD%97%E9%95%9C

http://www.chinesecj.com/newlearncj/cj5/cj3.php (accessed 19.06.2012)

http://www.chise.org/ (accessed 13.02.2011)

http://www.cns11643.gov.tw (accessed 03.03.2011)

http://www.cns11643.gov.tw/AIDB/encodings_en.do#encord1 (accessed 03.03.2011)

http://www.edu.tw/files/site_content/M0001/86news/86rest6.html?open (accessed 06.05.2012)

http://www.edu.tw/files/site_content/M0001/86news/ch2.html?open (accessed 06.05.2012)

http://www.hanglyph.com/en/hanglyph-index.shtml (accessed 19.06.2012)

http://www.hkpe.net/cj/cjtable.htm (accessed 24.02.2012)

http://www.meijigakuin.ac.jp/~pmjs/archive/2000/mojikyo.html (accessed 13.02.2011)

http://www.unicode.org/charts/PDF/U2FF0.pdf (accessed 27.08.2012)

http://www.unicode.org/charts/PDF/U31C0.pdf (accessed 27.08.2012)

http://www.unicode.org/versions/Unicode6.2.0/ch12.pdf (accessed 19.02.2013)

www.wenlin.com (accessed 10.09.2010)

http://www.wenlin.com/cdl/#stat (accessed 05.02.2012)

http://www.wenlin.com/cdl/cdl_strokes_2004_05_23.pdf (accessed 10.09.2010)
http://www.zsjy.gov.cn/yywz/yypg/gfwj/17.htm (accessed 19.06.2011)
https://zh.wikipedia.org/zh/%E5%AD%97%E9%A6%96 (accessed 15.10.2011)

WWW access to selected Chinese character standards and character sets

GF 3001-1997 信息处理用 GB 13000.01 字符集一汉字部件规范 [Chinese Character Component Standard of GB 13000.01 Character Set for Information Processing]: http://www.shyedu.gov.cn/level3.jsp?id=132 (accessed 15.02.2012)

GF 0014-2009 现代常用字部件及部件名称规范 [Specification of Common Modern Chinese Character Components and Component Names]: http://www.china-language.gov.cn/standard/%E6%B1%89%E5%AD%97%E9%83%A8%E4%BB%B6.pdf (accessed 19.08.2012)

通用规范汉字表 [List of Commonly Used Standardized Chinese Characters]: http://www.china-language.gov.cn/doc/zb2009.pdf (accessed 10.01.2013)

CNS 11643-2: http://www.spsp.gov.cn/DataCenter/Standard/PDFView.aspx?ca=7oYlA1SbRLs= (accessed 05.05.2012)

Appendix I – Chinese Documents Processing Lab (CDP) basic components list – ordered by frequency[348]

[348] Chuang & Teng 2009: 42-47.

No.	Reference N0.	Compo-nent	No. of characters	Frequency	No.	Reference No.	Compo-nent	No. of characters	Frequency
1	98	口	2526	6.29%	39	183	扌	515	0.64%
2	1	一	1732	6.03%	40	144	王	330	0.64%
3	38	人	955	2.56%	41	137	阝	306	0.63%
4	172	日	895	2.51%	42	184	攵	263	0.63%
5	39	亻	551	2.47%	43	73	厶	278	0.62%
6	5	丶	308	2.04%	44	27	丁	149	0.59%
7	82	土	795	1.82%	45	391	門	128	0.57%
8	281	白	191	1.72%	46	353	糸	324	0.57%
9	52	勹	292	1.64%	47	58	丷	398	0.56%
10	40	八	713	1.57%	48	68	刂	206	0.54%
11	18	十	572	1.48%	49	3	丨	90	0.54%
12	153	木	956	1.42%	50	83	士	188	0.54%
13	373	言	364	1.24%	51	35	𠂉	243	0.53%
14	55	亠	590	1.22%	52	323	月	453	0.53%
15	211	辶	236	1.11%	53	367	我	26	0.52%
16	99	囗	262	1.08%	54	158	不	39	0.52%
17	233	氵	855	1.00%	55	77	了	6	0.51%
18	70	又	483	1.00%	56	49	刀	158	0.50%
19	85	大	505	0.97%	57	60	冖	267	0.49%
20	171	龰	180	0.95%	58	149	𣥂	83	0.48%
21	22	匕	321	0.93%	59	118	䒑	229	0.48%
22	117	宀	450	0.92%	60	170	止	330	0.48%
23	198	月	190	0.91%	61	278	禾	264	0.48%
24	21	𠂇	138	0.89%	62	397	隹	271	0.48%
25	45	儿	408	0.85%	63	31	卜	135	0.46%
26	131	小	166	0.83%	64	256	田	529	0.46%
27	128	女	442	0.83%	65	113	夕	121	0.44%
28	92	寸	227	0.82%	66	366	貝	322	0.44%
29	4	丿	390	0.75%	67	116	广	264	0.44%
30	133	子	161	0.73%	68	44	乂	137	0.44%
31	214	心	274	0.72%	69	332	自	180	0.43%
32	140	幺	132	0.72%	70	69	力	152	0.42%
33	127	也	34	0.71%	71	42	入	42	0.41%
34	32	冂	350	0.68%	72	254	目	252	0.40%
35	166	戈	167	0.66%	73	132	⺌	170	0.39%
36	106	彳	137	0.66%	74	84	工	230	0.39%
37	168	艹	873	0.65%	75	261	罒	234	0.38%
38	288	立	242	0.64%	76	17	二	153	0.37%

77	209	⺍	554	0.37%
78	50	𠂉	121	0.34%
79	101	巾	150	0.34%
80	331	⺮	260	0.33%
81	10	𠃌	41	0.32%
82	155	朩	45	0.30%
83	311	至	53	0.30%
84	51	⺈	356	0.29%
85	362	里	50	0.28%
86	79	干	160	0.28%
87	67	刀	195	0.28%
88	109	夂	110	0.27%
89	192	斤	105	0.27%
90	301	耳	125	0.27%
91	81	丰	52	0.27%
92	208	火	359	0.27%
93	20	厂	245	0.27%
94	47	几	127	0.26%
95	348	艮	53	0.26%
96	216	小	282	0.26%
97	355	車	200	0.25%
98	344	𦍌	73	0.24%
99	334	𦥑	57	0.24%
100	304	覀	86	0.24%
101	276	生	38	0.24%
102	147	主	69	0.24%
103	365	見	69	0.24%
104	340	衤	153	0.23%
105	399	金	493	0.22%
106	318	虫	448	0.22%
107	119	尸	183	0.22%
108	110	攵	128	0.21%
109	181	手	38	0.21%
110	194	⺳	154	0.21%
111	177	中	45	0.21%
112	123	弓	90	0.21%
113	360	豕	97	0.20%
114	196	戶	73	0.20%
115	30	卜	25	0.20%
116	86	尢	36	0.20%
117	407	重	23	0.20%
118	225	巴	58	0.19%
119	253	以	5	0.19%
120	63	冂	58	0.19%
121	87	廾	125	0.18%
122	121	巳	61	0.18%
123	159	犬	86	0.18%
124	152	廿	113	0.18%
125	284	用	29	0.18%
126	342	米	148	0.18%
127	197	父	38	0.17%
128	411	首	7	0.17%
129	333	𦥑	105	0.17%
130	100	山	352	0.17%
131	74	マ	89	0.17%
132	251	戊	71	0.17%
133	205	勿	70	0.16%
134	66	凵	93	0.16%
135	269	业	43	0.16%
136	321	冎	66	0.16%
137	415	馬	163	0.16%
138	275	㠯	51	0.15%
139	188	生	60	0.15%
140	187	牛	90	0.15%
141	146	夫	109	0.14%
142	59	⺀	50	0.14%
143	352	糸	111	0.14%
144	314	虍	130	0.14%
145	41	儿	55	0.14%
146	169	丬	77	0.14%
147	315	且	48	0.14%
148	203	𧘇	44	0.14%
149	277	乍	31	0.13%
150	412	為	12	0.13%
151	105	千	30	0.13%
152	231	母	31	0.13%
153	107	彡	166	0.12%
154	298	癶	44	0.12%
155	126	屮	83	0.12%
156	163	五	27	0.12%

157	151	廿	77	0.12%	197	295	民	20	0.08%
158	46	几	89	0.12%	198	388	果	38	0.08%
159	260	皿	178	0.12%	199	14	乛	79	0.08%
160	339	衣	90	0.12%	200	303	臣	58	0.08%
161	383	事	3	0.12%	201	7	乙	37	0.08%
162	242	礻	82	0.12%	202	33	冂	84	0.08%
163	384	雨	133	0.12%	203	257	由	42	0.08%
164	13	𠃊	89	0.11%	204	435	黑	47	0.08%
165	223	爿	38	0.11%	205	160	犭	143	0.08%
166	255	且	55	0.11%	206	207	文	53	0.08%
167	327	年	3	0.11%	207	305	西	28	0.08%
168	222	弔	20	0.11%	208	75	九	24	0.08%
169	302	甚	40	0.11%	209	317	曲	24	0.08%
170	141	巛	79	0.11%	210	96	⺕	27	0.07%
171	241	示	90	0.11%	211	393	非	46	0.07%
172	248	石	219	0.11%	212	93	弋	38	0.07%
173	173	曰	79	0.11%	213	369	身	17	0.07%
174	239	未	43	0.11%	214	382	東	14	0.07%
175	290	永	8	0.10%	215	361	求	20	0.07%
176	377	長	15	0.10%	216	359	酉	134	0.07%
177	212	之	13	0.10%	217	221	夬	20	0.07%
178	232	水	47	0.10%	218	238	丑	65	0.07%
179	324	夕	44	0.10%	219	410	食	73	0.06%
180	349	艮	51	0.10%	220	26	七	7	0.06%
181	357	更	18	0.10%	221	29	与	27	0.06%
182	296	皮	37	0.10%	222	161	歹	70	0.06%
183	15	㇇	80	0.09%	223	394	無	18	0.06%
184	404	面	12	0.09%	224	90	⺧	91	0.06%
185	343	羊	55	0.09%	225	358	束	67	0.06%
186	65	丩	9	0.09%	226	122	已	1	0.06%
187	94	去	30	0.09%	227	347	圭	7	0.06%
188	265	冉	31	0.09%	228	186	气	32	0.06%
189	268	冊	59	0.09%	229	175	冃	31	0.06%
190	120	己	18	0.09%	230	62	工	14	0.06%
191	341	亥	27	0.08%	231	335	舟	72	0.06%
192	245	本	10	0.08%	232	287	疒	157	0.06%
193	124	彐	55	0.08%	233	264	兄	38	0.06%
194	200	氏	52	0.08%	234	230	毌	9	0.06%
195	28	丂	75	0.08%	235	114	丸	33	0.06%
196	309	而	67	0.08%	236	370	鳥	245	0.06%

