

Tereza Motalová, Lenka Matoušková

An Application of  
the Menzerath-Altmann  
Law to Contemporary  
Written Chinese

Edice Qfwfq

Olomouc  
2014

## **An Application of the Menzerath–Altmann Law to Contemporary Written Chinese**

Tereza Motalová

Lenka Matoušková

### **Recenzovali**

prof. RNDr. dr hab. Jan Andres, DSc.

Mgr. Martina Benešová, Ph.D.

Tato publikace vychází v rámci grantu Inovace studia obecné jazykovědy a teorie komunikace ve spolupráci s přírodními vědami. reg. č. CZ.1.07/2.2.00/28.0076.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

Neoprávněné užití tohoto díla je porušením autorských práv a může zakládat občanskoprávní, správněprávní, popř. trestněprávní odpovědnost.

1. vydání

© Tereza Motalová, Lenka Matoušková, 2014

© Univerzita Palackého v Olomouci, 2014

ISBN 978-80-244-4221-1

*We would like to thank Mgr. Ondřej Kučera not only for his support and help, but also for inspiration, discussions and providing valuable advice and comments to our publication.*

*Moreover, we would like to express our thanks to Mgr. Martina Benešová, Ph.D., for her helpfulness and patience, which she had over the course of consultations with us, and for contributive cooperation and support.*



# Contents

<b>Editorial note</b>	<b>9</b>
<b>Introduction</b>	<b>11</b>
<b>1. Quantitative linguistics</b>	<b>15</b>
1.1 Introduction	15
1.2 Menzerath–Altmann Law	16
<b>2. Methodology</b>	<b>20</b>
2.1 Determination of the criteria for the choice of the sample texts and their explanation	20
2.2 The choice of the sample texts	27
2.3 Determination and definition of the language units and language levels	28
2.4 Segmentation and quantification of the sample texts	42
2.5 Testing the model reliability by means of statistical methods	47
2.6 Interpretation of the the acquired data	48
<b>3. The application of the Menzerath–Altmann law to the scientific text</b>	<b>49</b>
3.1 Determination of the criteria for the choice of the sample text and their explanation	49
3.2 The choice of the sample text	50
3.3 Determination and definition of the language units and language levels	51

3.4 Segmentation and quantification of the sample text	52
3.5 Testing the model reliability by means of statistical methods	57
3.6 Interpretation of the acquired data	58
3.6.1 Language level L4	58
3.6.2 Language level L3	67
3.6.3 Language level L2	75
3.6.4 Language level L1	80
3.6.5 Conclusion	88
<b>4. The application of the Menzerath–Altmann law to the blog article</b>	<b>91</b>
4.1 Determination of the criteria for the choice of the sample text and their explanation	91
4.2 The choice of the sample text	92
4.3 Determination and definition of the language units and language levels	97
4.4 Segmentation and quantification of the sample text	98
4.5 Testing the model reliability by means of statistical methods	99
4.6 Interpretation of the the acquired data	99
4.6.1 Language level L4	99
4.6.2 Language level L3	104
4.6.3 Language level L2	119
4.6.4 Language level L1	124
4.6.5 Conclusion	129

<b>5. Comparison of the sample texts</b>	<b>133</b>
5.1 Comparison of the sample texts on language level L4	134
5.2 Comparison of the sample texts on language level L3	137
5.3 Comparison of the sample texts on language level L2	141
5.4 Comparison of the sample texts on language level L1	144
<b>Bibliographical references</b>	<b>155</b>
Monographies	155
Articles	157
Internet articles	158
Websites	159
Norm	164
Software	164
<b>Appendix</b>	<b>165</b>
Appendix 1: The scientific article – Sample A	165
Appendix 2: The blog article – Sample B	177
Appendix 3: The blog article – Sample C	181
Appendix 4: Tables containing the data empirically obtained from quantification of the newspaper article and the short story	192
<b>Index</b>	<b>197</b>





## Editorial note

A simplified form of Chinese characters is used throughout the whole work except for the examples which illustrate the traditional form of Chinese characters. For transcribing the sound of Chinese characters into the Latin script the Chinese official transcription pinyin (拼音字母) is used. Every term is first written in the form of pinyin without tones and is followed by pinyin with tones in parenthesis. For better lucidity, words in the form of pinyin with tones are distinguished by a different font – Courier New. Names of Chinese sources are accompanied by pinyin without tones and a translation. Names of authors are enclosed within the pinyin without tones.



## Introduction

The origin of quantitative linguistics is situated to Europe. *“Its development was the fastest one in Germany and Eastern Europe, where the majority of current representatives of this discipline come from. The most renowned of them is the professor of mathematical linguistics at the Ruhr-University in Bochum – Gabriel Altmann”*<sup>1</sup> (Liu and Huang, 2012, p. 180; translated by authors) who is considered to be one of the founders of the modern quantitative linguistics. In Luděk Hřebíček’s view, the real linguistics of the second half of the 20<sup>th</sup> century is Altmann’s linguistics (Hřebíček, 2008, p. 488).

*“Owing to the fact that the scientists focusing on the quantitative linguistic research are concentrated in Germany and Eastern Europe, the main fields of the research are Indo-European languages, Slavic languages and phonetic scripts. Research studies aimed at the Chinese language and its script are insufficient with regard to their scope as well as to their depth. Furthermore, the form of the Chinese language and its script is distinguished from Indo-European languages and Latin alphabet and for this reason it is required to perform subsequent research focusing on the question as to whether or not theories (regularities) and methods derived from Indo-European languages and Latin alphabet are applicable to quantitative linguistic research of the Chinese language”*<sup>2</sup> (Liu and Huang, 2012, p. 182; translated by authors).

Due to the previously mentioned circumstances, the object of this publication is quantitative linguistic research conducted on contemporary written Chinese. One of the fundamental contributions made by G. Altmann is the formulation of the Menzerath–Altmann law (MAL); therefore the publication is focused

- 
- 1 The original text: “此后，计量语言学在德国和东欧得到了快速发展。目前这一领域的主要代表人物大多来自德国、奥地利及东欧国家，其中最著名的是德国波鸿大学的 Gabriel Altmann 教授” (Liu and Huang, 2012, p. 180).
  - 2 The original text: “从事计量语言学研究的学者主要集中在德国与东欧地区，研究对象主要是印欧语、斯拉夫语和拼音文字。以汉语和汉字为对象的计量语言学研究无论从研究范围还是研究深度来说都还很不够。... 再加上汉语、汉字在形式上与印欧语、拉丁文字有较大差别，那些从印欧语和拉丁文字等语言材料中发现的计量语言学理论（定律）和方法是否适用于汉语和汉字的研究，仍需进一步检验” (Liu and Huang, 2012, p. 182).

on examining contemporary Chinese texts written in simplified Chinese characters and in different stylistic styles by means of this language law.

