The Chinese Script as a Self-Regulating System

Applying Köhler's Basic Model of Synergetic Linguistics to Simplified Chinese Characters

Cornelia Schindelin

Abstract. Köhler's basic model of synergetic linguistics endeavors to show language (sub-)systems as dynamic systems the units (of various levels) of which interrelate directly or indirectly. These relationships are controlled by needs or constraints which interact in complex ways. This study adapts and applies Köhler's basic model to modern simplified Chinese characters and tests the hypotheses it provides about the direct and indirect relationships between character frequency, graphical complexity, and functional complexity. The hypotheses are tested on data from a large corpus study published in the People's Republic of China in 1986.

Three hypotheses about direct relationships and three about indirect relationships between the three systemic features were operationalized and tested. While all three hypotheses about direct relationships could be accepted based on goodness of fit, this was not the case with all three hypotheses about indirect relationships. Here, the model or at least its adaptation—including the operationalization of "functional complexity"—seems to need improvement. Further study is needed.

1. Introduction

Modern Chinese characters seem to show some systemic features which correspond to those already examined on the lexical level of various languages.[1] For example, there are differences in text frequency (token frequency) among Chinese characters just as some words of any language are more frequently used than others. And just as more frequent words are shorter on average than less frequent ones, more frequent Chinese

Cornelia Schindelin 0009-0007-8724-4751
FTSK, Johannes Gutenberg-Universität Mainz, An der Hochschule 2, 76711 Germersheim, Germany. E-mail: schinc@uni-mainz.de

1. Including modern Chinese, see (L. Wang, 2011), (Lu Wang, 2014b), and (Lu Wang, 2014a).

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 10.
Fluxus Editions, Brest, 2024, pp. 739–770. https://doi.org/10.36824/2022-graf-schi
ISBN: 978-2-487055-06-3, e-ISBN: 978-2-487055-07-0

characters also seem to be, on average, graphically and/or structurally simpler than less frequent ones.[2]

Synergetic linguistics views language (sub-)systems as self-regulating systems[3], somewhat like ecosystems, and endeavors to model them as dynamic systems the units (of various levels) of which interrelate in certain ways, directly or indirectly. These relationships are controlled by needs or constraints which interact in complex ways.[4] For example, writers would want characters to be easy to write and thus tend toward a minimization of the coding effort while readers would want them to be easy to differentiate and thus prefer a minimization of the decoding effort. Both readers and writers would want the whole inventory to be limited in size so they would not need to learn endless numbers of graphical signs. However, at the same time, they would want each graphical form to be as unambiguous and specific as possible which would require graphically different forms for different morphemes of their language and thus result in an expansion of the character inventory. These conflicting needs and interests can push the system to develop in one or another direction.

Quantitative linguistic research has already identified and described, in mathematical or statistical terms, relationships between variables like unit length, its text frequency and frequency rank, its complexity, its breadth of usage, and so forth. These relationships or dependencies can be formulated as a "hypothesis" or even a "law" of quantitative linguistics and may be given a name which often honors the first person to describe the respective relationship, like "Zipf's law".[5] A further step would be to integrate the various hypotheses and "laws" into one model which considers the direct relationships already explored but also allows to derive, operationalize, and test indirect relationships.

Köhler's basic model of synergetic linguistics (Köhler, 1986) is such an attempt at formulating an integrated model.[6] The hypotheses that can be derived from it have been tested on data from various languages.[7]

In this study the attempt was made to apply the model to Chinese characters and find out if the relationships described by the model also hold for Chinese writing. So, a corresponding formulation of the model

2. For reasons of space we shall not go into the question of what a "word" is. For a discussion pertaining to modern Chinese, see (Duanmu, 2017). For a summary of studies on word length in Chinese see (Schindelin, 2017c).

3. This idea has been picked up in China as well, cf. (Wáng, 1995).

4. Cf. Altmann and Köhler (1996).

5. See Zipf (1932), for one of Zipf's seminal publications. For information on research done on the validity of Zipf's law for Chinese, see Schindelin (2017d).

6. For an introduction in English see Köhler (2005).

7. See Köhler (2004), for some examples.

was constructed and examined, and the results of this endeavor are presented here.[8]

If in the following simply the word "characters" is used, it is intended to mean "Chinese characters". After all, the principle of least effort presumably is an universal principle.

2. The Chinese Script

Modern Chinese writing, that is, the characters being used by the speech community to record modern Chinese language utterances, is best described as a morpho-syllabic writing system.[9]

Nearly 90 percent of the characters within the modern Chinese character inventory represent morphemes, approximately each half representing free and bound morphemes, respectively. The remaining 11 percent either stand for unique morphemes ("cranberry morphemes") or representations of submorphemic parts of disyllabic or polysyllabic morphemes which need two or more characters to be written down completely (cf. DeFrancis, 1984, p. 185). A certain number of Chinese characters may (as types) be employed to represent different morphemes and also have different pronunciations or readings which, however, does not mean that every morpheme has its own reading for the character concerned.[10]

A character in a text is read, when read out loud carefully, as one syllable, with a few minor exceptions which can be disregarded here.[11] The overwhelming majority of words in the lexicon (i.e., the word inventory) are disyllabic and thus are written down using two characters (tokens).[12] Chinese words, if the remark is allowed, do not have inflectional endings because grammatical relationships between words and between clauses are expressed mainly through their positioning within the sentence ("word order") and by lexical means.

---

8. This article is largely an English version of Menzel (2004).

9. For a well readable treatment, take DeFrancis (1984). Schindelin, 2007, pp. 6–7, presents the viewpoint of the Chinese scholar Qiú Xīguī 裘锡圭 who argues that the system should be described as one whose characters are made up of signific components, phonetic components and purely mnemotic components and thus argues for a Chinese term for it which can be translated as "semanto-phonetic writing" (意符音符文字 *yìfú yīnfú wénzì*, cf. Qiú, 1988, p. 18 and Qiú, 2000, p. 26).

10. For the distribution of number of readings per character among the commonly used characters, cf. Schindelin, 2007, p. 166.

11. The most obvious exception is the character 儿 which in many cases is used to write out the rhotacized version of a syllable.

12. Again I would like to refer the reader to DeFrancis, 1984, pp. 177–188, this time for his treatment of the "Monosyllabic Myth".

As for the size of the modern Chinese character inventory, various frequency counts conducted in the second half of the last century arrived at different numbers of currently used characters ranging from about 4,500 to more than 7,500 character types. The last number, however, was found by just one count which examined a corpus of nearly 12 million characters (tokens) in size. Two other research projects which examined corpora of around 21 million and around 40 million characters (tokens) in size, found 5,991 and 6,001 different characters (types), respectively. The character dictionary *Xīnhuá zìdiǎn* 新华字典 which up until the age of the smartphone could be found in nearly every household of the People's Republic of China (PRC) lists around 11,100 character entries, and when multiple entries resulting from characters having more than one reading are discounted, there are still over 8,000. People having mastered 1,500 frequently used characters are regarded to be officially "semi-literate" in the PRC while having mastered 3,000 frequently used characters makes one officially "literate". There is an official list of 2,500 most frequent characters and one of the next 1,000 frequent characters. Having mastered these, in sum, 3,500 frequent characters should enable one to recognize 99.48 percent of all the characters in ordinary texts, that is, non-specialized, everyday texts. The next 1,000 characters on the frequency list would add another 1.51 percent to that. If one were to take one thousand more characters, the added percentage contributed by these would be even lower. (Schindelin, 2005b; 2017a)

In the 1950s the government of the PRC implemented a language reform[13] aimed at making reading and writing easier for the broad masses of its people. During this reform 2,264 traditional character forms were replaced by 2,236 simplified ones. In most cases, character components occurring in a number of characters were simplified in (nearly) all characters they are a component of, which in effect simplified a lot of characters at once.[14] The simplified forms more often than not were forms which had been used in handwriting for a long time already, so they were familiar vulgar forms which now rose up to be standard ones.[15] Other methods of simplification were the renewed uptake of graphically simpler archaic forms,[16] the replacement of complicated components by simpler symbols[17] and sometimes by simpler phonetic components

13. For a short introduction to writing reform in the PRC see Chen, 1999, pp. 148–159.

14. For an introduction in English with further examples, see Yin and Rohsenow, 1994, pp. 103–112.

15. The traditional character 書 *shū*, book, was replaced by 书, a form already popular in handwriting. The component 言 (as in 說, *shuō*, to say) was replaced by the handwriting form 讠 (说) in all the characters containing it on their left, and so forth.

16. Trad. 雲 *yún*, cloud → 云.

17. Trad. 難 *nán*, difficult → 难.

which may or may not reflect current pronunciation better than the traditional ones,[18] the discarding of graphical components while keeping the overall contour or a salient component of the character,[19] and the creation of new associative compounds.[20] A combination of methods may have been applied to a traditional character in order to get a simpler form. The difference between the number of abolished characters and the number of simplified characters is the result of the reformers' merging characters for several morphemes which earlier had had their "own" character each to be represented by into just one resulting character with several meanings, i.e., able to represent more morphemes than before, although the morphemes were at least nearly homophonous and usually the original characters had had some similarity, like sharing a certain component.

Before the advent of the digital age, every printed text in China had had a hand-written original as its predecessor, so the need to reduce the required writing effort could understandably lead to differences between hand-written and printed versions of the same character, the latter conforming to the standard orthography. In a sample of 152 characters in their printed and hand-written form, about a third had the same number of strokes in both forms. 43 percent only had one stroke less in their hand-written form than in their printed form. So nearly three thirds of the characters examined were only very slightly or not at all "shorter" (counting their number of strokes) than their printed counterparts.[21] In other words, the number of strokes of printed simplified characters quite closely reflects the number of strokes of the handwritten form, which is helpful as we want to take number of strokes as an indicator of the effort it takes to write a character.