237	130	𠤎	71	0.06%	277	53	匸	34	0.03%
238	346	聿	31	0.05%	278	325	缶	51	0.03%
239	381	叀	34	0.05%	279	148	龶	16	0.03%
240	206	及	22	0.05%	280	210	⺍	46	0.03%
241	95	才	7	0.05%	281	263	央	33	0.03%
242	405	禺	37	0.05%	282	371	采	54	0.03%
243	154	朩	36	0.05%	283	437	齊	25	0.03%
244	228	尹	24	0.05%	284	428	婁	35	0.03%
245	351	羽	157	0.05%	285	220	尹	16	0.03%
246	108	㐅	14	0.05%	286	235	氺	26	0.03%
247	244	世	41	0.05%	287	328	缶	6	0.03%
248	423	莫	14	0.05%	288	218	聿	27	0.03%
249	129	廴	49	0.05%	289	61	コ	52	0.03%
250	167	牙	26	0.05%	290	199	夕	1	0.03%
251	338	豸	29	0.05%	291	185	毛	59	0.03%
252	289	必	33	0.05%	292	424	帶	12	0.03%
253	307	吏	2	0.05%	293	280	丘	9	0.03%
254	297	丱	4	0.05%	294	71	ㄟ	16	0.03%
255	433	單	39	0.05%	295	418	鬼	46	0.03%
256	226	尸	19	0.05%	296	202	小	21	0.03%
257	97	少	40	0.04%	297	189	丰	51	0.03%
258	57	冫	34	0.04%	298	345	州	8	0.03%
259	178	田	9	0.04%	299	413	飛	2	0.03%
260	409	食	20	0.04%	300	227	屮	73	0.03%
261	180	內	16	0.04%	301	138	乡	18	0.03%
262	64	巳	90	0.04%	302	299	矛	61	0.03%
263	25	匚	85	0.04%	303	135	彑	44	0.03%
264	165	旡	25	0.04%	304	219	戶	22	0.03%
265	271	电	1	0.04%	305	403	革	75	0.02%
266	236	夫	45	0.04%	306	380	亞	14	0.02%
267	56	刂	2	0.04%	307	23	匕	19	0.02%
268	143	川	20	0.04%	308	191	片	14	0.02%
269	9	亅	16	0.04%	309	294	弗	26	0.02%
270	356	甫	54	0.04%	310	190	丰	18	0.02%
271	72	乃	20	0.04%	311	247	丙	13	0.02%
272	259	申	16	0.04%	312	379	耳	21	0.02%
273	125	彑	45	0.03%	313	8	乚	24	0.02%
274	234	氺	50	0.03%	314	272	禸	26	0.02%
275	36	〢	22	0.03%	315	240	末	11	0.02%
276	243	甘	52	0.03%	316	204	反	12	0.02%

317	283	乎	11	0.02%	357	337	巛	13	0.01%
318	406	垂	19	0.02%	358	434	黽	17	0.01%
319	115	凡	19	0.02%	359	262	史	2	0.01%
320	376	眂	36	0.02%	360	6	乀	8	0.01%
321	37	厂	47	0.02%	361	145	井	6	0.01%
322	24	匚	34	0.02%	362	308	朿	15	0.01%
323	142	川	12	0.02%	363	392	妻	10	0.01%
324	293	巨	23	0.02%	364	104	乇	20	0.01%
325	111	久	7	0.02%	365	330	竹	1	0.01%
326	378	長	55	0.02%	366	80	于	24	0.01%
327	372	豸	35	0.02%	367	112	丂	17	0.01%
328	322	肉	10	0.02%	368	398	卑	36	0.01%
329	375	㡀	25	0.02%	369	368	冎	21	0.01%
330	416	鬥	6	0.02%	370	291	聿	15	0.01%
331	102	巾	17	0.02%	371	229	毋	6	0.01%
332	195	㠯	2	0.02%	372	201	丹	12	0.01%
333	34	冂	20	0.02%	373	420	隺	7	0.01%
334	439	龍	35	0.02%	374	12	乁	11	0.01%
335	385	丽	16	0.01%	375	438	鹵	38	0.01%
336	258	甲	16	0.01%	376	417	烏	9	0.01%
337	421	堇	19	0.01%	377	266	冊	7	0.01%
338	282	瓜	31	0.01%	378	425	曹	13	0.01%
339	174	曰	19	0.01%	379	286	尗	7	0.00%
340	326	乖	26	0.01%	380	88	丌	21	0.00%
341	414	鬲	27	0.01%	381	320	业	9	0.00%
342	400	隶	14	0.01%	382	224	丑	16	0.00%
343	252	戊	9	0.01%	383	390	典	12	0.00%
344	386	直	11	0.01%	384	363	串	5	0.00%
345	11	𠃊	12	0.01%	385	313	夷	16	0.00%
346	215	小	11	0.01%	386	426	棄	1	0.00%
347	427	畢	13	0.01%	387	162	歺	7	0.00%
348	401	承	1	0.01%	388	374	亥	2	0.00%
349	193	爪	5	0.01%	389	43	乂	7	0.00%
350	164	屯	25	0.01%	390	441	龜	2	0.00%
351	419	兼	30	0.01%	391	176	日	6	0.00%
352	250	犮	24	0.01%	392	440	羲	4	0.00%
353	279	𠂤	8	0.01%	393	270	电	20	0.00%
354	182	手	2	0.01%	394	91	巳	3	0.00%
355	89	丈	5	0.01%	395	213	冘	19	0.00%
356	237	瓦	34	0.01%	396	156	市	7	0.00%

397	150	丏	2	0.00%	420	54	七	1	0.00%
398	402	韭	19	0.00%	421	395	秉	1	0.00%
399	429	庸	10	0.00%	422	285	甩	1	0.00%
400	408	禹	9	0.00%	423	300	卍	1	0.00%
401	310	亙	5	0.00%	424	134	孑	1	0.00%
402	389	疌	14	0.00%	425	136	彐	1	0.00%
403	329	缶	7	0.00%	426	350	𦣞	3	0.00%
404	319	曳	6	0.00%	427	76	乜	1	0.00%
405	436	熏	11	0.00%	428	312	戋	1	0.00%
406	387	豖	10	0.00%	429	430	壺	1	0.00%
407	246	巿	9	0.00%	430	432	[illegible]	1	0.00%
408	354	巨	7	0.00%	431	16	乀	1	0.00%
409	273	凹	2	0.00%	432	78	巜	1	0.00%
410	48	八	3	0.00%	433	103	乇	1	0.00%
411	336	㐆	3	0.00%	434	139	孓	1	0.00%
412	274	凸	1	0.00%	435	217	[illegible]	1	0.00%
413	396	臾	9	0.00%	436	249	卌	1	0.00%
414	19	丅	7	0.00%	437	267	冊	1	0.00%
415	157	卅	1	0.00%	438	306	襾	1	0.00%
416	292	夬	4	0.00%	439	316	臾	1	0.00%
417	431	鼎	6	0.00%	440	364	弗	1	0.00%
418	179	冋	6	0.00%	441	422	重	1	0.00%
419	2	㇀	2	0.00%					

Appendix II – Big5 character set

一乙丁七乃九了二人儿入八几刀刁力匕十卜又三下丈上丫丸凡久么也乞于亡
兀刃勺千叉口土士夕大女子孑孓寸小尢尸山川工己已巳巾干廾弋弓才丑丐不中
丰丹之尹予云井互五亢仁什仃仆仇仍今介仄元允内六兮公冗凶分切刈匀勾勿化
匹午升卅卞厄友及反壬天夫太夭孔少尤尺屯巴幻廿弔引心戈户手扎支文斗斤方
日曰月木欠止歹毋比毛氏水火爪父爻片牙牛犬王丙世丕且丘主乍乏乎以付仔仕
他仗代令仙仞充兄冉冊冬凹出凸刊加功包匆北匝仟半卉卡占卯卮去可古右召叮
叩叨叼司叵叫另只史叱台句叭叻四囚外央失奴奶孕它尼巨巧左市布平幼弁弘弗
必戊打扔扒扑斥旦朮本未末札正母民氐永汁汀氾犯玄玉瓜瓦甘生用甩田由甲申
疋白皮皿目矛矢石示禾穴立丞丟乒乓乩亙交亦亥仿伉伙伊伕伍伐休伏仲件任仰
仳份企伋光兇兆先全共再冰列刑划刎刖劣匈匡匠印危吉吏同吊吐吁吋各向名合
吃后吆吒因回囝圳地在圭圬圯圩夙多夷夸妄奸妃好她如妁字存宇守宅安寺尖屹
州帆并年式弛忙忖戎戌戍成扣扛托收早旨旬旭曲曳有朽朴朱朵次此死氖汝汗汙
江池汐汕污汛汍汎灰牟牝百竹米糸缶羊羽老考而耒耳聿肉肋肌臣自至臼舌舛舟
艮色艾虫血行衣西阡串亨位住佇佗佞伴佛何估佐佑伽伺伸佃佔似但佣作你伯低
伶余佝佈佚兑克免兵冶冷别判利删刨劫助努劬匣即卵吝吭吞吾否呎吧呆呃吴呈
呂君吩告吹吻吸吮吵呐吠吼呀吱含吟听囱困囤囫坊坑址坍均坎圾坐坏圻壯夾妝
妒妨妞妣妙妖妍妤妓妊妥孝孜孚孛完宋宏尬局屁尿尾岐岑岔岌巫希序庇床廷弄
弟彤形彷役忘忌志忍忱快忸忪戒我抄抗抖技扶抉扭把扼找批扳抒扯折扮投抓抑
抆改攻攸旱更束李杏材村杜杖杞杉杆杠杓杗步每求汞沙沁沈沉沅沛汪決沐汰沌
汨沖沒汽沃汲汾汴沆汶沍沔沘沂灶灼災灸牢牡牠狄狂玖甬甫男甸皂盯矣私秀禿
究系罕肖肓肝肘肛肚育良芒芋芍見角言谷豆豕貝赤走足身車辛辰迂迆迅迄巡邑
邢邪邦那酉采里防阮阱阪阬並乖乳事些亞享京佯依侍佳使佬供例來侃佰併侈佩
佻侖佾侏侑佺兔兒兕兩具其典冽函刻券刷刺到刮制剁劾劻卒協卓卑卦卷卸卹取
叔受味呵咖呸咕咀呻呷咄咒咆呼咐呱呶和咚呢周咋命咎固垃坷坪坩坡坦坤坼夜
奉奇奈奄奔妾妻委妹妮姑姆姐姍始姓姊妯妳姒姅孟孤季宗定官宜宙宛尚屈居屆
岷岡岸岩岫岱岳帘帚帖帕帛帑幸庚店府底庖延弦弧弩往征彿彼忝忠忽念忿怏怔
怯怵怖怪怕怡性怩怫怛或戕房戾所承拉拌拄抿拂抹拒招披拓拔拋拈抨抽押拐拙
拇拍抵拚抱拘拖拗拆抬拎放斧於旺昔易昌昆昂明昀昏昕昊昇服朋杭枋枕東果杳
杷枇枝林杯杰板枉松析杵枚枓杼杪杲欣武歧歿氓氛泣注泳沱泌泥河沽沾沼波沫
法泓沸泄油況沮泗泅泱沿治泡泛泊沬泯泜泖泠炕炎炒炊炙爬爭爸版牧物狀狎狙
狗狐玩玨玟玫玥甽疝疙疚的盂盲直知矽社祀祁秉秈空穹竺糾罔羌羋者肺肥肢肱
股肫肩肴肪肯臥臾舍芳芝芙芭芽芟芹花芬芥芯芸芣芰芾芷虎虱初表軋迎返近邵
邸邱邶采金長門阜陀阿阻附陂隹雨青非亟亭亮信侵侯便俠俑俏保促侶俘俟俊俗
侮俐俄係俚俎俞侷兗冒冑冠剎剃削前剌剋則勇勉勃勁匍南卻厚叛咬哀咨哎哉咸
咦咳哇哂咽咪品哄哈咯咫咱咻咩咧咿囿垂型垠垣垢城垮垓奕契奏奎奐姜姘姿姣
姨娃姥姪姚姦威姻孩宣宦室客宥封屎屏屍屋峙峒巷帝帥帟幽庠度建弈弭彥很待
徊律徇後徉怒思怠急怎怨恍恰恨恢恆恃恬恫恪恤扁拜挖按拼拭持拮拽指拱拷拯
括拾拴挑挂政故斫施既春昭映昧是星昨昱昤曷柿染柱柔某柬架枯柵柩柯柄柑枴
柚查枸柏柞柳枰柙柢柝柒歪殃殆段毒毗氟泉洋洲洪流津洌洱洞洗活洽派洶洛泵
洹洧洸洩洮洵洎洫炫為炳炬炯炭炸炮炤爰牲牯牴狩狠狡玷珊玻玲珍珀玳甚甭畏
界畎畋疫疤疥疢疣癸皆皇皈盈盆盃盅省盹相眉看盾盼眇矜砂研砌砍祆祉祈祇禹