The research is grounded on the results which are obtained by the previous experiment and included in the article titled *An Application of the Menzerrath–Altmann Law to Contemporary Written Chinese* in the reviewed academic journal *Czech and Slovak Linguistic Review* 1/2013. The previous experiment tested the following hypothesis: if language units of the contemporary written Chinese are determined on the basis of the graphical principle, the mutual relationships between them exist on respective language levels and their validity is verified by means of the MAL. Owing to the need to compare stylistic styles, a newspaper article and a short story were chosen as sample texts. Individual language units were determined on the basis of precise criteria, i.e. Chinese texts were segmented according to the graphical criterion with minimal and inevitable consideration of the syntactic criterion. Over the course of conducting the experiment, the following language units were used: stroke – component – character – parcelate – sentence – paragraph. Linking these units into mutual relationships allowed obtaining four language levels. Results revealed that the existence of the mutual relationships between units determined primarily in accordance with the graphical principle are valid and the MAL showed itself as an adequate and well-fitting model on the following language levels: characters – component and sentence – parcelate. Although the tendency defined by this mathematical law also occurred on the level: parcelate – character, the agreement with the mathematical model of the MAL was minimal. Regarding the last language level paragraph – sentence, the assumption of the MAL was contradicted.

Due to the requirement to verify the tendencies indicated in the previous experiment, subsequent experiments testing a greater amount of sample texts and other stylistic styles will be performed. It is presumed that in the case of the language level character – component the MAL will prove itself as an adequate and well-fitting mathematical model and its agreement with empirically obtained observations will be also extremely high. Similar results are expected due to the same factor, i.e. a graphic field which exerts a strong influence

on the formation of the mutual relationship between the units on this level without regard to stylistic styles in which the sample texts are written.

As regards the next language level parcelate – character, our assumption is that the MAL will not be valid for this level; alternatively, empirically gained observations will show a minimal agreement with the mathematical model of the MAL because several factors could interfere in the relationship between these units. A different characteristic of the units could be regarded as the first of them: the parcelate represents a unit with a variable length; on the contrary, the component is a unit with invariable length due to its fixed unchanging structure. Another crucial influence could also be exerted by the reform of the Chinese script which reduced the number of strokes within 2,236 Chinese characters and consequently reduced the differences in numbers of components. These factors are probably not related to stylistic styles; hence we presume that the sample texts tested in this publication will show similar tendencies.

In contrast, the validity of the relationship between the sentence and the parcelate will be verified by means of the MAL. It seems on this language level that there is no factor which could adversely affect the mutual relationship between these units.

Last but not least, it is expected that the potential relationship between the language units on the last language level paragraph – sentence will not be revealed because the usage of punctuation does not have a long tradition in written Chinese texts. Punctuation marks which were established as borders of the sentence and the parcelate represented imported elements which meant innovation in Chinese texts. For this reason these language units do not have to be stable and unambiguous. Another cause could be a low frequency of obtained observations which has already been observed in the previous experiment. Owing to comparison of all selected samples it was necessary to choose texts with a similar length. Therefore it is expected that the low frequency will also have an influence on the validity of the MAL.

The publication is divided into five chapters. Let us begin by introducing the MAL which is the object of the first – theoretical – chapter. The following chapter describes the methodology and its individual steps. Results

of the experiments and their interpretations are presented in the third and fourth chapters. The fifth chapter is focused on a comparison of all the results obtained from the previous experiment and this experiment. And last, let us summarize the drawn conclusions at the end of this work.

# 1. Quantitative linguistics

## 1.1 INTRODUCTION

Quantitative linguistics is a sub-discipline of mathematical linguistics which originated at the turn of the 1950s and 1960s. The year 1957 is considered to be the official beginning of mathematical linguistics. At that time the Eighth international linguistics congress in Oslo, organized by the international committee (“Comité international permanent des Linguistes”, CIPL), took place. This information cf. (Černý, 1996, p. 248) and (Havránek, Horálek, 1958, pp. 47–52).

Mathematical linguistics studies natural and artificial language and uses mathematical, alternatively logical, methods (Těšitelová, 1987, p. 7). As Jiří Černý mentioned in his book, apart from quantitative linguistics this discipline also includes algebraic linguistics, which uses quantitative methods, and computational linguistics, which uses both quantitative and qualitative methods. Although the origin of mathematical linguistics is dated to the second half of the 20<sup>th</sup> century, mathematical methods have been penetrating into linguistics since the end of the previous century. Because these methods were related to quantitative methods in their characteristics, at present they are considered to be the origins of quantitative linguistics. The algebraic and the computational linguistics do not have such a long tradition because they originated about sixty years later in connection with the beginning of modern logic and computer technology (Černý, 1996, p. 248).

Quantitative linguistics deals with research of language phenomena and their relationships by means of suitable quantitative methods. It is important to point out that “... *quantitative data, formula etc. are not the aim of quantitative linguistics, ..., but they are an instrument, alternatively a verification, of our cognition. In term of linguistics it is important to use a linguistic interpretation of the ascertained statistical data, classifications, formulas and others*”<sup>3</sup> (Těšitelová, 1987, p. 9; translated by authors).

3 The original text: “... kvantitativní údaje, formule apod. nejsou cílem kvantitativní lingvistiky, ..., nýbrž prostředkem, popř. kontrolou našeho poznání. Z hlediska lingvistického

## 1.2 MENZERATH–ALTMANN LAW

The MAL represents a famous language law in the 20<sup>th</sup> century. This language law is considered as a milestone in the development of perception of language and it contributes to a rise of partial linguistic disciplines.

Whereas pre-Menzerath linguistics examined language phenomena only by themselves, a German linguist Paul Menzerath was the first man who noticed certain correlations between syllables and words. On the basis of analysing German words he found in 1928 that the longer a word, the shorter the average length of its syllables. However this observation remained unnoticed by linguists for a long time. This information cf. (Altmann, 1980, p. 124), (Andres et al., 2012, p. 1), (Hřebíček, 2002, p. 53), (Hřebíček, 2007, p. 84) and (Hřebíček, 2008, p. 490).

As late as in the 1980s Gabriel Altmann built on the work of P. Menzerath and introduced two terms: a construct and a constituent. The construct is a language unit at a higher language level and its constituent is a language unit on the nearest lower level. G. Altmann tested the relationship between these particular language units and proved that there is a correlation between them. This correlation is as follows (Altmann, 1980, p. 124):

*The longer a language construct is, the shorter its constituents are.*

By means of these two language terms G. Altmann generalized P. Menzerath's observation and promoted it to a language law. He suggested naming it Menzerath law. "*The hypothesis declared by Menzerath signifies what relationship between these two variables should be: The longer a construct is, the shorter its constituents are. This is a relationship of inverse proportionality also called reciprocal proportion. ... G. Altmann takes the view that the hypothesis first formulated by Menzerath is complemented by an assumption of continual proportion between infinitely small increment of size of the constituent and its given size: the longer a constituent, the longer*

---

je nezbytné podávat lingvistickou interpretaci zjištěných statistických dat, tříd, formulí apod." (Těšitelová, 1987, p. 9).



its increment is – and that all is considered a medium value or average<sup>4</sup> (Hřebíček, 2002, p. 54; translated by authors). G. Altmann formulated a mathematical model for this relationship in the form of power law. Its simple mathematical formula can be expressed as follows:

$$y = Ax^{-b}$$

where  $x$  is the length of a construct measured in its constituents,  $y$  is the average length of its constituents measured in units on the nearest lower language level, and  $A$ ,  $b$  are real parameters.