The first frequency count of characters of contemporary texts in China was done in 1927. In the last century the motivation for such research was mostly inspired by goals of writing reform or pedagogy. The size of the corpora examined has grown immensely with the development of modern computerized tools. Quantitative linguistics may not be a household name in China—and China in this respect is not different from other countries—, but quantitative research on language and writing, including on corpora, has grown quite a bit in recent decades. (Cf. Schindelin 2005a,b)

18. Trad. 畢 *bì*, to finish → 毕. However, to improve phoneticity obviously was not a priority.

19. Trad. 廠 *chǎng*, factory → 厂. Trad. 開 *kāi*, to open → 开.

20. Trad. 塵 *chén*, dust → 尘.

21. Unpublished study by this author.

3. The Corpus

The corpus at the bottom of the frequency data used for the present study consisted of texts written in simplified Chinese characters as used in the PRC. It encompassed texts of 1,808,114 character tokens or about 1.31 million running words altogether which turned out to use an inventory of 4,574 character types. The corpus had been put together by the original researchers with didactic purposes in mind. Their aim was to reflect contents and text types which an inhabitant of the PRC of average education would read. Thus, it consisted of factual prose (about 40 percent), drama, fictional prose and essays as well as folk-tales. The counting only considered Chinese characters while punctuation marks, non-Chinese numbers, Latin letters and such were ignored. The resulting data were compiled and published in a frequency dictionary[22].

The *Frequency Dictionary* contains word lists as well as a list containing each character found along with its absolute and relative frequency and its rank[23]. The list furthermore contains data on the number of words the respective character is part of in its written form in the corpus, how many different words it can be found in, in how many cases—in di- and polysyllabic words—it appears at the beginning, in the middle or at the end of the word or whether it can only be used to write monosyllabic words. The distribution of these cases is as follows:

- 217 characters (= 4.7 percent) only write monosyllabic words;
- 1,620 characters (= 35.5 percent) only occur in di- or polysyllabic words, of these 519 characters only ever occur at the beginning of words, 39 exclusively in the "middle" (which is not further specified) of words, 433 exclusively at the end of words, and 168 can appear in all three positions;
- 2,737 characters (= 59.8 percent) appear in texts as representations of monosyllabic words as well as parts of longer words.

This data set was chosen for the present study because it seemed sufficiently big in size and because the *Frequency Dictionary* provided more data than just frequencies and ranks. In light of the facts reported above about corpus studies and inventory sizes that have been variously published it appears that an inventory of 4,574 character types should be able to yield meaningful results.

---

22. Simply called *Frequency Dictionary* here which refers to 现代汉语频率词典 *[Frequency Dictionary of the Modern Chinese Language]*, Beijing, 1986.

23. More precisely: its ordering number, as characters of the same frequency and thus rank still have different numbers in this list.

4. The Basic Model and the Chinese Script

The underlying assumption of the following adaptation of Köhler's basic model for application to the Chinese character system is that this system has a structure with respect to its properties and processes which corresponds to that of the lexical system, which is why a similar behavior is expected for the relationships between corresponding variables. As far as the functional dependencies of the system variables are concerned, the same differential equation is used as a mathematical model which Köhler used for his basic model. The solution of the differential equation and its linearized form are taken over as well.

The "language" examined in the following sections is, to be clear, the Chinese character system and *not* the "Chinese language" or its lexicon.[24] Any findings or conclusions, therefore, should not simply be also applied to the "language" as a whole nor to its "lexicon" in the sense of its inventory of words.

"Inventory size" in the adapted model corresponds to Köhler's "lexicon size". The need[25] to encode a message (Cod) is the desire to graphically encode syllables of the Chinese language using characters which are different for each morpheme (as there are homophonous morphemes). The higher the number of syllables and morphemes which need to be written, the bigger the character inventory has to be.

There is another need running counter to the need to encode which is the need to minimize inventory size (minI) because the capacity of the brain to memorize characters is limited; this need is served by the fact that many character types can be used for various morphemes and their corresponding syllables. Inventory size is operationalized as the number of different characters (types) which were found in the corpus.

Number of components[26] in this adaptation of Köhler's basic model corresponds to number of phonemes in his original version. It is the number of character components or minimal component graphemes identified through minimal pair analysis. The size of the component inventory is influenced by the need to minimize the coding effort (minC) on part of the writer and the need to minimize the decoding effort (minD) on part of the reader. "minC" demands the inventory to be as small as possible and its elements to be as simple as possible, so the components can be executed swiftly without having to make many different

24. Lu Wang (2014b) undertook a study of Chinese word lengths confronting Köhler's model with a corpus of texts taken from the newspaper *People's Daily* (人民日报 *Rénmín Rìbào*). As the present study is concerned with writing and character complexity, Wang's study is not discussed here.

25. In Fig. 1 below, the needs which "pull" at the systemic features are represented by abbreviations in oval shapes.

26. Or: component graphemes.

movements. "minD" on the other hand demands the elements to be well distinguishable from one another in order to make characters easy to identify.

To test the hypotheses suggested by the adapted basic model, the component inventory which resulted from Bohn's minimal pair analysis was used (Bohn, 1998, pp. 12–14). So "number of components" in this study refers to the components Bohn found as they occur in the character types of the corpus used here.

"Graphical complexity" here corresponds to the length of lexical units in the original model. As elaborated above, Chinese characters when written by hand demand different amounts of effort. Characters consisting of more strokes require more effort than those with fewer strokes. Characters consisting of more components also require more effort to write down than those with fewer components even though the latter may in fact have fewer strokes than the former. As ibid., pp. 20–24 has shown, Menzerath's law holds for the relationship between the average number of character components and their average number of strokes, which means that characters which have more components on average consist of components with fewer strokes than those characters which consist of fewer components.[27] However, the arrangement of several components on paper within a small hypothetical rectangle is more difficult than having to arrange just two components which is why graphical complexity shall be measured both in number of strokes and number of components.[28]

Even finer measurements of the effort it needs to write Chinese characters by hand can be thought of but they would be relatively laborious to operationalize. What can be accomplished, though, is considering the different types of strokes which can be assigned different values of effort according to whether or not they change direction and if so, how many times. So Bohn's measures of stroke complexity (ibid., p. 15) were also used to measure graphical complexity with a finer grain.[29]

Inventory size and number of components affect character complexity in the same way lexicon size and number of phonemes affect word length in the original model.[30] The need for redundancy (Red) strives

27. For a summary of Bohn's study in English see Schindelin (2017b).

28. Wang and Chen, 2015, p. 238, also studied the question whether number of strokes or number of components is a better measure of character complexity. They come to the conclusion that both are "proper measurements".

29. For suggestions on how script complexity can be measured, see Altmann (2004).

30. Inventory size affects the microprocesses responsible for unit length globally, as Köhler, 1990, p. 184, points out; it does not determine the length of individual lexical units.

to avoid the appearance of characters which are too similar and thus has an effect on graphical complexity.

In this research we shall use the term "functional complexity" to refer to the fact that in many cases the same character can be used for various morphemes and words. There is a relationship between graphical complexity and functional complexity which is influenced by the need for specification (Spc). Diachronically speaking, "Spc" had the effect that characters which were used to represent different morphemes in different contexts were made more complex by adding a component to yield a more specific character for a certain meaning. To give an example: This process caused the character for the word *lái* "wheat", originally written 來, which was borrowed to write the homophonous word *lái* "to come" for a while, to be made more specific and at the same time more complex by adding the component 艸[31], called the "grass component", at its top to express that the resulting combination, the descendant of which is now written 萊[32], specifically meant the grain and not the motion verb. Thus, "Spc" has the effect of enlarging the inventory which in this modeling is contained within the need to encode, "Cod". Synchronically, "Spc" refers to the need to write more complex characters in order to achieve less ambiguous expressions, given the inventory at hand.[33]

The historical process which led to new characters lets us presume that within the present character inventory, characters with more components on average have a lesser functional complexity than characters consisting of fewer components.

The needs "minC" and "minD" also have effects on functional complexity and need to balance one another out. "minD" strives for lower functional complexity as readers would like to quickly and effectively decode which morpheme is represented by the character they are seeing. "minC" on the other hand strives for higher functional complexity because having fewer characters which can each be employed for more meanings or morphemes allows writers to comfortably utilize fewer character types.

Functional complexity helps the need "minI" as the inventory can be smaller when each of its elements, the characters, on average has a higher functional complexity.

A comparison of the adapted model with Köhler's original basic model shows functional complexity to be the integration of a part of the structure which models the relationship between properties Köhler called "polylexy" (the number of meanings of a word) and "polytexty"

31. Its modern simplified form is ⺾.

32. I give the traditional full form characters for the two words here because the addition happened long before the writing reform of the last century.

33. In China as elsewhere orthography is largely standardized and leaves the individual little space to choose signs according to their own whim.

(number of cotexts a word can be used in) and the needs affecting them in the original version. Simplifications like this are possible and allow for the calculation of more complicated systems (Köhler, 1986, pp. 48–49). In the original basic model, polytexty is a function of polylexy (ibid., p. 67), and frequency is a function of polytexty (ibid., p. 68). Apart from this, it could be shown for the basic model that frequency indirectly is a function of polylexy (ibid., p. 74). Thus, simplifying the model in the way done here should not damage it. The practical reason for this choice is that for the characters of the inventory we only know how many word types contain them in this corpus but we do not know in how many different texts they occur.

The relationship between graphical complexity and functional complexity in the linearized model can be expressed through the equation

$$\text{L-functional complexity} = Q_2 * \text{minC} - Q_1 * \text{minD} - T * \text{L-graphical complexity}.$$[34]

To test this hypothesis the functional complexity of each character was operationalized as the number of the various mono-, di- and polysyllabic words (i.e., word types) it appears in within the lexical inventory of the corpus.