禹科秒秋穿突竿竽籽紂紅紀紉紇約紆缸美羿耄耐耍耑耶胖胥胚胃胄背胡胛胎胞
胤胝致舢苧范茅苣苛苦茄若茂茉苒苗英茁苜苔苑苞苓苟苯茆虐虹虻虺衍衫要觔
計訂訃貞負赴赳趴軍軌述迦迢迪迥迭迫迤迨郊郎郁郃酋酊重閂限陋陌降面革韋
韭音頁風飛食首香乘亳倌倍倣俯倦倥俸倩倖倆值借倚倒們俺倀倔倨俱倡個候倘
俳修倭倪俾倫倉兼冤冥冢凍凌准凋剖剜剔剛剝匪卿原厝叟哨唐唁唷哼哥哲唆哺
唔哩哭員唉哮哪哦唧唇哽唏圃圄埂埔埋埃埆夏套奘奚娑娘娜娟娛娓姬娠娣娩娥
娌娉孫屘宰害家宴宮宵容宸射屑展屐峭峽峻峪峨峰島崁峴差席師庫庭座弱徒徑
徐恙恣恥恐恕恭恩息悄悟悚悍悔悌悅悖扇拳挈拿捎挾振捕捂捆捏捉挺捐挽挪挫
挨捍捌效敉料旁旅時晉晏晃晒晌晅晁書朔朕朗校核案框桓根桂桔栩梳栗桌桑栽
柴桐桀格桃株桅栓栘桁殊殉殷氣氧氨氦氤泰浪涕消涇浦浸海浙涓浬涉浮浚浴浩
涌涊浹涅浥涔烊烘烤烙烈烏爹特狼狹狽狸狷玆班琉珮珠珪珞畔畝畜畚留疾病症
疲疳疽疼疹痂疸皋皰益盍盎眩真眠眨矩砰砧砸砝破砷砥砭砠砟砲祕祐祠祟祖神
祝祇祚秤秣秧租秦秩秘窄窈站笆笑粉紡紗紋紊素索純紐紕級紜納紙紛缺罟羔翅
翁耆耘耕耙耗耽耿胱脂胰脅胭胴脆胸胳脈能脊胼胯臭臬舀舐航舫舨般芻茫荒荔
荊茸荐草茵茴荏茲茹茶茗荀茱茨荃虔蚊蚪蚓蚤蚩蚌蚣蚜衰衷袁袂衽衹記訐討訌
訕訊託訓訖訏訑豈豺豹財貢起躬軒軔軏辱送逆迷退迺迴逃追逅迸邕郡郝郢酒配
酌釘針釗釜釙閃院陣陡陛陝除陘陞隻飢馬骨高鬥鬲鬼乾偺偽停假偃偌做偉健偶
偎偕偵側偷偏倏偯偭兜冕凰剪副勒務勘動匐匏匙匿區匾參曼商啪啦啄啞啡啃啊
唱啖問啕唯啤唸售啜唬啣唳啁啗圈國圉域堅堊堆埠埤基堂堵執培夠奢娶婁婉婦
婪婀娼婢婚婆婊孰寇寅寄寂宿密尉專將屠屜屝崇崆崎崛崖崢崑崩崔崙崤崧崗巢
常帶帳帷康庸庶庵庾張強彗彬彩彫得徙從徘御徠徜恿患悉悠您惋悴惦悽情悻悵
惜悼惘惕惆惟悸惚惇戚戛扈掠控捲掖探接捷捧掘措捱掩掉掃掛捫推掄授掙採掬
排掏掀捻捩捨捺敝敖救教敗啟敏敘敕敔斜斛斬族旋旌旎晝晚晤晨晦晞曹勗望梁
梯梢梓梵桿桶梱梧梗械梃棄梭梆梅梔條梨梟梡梂欲殺毫毬氫涎涼淳淙液淡淌淤
添淺清淇淋涯淑涮淞淹涸混淵淅淒渚涵淚淫淘淪深淮淨淆淄涪淬涿淦烹焉焊烽
烯爽牽犁猜猛猖猓猙率琅琊球理現琍瓠瓶瓷甜產略畦畢異疏痔痕疵痊痍皎盔盒
盛眷眾眼眶眸眺硫硃硎祥票祭移窒窕笠笨笛第符笙笞笮粒粗粕絆絃統紮紹紼絀
細紳組累終紲紱缽羞羚翌翎習耜聊聆脯脖脣脫脩脰脤舂舵舷舶船莎莞莘荸莢莖
莽莫莒莊莓莉莠荷荻荼莆莧處彪蛇蛀蚶蛄蚵蛆蛋蚱蚯蛉術袞袈被袒袖袍袋覓規
訪訝訣訥許設訟訛訢豉豚販責貫貨貪貧赧赦趾趺軛軟這逍通逗連速逝逐逕逞造
透逢逖逛途部郭都酗野釵釦釣釧釭釩閉陪陵陳陸陰陴陶陷陬雀雪雯章竟頂頃魚
鳥鹵鹿麥麻傢傍傅備傑傀傖傘傚最凱割剴創剩勞勝勛博厥啻喀喧啼喊喝喘喂喜
喪喔喇喋喃喳單喟唾喲喚喻喬喱啾喉喫喙圍堯堪場堤堰報堡堝堠壹壺奠婷媚婿
媒媛媧孳孱寒富寓寐尊尋就嵌嵐崴嵇巽幅帽幀幃幾廊廁廂廄弼彭復循徨惑惡悲
悶惠愜愣惺愕惰惻惴慨惱愎惶愉愀愒戟扉掣掌描揀揩揉揆揍插揣提握揖揭揮捶
援揪換摒揚揹敞敦敢散斑斐斯普晰晴晶景暑智晾晷曾替期朝棺棕棠棘棗椅棟棵
森棧棹棒棲棣棋棍植椒椎棉棚楮棻款欺欽殘殖殼毯氮氯氬港游湔渡渲湧湊渠渥
渣減湛湘渤湖湮渭渦湯渴湍渺測湃渝渾滋溉渙湎湣湄湲湩湟焙焚焦焰無然煮焜
牌犄犀猶猥猴猩琺琪琳琢琥琵琶琴琯琛琦琨甥甦畫番痢痛痣痙痘痞痠登發皖皓
皴盜睏短硝硬硯稍稈程稅稀窘窗窖童竣等策筆筐筒答筍筋筏筑粟粥絞結絨絕紫
絮絲絡給絢絰絳善翔翕耋聒肅腕腔腋腑腎脹腆脾腌腓腴舒舜菩萃菸萍菠菅萋菁

華菱菴著萊菰萌菌菽菲菊萸萎萄菜萇菔菟虛蛟蛙蛭蛔蛛蛤蛐蛞街裁裂袱覃視註
詠評詞証詁詔詛詐詆訴診訶詖象貂貯貼貳貽賁費賀貴買貶貿貸越超趁跎距跋跚
跑跌跛跆軻軸軼辜逮逵週逸進逶鄂郵鄉鄖酣酥量鈔鈕鈣鈉鈞鈍鈴鈇鈑閔閏開閑
間閒閎隊階隋陽隅隆隍陲隄雁雅雄集雇雯雲韌項順須飧飪飯飩飲飭馮馭黃黍黑
亂傭債傲傳僅傾催傷傻傯僇剿剷剽募勦勤勢勣匯嗟嗨嗓嗦嗎嗜嗇嗑嗣嗤嗯嗚嗡
嗅嗆嗥嗉園圓塞塑塘塗塚塔填塌塭塊塢塒塋奧嫁嫉嫌媾媽媼媳嫂媲嵩嵯幌幹廉
廈弒彙徬微愚意慈感想愛惹愁愈慎慌慓慍愾愴愧愍愆愷戡戢搓搾搞搪搭搽搬搏
搜搔損搶搖搗搆敬斟新暗暉暇暈暖暄暘暍會榔業楚楷楠楔極椰概楊楨楫楞楓楹
榆楝楣楛歇歲毀殿毓毽溢溯滓溶滂源溝滇滅溥溘溼溺溫滑準溜滄滔溪溧溴煎煙
煩煤煉照煜煬煦煌煥煞煆煨煖爺牒猷獅猿猾瑯瑚瑕瑟瑞瑁琿瑙瑛瑜當畸瘀痰瘁
痲痱痺痿痴痳盞盟睛睫睦睞督睹睪睬睜睥睨睢矮碎碰碗碘碌碉硼碑碓硿祺祿禁
萬禽稜稚稠稔稟稞窟窠筷節筠筮筧粱粳粵經絹綑綁綏絛置罩罪署義羨群聖聘肆
肄腱腰腸腥腮腳腫腹腺腦舅艇蒂葷落萱葵葦葫葉葬葛萼萵葡董葩葭葆虞虜號蛹
蜓蜈蜇蜀蛾蛻蜂蜃蜆蜊衙裟裔裙補裘裝裡裊裕裒覜解詫該詳試詩詰誇詼詣誠話
誅詭詢詮詬詹詻訾詨豢貊貉賊資賈賄貲賃賂賅跡跟跨路跳跺跪跤跦躲較載軾輊
辟農運遊道遂達逼違遐遇遏過遍遑逾遁鄒鄗酬酪酩釉鈷鉗鈸鈽鉀鈾鉛鉋鉤鉑鈴
鉉鉍鉅鈹鈿鉚閘隘隔隕雍雋雉雊雷電雹零靖靴靶預頑頓頊頒頌飼飴飽飾馳馱馴
髡鳩麂鼎鼓鼠僧僮僥僖僭僚僕像僑僱僎僩兢凳劃劂匱厭嗾嘀嘛嘗嗽嘔嘆嘉嘍嘎
嗷嘖嘟嘈嘐嗶團圖塵塾境墓墊塹墅埉壽夥夢夤奪奩嫡嫦嫩嫗嫖嫘嫣孵寞寧寡寥
實寨寢寤察對屢嶄嶇幛幣幕幗幔廓廖弊彆彰徹愨愿態慷慢慣慟慚慘慵截撇摘摔
撤摸摟摺摑摧搴摭摻敲斡旗旖暢暨暝榜榨榕槁榮槓構榛榷榻榫榴槐槍榭槌榦槃
榣歉歌氳漳演滾漓滴漩漾漠漬漏漂漢滿滯漆漱漸漲漣漕漫漯漵漪滬漁滲滌滷熔
熙煽熊熄熒爾犒犖獄獐瑤瑣瑪瑰瑭甄疑瘧瘍瘋瘉瘓盡監瞄睽睿睡磁碟碧碳碩碣
禎福禍種稱窪窩竭端管箕箋筵算箝箔箏箸箇箄粹粽精綻綰綜綽綾綠緊綴網綱綺
綢綿綵綸維緒緇綬罰翠翡翟聞聚肇腐膀膏膈膊腿膂臧臺與舔舞艋蓉蒿蓆蓄蒙蒞
蒲蒜蓋蒸蓀蓓蒐蒼蓑蓊蜿蜜蜻蜢蜥蜴蜘蝕蜷蜩裳褂裴裹裸製裨褚裯誦誌語誣認
誡誓誤說誥誨誘誑誚誧豪貍貌賓賑賒赫趙趕跼輔輒輕輓辣遠遘遜遣遙遞遢遝遛
鄙鄘鄞酵酸酷酴鉸銀銅銘銖鉻銓銜銨鉼銑閡閨閩閣閥閤隙障際雌雒需靼鞅韶頗
領颯颱餃餅餌餉駁骯骰髦魁魂鳴鳶鳳麼鼻齊億儀僻僵價儂儈儉儅凜劇劈劉劍劊
勰厲嘮嘻嘹嘲嘿嘴嘩噓噎噗噴嘶嘯嘰墀墟增墳墜墮墩墦奭嬉嫻嬋嫵嬌嬈寮寬審
寫層履嶝嶔幢幟幡廢廚廟廝廣廠彈影德徵慶慧慮慝慕憂慼慰慫慾憧憐憫憎憬憚
憤憔憮戮摩摯摹撞撲撈撐撰撥撓撕撩撒撮播撫撚撬撙撣撳敵敷數暮暫暴暱樣樟
樨樁樞標槽模樓樊槳樂樅槭樑歐歎殤毅毆漿潼澄潑潦潔澆潭潛潸潮澎潺潰潤澗
潘滕潯潠潟熟熬熱熨牖犛獎獗瑩璋璃瑾璀畿瘠瘩瘟瘤瘦瘡瘢皚皺盤瞎瞇瞌瞑瞋
磋磅確磊碾磕碼磐稿稼穀稽稷稻窯窮箭箱範箴篆篇篁箠篌糊締練緯緻緘緬緝編
緣線緞緩綞緙緲緹罵罷羯翩耦膛膜膝膠膚膘蔗蔽蔚蓮蔬蔭蔓蔑蔣蔡蔔蓬蔥蓿蔆
螂蝴蝶蝠蝦蝸蝨蝙蝗蝌蝓衛衝褐複褒褓褕褊誼諒談諄誕請諸課諉諂調誰論諍誶
誹諛豌豎豬賠賞賦賤賬賭賢賣賜質賡赭趟趣踫踐踝踢踏踩踟踡踞躺輝輛輟輩輦
輪輜輞輥適遮遨遭遷鄰鄭鄧鄱醇醉醋醃鋅銻銷鋪銬鋤鋁銳銼鋒鋇鋰鋝閭閱霄霆
震霉靠鞍鞋鞏頡頫頜颳養餓餒餘駝駐駟駛駑駕駒駙骷髮髯鬧魅魄魷魯鴆鴉鴃麩
麾黎墨齒儒儘儔儐儕冀冪凝劑劓勳噙噫噹噩噤噸噪器噥噱噯噬噢噶壁墾壇壅奮