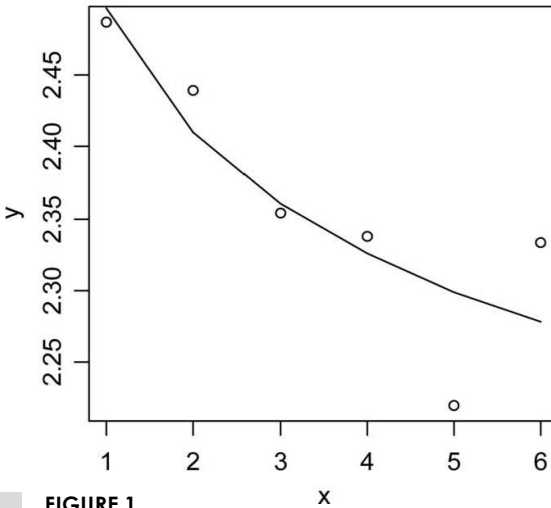
L. Hřebíček states that the length of a construct is measured in its constituents and it is always represented by an integer. Meanwhile the length of a constituent is measured in average values and most often it is represented by a decimal number. It is evident from the formula that the size of the constituent declines with an increasing size of the construct (Hřebíček, 2007, p. 84). In the case of a graphical visualization of the relationship between these two variables, the curve has a downward trend (cf. Figure 1).

To prove the validity of the MAL, the parameter  $b$  has to be a positive real number because it represents the degree of direct proportionality in this case. Parameter  $b$  takes a negative sign in the formula and it changes into a negative value which represents the reciprocal proportion. “... parameter  $b$  represents the above mentioned common degree of proportionality which interconnects two hypotheses about proportionality of considered values<sup>5</sup>” (Hřebíček, 2002, pp. 55–56; translated by authors). In the case of a graphical visualization it means that the curve is decreasing and convex, cf. Figure 2. “Parameter  $A$  determines the shift on the  $y$ -axis and can

4 The original text: “Hypotéza vyslovená Menzerathem naznačuje, jaký by měl být mezi těmito dvěma veličinami vztah: čím větší je konstrukt, tím menší je konstituent. To je vztah nepřímé úměrnosti čili nepřímé proporcionality. ... V Altmannově uvažování je tato hypotéza, původně formulovaná Menzerathem, doplněna předpokladem přímé úměrnosti mezi nekonečně malým přírůstkem velikosti konstituentů a danou velikostí konstituentů: čím větší je konstituent, tím větší je jeho přírůstek – to vše uvažováno jako nějaká střední hodnota čili průměr” (Hřebíček, 2002, p. 54).

5 The original text: “... parametr  $b$  představuje onu zmíněnou společnou míru proporcionality, která sjednocuje dvě hypotézy o proporcionalitě uvažovaných veličin” (Hřebíček, 2002, pp. 55–56).

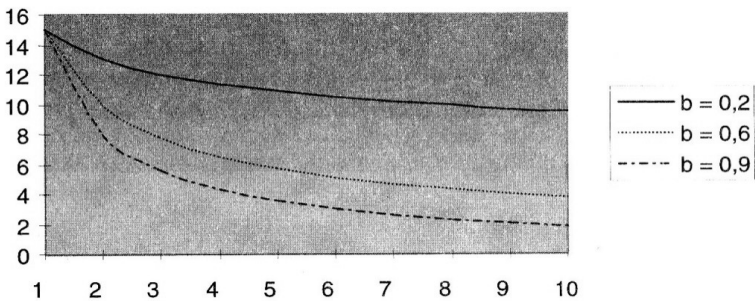
be understood as the ‘starting value’ of the fitting curve” (Kelih, 2010, p. 71). Values of both of the parameters are obtained by means of statistical software.



**FIGURE 1**

An illustration of the downward trend of curve which represents the relationship between the length of construct  $x$  and the length of constituent  $y$  defined by the MAL

Source: Andres et al., 2012, p. 15



**FIGURE 2**

An illustration of the relationship between construct  $x$  (the horizontal axis) and constituent  $y$  (the vertical axis) for three different absolute values of negative parameter  $b$  where  $A = 15$

Source: Hřebíček, 2002, p. 55

Because of the great contribution of G. Altmann to this law, this language law is presently known as Menzerath–Altmann law.

This and further information cf. (Altmann, 1980), (Andres et al., 2012), (Hřebíček, 1997), (Hřebíček, 2002), (Hřebíček, 2007), (Hřebíček, 2008) and (Kelih, 2010).

## 2. Methodology

The experiment was based on a quantitative method, namely Menzerath–Altmann law, applied to contemporary Chinese texts written in simplified characters and different styles and it was implemented in steps<sup>6</sup> listed as follows:

1. Determination of the criteria for the choice of the sample texts and their explanation
2. The choice of the sample texts
3. Determination and definition of the language units and language levels
4. Segmentation and quantification of the sample texts
5. Testing the model reliability by means of statistical methods
6. Interpretation of the acquired data

### 2.1 DETERMINATION OF THE CRITERIA FOR THE CHOICE OF THE SAMPLE TEXTS AND THEIR EXPLANATION

The choice of the appropriate sample texts is a fundamental step of this experiment; therefore, it was essential to determinate unambiguous criteria applicable to the choice of both texts:

- A. Contemporary written Chinese
- B. Simplified Chinese characters
- C. Stylistic styles
- D. Contemporarity
- E. Coherence
- F. Length
- G. Prestige of authors or institutions

Each of the selected criteria is described below in details.

---

<sup>6</sup> The experiment was implemented according to the steps which are suggested in Andres et al (2012), pp. 25–27.

