R O C Z N I K O R I E N T A L I S T Y C Z N Y, T. LXVII, Z. 1, 2014, (s. 169–179)

NORBERT KORDEK

Quantitative Analysis of the Structure of Chinese Characters in Terms of Cangjie Components

Abstract

The aim of this paper is to present the results of graphotactic analysis of the componential structure of Chinese Characters based on the Cangjie input method. The Cangjie method and the graphotactic framework are both briefly introduced in the initial part of the paper. The results of the quantitative analysis of Cangjie representations of Chinese characters are presented in the main part of the study.

Keywords: Chinese characters, Cangjie method, graphotactic analysis, componential structure

1. Cangjie input method (CIM)[1]

The term ‘input methods’ pertains to the devices for sorting and categorizing Chinese characters supporting computer technology. In this paper, only the Cangjie method will be introduced in more details; the reason for this is the fact that only CIM provides an opportunity to conduct graphotactic analysis, while the most popular input methods are based on pronunciation.

The Cangjie input method (倉頡輸入法 *Cāng Jié shūrùfǎ*) was invented in the 1970s by Chu Bong-Foo and is still considered a popular typing tool. CIM provides the means of conducting a quantitative analysis of the componential structure of Chinese characters that can be treated as a basis for a direct comparison with.

The input strategy in CIM is based on a rather simplistic geometric decomposition of characters. For practical reasons the CIM components (that are symbols representing different compositional and graphical features of characters) are classified into 24

[1] The introduction of CIM is in large part based on Kordek 2013: 177–184.

categories represented by Cangjie 'letters' (倉頡字碼 *Cāng Jié zìmǎ* or 倉頡碼 *Cāng Jié mǎ*). The choice of components was based on their frequency and capability to represent the composition of different types of character structures in a more or less abstract way. The 24 Cangjie letters correspond to the English alphabet, which allows their input using a standard keyboard.

The graphical representation system in CIM is an unique one – it takes a lot of training to correctly code and decode the structure of characters in terms of Cangjie letters. The reason for this is the fact that in many cases the CIM components represent shapes that graphically are only remotely related to them, moreover, the choice of graphical fragments of characters that are represented in CIM codes is not determined by the componential structure of characters. This results in significant discrepancies between the traditionally understood componential structure and the CIM representation. Due to the non-orthodox representation of the structure of characters a short introduction of CIM is required.

For practical reasons the CIM representation of a character consists of English alphabet letters, and the Cangjie symbols corresponding to English letters may also be used for coding. The symbols/letters represent primary shapes and a set of 76 auxiliary shapes (輔助字形 *zhuǎnzhù zìxíng*). The full CIM component system is introduced in Table 1.1.

CIM was designed to serve as an efficient typing tool, but in fact it is also a character description language. The learnability, the adequacy and relatively low ambiguity are strong indications that CIM provides viable and analyzable structural information.

Table 1.1. The Cangjie system symbolic structural representation[2]

Cangjie Symbol	Corresponding Letter	Pithy Name	Auxiliary Shapes
日	A	日	rotations of 日, 曰
月	B	月	first four strokes of 目, also 冂, 𠂉, 冖, 夕, 𠂊, first four strokes of 骨
金	C	金	丷, 八, 儿, 亻
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	
竹	H	斜	𥫱, 丿 and the short slant 丿
戈	I	點	丶, 广, 厶

[2] The contents of the table are based on various sources, for more details see Kordek 2013: 181.

Cangjie Symbol	Corresponding Letter	Pithy Name	Auxiliary Shapes
十	J	交	crossing strokes shape, 亠
大	K	叉	乂, 𠂇, 𠂆, X shaped elements
中	L	縦	丨, 礻, 丰
一	M	横	𠂉, 厂, 工
弓	N	鉤	亅
人	O	人	乀, 厂, 𠂉, last two strokes of 兆
心	P	心	忄, 灬, second stroke in 心, also 匕, 七, 勹, last two strokes in 代
手	Q	手	
口	R	口	
尸	S	側	first two strokes of 己, 刁, 𠃌, 𠕁
廿	T	並	廾, 廾, 艹 (also in broken form)
山	U	仰	an enclosing structure with an open top
女	V	紐	right hook, V shaped elements, 𠂢
田	W	方	冂, enclosed shapes also with elements inside
卜	Y	卜	辶, 辶, 丷
Disambiguation symbol	X		
Special symbol	Z		

The controversial issues pertaining to the CIM decomposition procedure and its results are discussed in more details in Kordek 2013.[3]

The following examples illustrate the strategies of Cangjie encoding, including its peculiarity, idiosyncrasy and controversies. Especially a point concerning the inability, in many cases, to distinguish the characters similar in shape and characters having identical components composed in a different way should be made.