Characters with higher functional complexity presumably appeared more often in the corpus than those with lower functional complexity. This means, their frequency would be a function of their functional complexity, as the need to use a certain character (Use) would have an effect on its frequency. This relationship is modeled as a directly proportional dependency in the model:

$$\text{L-frequency} = R + \text{Use} + K * \text{L-functional complexity}.$$

It is known for each character how many times it was used in the corpus.

Back to graphical complexity once more. It is known that within the character system there is a relationship between the text frequency of characters and their graphical complexity similar to that between the text frequency of words and their length. The need to minimize the effort of production (minP) of each character can be seen to manifest itself in simplifications and abbreviations of characters that need to be written often. One can even say that "minP" has driven Chinese characters to evolve from their ancient forms to their modern forms as well as various swiftly executable handwritten forms which are still in use today for private or semi-official use. When the people responsible for script reform in the 1950's declared a large number of simplified characters already in use to be the new standard forms, they acknowledged

34. An "L-" signals the use of the logarithmized (linearized) form of the equation and its variables.

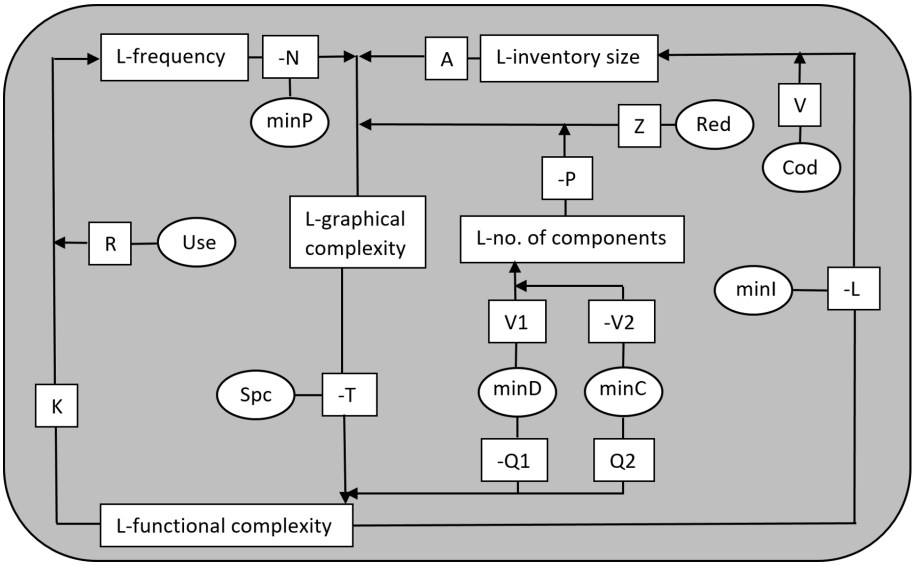

FIGURE 1. Adapted basic synergetic model of the Chinese character system, linearized (i.e., logarithmized)

the results of this "natural" development. And when one examines Chinese characters synchronically, it is quite evident that frequently used characters are "shorter," that is, less complex than characters which are used more rarely.[35]

Graphical complexity is modeled in this equation:

$$\begin{aligned}\text{L-graphical complexity} = {} & A * \text{L-inventory size} + Z * \text{Red} \\ & - P * \text{L-number of components} \\ & - N * \text{L-frequency}.\end{aligned}$$

The complete adapted model is shown in Figure 1.[36]

4.1. The Hypotheses

The adapted model enables us to derive three hypotheses about direct relationships and another three hypotheses about indirect relationships

35. Cf. Schindelin (2017a).

36. Of course, Köhler's model has not remained without criticism. See Hammerl and Maj (1988), and Maj (1990), for instance. The debate, during which replies by Köhler were also published and discussed in turn, somewhat continued through the volumes of *Glottometrika* in the following years.

between systemic features. The direct functional relationships are (in their non-linearized forms):

$$H_1 : \text{functional complexity} = A_1 * \text{graphical complexity}^{B_1}$$
$$H_2 : \text{frequency} = A_2 * \text{functional complexity}^{B_2}$$
$$H_3 : \text{graphical complexity} = A_3 * \text{frequency}^{B_3}.$$

Insertion yields the following three hypotheses about indirect functional relationships:

$$H_4 : \text{graphical complexity} = A_4 * \text{functional complexity}^{B_4}$$
$$H_5 : \text{functional complexity} = A_5 * \text{frequency}^{B_5}$$
$$H_6 : \text{frequency} = A_6 * \text{graphical complexity}^{B_6}.$$

The six hypotheses were verified using the linearized model with the help of the statistics software package SPSS. Multiple linear regression was performed using the method of least squares fit. For all relationships examined there were replicated responses. Therefore, the means of the replicated responses weighted with the number of values was used for the independent variable. Data points with a weight of 5 or less were in general excluded from regression. When no such exclusion was made, it shall be mentioned below. To measure the quality of the fit, the determination coefficient R^2 was used, below abbreviated as D (for coefficient of determination). A fit was considered good when D reached at least the value 0.9.

4.2. Direct Functional Dependencies (H_1–H_3)

4.2.1. *Functional complexity as a function of graphical complexity*

The functional complexity of Chinese characters is directly a function of their graphical complexity. It is lower when their graphical complexity is higher. In the linear model, the equation is

$$\text{L-functional complexity} = \ln A + B * \text{L-graphical complexity}, \qquad (H_1)$$

where B is expected to be negative.

Graphical complexity was measured in three ways: (a) number of strokes, (b) number of component graphemes, and (c) sum of the effort values of each stroke of the character counting their change of direction when being executed manually, called "writing effort" below. Regression was applied to each of these data sets including the data of all characters.

(a)	No. of strokes:	$D = 0.956$	$A = e^{5.59} = 268.12$	$B = -1.373$
(b)	No. of components:	$D = 0.953$	$A = e^{3.666} = 39.09$	$B = -1.133$
(c)	Writing effort:	$D = 0.95$	$A = e^{6.086} = 439.72$	$B = -1.44$.

As expected, the value of B is negative and in addition has very similar values in all three kinds of measuring. The values of A differ as the values of the entities counted are very different in absolute numbers.

Figures 2 through 4 show the data points as well as the curve of the respective function in non-logarithmic form.

The quality of the fit as well as the visual appearance of the curves in relation to the data points suggest that the first hypothesis can be accepted.

4.2.2. *Frequency as a direct function of functional complexity*

The text frequency of Chinese characters is a function of their functional complexity. It is higher for characters with a higher functional complexity. In its linearized form, H_2 is expressed as

$$\text{L-frequency} = \ln A + B * \text{L-functional complexity}, \qquad (H_2)$$

and a positive value is expected for B.

Regression was performed on the complete data set. The fit was very good: D = 0.958. Figure 5 shows the data points and the curve of the derived function in non-linearized form. For he sake of graphical resolution, only data points with weights >5 were included in drawing the figure. As functional complexity gets higher, the data points are more widely scattered around the curve. The curve seems to reflect the tendency of the relationship nonetheless. This and the quality of the fit leads us to accept the second hypothesis as well.

4.2.3. *Graphical complexity as a function of text frequency*

The graphical complexity of Chinese characters is a function of their text frequency. Graphical complexity is on average lower when text frequency is higher. In linearized form the equation is

$$\text{L-graphical complexity} = \ln A + B * \text{L-frequency}. \qquad (H_3)$$

Again, a negative value for B is expected.

Text frequency is measured as the absolute number of occurences of each character in the corpus. Possible values are very disparate. Especially among very frequent characters there are hardly any two with the same frequency. So in order to use frequency values feasibly as the independent variable they were condensed into frequency classes. The resulting classes were weighted with the number of data points in them. Two class widths were chosen: 50 and 100. Frequency classes which contained only five data points or less were excluded from regression. The central value of the class was chosen as the value of the independent

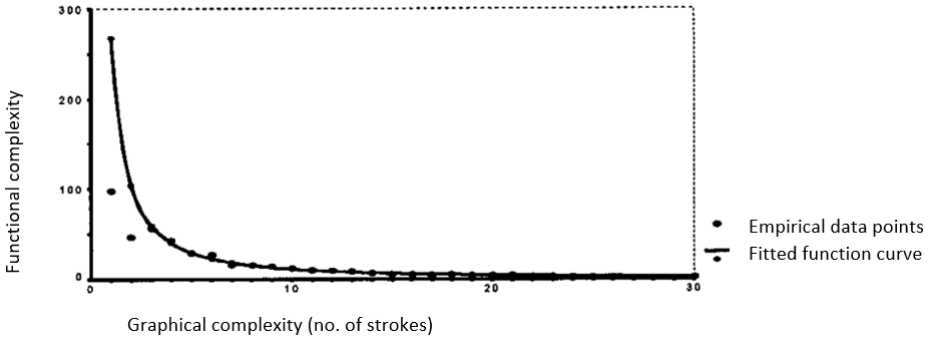

FIGURE 2. Functional complexity as a function of graphical complexity, measured in number of strokes

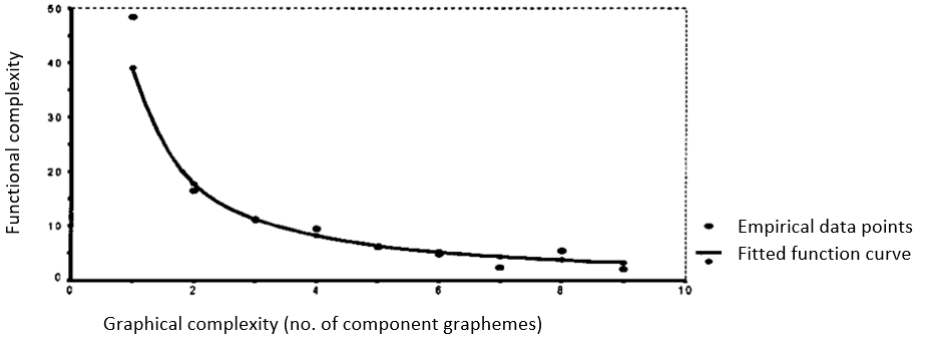

FIGURE 3. Functional complexity as a function of graphical complexity, measured in number of component graphemes