## A. Contemporary written Chinese

“Chinese is only one of a very few contemporary languages whose history is documented in an unbroken tradition extending back to the second millennium BC” (Norman, 2012, p. ix). Over the course of this long-standing development the Chinese language underwent several language transformations. On the basis of analysed extant written sources the development of Chinese can be divided into several periods, but it remains that it is not possible to clearly specify their periodization, largely as a result of its complicated development, cf. (Švarný, 1967) and (Vochala, Hrdličková, 1985). To take one example, one of alternatives is a periodization suggested by Chinese linguist Wang Li, cf. (Wang Li, 2004). Wang Li based his periodization primarily upon grammatical and phonological changes and partially upon changes in lexicon. He was of the opinion that the development of the Chinese language could be divided into the following periods:

**TABLE 1**

Periodization of the Chinese language

Source: Wang Li, 2004, p. 43–44; created by authors

Period	Characters	Pinyin	Dating	Note
Ancient	上古期	Shànggǔqī	till 3rd century AD	3rd–4th century period of transition
Medieval	中古期	Zhōnggǔqī	4th–12th century	12th–13th century period of transition
Pre-Modern	近代	Jìndài	13th–19th century	1840 (Opium Wars) – 1919 (May Fourth Movement) period of transition
Modern	现代	Xiàndài	20th century	Since May Fourth Movement

With the exception of Wang Li there are also other suggestions which differ from each other in dating of the individual periods due to different

approaches. Professor Ping Chen, a Chinese linguist, and entirely based his periodization upon grammar. Although he uses identical terminology to Wang Li, dating of the individual periods diverges, cf. (Chen, 1999). Jerry Norman, an American sinologist and linguist, introduces periodization suggested by Bernhard Karlgren, a Swedish sinologist, in his book *Chinese*. Karlgren built his periodization of Chinese entirely on phonology and he used his own dating as well as terminology. The theory was published from 1915 to 1916; cf. (Norman, 2012).

On account of differences between the individual development periods of Chinese we can talk about several different languages. For this reason it was necessary to unambiguously determine which period will be analysed. This experiment was focused on the last development period – modern Chinese. Over the course of this relatively short period Chinese has also dramatically changed. Consequently, there was again an urgent need to specify to which language phase the sample texts will belong. Owing to the emphasis on the synchronous aspect of this experiment the first criterion was to choose texts written in a standard form of contemporary Chinese known as Putonghua (pǔtōnghuà, 普通话).

## B. Simplified Chinese characters

The earliest written historical sources are dated back to the Shang Dynasty (17<sup>th</sup>–11<sup>th</sup> century BC) and refer to the oracle bone script (jiǎgǔwén, 甲骨文) from the period between the 14<sup>th</sup> and 11<sup>th</sup> century BC. In parallel with jiaguwen another kind of writing system emerged; in this instance we talk about the bronze script (jīnwén, 金文). Both the systems of writing developed from pictograms and “*The maturity of this early script has suggested to many scholars that it must have passed through a fairly long period of development before reaching this stage, but the few examples of writing which precede the fourteenth century are unfortunately too sparse to allow any sort of reconstruction of this development*” (Norman, 2012, p. 58).

The bronze script appeared, to a large extent, under the reign of the Western Zhou Dynasty (1046–771 BC) and in the first half of the Eastern Zhou Dynasty,

i.e. in the Spring and Autumn period (771–476 BC). The writing system of that time was very similar to the oracle bone script because it still maintained its pictographical form.

In the Spring and Autumn period and particularly in the Era of Warring States (475–221 BC), when the kingdom disintegrated into seven independent states, the development of the Chinese writing started to diversify. The first script which gradually began emerging was the great seal script (*dàzhuàn*, 大篆) written on silk and bamboo slips. As L. Zádrapa mentions in his book, to define this kind of system of writing is difficult. Some researchers regard it as all scripts appeared in the Eastern Zhou period and others as the script of the Qin state and the predecessor of the small seal script (Zádrapa, Pejčochová, 2009, pp. 138–9). Apart from these suggestions, according to J. Vochala and J. Norman the great seal script also represents the script of the Western Zhou period, known as *zhouwen* (*zhòuwén*, 籀文), cf. (Vochala et al., 1975) and (Norman, 2012).
















In 221 BC the first sovereign emperor of Qin came to power and unified China for the very first time. Over the course of his reign he made a series of reforms; one of them involved the Chinese writing system and its aims were the simplification of Chinese characters and their implementation in the whole empire. Qin Shi Huangdi unified not only China, but also the system of writing which is presently known as the small seal script (*xiǎozhuàn*, 小篆). Simultaneously the simplified form of the great seal script emerged. This kind of writing system is known as the clerical script (*lìshū*, 隶书) and it was used by inferior officials. Although this system of writing appeared in the Qin Dynasty period, it is traditionally associated with the following Han dynasty (206 BC – 220 AD) which raised this writing to the official written form of Chinese language. For this reason the clerical script is divided into two kinds: old clerical script (also known as Qin clerical script; *qín lì*, 秦隶) and new clerical script (also known as Han clerical script; *hàn lì*, 汉隶). Towards the end of the Han dynasty a new kind of writing system emerged – the regular script (*kǎishū*, 楷书) which developed from the new clerical script and it is used up to these days. By way of illustration we cite in Table 2 several selected characters which demonstrate the individual development phases of the Chinese writing system.

For further information cf. (Kučera et al., 2005), (Norman, 2012), (Qiu, 2000), (Vochala et al., 1975), (Vochala, Hrdličková, 1985) and (Zádrapa, Pejščochová, 2009).

**TABLE 2**

Illustration of the selected Chinese characters development

Source: (Xiang de shufa, © 2004–2012), (“Xiang” zi de jiben xinxi, © 2004–2012), (Xin de shufa, © 2004–2012), (“Xin” zi de jiben xinxi, © 2004–2012), (Yu de shufa, © 2004–2012), (“Yu” zi de jiben xinxi, © 2004–2012); created by authors

The kind of Chinese script	Elephant	Rain	Heart
Oracle bone script 甲骨文			
Bronze script 金文			
Small seal script 小篆			
Clerical script 隶书			
Regular script 楷书			

*“From the oracle bone script up to the regular script and its italic modifications a predominant tendency in the development of Chinese characters has been apparent. The tendency manifested itself in graphical formalization which culminated in simplification*



of Chinese characters. This phenomenon was caused by the demand for easier writing of these graphically too complicated scripts and it became a constant factor which exerted a strong influence on the Chinese characters' development"<sup>7</sup> (Vochala et al., 1975, p. 26; translated by authors). The tendency towards simplification is mainly typical for the unofficial sphere where the non-standard simplified characters, known as *suti* (sútǐ, 俗体), gained in a widespread popularity. "... A large number of popular simplified characters were created and used widely among the common people for writing such things as account books, pawn tickets, medicinal prescriptions, operatic scripts and certain forms of vernacular literature. Even members of the literati employed these non-official but convenient forms in personal correspondence and for copying materials for private use; ..." (Norman, 2012, p. 80). Contrary to the unofficial sphere, these simplified characters were regarded by the official sphere as vulgar and particularly in the period from Tang dynasty to the beginning of the 20<sup>th</sup> century there existed a conservative approach which opposed any innovation; therefore, the graphic form of the regular script did not change too much. This and further information cf. (Norman, 2012) and (Vochala et al., 1975).

The process of simplification culminated in the mid 20<sup>th</sup> century when the governmental authority of the People's Republic of China (hereafter referred to as PRC) decided to simplify the traditional Chinese script in order to make it accessible to a great number of people. This reform of script was implemented in 1956 and 1964 and it had two phases. Over the course of the first phase the simplification involved more than 500 characters and in the second phase more than 2,000 characters. Both the phases were based upon the natural process of Chinese characters simplification which appeared in the previous centuries. This and further information cf. (Chen, 1999), (Motalová et al., 2013), (Norman, 2012) and (Zádrapa, Pejčochová, 2009).