[3] Ibid.: 180–184.

A. Cases where the CIM manages to differentiate between similar characters:

Character	Cangjie Description	Character	Cangjie Description
士	十一	土	土
匕	山竹	七	十山
天	一大	夭	竹大

B. Examples of assigning the same codes to similar, but different characters:

Character	Cangjie Description	Character	Cangjie Description
己	尸山	已	尸山
曄	日廿一十	暴	日廿一十
晶	日日日	晶	日日日

The Cangjie code OM (人一) may serve as an extreme example of the coding ambiguity – it describes five different characters: 丘, 仝, 仜, 厶, 全.

The idiosyncrasy of CIM coding and the sources of ambiguity will not be discussed here, but it should be pointed out that the peculiar approach to the notion of a 'component' results in treating the traditionally distinctive elements as belonging to the same symbol category.

The Cangjie Input Method is a coherent system, as farfetched as it may seem, the CIM codes may serve as an structural approximation to alphabetical writing systems, but it should be remembered that the componential structure of characters is not reflected in a straightforward way in CIM.

2. Graphotactics

The quantitative properties of the componential structure of Chinese characters in terms of CIM components are presented in Section 3 from the graphotactic perspective. At this point it is necessary to introduce the terminology and theoretical premises the quantitative analysis is based on. The abridged introduction presented below is based on Bańczerowski (2009) and Kordek (2012 and 2013).

The idea of graphotactics originates from the new concept of phonotactics proposed by Bańczerowski (2009). The affiliation of graphotactics with Bańczerowski's proposal is discussed in Kordek 2012 and 2013, therefore it is not necessary to go into the details here. Instead, this section introduces a fragment[4] of basic graphotactic terminology and

[4] The introduction excludes, i.a., distributional properties of graphemes.

provides an exemplary graphotactic analysis of a small sample of Chinese characters to facilitate an understanding of the proper, large scale analysis performed in Section 3.[5]

The ***grapheme*** is a component part of a character at any layer of decomposition, excluding the strokes. The ***graphotacteme*** is the spatial representation of minimal graphical units of script in terms of graphemes (a character). 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) ***graphotactemic efficiency*** – the ratio between the graphotactemicity and the graphemicity of a given tactographeme.

2.1. 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 the next section. The introduction is limited to the quantitative properties of CIM graphotactemes and tactographemes.

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 Section 3.

Table 2.1. contains the exemplary sample of Chinese characters together with corresponding tactographemes.

Table 2.1. Sample set of tactographemes and characters generated out of it

Tactographeme	Graphotactemic range (characters)
{木}	{木 *mù* ‘tree’, 林 *lín* ‘woods’, 森*sēn* ‘forest’};
{一, 日}	{旦 *dàn* ‘dawn’, 亘 *gèn* ‘continuous’};
{一, 丨}	{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’}

[5] The exemplary analysis is more or less directly quoted from Kordek 2013.

The first set of data shown in Table 2.2. concerns the graphemicity of each tactographeme.

Table 2.2. Graphemicity and graphotactemicity of tactographemes

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

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 graphotactemes, 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:

Table 2.3. Graphotactemic range and efficiency of tactographemes

Tactographeme	Graphotactemic range	Graphotatemic efficiency
{木}	{木, 林, 森}	3
{一, 日}	{旦, 亘}	2
{一, 丨}	{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 graphotactemic 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, 丨}, {句, 多}, {木, 日}, {一, 大}}.

Tactographemicity is the number of tactographemes 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 Table 2.4.