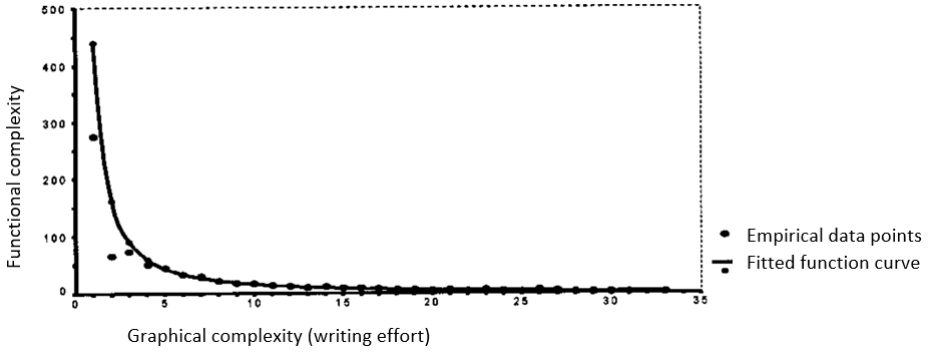

FIGURE 4. Functional complexity as a function of graphical complexity, measured in writing effort

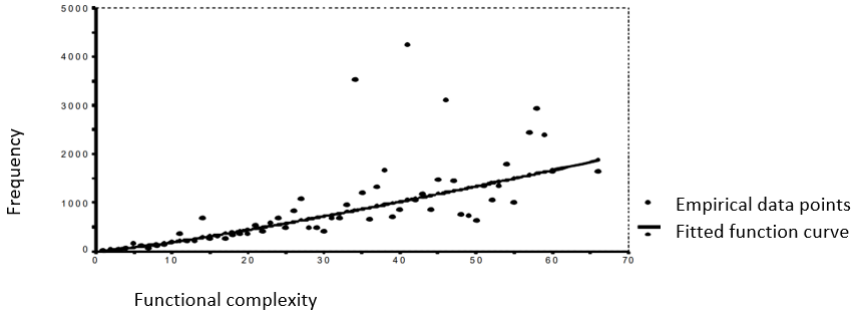

FIGURE 5. Text frequency as a direct function of functional complexity

TABLE 1. Results for H_3 for both class widths and three ways for measuring complexity

	Measure	Class width 100	Class width 50
(a)	No. of strokes	$D = 0.94$	$D = 0.93$
		$A = e^{2.846} = 17.22$	$A = e^{2.72} = 15.18$
		$B = -0.114$	$B = -0.094$
(b)	No. of components	$D = 0.95$	$D = 0.897$
		$A = e^{1.51} = 4.53$	$A = e^{1.4} = 4.066$
		$B = -0.0958$	$B = -0.078$
(c)	Writing effort	$D = 0.946$	$D = 0.92$
		$A = e^{3.057} = 21.28$	$A = e^{2.94} = 18.88$
		$B = -0.11$	$B = -0.09$

variable. Regression was performed for each of the three ways graphical complexity was measured in this study. The results are shown in Table 1.

B has, as expected, a negative value and is quite close to -0.1 in five out of six cases. The differences between the values is a little bigger for the grouping in classes of width 50. Again, as the absolute values of the three kinds of measures vary quite a bit, so do the values of A, but variation for the same kind of measurement is only very small between the two classes.

Figures 6 through 11 show the data and curves for class width 100 and class width 50, respectively, in non-linearized form. In each case the data points scatter more widely around the curves as frequency gets higher.

The fits were very good and the curves do seem to reflect the relationship quite convincingly, so this hypothesis is also accepted. The criterion to include only classes with more than five data points, however, led to a substantial reduction of the data points to be considered in the regression. Especially characters with very high frequencies were

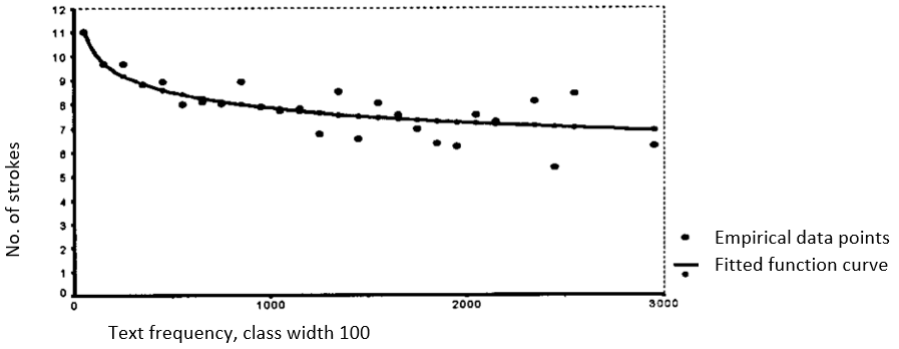

FIGURE 6. Graphical complexity measured in number of strokes as a function of text frequency

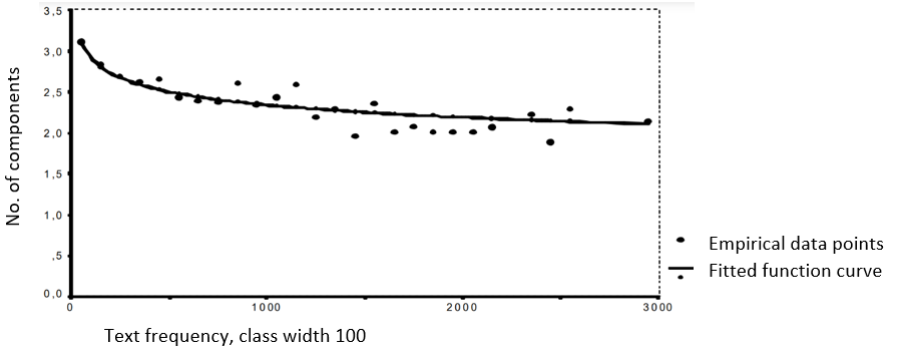

FIGURE 7. Graphical complexity measured in number of component graphemes as a function of text frequency

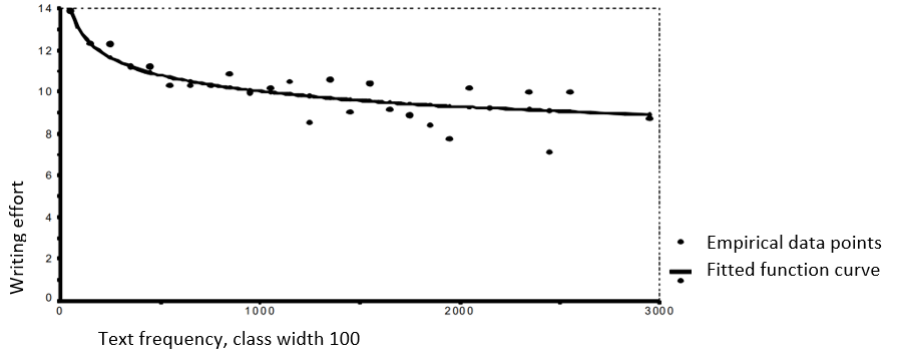

FIGURE 8. Graphical complexity measured in writing effort as a function of text frequency

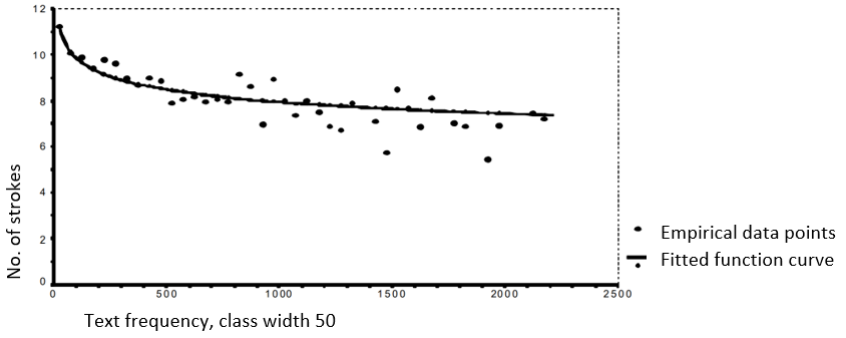

FIGURE 9. Graphical complexity measured in number of strokes as a function of text frequency

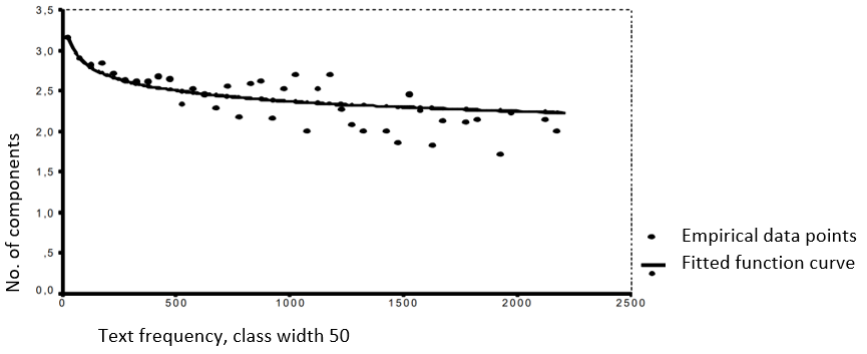

FIGURE 10. Graphical complexity measured in number of component graphemes as a function of text frequency

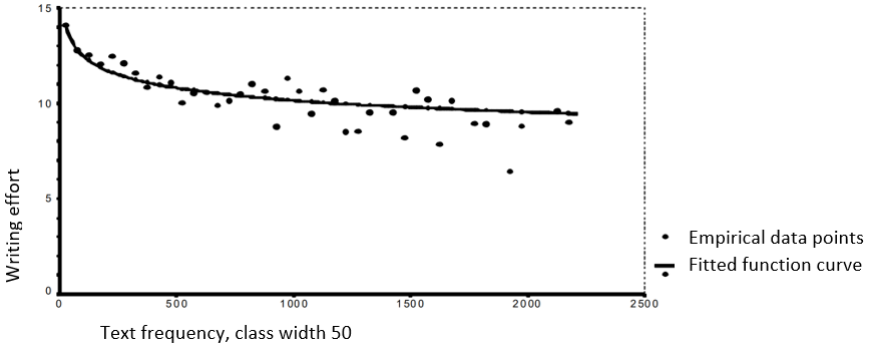

FIGURE 11. Graphical complexity measured in writing effort as a function of text frequency

excluded because among the very frequent characters, each class contained only very few data points. Thus, this relationship mainly is valid for medium and low frequency characters which, on the other hand, make up the vast bulk of the entire inventory.