嬝嬴學寰導彊憲憑憩憊懍憶憾懊懈戰擅擁擋撻撼據擄擇擂操撿擒擔撾整曆曉暹
曄曇暸樽樸樺橙橫橘樹橄橢橡橋橇樵機橈歙歷氅濂澱澡濃澤濁澧澳激澹澶澦澠
澴熾燉燐燒燈燕熹燎燙燜燃燄獨璜璣璘璟璞瓢甌甍瘴瘸瘺盧盥瞠瞞瞟瞥磨磚磬
磧禦積穎穆穌穋窺篙簑築篤篛篡篩篦糕糖縊縑縈縛縣縞縝縉縐罹羲翰翱翮耨膳
膩膨臻興艘艙蕊蕙蕈蕨蕩蕃蕉蕭蕪蕞螃螟螞螢融衡褪褲褥褫褡親覦諦諺諫諱謀
諜諧諮諾謁謂諷諭諳諶諼豫豭貓賴蹄踱踴蹂踹踵輻輯輸輳辨辦遵遴選遲遼遺鄴
醒錠錶鋸錳錯錢鋼錫錄錚錐錦錡錕錮錙閻隧隨險雕霎霑霖霍霓霏靛靜靦鞘頰頸
頻頷頭頹頤餐館餞餛餡餚駭駢駱骸骼髻髭鬨鮑鴕鴣鴦鴨鴒鴛默黔龍龜優償儡儲
勵嚎嚀嚐嚅嚇嚏壕壓壑壎嬰嬪嬤孺尷屨嶼嶺嶽嶸幫彌徽應懂懇懦懋戲戴擎擊擘
擠擰擦擬擱擢擭斂斃曙曖檀檔檄檢檜櫛檣橾檗檐檠歜殮毚氈濘濱濟濠濛濤濫濯
澀濬濡濩濕濮濰燧營燮燦燥燭燬燴燠爵牆獰獲璩環璦璨癆療癌盪瞳瞪瞰瞬瞧瞭
矯磷磺磴磯礁禧禪穗窿簇簍篾篷簌篠糠糜糞糢糟糙糝縮績繆縷縲繃縫總縱繅繁
縴縹繈縵縿縯罄翳翼聱聲聰聯聳臆臃膺臂臀膿膽臉膾臨舉艱薪薄蕾薜薑薔薯薛
薇薨薊虧蟀蟑螳蟒蟆螫螻螺蟈蟋褻褶襄褸褽覬謎謗謙講謊謠謝謄謐豁谿豳賺賽
購賸賻趨蹉蹋蹈蹊轄輾轂轅輿避遽還邁邂邀鄹醣醞醜鍍鎂錨鍵鍊鍥鍋錘鍾鍬鍛
鍰鍚鍔闊闋闌闈闆隱隸雖霜霞鞠韓顆颶餵騁駿鮮鮫鮪鮭鴻鴿麋黏點黜黝黛鼾齋
叢嚕嚮壙壘嬸彝懣戳擴擲擾攆擺擻擷斷曜朦檳檬櫃檻檸櫂檮檯歟歸殯瀉瀋濾瀆
濺瀑瀏燻燼燾燸獷獵璧璿甕癖癘癒瞽瞿瞻瞼礎禮穡穢穠竄竅簫簧簪簞簣簡糧織
繕繞繚繡繒繙罈翹翻職聶臍臏舊藏薩藍藐藉薰薺薹薦蟯蟬蟲蟠覆覲觴謨謹謬謫
豐贅蹙蹣蹦蹤蹟蹕軀轉轍邇邃邈醫醬釐鎔鎊鎖鎢鎳鎮鎬鎰鎘鎚鎗闔闖闐闕離雜
雙雛雞霤鞣鞦鞭韹額顏題顎顓颺餾餿餽餮馥騎髁鬃鬆魏魍魎鯊鯉鯽鯈鯀鵑鵝鵠
黠鬈鼬儳嚥壞壟壢寵龐廬懲懷懶懵攀攏曠曝櫥櫝櫚櫓瀛瀟瀨瀚瀝瀕瀘爆爍牘犢
獸獺璽瓊瓣疇疆癟癡矇礙禱穫穩簾簿簸簽簷籀繫繭繹繩繪羅繳羶羹羸臘藩藝藪
藕藤藥藷蟻蠅蠍蟹蟾襠襟襖襞譁譜識證譚譎譏譆譙贈贊蹼蹲躇蹶蹬蹺蹴轔轎辭
邊邋醱醮鏡鏑鏟鏃鏈鏜鏝鏖鏢鏍鏘鏤鏗鏨關隴難霪霧靡韜韻類願顛颼饅饉鶩騙
鬍鯨鯧鯖鯛鶉鵡鵲鶴鵬麒麗麓麴勸嚨嚷嚶嚴嚼壤孀孃孽寶巉懸懺攘攔攙曦朧櫬
瀾瀰瀲爐獻瓏癢癥礦礪礬礫竇競籌籃籍糯糰辮繽繼纂罌耀臚艦藻藹蘑藺蘆蘋蘇
蘊蠔蠕襤覺觸議譬警譯譟譫贏贍躉躁躅躂醴釋鐘鐃鏽闡霰飄饒饑馨騫騰騷騵鰓
鰍鹹麵黨鼯齟齣齡儷儸囁囀囂夔屬巍懼懾攝攜斕囊櫻欄櫺殲灌爛犧瓖瓔癩矓

籐纏續羼蘗蘭蘚蠣蠢蠡蠟襪襬覽譴護譽贓躊躍躋轟辯醺鐮鐳鐵鐺鐸鐲鐫闢霸
霹露響顧顥饗驅驃驀騾髏魔魑鰭鰥鶯鶴鷂鶸麝黯鼙齜齦齧儼儻囈囊囉孿巔巒彎
懿攤權歡灑灘玀瓤疊癮癬禳籠籟聾聽臟襲襯艫讀贖贗躑躓轡酈鑄鑑鑒霽霾韃韁
顫饕驕驍髒鬚鰲鰱鰾鰻鷓鷗鼴齬齪龔囌巖戀攣攫攪曬欏瓚竊籤籣籥纓纖纔臢蘸
蘿蠱變邐邏鑣鑠鑤靨顯饜驚驛驗髓體髑鱔鱗鱖鷥麟黴囑壩攬灞癱癲矗罐羈蠶蠹
衢讓讒讖艷贛釀鑪靂靈靄韆韉驟鬢魘黶鷹鷺鹼鹽鼇齷齲廳欖灣籬籮蠻觀躡釁鑲
鑰顱饞髖鬣鬻灤矚讚鑷韉驢驥纜讜躪釅鑽鑾鑼鱷鱸黷豔鑿鸚爨驪鬱鸛鸞籲乂乜
凵匸厂万丌乇亍囗屮彳丏冇与丮亓仂仉仈冘勼卬厹圠夃夬尐巿旡殳毌气爿丱丼
仨仜仕仡仝仚刌匜卌圢圣夗夯宁宄尒尻屴屳帄庀庆忉戉扐氕氶汃氿汈犮犰王内
肊阞伎优伬仵伉伶伀价伈伝伂伅伢伓伄仴伒冱刑刉刖劦匢匠卍厊吘囝囟圮圪圴
夰改奼妅妏妉奷尥孖尕尦屼屺屶屾宂秆庄异弚彴忕忔忏扜扞扤扡扦扢扙扠扚托
旯旮朾朹朸朻机朿杔朳氘汆汒汜汏汊汔汋汌灱牞犴犵玎甪癿穵网艸芋芳芁芀虍