In spite of the fact that the Chinese government and academic circles promote the usage of simplified characters, the traditional character set is still

7 The original text: "Od písma na věštebných destičkách až po vzorové písmo a jeho kursivní modifikace je převládající tendencí ve vývoji čínského znakového písma jeho grafická formalizace, která se projevuje ve zjednodušování tvaru znaků. Permanentním činitelem, který v tomto směru ovlivňoval vývoj znaků, byl požadavek snadnějšího psaní tohoto graficky příliš náročného písma" (Vochala et al., 1975, p. 26).

employed, particularly in publications focused on the history of Chinese language and script and in the unofficial sphere for example on various signs (Zádrapa, Pejčochová, 2009, pp. 33–34). In the Republic of China (Taiwan) and in other areas which do not fall under the administration of the PRC (e.g. Hong Kong, Macau), the usage of traditional characters also still persists. This and further information cf. (Chen, 1999), (Motalová, 2013), (Norman, 2012) and (Zádrapa, Pejčochová, 2009).

According to the above described development, it was necessary to determine which type of Chinese scripts will be analysed. With regard to the above mentioned synchronous aspect the experiment was aimed at the simplified form of the latest developmental stage of Chinese script, i.e. kaishu (the regular script).

### **C. Stylistic styles**

In view of the fact that the previous experiment was focused on the quantitative analysis of the newspaper and literary styles, the third requirement was the choice of the sample texts written in two other different stylistic styles due to comparison between the data acquired from both experiments. Therefore we decided that in this case the analysis will be aimed at the scientific and artistic styles.

### **D. Contemporarity**

The fourth criterion refers to the demand of the sample texts written in the standard form of contemporary Chinese. Due to the reflexion of the required form of Chinese language it was necessary to choose samples which were published in recent years. In accordance with this criterion it was determined that the sample texts would have to be published after year 2002, in other words publication date of the sample texts cannot be older than eleven years.

### **E. Coherence**

Coherence of the text was also included into the criteria of the sample texts choice. L. Hřebíček mentions in his publication that an analysed text has to be

a coherent structure, in other words an uninterrupted sequence of language units on all various language levels. This coherent structure is unambiguously defined by a distinct beginning and end and it is not interrupted for example by a picture, a graph etc. This information cf. (Hřebíček, 2002, p. 43).

## **F. Length**

The fifth criterion was determined with respect to the representativeness of the sample texts length. Over the course of its determination it was necessary to take into consideration that the sample texts cannot be either overly short or too long. In overly short sample texts mutual relationships of respective language units defined by the MAL may not be detected, on the other side in overly long sample texts certain relationships may disappear. The length scale of the sample texts was in this instance empirically determined on the basis of the previous experiment. It means that the number of Chinese characters used in the text can fluctuate between 2,500 and 3,500.

## **G. Prestige of authors or institutions**

Prestige of an author or an institution where the texts have been published is the last requirement. The formation of contemporary Chinese language is also influenced by writers as they are popular with and read by the general public. Words and idioms which are chosen by authors have an impact on the vocabulary structure of readers and thereby on the frequency of the Chinese words and characters. In other words writer's lexicon represents one of the factors modifying the contemporary form of Chinese language. The prestige of the institutions can potentially guarantee a greater awareness of the public and thereby a much wider readership of works which are published under their patronage.

## **2.2 THE CHOICE OF THE SAMPLE TEXTS**

On the basis of the mentioned criteria two sample texts representing two different stylistic styles were chosen. The first one – a scientific article – represents the scientific style and the second one – a blog article – the artistic style.

### 2.3 DETERMINATION AND DEFINITION OF THE LANGUAGE UNITS AND LANGUAGE LEVELS

The next step after finishing the choice of the sample texts was determination and definition of the language units. A segmentation on the basis of the exactly determined units is a fundamental step in every successful experiment. Therefore the language units need to be determined precisely.

In case of this experiment a selection of language units did not follow common linguistics definitions but an alternative approach was chosen. Because of the fact that these experiments concern written language, the most appropriate alternative in determining language units was to apply a graphic principle. In addition to this main criterion, in case of language units whose borders were determined by punctuation, it was necessary to take a syntactic principle into consideration. When selecting these principles, the graphic principle was chosen as the main criterion as in the previous experiment. This approach was chosen on account of the request to compare results from these experiments with the result gained from the previous experiment. All of the experiments have to be carried out under identical conditions so that the results can be compared.

In compliance with these principles the six following language units were unambiguously defined and used:

*stroke – component – character – parcellate – sentence – paragraph.*

The stroke represents the lowest language unit and the paragraph represents the highest language unit.

The language units and their definitions will be discussed below. Description will proceed from the lowest to the highest language unit, i. e. from the stroke to the paragraph.

#### **The stroke** (bǐhuà, 笔画)

According to J. Vochala the stroke is the minimal graphic unit of the Chinese writing system. The stroke represents an uninterrupted line which is written ‘at one go’. Works which deal with this unit diverge in many cases not only in the classification

and number of strokes, but also in their terminology. Various approaches can be found in both Chinese and foreign publications, cf. (Vochala, 1986, p. 14).


The aim of this subchapter is not to offer an exhaustive overview of the approaches dealing with strokes' classification, but only to present the basic forms of the language unit and to illustrate several ways of their classification. The second category – compound strokes – will not be discussed in this work since they are usually only various modifications of the basic strokes (Zádrapa and Pejšochová, 2009, p. 67).

Calligraphy handbooks usually distinguish eight basic strokes which are traditionally illustrated on the Chinese character meaning “eternity”, “forever”, “permanence” (yǒng, 永), cf. Table 3. For example, this classification is used in the books *Empire of Chinese Characters* (Cecilia Lindqvist, 2010), *The Chinese Language: Its History and Current Usage* (Kane, 2009), *Chinese Calligraphy* (Qu Leilei, 2005) or *Chinese Running Script Calligraphy for Beginners* (Wang Xianchun, 2007).

**TABLE 3**

Classification of basic strokes illustrated on the Chinese character 永 according to the calligraphic handbooks

Source: Kane, 2009, p. 77; Chinese terminology taken from: Eight principles of Yong, 2013; created by authors

Character	The strokes occurred in this character are as follows:
	1 Dot: cè, 側 or diǎn, 点
	2 Rightward stroke: lēi, 勒 or héng, 横
	3 Downward stroke: nǔ, 努 (弩) or shù, 竖
	4 Hook: tí, 趯 or gōu, 钩
	5 Flick up and rightwards: cè, 策 or tiāo, 挑
	6 A tapering thinning curve, usually concave left: lüè, 掠 or wān, 弯
	7 Falling leftwards: piē, 撇
	8 Falling rightwards: zhé, 磔 or nà, 捺

Regarding educational materials, classification of basic strokes is not unified. For example *Textbook of Chinese Characters I*, written by Ondřej Kučera et al. (2005) distinguishes five basic strokes. Other types of strokes are described as their modifications so they are classified as compound strokes. In contrast, the *Integrated Chinese* textbook written by Yuehua Liu et al. (2009) states that there are 11 basic strokes, compound strokes are not listed here.