Table 2.4. Properties of tactographons

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

3. Graphotactic analysis of Cangjie codes

The analysis of CIM encoding is based on the corpus of 27,607 codes representing 30,301 characters[6], the difference being the result of ambiguity in coding. The calculations revealed following quantitative properties of the investigated set:

- 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.

The average tactographemic efficiency is significantly higher than calculated by Bańczerowski and Wierzchoń for Polish letters (1.36) and Chinese *pīnyīn* transliteration (1.11).[7]

---

[6] http://en.wiktionary.org/wiki/Wiktionary:Chinese_Cangjie_index.

[7] Bańczerowski 2009: 15–19.

The average length of CIM code sequences may be viewed as a measure of the average complexity of CIM graphotactemes in terms of CIM graphemes per character.

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. Table 3.1. below presents more detailed quantitative data pertaining to the categories of graphemicity.

Table 3.1. 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 Chinese characters will be presented with the use of diagrams, with comments when necessary.[8]

3.1. CIM tactographemicity and t-graphotactemicity

In Fig. 3.1. the tactographons are plotted on the x-axis according to their graphemicity, while tactographemicity is plotted on the y-axis. It 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.

[8] The quantitative data presented here were obtained prior to the publication of Kordek 2013, but because of the publication dates and the extensiveness of presentation it should be acknowledged that this section of the paper is based on Kordek 2013, some parts being a direct quote.

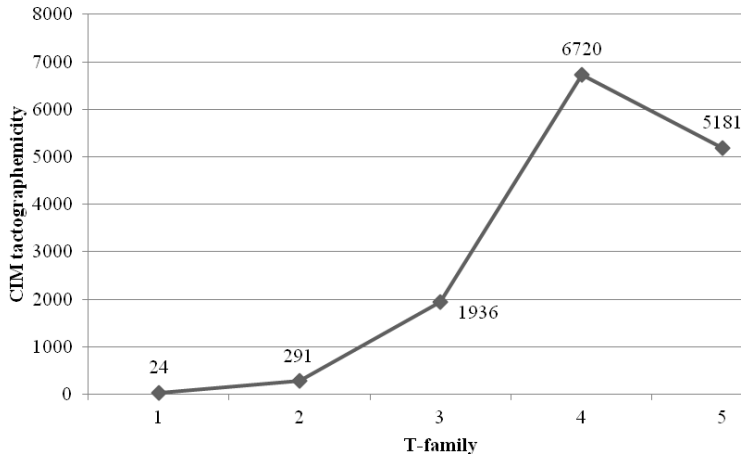

Fig. 3.1. CIM tactographemicity by tactographons (t-families)

In Fig. 3.2. the tactographons are plotted on the x-axis according to their graphemicity, while t-graphotactemicity is plotted on the y-axis.

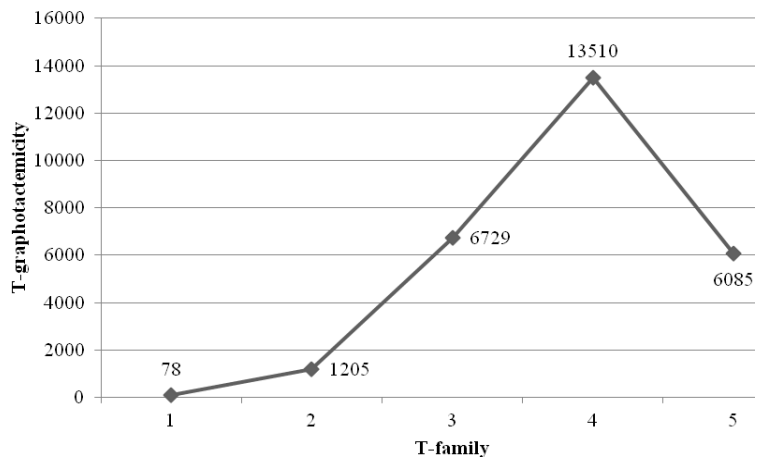

Fig. 3.2. CIM t-graphotactemicity by tactographons (t-families)

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 graphotacteme.

3.2. T-efficiency

T-efficiency pertains to the graphotactemic efficiencies of individual tactographons (members of t-family).