襾邙邗邦邛邵阮阤阢阣佖伻佢佉体佤伾伬佒佟佁佘伭伳伿㑆冏冹刜刞刡劭劮匑
卣卲底厏呟咉吡呔呅吹哑吥吘吽呏呁吨吤杏囮囲园坻坽坌坉坋坒夆奀妦妘妠妗
妎妢妐妏妧妡宎宅尨尪岍岏岈岋岉岭岜岆岓岕巠帊帎庋庉庌庈庍弅弝彸彶忒忑
忐忭忨忮忳忡忤忣忺忯忷忻怀忴戺抃抌抎抏抔抇扱抆抵扰抁抈扶扽扲扴攷旰旴
旳旲昆杅杇杙杕杌杈杝杍杚杋毐氙氚汸汧汫沄沋沏汱汯汩沚汭沇沕汌汦汳汥汻
沘灴炖牣犴犽狃狆犺犹狂玕玗玓玔玒町甹疔疕皁礽耴肕肙肐肒肜芐芏芅芎芑芓
芊芃芄豸迉辿邟邡邩邫祁邠阰阸阯阬弗侂佼侅佽侀侇佶佴侉侄佷佌侗佪侚佹侁
佸侐侜侔侞侒侘侕佫佮釆冼冾刵刲刳剆劺劼匊匋匼厗厔咇呿咁咑哂咈呫呺呾呥
呬呴呦咍呯呡味咘呣呧吟囷囹坯坲坭坫坱坰坶坪坵坘坳坴坢坨坽麦奓姓妹姏姎
姐姌姁妶妼姃姖妱妽妺姈娑妥孢孥宓宕屄届岮岴岠岵岯岨岬岟岣岭岢岪岧岸岥
岶岰岦帗帔帙弨弢弣弤彔徂彾彽忞忥怭怦怙怲怋怴怊怗怳怚怞怬怢怍怐怮怓怑
怌怉怜戔戽抗抴拑抾抪抶拊抮抳抯抻抩抰抸攽斨斻昉旼昄昒昈旻昃香田昅旽昑
昐曶朊枅杬枎枒杶杻枘枆构杴枍枌杺枟枑枙杓枰极杸枇柃欥殀歾毞氛沓沉泫泮
泙沶泔沭泧泼泑泂沺泃泆泭沸泒泝沴炏秝杼泞泀洰沐泇沰泹泏泩泑炔炘炅炓炆
炄炢炖炂炚炃牪狖狋狘狉狜狒狔狚狌狑玤玡玭玦玢玠玥玝瓝瓨甿畀畄疌疘皯盳
盱盰盵矸矼矹矻矺矷祂礿秅穸穻竻籵紅耵肏肮肣肸肵肭舠芠芫芜芚芘芛芵芧芮
芼芞芺芴芨芡芩芠芤芟芍芢虰虬虭虮豖迒迋迓迍迖迕迗邲邴邯邳邰阹阽阼阺陃
俍俅俓侲俉俋俁俔俜俙侻侳俛俇俖侾俀侹俬剄剉勍勂匽卼厗厖厙厘咺咡咭咥哏
哃苟咷咮哖咶哅哆咠呰咼咢咾呲哞响垵垞垟垤垌垗垝垛垔垘垏垙垥垚垕壴复奓
姡姞姮娀姱姝姺姽姼姶姤姲姷姛姩姳姵姠姾姴婺宨屌峐峘峌峗峋峛峞峑峉峇峊
峖峡峔峏峈峆峎峟峸巹帡帢帣帠帤庰庤庢庛庣庥弇弮彖徆怸怹恔恲恞恅恓恇恉
恛恌恀恂恟怤恄恷恦恮扂扃拏挍挋拵挎挃拫拹挏挌拸拶振挓挖捒挕挾挻敁敀斪
旂昶昡昲昵易昦昢昳昫昺昝昴昹昜朏朐柁柲柈枺柜柑柸柘柀枷柅柫柤柟枵柍枳
柷柶柮柣柂枹柎柧柰枲柼柆柭柌枯柦柛枴柉柊柃柪柋欨殂殄殶毖毘毠氠氜洨洴
洭洟洼洿洒洊泚洳洄洙洺洚洑洀洝浂洁洘洷洃洏浀洇洋渢洈洢洉洐炷炟炾炱炰
炡炴炵炩牁牉牊牬牰牳牮狊狤狨狫狟狪狦狣玅珌珂珈珅玹玶玵玴珫玿珇玾珃珆
玸珋瓬瓮甮畇畈疧疪癹盄眈眃眄眅眊盷盻盺矧矨砆砑砒砅砐砏砎砉研砓祊神祋
祅祄秕种秏秖秎窀穾竑笆笂籺籸籹籿粀粁紃紈紁罘羑羍羾耇耎耏耔耷胘胇胠胑
胈胂胐胅胣胙胜胊胕胉胏胗胦胍臿舡芔苙苾苹苃苊苐苕茺苫苗苴茵苡苲苵茌苻
茶苰苪茎苠苺苳茘虷虴虼虳衁衎衦衪衩觓訄訇赲迣迡迮迠郱邽邿郕郅邾郇郎郈
釔釓陔陏陑陓陊陎倞倅倇倓倢倰倛俵俴倳倷倬俶俷倗倜倠倧倵倯倱倎党冔冓凊
凄凅凈凃剡剚剒剞剟剕剢勍匎厞唦哢哾唒哧哳哤唚哿唄唈哫唑唅哱唊哻呼哑唂
唎唤唋圁圂埌堲埕埒垺埆埑垈垸埣埍埇埐埗夎奊娙娖娭娮娕娏娗娊娞娳孬宧
宭宬尃屖屔峬峿峮峱崋峷峹帩帨庨庪庩弳弰彧恝恚恧恁悢悈悀悒悁悝悃悕悛
悗悇悜悎戙扆拲挐捖挬捄捅挶捃揤挹捋捊挼挩捁挴捘捔捙挭捇挳捚捑挸捗捀捈
敊敆旆旃旄旅晊晟晇晑朒朓栟栚桉栲栳栻桋椰栖栱梀栵栫栭栯桎桄梅栝栒栔栦
栨栮桍栺桼枽欬欯欭欱欴歭肂殈毦毨毤毣毢毧氤沖浣浤浶洍浡涒浘浢浭浯涑涍
淯浿涆浞浧浠涗浰浼浟涂涘洯浨涋浾涀渟湨涃淍浽浵涐烜烓烑烝烋焦烢烗烒烞
焍烔烧烅烆烇烚烎粪牂牸牷牶狲狺狴狾狶狳狻猁珓珙珥珖玼珧珣珩珜珒珛珶珝
珚珗珛珨瓞瓟瓴瓵甡畛畟疰痁疻痄疴疿疶疺皊盉眝眛眐眓眒眣眑眕眙眚眢眧砣
砬砢砵砯砨砮砫砡砩砙砪砱祔祛祏祜祓祒祑秫秬秠秮秭秪秜秞秝窆窉窅窑窌窋

窇竘笐笄笓笅笏笈笊笎笉笒粄粑粊粌粈粍粅紞紝紑紎紘紖紓紟紒紏紌罜罡罞罠
罝罛羖羒翃翂翀耖耾耹胺胲胹胵脁胻脀舁舯舥茳茭荄茙荑茥荖茿荁茦茜茢荂荎
茛茪茈茼荍茖茤茠茷茯茩荇荅荌荓茞茬荋茧荈虓虒蚢蚨蚖蚍蚑蚞蚇蚗蚆蚋蚚蚅
蚥蚙蚡蚧蚕蚘蚎蚝蚐蚔衃衄衭衵衶衲袀衱衿衯袃衾衴衼訒豇豗豻貤貣赶赸趵趷
趶軑軓迾迵适迿迻逄迼迶郖郠郙郚郣郟郥郘郛郗郜郤酐酎酏釕釢釚陜陟隼飣髟
鬯乿偰偪偡偞偠偓偋偝偲偈偍偁偛偊偢倕偅偟偩偫偣偤偆偀偮偳偗偑凐剫剭剬
剮勖勓匭厜啵啶唼啍啐唴唪啑啢唶唵唰啒啅唌唲啥啎唹啈唭唻啀啋圊圇埻堔埢
埶埜埴堀埭埽堈埸堋埳埏堇埮埣埲埥埬埡堎埼堐埧堁堌埱埩埰堍堄奜婠婘婕婧
婞娸娵婭婐婟婥婬婓婤婗婃婝婒婄婛婈媎娾婍娹婌婰婩婇婑婖婂婜孲孮寁寀屙
崞崋崝崚崠崌崨崍崦崥崏崰崒崣崟崮帾帴庱庴庹庲庳弶弸徛徖徟悊悐悆悾悰悺
惓惔惏惤惙惝惈悱惛悷惊悿惃惍惀挲捥掊掂捽掽掞掭掝掗掫掎捯掇掐据掯捵掜
捭掮捼掤挻掟捸掅掁掑掍捰敓旍晥晡晛晙晜晢朘桹梇梐梜桭桮梮梫楖桯梣梬梩
桵桴梲梏桷梒桼桫桲梪梀桱桾梛梖梋梠梉梤桸桻梑梌梊桽欶欳欷欸殑殏殍殎殌
氪淀涫涴涳湴涬淩淢涷淶淔渀淈淠淟淖涾淥淜淝淛淴淊涽淭淰涺淕淂淏淉淐淲
淓淽淗淍淣涻烺焍烷焗烴焌烰焄烳焐烼烿焆焓焀烸烶焋焂焎牾牻牼牿猝猗猇猑
猘猊猈狿猏猞玈珶珸珵琄琁珽琇琀珺珼珿琌琋珴琈畤畣痎痒痏痋痌痑痐皏皉盓
眹眯眭眱眲眴眳眽眥眻眵硈硒硉硍硊硌砦硅硐祤祧祩祪祣祫祡离秺秸秶秷窏窔
窐笵筇笴笥笰笢笤笳笘笪笝笱笫笭笯笲笸笚笣粔粘粖粣紵紽紸紶紺絅紬紩絁絇
紾紿絊紻紨罣羕羜羝羛翊翋翍翐翑翇翏翉耟耞耛聇聃聈脘脥脙脛脭脟脬脞脡脕
脧脝脢舑舸舳舺舴舲艴莐莣莨莍荺荳莤荴莏莁莕莙荵莔莩荽莃莌莝莛莪莋荾莥
莯莈莗莰荿莦莇莮荶莚虙虖蚿蚷蛂蛁蛅蚺蚰蛈蚹蚳蚸蛌蚴蚻蚼蛃蚽蚾衒袉袕袨
袢袪袚袑袡袟袘袧袙袛袗袤袬袌袓袎覂觖觙觕訰訧訬訞谹谻豜豝豽貥赽赻赹趼
跂趹趿跁軘軞軝軜軗軠軡逤逋逑逜逌逡郯郪郰郴郲郳郔郫郬郩酖酘酚酓酕釬釴
釱釳釸釤釹釪釫釷釨釮镺閆閈陼陭陫陱陯隿靪頄飥馗傛傕傔傞傋傣傃傌傎傝偨
傜傒傂傇兟凔匒匑厤厧喑喨喥喭啷噅喢喓喈喏喵喁喣喒喤啽喌喦啿喕喡喎圌堩
堷堙堞堧堣堨埵塈堥堜堛堳堿堶堮堹堸堭堬堻奡媯媔媟婺媢媞婸媦婼媥媬媕媮
娷媄媊媗媃媋媩婻婽媌媜媏媓媝寪寍寋寔寑寊寎尌尰崷嵃嵫嵁嵋崿崵嵑嵎嵕崳
崺嵒崽崱嵙嵂崹嵉崸崼崲崶嵀嵅幄幁彘徦徥徫惉悹惌惢惎惄愔惲愊愖愅惵愓惸
惼惾惁愃愘愝愐惿愄愋扊掔掱掰揎揥揨揯揃撝揳揊揠揶揕揲揵摡揟掾揝揜揄揘
揓揂揇揌揋揈揰揗揙攲敧敪敤敜敨敥斌斝斞斮旐旒晼晬晻暀晱晹晪晲朁椌棓椄
棜椪棬棪棱椏棖棷棫棤棶椓椐棳棡椇棌椈楰梴椑棯棆椔棸棐棽棼棨椋椊椗棎棈
棝棞棦棴棑椆棔棩椕椥棇欹欻欿欼殔殗殙殕殽毰毲毳氰淼湆湇渟湉溈渼渽湅湢
渫渿湁湝湳渜渳湋湀湑渻渃渮湞湨湜湡渱渨湠湱湫渹渢渰湓湥渧湸湤湷湕湹湒
湦渵渶湚焠焞焯烻焮焱焣焥焢焲焟焨焺焛牋牚犈犉犆犅犋猒猋猰猢猱猳猧猲猭
猦猣猵猌琮琬琰琫琖琚琡琭琱琤琣琝琩琠琲瓻甯畯畬痧痚痡痦痝痟痤痗皕皒盚
睆睇睄睍睅睊睎睋睌矞矬硠硤硥硜硭硱硪确硰硩硨硞硢祴祳祲祰稂稊稃稌稄窙
竦竤筊笻筄筈筌筎筀筘筅粢粞粨粡絘絯絣絓絖絧絪絏絭絜絫絒絔絩絑絟絎缾缿
罥罦羢羠羡翗聑聏聐胾胔腃腊腒腏腇脽腍脺臦臮臷臸臹舄舼舽舿艵茻菏菹萣菀
菨萒菧菤菼菶萐菆菈菫菣莿萁菝菥菘菿菡菋菎菖菵菉萉萏菞萑萆菂菳菕菺菇菑
菪萓菃菬菮菄菻菗菢萛菛菾蛘蛢蛦蛓蛣蛚蛪蛝蛫蛜蛬蛩蛗蛨蛑衈衖衕袺裗袹袸
裀袾袶袼袷袽袲褁裉覕覘覗觝觚觛詎詍訹詙詀詗詘詄詅詒詈詑詊詌詏豟貁貀貺