Regression was also performed on the unfiltered data, that is, including all data points. The fit was not very good: $D = 0.76$ through $D = 0.85$. The parameters A and B estimated from the unfiltered data were very close to those reported above.

4.2.4. *Conclusion for the Direct Functional Dependencies*

All three hypotheses about direct functional relationships can preliminarily be accepted as their verification yielded good to excellent results. The non-linearized curves among the data points also seem very reasonable to the eye. On this basis, the indirect hypotheses are tackled next.

4.3. Indirect Functional Dependencies (H_4–H_6)

Regression on the data for the direct functional relationships have yielded estimates for the parameters. By inserting them into the equations, the indirect functional dependencies can now be modeled theoretically. Thus, it is possible to compute what the curve should theoretically look like and compare it with the curve arrived at by regression on the data. Statistical testing is applied to find out whether differences between the theoretical model and the data are statistically significant. If such a difference is not statistically significant, an indirect hypothesis can also be accepted. If the difference is statistically significant, however, it has to be rejected.

4.3.1. *Graphical complexity as an indirect function of functional complexity*

The graphical complexity of Chinese characters is indirectly a function of their functional complexity, mediated by frequency. The linearized equation is:

$$\text{L-graphical complexity} = \ln X + Y * \text{L-functional complexity}. \qquad (H_4)$$

As graphical complexity was measured in three ways and there were two class widths for frequency, we get six theoretical models:

a) Graphical complexity measured in number of strokes

$$\begin{aligned}\text{L-graphical complexity}_{a_1} &= 2.72 - 0.094 * (2.444 + 1.215 * \text{L-functional complexity}) \\ &= 2.49 - 0.114 * \text{L-functional complexity}\end{aligned}$$

and

$$\begin{aligned}&\text{L-graphical complexity}_{a_2}\\&\quad = 2.85 - 0.114 * (2.444 + 1.215 * \text{L-functional complexity})\\&\quad = 2.57 - 0.138 * \text{L-functional complexity.}\end{aligned}$$

b) Graphical complexity measured in number of component graphemes

$$\begin{aligned}&\text{L-graphical complexity}_{b_1}\\&\quad = 1.4 - 0.078 * (2.444 + 1.215 * \text{L-functional complexity})\\&\quad = 1.2 - 0.095 * \text{L-Functional complexity}\end{aligned}$$

and

$$\begin{aligned}&\text{L-graphical complexity}_{b_2}\\&\quad = 1.51 - 0.096 * (2.444 + 1.215 * \text{L-functional complexity})\\&\quad = 1.277 - 0.116 * \text{L-functional complexity.}\end{aligned}$$

c) Graphical complexity measured in effort of execution (writing effort)

$$\begin{aligned}&\text{L-graphical complexity}_{c_1}\\&\quad = 2.94 - 0.09 * (2,444 + 1,215 * \text{L-functional complexity})\\&\quad = 2.72 - 0.109 * \text{L-functional complexity}\end{aligned}$$

and

$$\begin{aligned}&\text{L-graphical complexity}_{c_2}\\&\quad = 3.06 - 0,109 * (2,444 + 1,215 * \text{L-functional complexity})\\&\quad = 2.79 - 0.13 * \text{L-functional complexity.}\end{aligned}$$

The results of regression on the actual data were:

(a)	No. of strokes:	$D = 0.73$	$A = e^{2.55} = 12.82$	$B = -0.116$
(b)	No. of components:	$D = 0.60$	$A = e^{1.25} = 3.49$	$B = -0.092$
(c)	Writing effort:	$D = 0.75$	$A = e^{2.78} = 16.19$	$B = -0.114.$

The following figures (Fig. 12 through Fig. 14) show the curves estimated from the data and the data points.

D did not come out very good. The values of the parameters arrived at by regression seem, at first glance, to get quite close to those expected on the basis of the theoretical equations, as Table 2 shows.

It seems that the parameters of the first theoretical function in each case agrees better with the function parameters arrived at through regression. As for case (a), Figure 15 shows that the two curves intersect at about functional complexity = 24. The first theoretical function generally keeps the same distance from beginning to end from the empirically estimated function curve while the curve for graphical complexity$_{a_2}$

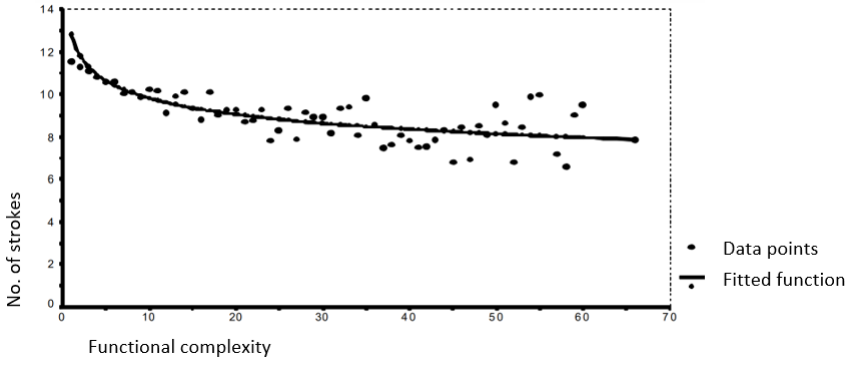

FIGURE 12. Graphical complexity measured in number of strokes as an indirect function of functional complexity, empirical fit.

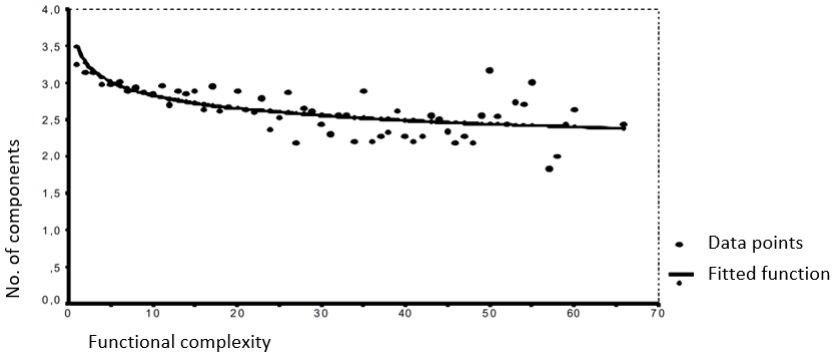

FIGURE 13. Graphical complexity measured in number of component graphemes as an indirect function of functional complexity, empirical fit.

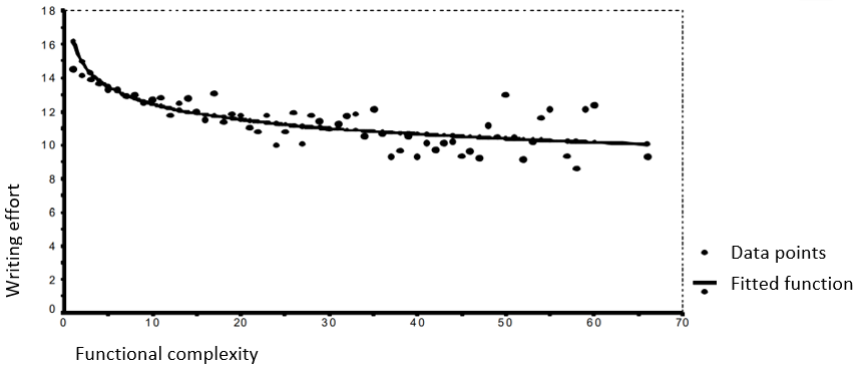

FIGURE 14. Graphical complexity measured in writing effort as an indirect function of functional complexity, empirical fit.

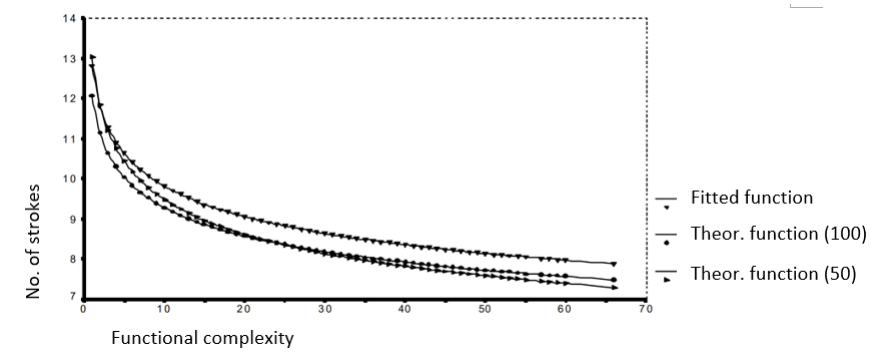

FIGURE 15. Graphical complexity measured in number of strokes as an indirect function of functional complexity, fitted function curve and theoretical function curves.