Fig. 3.3. shows that the tactographemes consisting of a smaller number of CIM graphemes (graphemicity 1 to 3) on average produce most CIM graphotactemes.

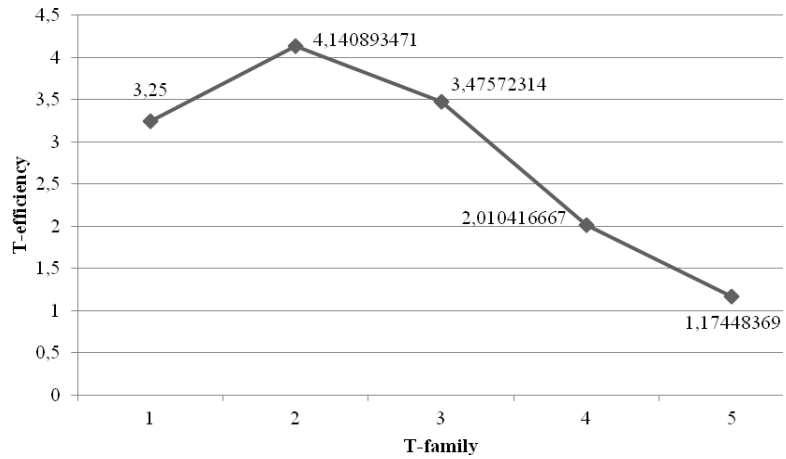

Fig. 3.3. CIM t-efficiency

3.3. CIM categorial graphotactemic efficiency

Another type of graphotactic data that can be extracted from the corpus is the cardinality of individual CIM tactographons.

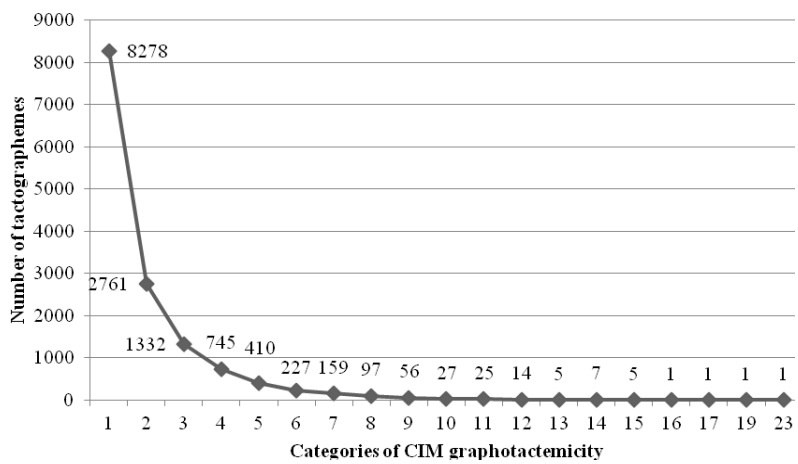

Fig. 3.4. The number of CIM tactographemes generating a given number of graphotactemes

Fig. 3.4. illustrates the numerosity of categories of graphotactemicity. A category of graphotactemicity is a set of all tactographemes having the same graphotactemic load (graphotactemicity), or, in the other words, generating the same number of graphotactemes. There are graphotactemic 19 categories of graphemicity ranging from 1 to 23. As expected, 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.

The author hopes that investigation presented in this paper contribute to the establishment of the analytic model outlined in Bańczerowski (2009) as legitimate research tool. More complete results in Kordek 2013 that include graphotactic analysis of Chinese characters in terms of traditional constituents provide further, even more compelling evidence.

Bibliography

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).

Bańczerowski, J. 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.

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).

Coulmas, F. 2003. *Writing systems. An introduction to their linguistic analysis.* Cambridge University Press.

Kordek, N. 2012. "Segmentotactics of Mandarin Chinese". *Rocznik Orientalistyczny* LXV.1: 107–119.

Kordek, N. 2013. *On Some Quantitative Aspects of the Componential Structure of Chinese Characters.* Poznań: Wydawnictwo Rys.

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).

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).

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).

Yin, B., Rohsenow, J.S. 1994. *Modern Chinese Characters.* Beijing: Sinolingua.