賬貰賍賨趄趑趉跘跓跍跇跖跜跏跕跙跈跗跅軯軙軺軹軦軮軥軵軧軨軶軫軱軬軭
軩逭逴逯郰郲郢郿郼郑郹鄍郶郪郷郿鄃酡酤酟酢酠鈁鈊鈙鈃鈚鈦鈏鈌鈀鈒鈄鈲
鈆鈄鈧鈂鈜鉅鈘鈗鈅鈖鉄閍閌閐隇陝隈隍隃隀隹隺雁雺雾靬靰靮頊颩飫馻黹黺
亄亶偉傿僆僒僄僊傴僈僂僃僁僆僈僉僉僋僀傕勥剸剻劉嗃嗛嗌嗐嗋嗊嗝嗀嗔嗄
嗩嗓嗒嗏嗏嗕嘔嗖嗢嗲嗍嗙嗂圇塓塨塤塏塍塉塯塕塎塝塙塥塛塭塣塱壺媓媛媰
媺媸媱媵媹媿媺媻媬媷嫀嫊嫃媮嫍媰寖寘寙尠尟嵱嵣嵊嵥嵲嵬嵞嵨嵧嵢嵰嵺
幎幀幍幣廅廌廆廋廇彀徯徭慂慉慊慀慅慒慆慆慆慅慏慐慀戠戡戣戥戤摰摮摯搐
搒搉搠搤搳搷搟搕搢搞搷搘搣搌搦搰搨搮搵搯搥搗搥搴搨搙搦搒搕搎搮搋毀斒旓
暆暌暕暐暋暊暙暔晸暎榅榠椸樟楢楱椿楅楪楛楂楗楙楺楫楉楒椵楬楰橡楥楳楸榝
梗楀楯楄稾楘楁楴楔楿棟楲楖楗椵楎楒楐楻楦歆歈歆歒歁歌殛毷毹毻毼毸滖滾
滈溏溦溟溓溔溠溱溹溡滒溽滁溞溷溷溰溎滋溲溾滃滉溛溱溒溦溍溼溳溿溸溴漂
熗溹澌淪煇煔煒煣煠煁煝煢煖煸煪煡煒煘煌煜煰煟煐煗煄煍煓煚煠煐牏犍猳猣犗猼
源猻猺猵猹猿瑄瑊瑋瑒瑑瑗瑀瑈瑜瑎瑂瑆瑍瑏瓡瓿甀甃皊皗皘畽瘏瘃瘖瘛瘍瘑
痹瘌瘈痻瘓瘋瘃瘐晢睶盝睕睟睠睒睖睚睩睧睔睤睶睤稓碇碚碱碏碄碕碅碒碡碤
碉碀碖碞裸裯裨裪稑稘稙稒稗稃稪稓稛稐窣窢窞竫筦筤筭筴筩筲筶筳筰筡箂
箵箣粲粴粷綈綆綀綒綍綅綈綎綄綃綊綌綍綄綂罬罭罧罦罳羦羥羧翛翙耡腤膆
腷腜腩腥腢腮腰腯腶腧腯腨腡腷艀艉艄艅艆艓荓菭葵葶葹葽蒍葪葑葀葽葧葉萋
葚葙葳葳葝葃葅葥葍葥葺葄葸萲葅葁萂葪葍葌葿葔葠萹葏菭葯葂葌葥葇萴葸
葨葼葃葃葎葲葭葢蜋蜄蛷蜉蛺蜂蜁蜐蜎蜛蜉蜓蜍蜅蛷裬裋裍裎裞裛裴裓裥觌
觟觤觥觡觠觢觧触訕詎詿詡詵詷誂詸詵誃詳誒詺詌豋豊豥豤豦豠貄貅賁赩赨赵
趌趎趏趍趚趔趐跰跦跬跱跬跳跮跦跧跰跲趯跴輇輧輆輍輅輇輖輂輦遄遏遄遉
遖鄁鄍郭鄌鄖鄔鄕鄎酮酯鉈鉒鉘鉔鉦鈳鉥鉞鉡鉯鉊鉆鉭鉙鉭鉄鉧鉝鉛鉼鉰鈮鉟
鉣鉐鉌鉎鉓鉎鉥鉯閟閜閕閗隒陰隑隗隉雽雩雷雵靳靷靲靲頎頍頏颬飺飹馯馲駒
鼏鼾鼿鼽鳪鳭鳧麀黽傶傳傍債僄傲僑僇僤僓僬僰僯僣僠厮劅劇勭勦匯厬嘧嗎嘌
嘒嘆嘏嘍嘁嘓嗢嘊嘞嗚嘖嘕塳塼墐墘墆墁塿塴墋墌墇墑墎墭墛墂墫墁墶嫞嫜
嫮嫭嫣嫪嫚嫦嫫嫳嫯嫠嫛嫬嫝嫙嫨嫟嫴嫈屜嶂嶀嵽嶆嶁嶑嶉嶊嶉嶛嶒嶝
嶯嵊幘幙幓廘廑廗廎廔廕廙廄廎彄彈彰徹慝慤慁慞慱慴慳慲慞慬慪慮慓慒慒
慫慪慪慖慱慽截摞擎摍摘摝摴摶撕摳摽摵摫摰摎摻摷摜摬摠摗摓摥摙撖撗撣擖撐
敳斠斡暠暟暍暕朢榱榶槉榠榹穀榗榬榼槫榙榎榧榢榛榾榯榿榴楒楬榤樂橰槁槊槎槏
槙榿榪榬榞槇榴榢槁槔榥槆歊歍歉殞殟殠毃毄毿滎滵滽漵漥滸滰漧潲漉漵漙漚
漗漘漻漒漳漊漐漈漊潉漭漮漻滯漧漙漼漦溧潻漃漜潐漶滍漡漸漰熇熐熉
熄熅熆熏煻熆熣熗熗膀犒犕犓獃獍獑獿瑢瑳瑱瑤瑲瑧瑮甄甂瓺瞳瘇瘖瘱瘌瘕瘕
瘢瘩皸瞑瞌瞅瞀睮睯瞖睾瞄碲碪磋碣碬碶碫碸碥碢碨碬碿禘禊禋禖禕禔禓禐禪
稼稷稫概稰稯稨稶窨窫窬窮箈箜箊箑箐箖箍箌箛箎箄箘箙箤箂粻粿粿糁縓綷
緂綣緀緁緮緅緋緎緄緆緋緌緅緇綖緈緉緊緌緘緛罳翢翣翥翞翵聝聜腦膝膃膇
膍膌膂膍蒗蒁蒡蒟蒢蒫蒖蒬蒮蒩蒹蒴蒭蒨蒔蒿蒪蒲蒱蒝蒧蒡蓧蒔蒨蒰蒛蒩蒨蒄
蓻蒘蒤蒠蒤蒸蒽蒚蒰蒗蒝虡蜳蜣蜨蜭蜴蜮蜝蜡蜙蜛蜚蜛蜼蜾蜽蝀蜠蜿蜺蜴蜷蜱
蝸蜧蝓蜧蜸蜇蜚蜛蜱裷裧裭褹裺裾褍褟裵裰裺褑褌褀褂褁褂觩觫觭誏誙誋誒誏
誟誋豨豩賕賏賗趚踉踂踀踆踠踊踃踄踆踂踀踒踄踊踆踄鄗鄘鄝鄠鄚鄝鄟鄢鄡
鄞醐醒醊醎銶銖銂銛銰銠銩銪銍銦銚銫銛銗銯銣銇銘銅銙銍銡銭銙銙銙銯銓銚
銔銖銠銥銋銉隞隥霃艴鞊鞋鞁鞃鞀鞂鞄鞃鞈鞂鞁鞏頪颭颮餂餀餇馝馜駃駉駃駁