Materials dealing with the Chinese language in general also diverge in the number of basic strokes. Some authors distinguish six basic strokes, for example Lukáš Zádrapa and Michaela Pejčochová in *Chinese Script* (2009) and Oldřich Švarný in *Introduction to Spoken Chinese I*. (1967). Wang Ning (2002) mentions seven basic strokes in his book *Hànzì gòuxíng xué jiǎngzuò*<sup>8</sup> (汉字构形学讲座) *Lectures on Structure of Chinese Characters*. On the other hand, Leon Wieger (1965) states nine basic strokes in his work *Chinese Characters: Their origin, etymology, history, classification and signification*.

As well as in the previous experiment, let us introduce the graphic characterization of elementary strokes suggested by J. Vochala (1986). This Czech sinologist determines 11 elementary strokes in his work *Chinese Writing System*. In addition, he divides these elementary strokes into two subcategories, simple strokes and hooked simple strokes, cf. Table 4, 5:

**TABLE 4**

The classification of strokes: The elementary strokes – simple strokes  
Source: (Vochala, 1986, p. 30), (Bihua, © 2013); created by authors

Stroke	Pinyin	Characters	English terminology
一	héng	横	Horizontal Stroke
丨	shù	竖	Vertical Stroke
㇇	piě	撇	Left Skew Stroke
㇇	nà	捺	Right Skew Stroke
㇇	tí	提	Ascending Stroke

8 In English: *Studies on the Chinese characters formation*.

Stroke	Pinyin	Characters	English terminology
㇇	diǎn	点	Left Skewed Point Stroke
㇏	diǎn	点	Right Skewed Point Stroke

**TABLE 5**

The classification of strokes: The elementary strokes – hooked simple strokes

Source: (Vochala, 1986, p. 30), (Bihua, © 2013); created by authors

Stroke	Pinyin	Characters	English terminology
→	hénggōu	横钩	Horizontal Hook Stroke
↓	shùgōu	竖钩	Vertical Hook Stroke
㇇	wāngōu	弯钩	Curved Vertical Hook Stroke
㇏	xiégōu	斜钩	Right Hook Stroke

Other types of strokes and their modification are published in the work of Jaromír Vochala, cf. (Vochala, 1986).

The stroke or combination of strokes creates the next higher language unit, the *component*.

### **The component** (bùjiàn, 部件)

The component is a language unit which consists of a certain number of strokes and which is involved in creating a structure of a character. The component is set by various definitions which are not unified. Although there are different definitions which operate within a component and define it, it is difficult to apply them on account of their indefinite formulations. In general terms, the component is interpreted as a structural unit which is on a higher linguistic level than the stroke and on a lower linguistic level than the character. But it is difficult to determine unambiguously which combinations of strokes create

a component and which do not. Over the course of segmenting character into components, lots of exceptions may arise, e.g. the existence of characters which are composed of one stroke does not correspond to the general conception which claims that the component is a higher language unit than the stroke. Similarly, certain components can simultaneously create individual characters so that the second part of this general definition which states that the component is lower language unit than the character is also incomplete.

As it has been said before, the definitions of components diverge. For example in the Chinese *Baidu Baike* encyclopedia, the following criteria are mentioned:

1. The first criterion is based on the number of strokes. In this case the components can be divided into two categories – the first category includes components comprised of one stroke (e.g. 一 and 乙), the second category includes components comprised of two or more strokes (e.g. 士 and 重).
2. The second criterion categorizes components into two groups on the basis of their independence. The first group contains those components which create a character in combination with other components and which also create independent characters at the same time (e.g. 吉 → 口 and 和 → 可). The second group includes components which are only parts of characters and do not create individual characters (e.g. 同 → 冂 and 病 → 疒).
3. The third criterion divides components into two groups on the basis of their decomposition. The first group includes components which represent the smallest inseparable part of a character (e.g. 男 → 田 and 力). On the contrary, the second group includes compound components, i.e. those components which can be decomposed into two or more simpler components (e.g. 想 → 相 and 心 are here the first “layer” and 相 → 木 and 目 are here the second “layer”).

Since the definitions of components diverge, it was necessary in this experiment to choose the method segmenting characters into components which



exclusively follows a graphical aspect and an aspect of formalization. For this reason, the characters are divided into components based on the contacts of strokes. On the basis of this conception, the component is regarded here as a so-called ‘island’, “i.e. as a separate part of the character which is composed of one stroke or a group of strokes connected to one another and obviously separated from other parts (i.e. components) of the character” (Motalová et al., 2013).

Various combinations of these units create the next language unit, the *character*.

The application of this conception revealed in the previous experiment that various fonts determine the borders of components in different ways. The total number of components consequently varies within an identical character according to a used font. For an illustration, Table 6 cites the examples of those characters whose numbers of components fluctuate most significantly depending on the used fonts (cf. Table 6).

Table 6 is divided into five columns. The first column contains the eight selected fonts which demonstrate the variation of the number of components within a character. The remaining four columns comprise four characters whose numbers of components vary according to the selected font. The first line shows examples of selected characters in which the borders of components vary and these problematic parts are highlighted in red. The remaining lines show the selected characters in particular fonts and the number of components ( $N_c$ ) is placed at the bottom of the cell.

**TABLE 6**

Comparison of the characters demonstrating an oscillation of numbers of components ( $N_c$ ) depending on the fonts

Source: (Motalová et al., 2013)

Fonts		Chinese characters							
		翻		各		麻		新	
1.	SimSun	翻		各		麻		新	
		$N_c$	6	$N_c$	1	$N_c$	3	$N_c$	5

Fonts		Chinese characters							
		翻		各		麻		新	
2.	DF Kai-SB	翻		各		麻		新	
		N <sub>c</sub>	10	N <sub>c</sub>	2	N <sub>c</sub>	6	N <sub>c</sub>	4
3.	Han ding jiankai (汉鼎简楷)	翻		各		麻		新	
		N <sub>c</sub>	10	N <sub>c</sub>	2	N <sub>c</sub>	4	N <sub>c</sub>	7
4.	Mingliu	翻		各		麻		新	
		N <sub>c</sub>	8	N <sub>c</sub>	1	N <sub>c</sub>	7	N <sub>c</sub>	4
5.	Fangsong	翻		各		麻		新	
		N <sub>c</sub>	5	N <sub>c</sub>	2	N <sub>c</sub>	3	N <sub>c</sub>	6
6.	Meiryo	翻		各		麻		新	
		N <sub>c</sub>	6	N <sub>c</sub>	1	N <sub>c</sub>	3	N <sub>c</sub>	2
7.	Jhenghei	翻		各		麻		新	
		N <sub>c</sub>	9	N <sub>c</sub>	2	N <sub>c</sub>	7	N <sub>c</sub>	4
8.	SimHei	翻		各		麻		新	
		N <sub>c</sub>	6	N <sub>c</sub>	2	N <sub>c</sub>	4	N <sub>c</sub>	6

Due to this fact, only one font has to be applied to both sample texts in this experiment. With regard to the previous experiment, a crucial aspect for this selection was the font used in the newspaper article and in the short story, i.e. SimSun. In this experiment, the same segmentation approach is preserved in order to compare results of both experiments. However it was not necessary

to convert the sample texts into the required font because both the sample texts use the same font, i.e. SimSun.