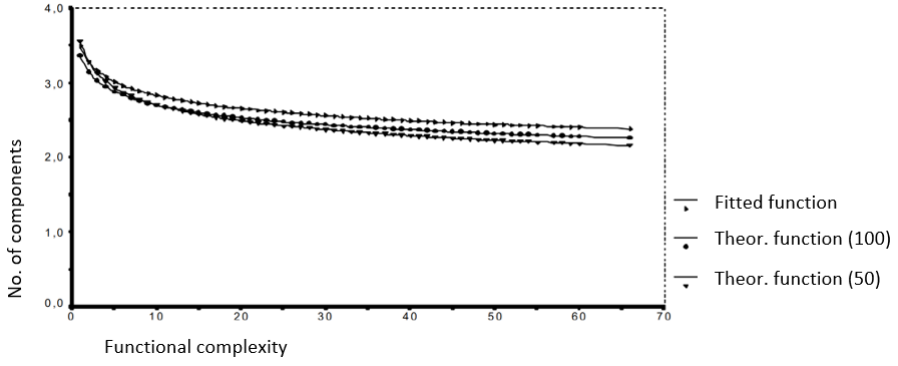

FIGURE 16. Graphical complexity measured in number of component graphemes as an indirect function of functional complexity, fitted function curve and theoretical function curves.

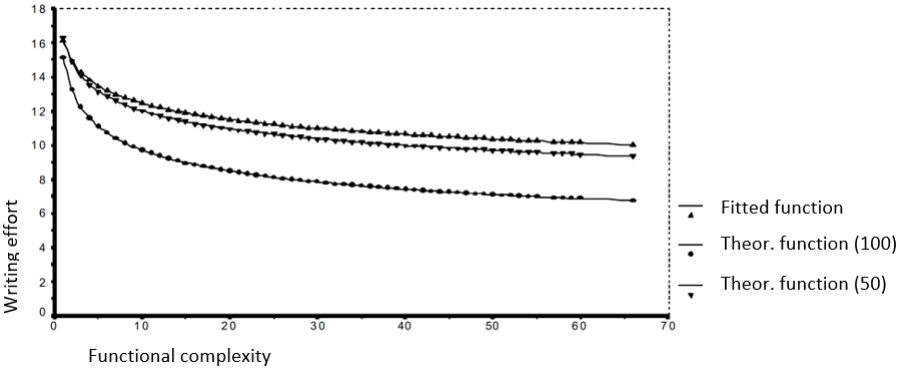

FIGURE 17. Graphical complexity measured in writing effort as an indirect function of functional complexity, fitted function curve and theoretical function curves.

TABLE 2. Results of the regressions for H_3. "fc" stands for "functional complexity."

	Measure	Function Type	Function
(a)	No. of strokes	Theoretical	$\text{Graph.comp.}_{a_1} = 12.06 * \text{fc} - 0.114$
		Theoretical	$\text{Graph.comp.}_{a_2} = 13.04 * \text{fc} - 0.138$
		Empirical	$\text{Graph.comp.}_{a_e} = 12.82 * \text{fc} - 0.116$
(b)	No. of components	Theoretical	$\text{Graph.comp.}_{b_1} = 3.36 * \text{fc} - 0.095$
		Theoretical	$\text{Graph.comp.}_{b_2} = 3.59 * \text{fc} - 0.116$
		Empirical	$\text{Graph.comp.}_{b_e} = 3.49 * \text{fc} - 0.092$
(c)	Writing effort	Theoretical	$\text{Graph.comp.}_{c_1} = 15.16 * \text{fc} - 0.109$
		Theoretical	$\text{Graph.comp.}_{c_2} = 16.3 * \text{fc} - 0.13$
		Empirical	$\text{Graph.comp.}_{c_e} = 16.19 * \text{fc} - 0.114$

gradually swerves away as functional complexity gets higher. Similar observations can be made for case (b) in Figure 16. Figure 17 shows that the curve for graphical complexity$_{c_2}$ is farther away from the empirically estimated function curve than that of graphical complexity$_{c_1}$.

Köhler (1986) used the *t-test* to evaluate the differences between the theoretically expected function values and the empirical function that was fitted to the empirical data. The *t-test* is a statistical test to compare means. This study follows his choice.[37]

The individual results of the *t-tests* shall not be reported here. They showed significant differences for all six comparisons, so the hypothesis has to be rejected on these grounds for the time being. The fit of the function to the empirical data was not satisfactory, so this subsystem of the model seems to require improvement.

The graphs of the linearized function fit to the logarithmized data (not shown here) showed the straight line suggested for the second theoretical function (lower index 2) in all three cases to run nearly parallel to the function fit to the data which was not the case for the first suggested theoretical function (lower index 1). This seems to indicate that grouping the data in frequency classes of width 50 yields better results than grouping them in classes of width 100. In addition, the nearly parallel run of both lines may indicate that there is a factor (maybe a constant?) not yet considered in the model which is responsible for the discrepancy between the theoretically expected and empirically determined parameters.

Although this hypothesis has to be rejected in its present form, a future revision of it will contain it to some degree, which is why it is seen as a step into a promising direction at this point.

37. Köhler's choice of this statistical test has also been criticized, cf. Grotjahn (1992), and more sophisticated testing would certainly be desirable.

4.3.2. *Functional complexity as an indirect function of text frequency*

Functional complexity indirectly is a function of text frequency, where graphical complexity mediates the dependency. The equation in linearized form is

$$\text{L-functional complexity} = \ln X + Y * \text{L-frequency}. \qquad (H_5)$$

There were three ways employed to measure graphical complexity and frequencies were grouped into classes of two widths in order to make regression feasible, so there are once more six theoretical functions possible. The same types of abbreviations and indices as above are used here again.

(a)

$$\begin{aligned}\text{L-functional complexity}_{a_1} &= 5.59 - 1.373 * (2.85 - 0.114 * \text{L-frequency}) \\ &= 1.68 + 0.156 * \text{L-frequency}\end{aligned}$$

and

$$\begin{aligned}\text{L-functional complexity}_{a_2} &= 5.59 - 1.373 * (2.72 - 0.094 * \text{L-frequency}) \\ &= 1.85 + 0.13 * \text{L-frequency}.\end{aligned}$$

(b)

$$\begin{aligned}\text{L-functional complexity}_{b_1} &= 3.666 - 1.133 * (1.51 - 0.096 * \text{L-frequency}) \\ &= 1.95 + 0.108 * \text{L-frequency}\end{aligned}$$

and

$$\begin{aligned}\text{L-functional complexity}_{b_2} &= 3.666 - 1.133 * (1.4 - 0.078 * \text{L-frequency}) \\ &= 2.076 + 0.088 * \text{L-frequency}.\end{aligned}$$

(c)

$$\begin{aligned}\text{L-functional complexity}_{c_1} &= 6.086 - 1.441 * (3.06 - 0.109 * \text{L-frequency}) \\ &= 1.68 + 0.157 * \text{L-frequency}\end{aligned}$$

and

$$\begin{aligned}\text{L-functional complexity}_{c_2} &= 6.086 - 1.441 * (2.94 - 0.09 * \text{L-frequency}) \\ &= 1.85 + 0.13 * \text{L-frequency}.\end{aligned}$$

Frequency was the independent variable, so the data was once more grouped into frequency classes of widths 100 and 50, respectively, and the center of the class was used to compute the regression.

The results of the fits were as follows:

Class width 100:	$D = 0.969$	$A = e^{-1.649} = 0.192$	$B = 0.804$
Class width 50:	$D = 0.97$	$A = e^{-1.173} = 0.31$	$B = 0.74.$

The fitted curves are shown in Figures 18 and 19. The empirical data points begin to scatter below the curves at about the frequency of 800.

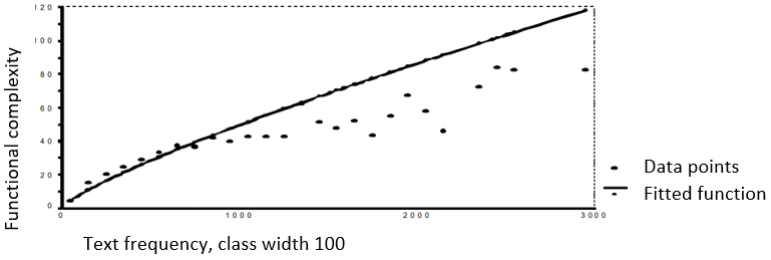

FIGURE 18. Functional complexity as an indirect function of text frequency, class width 100.

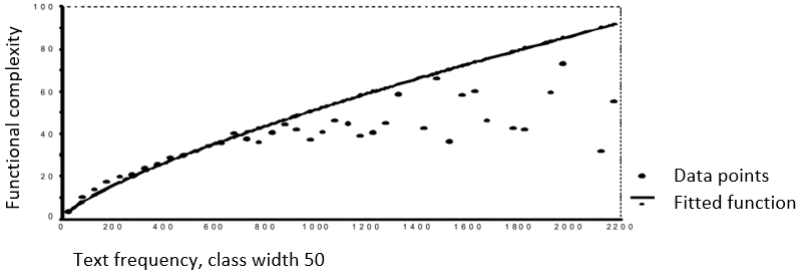

FIGURE 19. Functional complexity as an indirect function of text frequency, class width 50.

For greater ease of comparison, the theoretically expected and empirically estimated parameters are shown in Table (3).

While the parameters of the theoretical functions look similar as to their magnitude, there is still a visible discrepancy between the expected values and the empirically determined parameters of the functions.

Figures 20 and 21 show the curves of the function fits and the three curves of the theoretically expected functions. The curves of the three theoretical functions in both cases are all close together while the curve resulting from the fit to the empirical data is much steeper.

TABLE 3. Functional complexity as an indirect function of text frequency (H_5), comparison of theoretically expected and empirically estimated parameters.