鎷䍧駇骱髣髿魃魆魠魡魟鴇鳵鳲麧僿儃僞僸儆儇僶僾儋傲僁儊剸劌勱勯噈噂噌
嘵噁嘺噋噆噘噚噀嘳嘽嘬嘾嘸噴噊圚墫墝墱墠墣墯墜墥墡墫嫿嫜嫽嫡嫶嬃嬉嬂
嫹嬁嬇嬅嬏屧嶙嶗嶟嶒嶢嶓嶕嶠嶜嶡嶑隓幩幝幠幜廔廛廞廡彉徲憋憃慹憱憍憢
憉憛憓憯憭憟憒憪憡憍憝憳戭擊摯撖撠撅撗撜撏撊撋撌撣撟撨撱搭敶敺敹敻斳
斮暵暰暩暲暷暪暯樀樆樗槥槸樕標樋樠槿㯷槢樛槙槾樧槲槮樸槷槧榼樈槦槻樍
槼槫樉樄樘樥樏槶樦樇樞翕歔殥殣殢殦毾氀毿氂潁槃漭澇潰澒澍澉澌潢潏澅潚
澗漂潬潡潕潲潒潐潗澔澓潝漀激潫潽潧澐潓澋潩潿澕潤漨潪潻熲熯熛熰熠熚熩
熵熝熥熞燈熡熪熜熒熳犘犚獘獒獞獟獠獝獛獡獚獥獢璇璉璊璆璁瑽璅璈瑼㻊甂
甇畾瘥瘞瘙瘝瘜瘣瘚瘨瘛皜皝皞皛瞍瞏瞉瞈磋碻磏磌磑磎磔磈磃磄磟禚禡禠禜
禓禛禩稹窫窴窳箎篋箾箬篎箯箹箕箵糅糈糌糋緷緛緬緧緗緡縃緺緦緶緱緰緮緟
罶羬羰羭翭翫翪翬翦翣聤聧腟膵膞膕膢膙膗舖艏艓艒艐艎艑蔤蔻蔏蔀蔶蔎蓑蔍
蔟蔊蔧蔜蓻蔫蓺蔈蔌蓴蔪蓲蔕蓷蓫蓳蓼蔒蓪蓩蔖蔮蔰蔝蔮蔂蓽蔞蓶蔱蔦蓧蓨蓰
蓯蓹蔘蔠蔖蔋蔙蔯虢蝖蝣蝤蝷蝺蝳蝘蝔蝛蝒蝡蝚蝑蝞蝭蝪蝐蝎蝟蝝蝯蝬蝺蝮蝜
螌蝏蝻螢螝蝧蝩衚褅褌褔褋複褘褙褆褖褑褎褉覝覞覭觭觰觬諏諆諡諓諑諔諕諧
諗誾諀諅諘諃諌說諳谾豍貏賥賟賙賓賚賝賧趠趜趡趛踠踣踥踤踮踕踛踖踑踙踦
踧踔踒踘踓踜踗踚輬輤輘輚輠輣輖輗遳遰遯違遨鄯鄫鄣鄭鄲鄦鄧醅醆醊醁醂醄
醀鋐鋃鋖鋀鋙錄鋏鉽鋟鋘鋩鋗鋝鋌鋯鋂鋨鋊鋈鋎鋦鋍鋕鋉鋠鋞鋧鋑鋓鋥鋡鋆鋬
䦵閬閫閮閭隤隢雂雺霈霂靚鞊鞎鞈韐鞏頞頝頫頩頠頯頣頮颲餈餐餑餔餖餗餕駜
駍駏駓駔駎駉駖駘駋駗駕骳髶髫髲髮髦魅魃魧魴魱魦魶魵魰魨魮魬鴆鴂鴈鳿鴃
鴇鴒鴝鴋鴈鴊鴠麃默鼏鼐儜儓儗儚儑凞匴叡噰噠噮噳噦噣噭噲噞噷圜圛壈墩壉
墿墺壂墼壆嬗嬙嬛嬡嬔嬓嬐嬖嬨嬚嬠嬞寯嶬嶱嶩嶧嶵嶰嶮嶪嶨嶲嶭嶯嶴幧幨幦
幱廩廧廦廨廥彋徼憝憨憖懅憴懆懁懌憺憿憸憌擗擖擐擏擉撽擎擃擛擳擙趫斀鼓
斢曈暾曀曊曋曏暽曍曄曌朣樴橦橉橧橨樍樾橝橖機橛橑樨橚橨樿橢橪橤橐橲橔
橯橷橠橩橞橒橗橍橎橆歕歔歖殧殪殫毈毇氄氃氆澭濋濢澮澼濎濈潞濄澽澞濊澨
澩澥澮澺濆澪澁澿澸澢濉澫澬澯澲澰燚燂燌熷燖燀燁燋燔燊燇燏燣燗熼燆燚燛
犝犞獩獨獧獬獥獫獪璧璚璠璔璒璕璡甋鏈瘯瘭瘱瘽瘳瘼瘵瘲瘰皻盦瞚瞝瞡瞜瞛
瞢瞣瞕瞙瞗磝磩磥磪磞磣磛磡磢磭磟磠禤穄穈穇窶窸窵窱窷篁篣篧篝篕篥篚篠
篡篔篪篢篜篫篘篟糒糔糗糐糑縒縡縗縌縟縠縓縎縜縕縚縢縋縏縖縍縩縥縤罃罻
罼罺羱翯耪耩聬膱膦膮膹臂膫膰膬膴膲膷膧艗艕艖艦蕖蕅蕫蕍蕓蕢蕘蕀蕆蕔蕁
蕡蕄蕑蕇蕣蔾蕛蕏蕎蕮蕵蕕蕧蕠薌蕦蕝蕔蕥蕬虤虥虦螛螏螗螓螒螈螁螖螘蝹螇
螣螅螐螑螝螄螔螌螜螉褞褫褰褢褮褧褱褻褩褣褯褬褟觱諠諢諲諴諵諝謔諤諟諰
諈諞諡諨諿諯諻貑貒貐賵賮賱賰賹赬赮趥趧踳踾踸蹀蹅踶踼踽蹁踰踿軀輶輮輵
輲輹輷輴遶遹遻遾鼂鄹鄡鄶醓醐醑醍醏錧錞錈錟錆錏鍺錸錝錛錣錒錁鍆錭錎錍
鋋錝錽錡錵鋹鋷錴錂錤鍋錩錆錵錪錔錋錭錉錀鋻鋬閍闍閾閹閿閶閺闈闇隩雔
霒黔霘鞙鞗鞔鞓韸頵頯頸餤餟餧餩馞駮駬駥駤駰駣駪駩駧骹骿骴骻髼髾髹髺鬳
鮀鮅鮇鮕魾鮃鮂鮓鮒鮐鮣鮎魽鮈鮖鴪鴡鴺鴔鴸鴝鴒鴭鴐鴗鴎麈麆麌麩麭黕黓黺
鼒鼽儦儥儢儤儠儩勴嚓嚌嚧嚍嚄嚃嚐嚂嚀嚁壖壔壏壒嬭嬥嬲嬣嬬嬧嬦嬘嬮嬯寱
寲嶷嶹幪徾徻懃懕懘懧懠懛懤懨懞擯擩擣擫擤擨斁斀斶旚曒檍檖檁檥檉檟檛檡
檞檇檓檎檕檃檨檤檑橿檒檚檅檌藏歛殭氈黏澩濚濔濣濜濫濧濦濞濲濭澤濨燡燱
燨燪燤燰燢獳獮獯璗璲璫璐璪璭璱璥璯甐甑甒鬆暽癃癈癉癇瞂盩瞵瞫瞲瞯瞶瞴
瞱瞨矰磳磽磶磻磼磲礅磹磾礄禫禨穜穛穖穘穔穚窾窺竈篷簏篲簣篿篻簎篴篕篳

籄簉簃簁篸篽篾篰篱篳簺糨縭縼繂縛顈縸縪縨縲縿縷縱縻縶繄縺縳罿罾罽羄
翷耬膻臄臌臊臅臇膼臩艛艚艜薃薀薏薧薕薠薋薣蘄薙薚薞蕷薛薉蕭薡蕸蕗葭薖
薆薍薙薝薁薢薂薈薅蕹薒薘薐薟虨螾螭蟎蟅螰螬螹螵螼螮蟉蟃螅蟌蠯螫蟄蝨螴
蟂螿螢螽螌螳褵褳褼褾襁襒褷襂覬覯覲觲觳謞謘謖謑謅謋謢謏謒謕謇謍謈謆謜
謓謚豏豰豲豱豯貕貔賹趟蹎蹍蹓蹐蹌蹇轃轀邅遾鄸醚醢醛醙醟醡醝醠鎡鎃鎯鍤
鍖鍇鍼鍘鍜鍶鍉鍐鍑鍠鍭鎏鍌鍪鍹鍗鍕鍒鍏鍱鍷鍻鍡鍞鍣鍧鎀鍎鍙闇闀闉闃闅
罽隮隰隬霚霟霨霝霙鞚鞡鞜鞞鞝鞫韔韱頏頟頔頍頎頖餈餫餬餪餳餲餘餭餱餰馘
馣馡騂駺駴駷駹駸駶駻駽駾駼騃骾髾髽鬄鬀魈鮚鮨鮞鮛鮦鮡鮥鮤鮆鮢鮠鮯鴳鵁
鵧鴶鴮鴯鴱鴸鴰鵅鵂鵃鴾鴷鵀鴽翵鴭麊麉麍麰黈黚黻黿鼤鼣鼢齔龠儱儭儮嚘嚜
嚗嚚嚝嚙嬹嬼屩屫巀幭幮懘懟懭懮懱懪懰懫懖懩擿攄擽擸攁擃擼斔旛曚曛曘櫅
檹檽櫡櫆檺檶檷櫇檴檭歞毉氋瀇瀌瀍瀁瀅瀔瀎濿瀀濻瀦濼濷瀊爁燿燹爃燽獶璸
瓀璵瓁璾璶璛瓂甒甓癜癤癙癐癓癗癚皦皽盬矂瞺磿礌礓礔礉礐礒礑禭禬穟簜簩
簙簠簟簭簝簦簨簢簥簰繜繐繖繣繘繢繟繑繠繗繓羵羳翷翸聵臑臒臐艟艞薴藆藀
藃藂薳藸藬藇藄薿藋藎藈藅藯藭藒藹薸薷薾虩蟧蟦蟢蟛蟫蟪蟥蟟蟳蟤蟔蟜蟓蟭
蟘蟣蟩蟗蟙蠁蟴蟨蟝襓襋襏襌襆襐襑襉謪謧謣謳謰謵譇謯謼謾謱謥謷謦謶謮謤
謻謽謺謬豵貙貘貗賾贄賸贀蹜蹢蹠蹗蹖蹞蹥蹧蹛蹚蹡蹝蹵蹔轆轇轈轋鄨鄺鄻鄾
醨醥醧醘醪鎵鎌鎒鎷鎛鎝鎉鎧鎎鎪鎞鎦鎕鎈鎙鎟鎍鎱鎑鎲鎤鎨鎴鎣鎥闒闓闑隳
雗雚嶲巂雘雝霣霢霥鞬鞮鞨鞫鞪鞏鞢鞥韗韙韖韘韺顐顑顒颸饁餼餺騏騋騉騍騄
騑騊騅騇騆髀髜鬈鬄鬅鬩鬵魊魌魋鯇鯆鯃鮿鯁鮵鮸鯓鮶鯄鮹鮽鵜鵓鵏鵊鵛鵋鵙
鵖鵌鵗鵒鵔鵟鵘鵚麎麌黟鼁鼀鼖鼥鼫鼪鼩鼨齌齕儴儵劖勷厴嚫嚭嚦嚧嚪嚬壚壝
壛夒嬽嬾嬿巃幰徿懻攇攐攍攉攌攎斄旞旝曞櫧櫠櫌櫑櫙櫋櫟櫜櫐櫫櫏櫍櫞歠殰
氌瀙瀧瀠瀖瀫瀡瀢瀣瀩瀗瀤瀜瀪爌爊爇爂爅犥犦犤犣犡瓋瓅璷瓃甖癠矉矊矄矱
礝礛礡礜礗礞禰穧穨簳簼簹簬簻糬糪繶繵繸繰繷繯繺繲繴繨罋罊羃羆羷翽翾聸
臗臕艤艡艣藫藱藭藙藡藨藚藗藬藲藸藘藟藣藜藑藰藦藯藞藢蠀蟺蠃蟶蟷蠉蠌蠋
蠆螢蠈蟿蠊蠂襢襚襛襗襡襜襘襝襙覈覷覶觶譐譈譊譀譓譖譔譋譕譑譂譒譗豃豷
豶貚贎贇贉趬趪趭趫蹭蹸蹳蹪蹯蹻軂轒轑轏轐轓辴酀鄿醰醭鏞鏇鏏鏂鏚鏐鏹鏬
鏌鏙錣鏦鏊鏔鏮鏕鏖鏄鏎鏓鏒鏧蹽闚闛雡霩霫霬霨霦鞳鞶韝韞韟顜顙顝顗颼
颽颻颾饈饇饃馦馧騚騕騥騝騤騛騢騠騧騣騞騜騔骼鬋鬊鬎鬌鬷鯪鯫鯠鯞鯤鯦鯢
鯰鯔鯗鯊鯘鯙鯥鯕鯡鯚鵷鶇鶊鶄鶈鵱鶀鵸鶆鶋鶌鵽鵫鵴鵵鵰鵩鶅鵳鵻鶂鵯鶩鳽
鶇鵭麐麑麙黼鼭鼝齁齍齖齗齘匷嚲嚵嚳壣孅巆巇廮廯忀忁懹攗攖攕攓旟曨曣曤
櫳櫰櫪櫨櫹櫱櫮櫯瀼瀵瀯瀷瀴瀱灂瀸瀿瀺瀹灀瀻瀳灁爓爔犨獽獼璺皫皪皾盭矌
矎矏矍矲礥礣礧礨礤礩禲穮穬穭竷籉籈籊籇籅糮繻繾纁纀羺翿聹臛臙舋艨艩蘢
藿蘁藾蘛蘀藶蘄蘉蘅蘌藽蠙蠐蠑蠗蠓蠖襣襦覹觷譠譪譝譨譣譥譧譭趮躆躈躄轙
轖轗轕轘轚邍酄酁醷醵醲醳鐋鐓鏻鐠鐏鐔鐝鐕鐐鐨鐙鐍鏵鐀鏷鐇鐎鐖鐒鏺鐉鏸
鐊鐆鐃鐌鏶鐑鐩闞闠闟霳霵鞹韣韡顤顟顠顣顢颾飂饐饎饙饌饋饓騵騴騱騬騪驀
騩騮騸騺髝髊髆鬐鬒鬕鰋鰈鯷鰅鰒鯸鱀鰇鰎鰆鰗鰔鰉鶟鶙鶤鶝鶒鶘鶐鶛鶠鶔鶜
鶪鶗鶡鶚鶢鶨鶞鶣鶿鷔鶩鷌麜麛麚黥黫黨鼳鼰鼲鼵齣齠齞齝齙龑儺儹劘劗囃嚽
嚾孎孇巋巏廱懽攛欂櫼欃櫸欀灃灄灊灈灉灅灆爝爚爙獾甗癪矐礭礱礸籔籓糲纊
纇纈纋纆纍罍攫耰臝蘘蘪蘦蘟蘣蘜蘙蘧蘮蘡蘠蘩蘞蘥蠩蠝蠛蠠蠤蠜蠫衊襭襩襮
襫覼譹譸譅譺譻贐贔趯躎躌轞轛轝酆酄酅醹鐿鐻鐶鐩鐽鐼鐰鐹鐪鐷鐬鑀鐱闥闤
闣霰霺鞿韡顤飋飆飀饘饖騹騽驆驄驂驁騺騲髍鬤鬡鬘鬖鬺魒鰫鰝鰜鰬鰣鰨鰩鰤