### **The character** (hànzì, 汉字)

The character represents the basic graphic unit of the Chinese writing system that corresponds to one syllable in the spoken language.<sup>9</sup> According to Oldřich Švarný the characters represent the basic graphic units which are approximately equal in size regardless of the number of strokes composing a character. The strokes are arranged into a square or into a rectangle (whose height is not much bigger than its width), cf. (Švarný, 1967, p. 31). Individual graphic fields adhere to one another, they are not separated by a space. Consequently, Chinese written texts do not determine the borders of the written word.

Apart from the Chinese characters, Chinese texts also operate with Arabic numerals which have two different formats. In one case each Arabic numeral occupies an individual graphic field therefore is considered to be a character. However, if Arabic numerals are formatted in the other way, individual numerals adjoin tightly each other, therefore, the numerals or numbers (i.e. both integer and rational numbers) are both considered to be characters.

A group of characters creates the next language unit, the *parcelate*.

### **The parcelate**

The parcelate is a language unit which is determined by specific punctuation marks. This language unit was delimited in the previous experiment in which it was crucial to find a language unit higher than a character and lower than a sentence. Contemporary Chinese written texts are graphically structured into partial segments by the punctuation. For this reason, punctuation marks were considered as the borders of this language unit. For the purpose of the previous experiment this language unit was called the parcelate.

With regard to the variability of Chinese punctuation marks and their various functions, it became necessary to unambiguously define which of them

9 In Chinese texts there is only one exception when two characters represent one syllable, it is the case of er-coloring, for instance the word "painting" *huar* (huà 画, 画儿).

create the borders of the parcelates (cf. Table 7). Over the process of selecting the punctuation marks valid for this language unit, their hierarchization was created and apart from applying the graphical principal it was also inevitable to take the syntactic criterion into consideration.

**TABLE 7**

The selected punctuation marks valid for the borders of parcelates  
Source: Chinese terminology from (Biaodian fuhao, © 2013); created by authors

Punctuation marks	Pinyin	Characters	English terminology
。	jùhào	句号	full stop
？	wèn hào	问号	question mark
！	tàn hào	叹号	exclamation mark
，	dòu hào	逗号	comma
；	fēn hào	分号	semicolon
：	mào hào	冒号	colon

Apart from the above-mentioned punctuation marks, Chinese texts also operate with other punctuation marks with specific functions. Because their usage in Chinese sentence does not effect a syntactic structure S-V-O (subject – verb – object), they are not considered as borders of parcelates. These punctuation marks are as follows:

1. Enumeration comma (dùn hào, 顿号): 、

This kind of comma separates parts of a sentence in a coordinate relationship usually in the instance of enumeration or it is written after a number in a numerical list, cf. (GB/T 15834 – 2011, 2012) and (Dunhao, © 2013).

2. Quotation marks (yǐn hào, 引号): “ ” [ ] ‘ ’

Quotation marks are used for quoted speech, direct speech and for the parts of texts which should be emphasised. They can be divided into two categories:

double quote marks and single quote marks. They both have a Western-style and traditional Chinese alternative of notation, cf. (GB/T 15834 – 2011, 2012) and (Yinhao, © 2013).

3. Ellipsis (shěnglüèhào, shānjié hào, 省略号, 删节号): .....

Ellipsis is used for omission of certain parts of a cited text, for intermittent speech or it can substitute an omission of an item in a list, cf. (GB/T 15834 – 2011, 2012) and (Shengluèhao, © 2013).

4. Emphasizing dot (zhuózhòng hào, 着重号): •

This dot emphasizes important parts of a text (character, word or sentence) to which readers should pay attention. The emphasizing dot is usually placed below the text, cf. (GB/T 15834 – 2011, 2012) and (Zhuozhonghao, © 2013).

5. Titles marks (shūmíng hào, 书名号): 《》〈〉

This kind of parentheses is used for titles of books, periodicals, articles, documents, and other literary works, also for dramas, songs, films, and so on. This punctuation mark also has double and simple alternatives, cf. (GB/T 15834 – 2011, 2012) and (Shuminghao, © 2013).

6. Middle dot (jiàngé hào, 间隔号): •

Middle dot separates words which are in some correlation. For example it is used in proper names of persons and books and so on. They define individual entities of respective names, cf. (GB/T 15834 – 2011, 2012) and (Jiangèhao, © 2013).

7. Dashes: En, em, wavy dash (liánjiē hào, 连接号): — —— ~

These dashes are used for connection between words which are in some correlation. As in Western languages, there are also two variants of them – short and long one, cf. (GB/T 15834 – 2011, 2012) and (Lianjiehao, © 2013).

## 8. Proper name mark (zhuānmíng hào, sī míng hào, 专名号, 私名号):

Proper name mark highlights a special category of proper names which occur in historical works. It is used for marking personal names, place names, names of dynasties, names of ethnic groups, states, and institutions, cf. (GB/T 15834 – 2011, 2012) and (Zhuanminghao, © 2013).

**TABLE 8**

Examples of the usage of the selected punctuation marks which are not valid for borders of parcellates

Source: Chinese terminology taken from (Biaodian fuhao, © 2013), examples taken from (GB/T 15834 – 2011, 2012); created by authors

English terminology	Punctuation marks	Examples of the usage
Enumeration comma	、	1. 这里有自由、民主、平等、开放的风气和氛围。 2. 我准备讲两个问题：一、逻辑学是什么？ 二、怎样学好逻辑学？
Quotation marks	“ ” 『 』	1. 这里所谓的“文”，并不是指文字，而是指文采。
	， 「 」	2. 他问：“老师，“七月流火”是什么意思？”
Ellipsis	……	1. 他气得连声说：“好，好……算我没说。” 2. 她磕磕巴巴地说：“可是……太太……我不知道……你一定是认错了。”
Emphasizing dot	·	1. 诗人需要表现，而不是证明。 2. 下面对本文的理解，不正确的一项是：……
The titles marks	《 》	1. 《红楼梦》
	〈 〉	2. 科研人员正在研制《电脑卫士》杀毒软件。 3. 《教育部关于提请审议〈高等教育自学考试试行办法〉的报告》
Middle dot	·	1. 克里斯蒂娜·罗塞蒂 2. “一·二八”事变 “一·二·九”运动
Dash	— ~	1. 参见下页表2–8、表2–9。 2. 2011年2月3日–10日 3. 25~30g

English terminology	Punctuation marks	Examples of the usage
Proper name mark	——	1. 孙坚人马被刘表率军围得水泄不通。 (人名) 2. 于是聚集冀、兖、幽、并四州兵马七十多万准备决一死战。(地名)

Between these two groups there are transitional punctuation marks, namely bracket () [ ] { } (kuòhào, 括号) and a special type of dash —— (pòzhéhào, 破折号). With regard to the syntax, it has to be determined in which cases these punctuation marks are considered as borders of parcellate and in which cases they are not.