C.w.	Theoretical functions	Empirical function
100	$\text{funct. comp.}_{a1} = 5.37 * \text{freq.}^{0.156}$ $\text{funct. comp.}_{b1} = 7.05 * \text{freq.}^{0.108}$ $\text{funct. comp.}_{c1} = 5.36 * \text{freq.}^{0.157}$	$\text{funct. comp.}_{e1} = 0.192 * \text{freq.}^{0.804}$
50	$\text{funct. comp.}_{a2} = 6.36 * \text{freq.}^{0.13}$ $\text{funct. comp.}_{b2} = 7.98 * \text{freq.}^{0.088}$ $\text{funct. comp.}_{c2} = 6.36 * \text{freq.}^{0.13}$	$\text{funct. comp.}_{e2} = 0.31 * \text{freq.}^{0.74}$

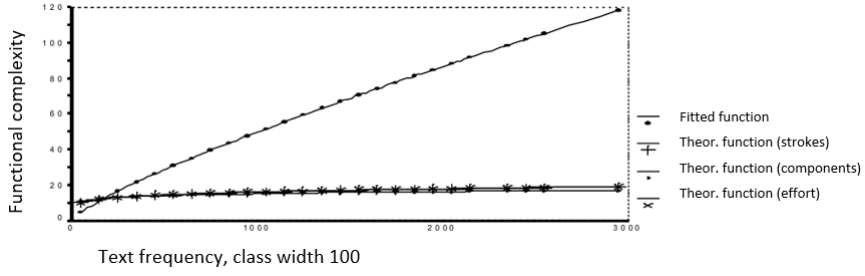

FIGURE 20. Functional complexity as an indirect function of text frequency, class width 100, fitted function curve and theoretical curves.

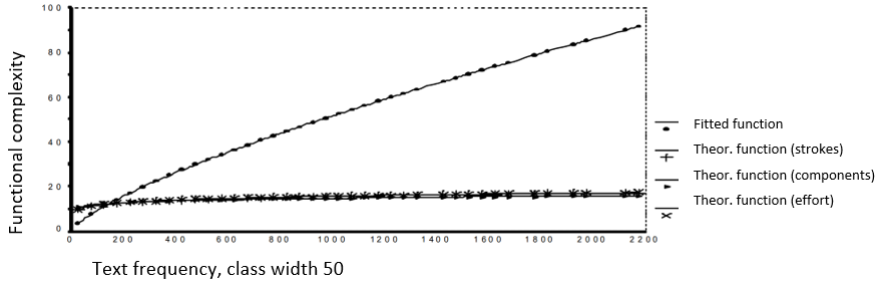

FIGURE 21. Functional complexity as an indirect function of text frequency, class width 50, fitted function curve and theoretical curves.

The *t-test* was employed again and, unsurprisingly, once more showed significant differences between the means of the various functions. The empirical fit to the data had been very good, so there seems to be a functional relationship, however, the theoretical model does not describe it well, so the hypothesis in its present form has to be rejected.

The data points had been filtered according to their weight when the dependency of graphical complexity from text frequency (hypothesis

H_3) had been modelled, but this should not be responsible for the enormous discrepancy between the expected function and the one estimated from the data because, as mentioned above, when all data points were included to test H_3 the fit had not been as good, but the parameters arrived at had varied only very little from those found when performing regression on only the data points carrying enough weight.

4.3.3. *Text Frequency as an Indirect Function of Graphical Complexity (H_6)*

The model allows the text frequency of Chinese characters to be seen as being indirectly a function of their graphical complexity, functional complexity mediating the dependency:

$$\text{L-frequency} = \ln X + Y * \text{L-graphical complexity}. \qquad (H_6)$$

As graphical complexity had been measured in three different ways, three theoretical functions are possible:

(a) No. of strokes
L-freq$_a$ = 2.444 + 1.215 * (5.59 - 1.373 * L-graph. comp.)
= 9.24 - 1.67* L-graph. comp.
(b) No. of comp.
L-freq$_b$ = 2.444 + 1.215 * (3.666 - 1.133* L-graph. comp.)
= 6.9 - 1.377* L-graph. comp.
(c) Writing effort
L-freq$_c$ = 2.444 + 1.215 * (6.086 -1.441* L-graph. comp.)
= 9.84 - 1.75* L-graph. comp.

Regression on the empirical data yielded the following results:

(a) Number of strokes
$D = 0.93 \quad A = e^{11.077} = 64{,}690.26 \quad B = -2.466.$
(b) Number of component graphemes
$D = 0.955 \quad A = e^{7.63} = 2{,}058.5 \quad B = -1.98.$
(c) Writing effort
$D = 0.88 \quad A = e^{11.675} = 117{,}557.75 \quad B = -2.47.$

In case (c) the fit was not entirely satisfying.

Figures 22 through 24 show the data points and the curves fit to them. The deviations where stroke numbers, component numbers and effort weights are low can be explained by their low weights.

The following overview shows the power functions with the theoretically expected parameters and the parameters estimated from the data (*Freq* stands for text frequency and *Comp* for graphical complexity):

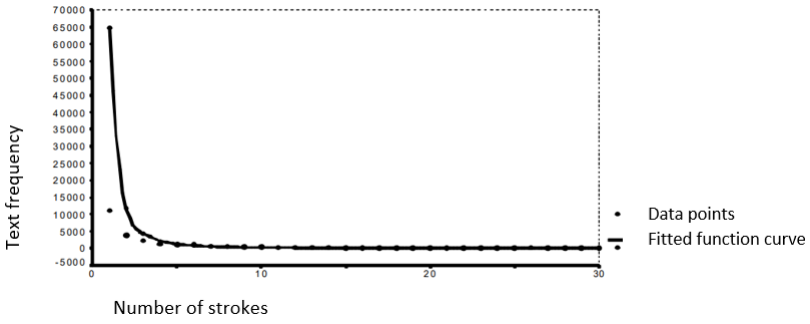

FIGURE 22. Text frequency as a function of graphical complexity, measured in number of strokes.

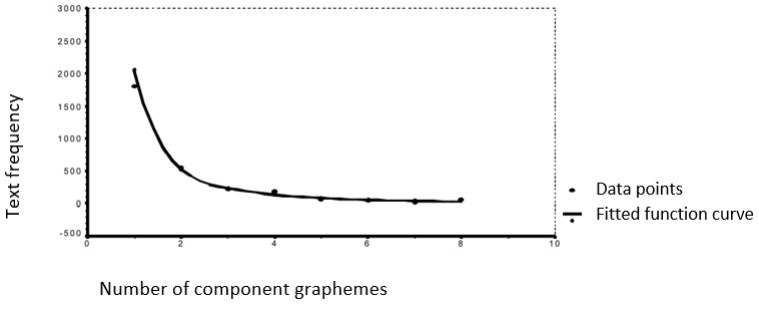

FIGURE 23. Text frequency as a function of graphical complexity, measured in number of component graphemes.

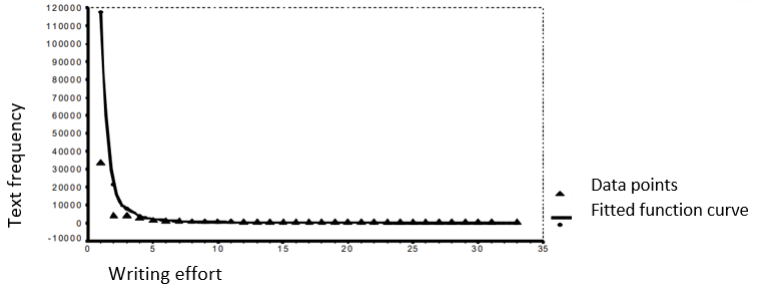

FIGURE 24. Text frequency as a function of graphical complexity, measured in writing effort.

	Theoretically	Empirically
(a)	$\text{Freq}_a = 10,287.14 * \text{Comp}^{-1.67}$	$\text{Freq}_{e_a} = 64,690.26 * \text{Comp}^{-2.466}$
(b)	$\text{Freq}_b = 992.27 * \text{Comp}^{-1.377}$	$\text{Freq}_{e_b} = 2,058.5 * \text{Comp}^{-1.98}$
(c)	$\text{Freq}_c = 18,797.89 * \text{Comp}^{-1.75}$	$\text{Freq}_{e_c} = 117,557.75 * \text{Comp}^{-2.47}$

Theoretically expected and empirically estimated values of the exponents are similar to one another to a certain degree but the empirical values are higher than the theoretical ones by about 0.6 to 0.7. The differences in magnitude between the values of the multipliers are especially eye-catching. The differences between the various theoretically expected multipliers are much smaller than those among the empirically estimated ones. The latter deviate from the theoretical values by magnitudes. However, when graphical complexity is measured in number of component graphemes, this discrepancy is lowest (case b).

The quality of the fit for graphical complexity measured in number of strokes (a) and number of component graphemes (b) suggests that frequency indeed is dependent on graphical complexity. It is possible, though, that the model needs to be refined here as this dependency perhaps should not be modelled as mediated by functional complexity or possibly other sources of influence need to be considered as well which are not yet contained in the model.

Figures 25 through 27 show the function curves. The theoretical curve and the empirical one start to overlap very early. Comparisons of cases (a) and (c) with Figures 10a and 10b show that the theoretical curves reflect the data points better than the empirical ones which can be explained by the fact that the data points for low graphical complexity carry only little weight.

In this case, as above, the *t-test* was be used to evaluate the differences between theoretically expected and empirically estimated parameters. It showed for all three comparisons that there were no significant differences. Thus, this hypothesis can be accepted.

4.4. Some Conclusions

In contrast to the three hypotheses about direct relationships two of the three hypotheses about indirect dependencies have to be rejected. Only hypothesis H_6 withstood testing.