鰡鶛鶌鶼鷁鷇鷊鷏鶾鷕鷜鶺鶵鶹鵻鶬騬鶱鷟騗騺鶂鷂鹺麠黫黮黥鼇鼛鼜齆齎
齱齹龔疊嚾囅囋奱孋孌巏巑廲攤攠攦攢欋欈欉氍灘灖灗灒爞爟犩玃瓘瓕瓌瓗癭
矏礝禴穰穱籗籜籙籛籚纙糱纑罏羇臞艫蘴蘵蘳蘬蘁蘶蠬蠨蠦蠪蠥襱覿觀觻譾讄
謏讏譒譿贖躕躔躚躒躐躖躛轀轢酇鑌鑐鑊鑿鑏鑇鑅鑈鑉鑆霿韣顛顩颸饔饛驎驓
驔驌驏驕驊驉驒驐髐鬙鬫鬻魖魕鱆鱈鰿鱄鰹鰳鱁鰼鰷黴鰲鱛鰶鷉鷒鷞鷚鷎鷐鷜
鷂鷟鷩鷙鷔鷖鷗鷕鷤麴黰鼵鼶鼸齂齫龕龢儽劙壨壧奲孍巘廱彏戁懹戄攩攥蠱讐
欑欒欏欝瀾灚爢獑玁玃癰矔籧鐘纕艬蘺虀蘹蘼蘱蘻瓖蠰蠲蠿蠳襶襴襳艥讌讎讋
讈礲贙躘轤轣醼鑢鑕鑝鑗鑞韄韅護驖驙鬞鬟鬠鱒鱘鱐鱊鱍鱋鱕鱙鱌鱎驐鷷鷲鷤
鷫鷸鷤鷶鷡鷮鷦鷲鷰鷹鷬鷳鷴鷨鷭麡黐黲黳鼆鼜鼸鼷鼶齃齏齱齰齮齯囓囍孎屭
攭曭曮欓灟灡灝灠爣瓛瓥矕礸禷禶籪纗羉艭虃蠸蠷蠵蠹讔讕躞躟躠躝醾醹釂鑫
鑨鑩雥霴靃靇韇韥驞髕鬹鱣鱧鱦鱢鱞鱠鸂鷾鸇鸃鸆鸅鸄鸁鸀鸊鸒鷩麢鼟鼲齴齵
齶囔攮斸欘欙欗欚灢爦犪矘矙礹籩籫糶纚纘纛纙臠臡虆虇虈襹襺襼襻艬讘讙躣
躤躍鑮鑭鐵鑱鑳霿顩饟鱨鱮鱭鸋鸍鸐鸏鸒鸑麙黵鼉齇齸齻齺齹圞灦籯蠼趲躦釃
鑴鑸鑶鑵驠鱴鱳鱱鱵鱲鸓黶鼊龤灨灥糷虪蠾蠽蠿讞貜躩軉靋顳顴飌饡馫驤驦驧
鬤鸕鸗齈齆欞爧虌躨钂钀钁驩驨鬮鸙爩虋讟钃鱹麷癵驫鱺鸝灩灪麤齾齉龘

Appendix III – Unihan inventory of basic components[349]

[349] For technical reasons only 588 out of 593 components are presented. The graphical represenation of CDP components can be checked at http://glyphwiki.org.

CDP-854ECDP-856ACDP-85BFCDP-85C0CDP-85C6 CDP-85CE CDP-85D6 CDP-85E0
CDP-85E2 CDP-85E4 CDP-85E6 CDP-85E9 CDP-85EA CDP-85EF CDP-85F0 CDP-85F2
CDP-85F3 CDP-85F6 CDP-8642 CDP-8643 CDP-8647 CDP-864B CDP-8657 CDP-8658
CDP-8659 CDP-865B CDP-865D CDP-865F CDP-8660 CDP-8661 CDP-8664 CDP-8665
CDP-8667 CDP-866F CDP-8670 CDP-8671 CDP-867A CDP-867D CDP-86A1 CDP-86AC
CDP-86B2 CDP-86B9 CDP-86C4 CDP-86C5 CDP-86C6 CDP-86CC CDP-86CD CDP-86CF
CDP-86D6 CDP-86DB CDP-86DF CDP-86E1 CDP-86E2 CDP-86E5 CDP-86E6 CDP-86E7
CDP-86E8 CDP-86EB CDP-86EC CDP-86F1 CDP-86F3 CDP-86F5 CDP-86F6 CDP-86F7
CDP-86FD CDP-86FE CDP-8740 CDP-8744 CDP-8747 CDP-8748 CDP-874A CDP-874B
CDP-874C CDP-874F CDP-8750 CDP-8751 CDP-8758 CDP-8759 CDP-8766 CDP-876E
CDP-876F CDP-8773 CDP-8776 CDP-87A2 CDP-87A8 CDP-87A9 CDP-87AC CDP-87AF
CDP-87B0 CDP-87B2 CDP-87B3 CDP-87B6 CDP-87BC CDP-87BD CDP-87C5 CDP-87C8
CDP-87CB CDP-87E2 CDP-87EA CDP-87EC CDP-87F9 CDP-87FD CDP-87FE CDP-8842
CDP-8843 CDP-8846 CDP-8847 CDP-884A CDP-884B CDP-8851 CDP-8854 CDP-8859
CDP-885A CDP-885B CDP-885E CDP-8865 CDP-886B CDP-88A1 CDP-88B1 CDP-88B7
CDP-88B9 CDP-88BB CDP-88BC CDP-88C0 CDP-88C8 CDP-88CB CDP-88D3 CDP-88D4
CDP-88D5 CDP-88DB CDP-88E2 CDP-88EE CDP-88F1 CDP-8954 CDP-8959 CDP-895C
CDP-8961 CDP-8962 CDP-8964 CDP-896A CDP-8977 CDP-89B0 CDP-89B9 CDP-89BA
CDP-89BB CDP-89C5 CDP-89CA CDP-89CC CDP-89D5 CDP-89D9 CDP-89DE CDP-89DF
CDP-89E0 CDP-89E1 CDP-89E3 CDP-89E4 CDP-89EE CDP-8A41 CDP-8A49 CDP-8A4D
CDP-8A4E CDP-8A4F CDP-8A77 CDP-8AEA CDP-8AEB CDP-8AF2 CDP-8B7C CDP-8BBF
CDP-8BC0 CDP-8BC5 CDP-8BD0 CDP-8BEA CDP-8BF8 CDP-8C4B CDP-8C4E CDP-8C66
CDP-8C78 CDP-8C7A CDP-8CAC CDP-8CB5 CDP-8CBB CDP-8CBD CDP-8CD4 CDP-8CE4
CDP-8D46 CDP-8D6B CDP-8DBA CDP-8DC1CDP-8DDFCDP-8DE4CDP-8DEA

⺄冂⺊⺌⺍⺶〇⻄𠆢⺧㐅肙⺘㔾㠯𢦏𤴔冄一丈丏丐丑专且世丘东丣严丨丩丬丱丶丷丹为丿乀乁乂久乍乎乐乑乗乙乚乛乜九也乡书亅亊事于井亜人亻以先兆入八冂円冉冊冖冘冫几凵凸凹刂勹匚匸十卅卍卐卜卝卩厂厶又及发口史囗夊夂央头女子孑孒孓宀寸小尢尸尺山川州工巨己已巳巴巿干年广廴廿弋弓彐彑彡彳心忄戊戌我戼手扌才承攵斗旡日曰甲曲曳月木朩未末本朱朿束来東柬欠止毋毌母比毛氏民氵永灬為熏爪爲爿片牙牛牜犬犭瓜瓦甘生田由甲申疋疌疒白皮皿目示礻禹禺米肃纟缶罒耂耳肀肅肉臣自臼舟艮虫衤襾见角言訁讠谷豆豸贝身車车辶酉重金釒钅长门阝隶隹非革韦頁页飛飞饣黽齿龜龟⻊龴龵龶卑既者龷⻌㔾𠀉𠀎丠丗覀東𦉼𠙴⺊⺈⺆𠃜𠀃𠁷𡗗𠃑𡳿𠂇𠂇𠄌𠂉𠂊𠀟𠨍生𦙶𠂹𠂢乙乚𠃊𠃌𠄎亅𠃑𠃒𠂎𠃛𠂆𠃓乚乃𠃉𠁁𠫓事𠆢兆冋𠘧几𠀁𠚍凹匠𠔾㔿𣎳華离𤰔史𠘨小尢𡰣肙𠂢垂𡿨丕巳𠀚𠄂𠀬亐亏号𢎘𠃟𢆉朩𣎴東少𠀃毋𠂋日王𢆶甘用𦘒𤴓𡨄𢁉𥝌⺮兕𠥓𠥓凹𦥔𡦹廾𠈌豕豖𠂣身𨈐邑門𠂤非𣍭𣍳飛𩙿𣎆𠥧𪛆𪛋𪚦龟𠂒U+2B793U+2F82F