H_4 had to be rejected but there are indications that an improvement of the model and thus a refinement of the hypothesis could yield better results. For this reason, the model is seen here as a step into the right direction.

For H_5, which modeled the indirect dependency of functional complexity from text frequency, the theoretical expectations and empirical

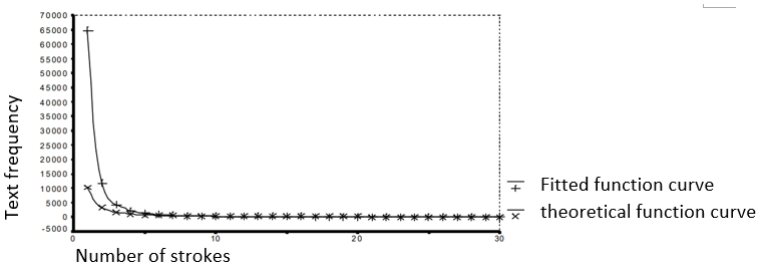

FIGURE 25. Text frequency as a function of graphical complexity measured in number of strokes; fitted function curve and theoretical function curve.

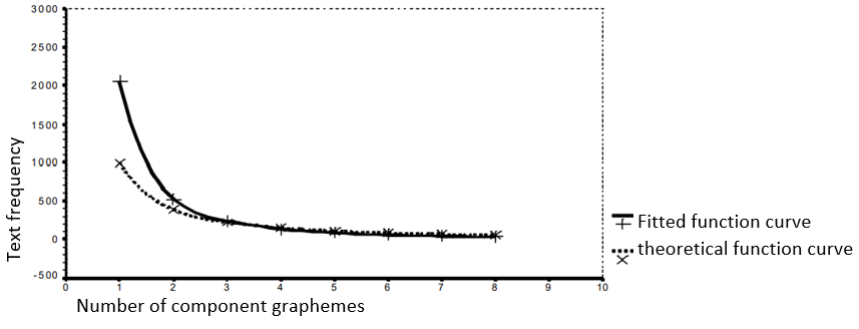

FIGURE 26. Text frequency as a function of graphical complexity measured in number of component graphemes; fitted function curve and theoretical function curve.

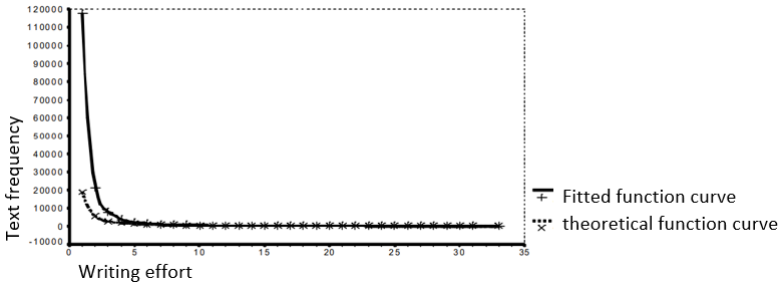

FIGURE 27. Text frequency as a function of graphical complexity measured in writing effort; fitted function curve and theoretical function curve.

data deviated widely from one another. Here, the model has definitely to be improved and there is a certain possibility that this will not be limited to additionally considering more factors.

5. Closing Remarks

The adaptation of Köhler's basic model has been, at least in the eyes of this author, a rewarding experiment. The verification of the three hypotheses about direct dependencies has shown that these dependencies also exist in the Chinese character system. For these the model seems to be adequate.

For two of the three hypotheses about indirect dependencies the predictions based on the model were not sufficiently adequate. The way functional complexity was operationalized may be one of the sources of the problem because the number of lexemes for which a character is used in the texts of the corpus may be too inaccurate a measure for the answer to the question how many different morphemes a given character may actually serve to represent. The results for hypothesis H_5 ("functional complexity is indirectly a function of text frequency"), however, probably did not just only for this reason deviate so evidently from the predicted values. The model may be incomplete here.

As a specific manifestation of human linguistic ability and behavior, the Chinese character system shows certain relationships between its systemic features which also can be found in other subsystems of language, like the vocabulary. To show that this is the case was the aim of the present endeavor.

Furthermore, the author hopes to have demonstrated, by examining aspects of a language that seems distant to many and its script rather inaccessible, that Köhler's basic model can be employed to examine other levels of analysis and that by doing so new discoveries about the model itself as about the object domain under inspection can be gained

Acknowledgements

I would like to thank Dr. Hartmut Bohn for sharing his component analysis and Prof. Dr. Reinhard Köhler for letting us try new things with his model, and both for all our discussions around it.

References

Altmann, G. (2004). "Script Complexity." In: *Glottometrics* 8, pp. 68–74.

Altmann, G. and R. Köhler (1996). "'Language Forces' and Synergetic Modelling of Language Phenomena." In: *Issues in General Linguistic Theory and the Theory of Word Length*. Ed. by P. Schmidt. Vol. 15. Glottometrika. Trier: WVT, pp. 62–76.

Bohn, Hartmut (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Hamburg: Dr. Kovač.

Chen, Ping (1999). *Modern Chinese. History and Sociolinguistics*. Cambridge: Cambridge University Press.

DeFrancis, John (1984). *The Chinese Language. Fact and Fantasy*. Honolulu: University of Hawai'i Press.

Duanmu, San (2017). "Word and Wordhood, Modern." In: *Encyclopedia of Chinese Language and Linguistics*. Ed. by Rint Sybesma et al. Vol. 4. Leiden: Brill, pp. 543–549.

Grotjahn, R. (1992). "Evaluating the adequacy of regression models. Some potential pitfalls." In: *Glottometrika* 13, pp. 121–172.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik. Aufbau eines lexikalischen Basismodells*. Trier: WVT.

Hammerl, R. and J. Maj (1988). "Ein Beitrag zu Köhler's Modell der sprachlichen Selbstregulation." In: *Glottometrika* 10, pp. 1–31.

Köhler, Reinhard (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

———— (1990). "Elemente der synergetischen Linguistik." In: *Glottometrika* 12, pp. 179–188.

———— ed. (2004). *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*. Trier: Universität Trier.

———— (2005). "Synergetic Linguistics." In: *Quantitative Linguistics. An International Handbook*. Ed. by Gabriel Altmann and Rajmund G. Piotrowski. Berlin. New York: Walter de Gruyter, pp. 760–775.

Maj, J. (1990). "Kybernetische Aspekte des synergetischen Modells von R. Köhler. Diskussion." In: *Glottometrika* 12, pp. 175–177.

Menzel, Cornelia (2004). "Das synergetische Basismodell der Lexik und die chinesische Schrift." In: *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*. Ed. by Reinhard Köhler. https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/deliver/index/docId/146/file/06_menzel.pdf. Universität Trier, pp. 178–205.

Qiú, Xīguī 裘锡圭 (1988). 文字学概要 *[Chinese writing]*. 3rd printing 1996. Beijing: 商务印书馆 [Shangwu yinshuguan].

———— (2000). *Chinese Writing*. Trans. by Gilbert L. Mattos and Jerry Norman. Vol. 4. Early China special monograph series. Berkeley, CA: The Society for the Study of Early China and The Institute of East Asian Studies, University of California.

Schindelin, Cornelia (2005a). "Die quantitative Erforschung der chinesischen Sprache und Schrift." In: *Quantitative Linguistik. Ein internationales Handbuch / Quantitative Linguistics. An International Handbook*. Ed.

by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: Walter de Gruyter, pp. 947–970.

Schindelin, Cornelia (2005b). "Zur Geschichte quantitativ-linguistischer Forschungen in China." In: *Quantitative Linguistik. Ein internationales Handbuch / Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: Walter de Gruyter, pp. 96–115.

——— (2007). *Zur Phonetizität chinesischer Schriftzeichen in der Didaktik des Chinesischen als Fremdsprache. Eine synchronische Analyse von 6.535 in der Volksrepublik China gebräuchlichen Schriftzeichen*. Vol. 13. SinoLinguistica. München: iudicium.

——— (2017a). "Character Frequency." In: *Encyclopedia of Chinese Language and Linguistics*. Vol. 1. Leiden: Brill, pp. 358–362.

——— (2017b). "Menzerath's Law." In: *Encyclopedia of Chinese Language and Linguistics*. Vol. 1. Leiden: Brill, pp. 1–3.

——— (2017c). "Word Length." In: *Encyclopedia of Chinese Language and Linguistics*. Vol. 4. Leiden: Brill, pp. 584–589.

——— (2017d). "Zipf's Law." In: *Encyclopedia of Chinese Language and Linguistics*. Vol. 4. Leiden: Brill, pp. 723–724.

Wang, L. (2011). "Polysemy and Word Length in Chinese." In: *Glottometrics* 22, pp. 73–84.

Wang, Lu (2014a). "Quantitative and Synergetic Studies on Lexical Units in Chinese." PhD thesis. Trier University.

——— (2014b). "Synergetic Studies on Some Properties of Lexical Structures in Chinese." In: *Journal of Quantitative Linguistics* 21.2, pp. 177–197.

Wang, Yanru and Xinying Chen (2015). "Structural Complexity of Simplified Chinese Characters." In: *Recent Contributions to Quantitative Linguistics*. Ed. by Arjuna Tuzzi, Martina Benešová, and Ján Mačutek. Berlin: de Gruyter Mouton, pp. 229–240.

Wáng, Xījié 王希杰 (1995). "汉语的规范化问题和语言的自我调节功能 [The standardization of Chinese characters and the self-regulating functions of languages]." In: 语言文字应用 *[Applied Linguistics]* 3.10, pp. 9–15.

Yin, Binyong and John S. Rohsenow (1994). *Modern Chinese Characters*. Beijing: Sinolingua.

Zipf, George Kingsley (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

现代汉语频率词典 *[Frequency Dictionary of the Modern Chinese Language]* (1986). Beijing: 语言学院出版社 [Yuyan Xueyuan chubanshe].