#### 1. Bracket (kuòhào, 括号): () [ ] { }

Brackets contain a note inserted directly into the text. Additional information about the topic inserted in them can be expressed as a part of a sentence or as a whole sentence, cf. (GB/T 15834 – 2011, 2012) and (Kuohao, © 2013). In the former case, the brackets are not considered as borders of a parcellate, on the contrary, in the latter case, the brackets create borders of a parcellate.

#### 2. Dash (pòzhéhào, 破折号): ——

This type of dash precedes explanation and additional information about a topic, and it also signifies a change of conversation topic, meaning or way of expression. When used after onomatopoeia it indicates that the sound continues. It also serves as a bullet before each item in a list. This dash can also be followed by a part of a sentence or the whole sentence, cf. (GB/T 15834 – 2011, 2012) and (Pozhehao, © 2013). For the reason of its wide usage, it was necessary to establish rules which determine in which cases the dash is regarded as borders of parcellates:

The dash is not taken into consideration as borders of the parcellates in these cases:

- a) before an explanation and additional information in the form of parts of a sentence or in the form of their enumeration
- b) after an onomatopoeia
- c) before every item in a list which is expressed by one part of a sentence

The dash is taken into consideration as borders of the parcelates in these cases:

- a) before an explanation and additional information in the form of a sentence, i. e. the sentence composed of a subject, a verb and an object
- b) before a change of conversation topic, meaning or way of expression
- c) before every item in a list which is expressed by a sentence

Western punctuation marks can also occur in Chinese texts, in the cases when the author uses Arabic numerals in place of Chinese characters. The signs that accompany these numbers (e.g. a decimal point) do not compose any language unit.

A group of parcelates composes the next language unit, the *sentence*.

### **The sentence** (fùjù, 复句)

As well as the previous language unit, borders of this language unit are formed by punctuation marks. Unlike the parcelate, the sentence is only separated by a full stop 。 (jùhào; 句号), a question mark ? (wèn hào; 问号) and an exclamation mark ! (gǎntànhào; 感叹号). Other punctuation marks are only valid for the lower language unit, i.e. parcelate. By this selection, the syntactic aspect was again taken into account.

A sentence or group of sentences constitute the next language unit, the *paragraph*.

### **The paragraph** (duànluò, 段落)

Borders of paragraphs are determined entirely according to the graphic principle. There are three ways how the texts are divided into paragraphs. Firstly, each



paragraph begins on a new line and it is separated through an inserted blank line (namely the scientific article). Secondly, the beginning of the paragraph begins on a new line and it is formed by the indentation at the side of the paper (the short story). Lastly, the beginning of the paragraph is not only formed by the indentation at the side of the paper, but also each paragraphs is separated through an inserted blank line (the newspaper article and the blog article).

### Language levels

As L. Hřebíček mentions “*two language units of various types can be considered to constitute a relationship between two mutually hierarchical levels (...), in other words they can be considered a construct and a constituent. But when we think about a text, the limitation to a relationship between just two levels is insufficient; we immediately ask what is happening with other levels in the text. ... In other words, each language entity is a constituent with regard to all higher language levels and it is a construct with regard to all lower levels*”<sup>10</sup> (Hřebíček, 2002, p. 59; translated by authors).

By linking the above mentioned language units, which are immediately contiguous in the hierarchy of the selected language units, into a relationship, four language levels are acquired. On the lowest language level L4, the construct is represented by the character (measured in components), and the constituent is represented by the component (measured in the average number of its strokes). Language level L3 represents both the parcelate (measured in characters), which is the construct on this level, and the character (measured in the average number of its components), which is the constituent. Language level L2 represents both the sentence (measured in parcelates), which is the construct on this level, and the parcelate (measured in the average number of its characters) which is the constituent. On the highest language level L1, the construct is represented by the paragraph (measured in sentences), and the constituent is represented by

---

10 The original text: „dvě jazykové jednotky různého druhu mohou být uvažovány jako vztah dvou vzájemně hierarchizovaných úrovní ..., čili mohou být uvažovány jako konstrukt a konstituent. Když však přemýšlíme o textu, je omezení na vztah právě dvou úrovní nedostatečné; okamžitě se ptáme, co se v textu děje s ostatními úrovněmi. ... Jinak řečeno, každá jazyková entita vůči všem vyšším jazykovým úrovním je konstituentem a vůči všem nižším úrovním je konstruktem“ (Hřebíček, 2002, p. 59).

the sentence (measured in the average number of its parcelates). For easier reference, the language levels and units are listed in Table 9:

**TABLE 9**  
Language levels  $L_i$ ,  $x_i$  construct,  $y_i$  constituent ( $i = 1, 2, 3, 4$ )  
Source: created by authors

Language level	Construct $x_i$ ; constituent $y_i$		Length
L4	$x_4$	character	in components
	$y_4$	component	in the average number of its strokes
L3	$x_3$	parcelate	in characters
	$y_3$	character	in the average number of its components
L2	$x_2$	sentence	in parcelates
	$y_2$	parcelate	in the average number of its characters
L1	$x_1$	paragraph	in sentences
	$y_1$	sentence	in the average number of its parcelates

## 2.4 SEGMENTATION AND QUANTIFICATION OF THE SAMPLE TEXTS

Determination and definition of the language units and language levels allowed proceeding to a next step – segmentation of the sample texts.

The following four tables present examples of segmentation on the respective language levels. The part of scientific article was chosen and segmented according to the above mentioned definitions of the selected language units. The first table shows division of three characters into their components and subsequently components into their strokes. Table 11 shows the division of one parcelate into its characters and subsequently these characters into their

components. The third table shows the division of one sentence into its parcelates and subsequently the parcelates into their characters. Table 13 presents the division of one paragraph into its sentences and subsequently these sentences into their parcelates. Cf. Table 10, 11, 12 and 13.

**TABLE 10**  
 An example of segmentation on the language level character – component  
 Source: created by authors

Character 1		Character 2	
场		值	
Component 1	Component 2	Component 1	Component 2
土	勹	亻	直
Strokes 1-3	Strokes 1-3	Strokes 1-2	Strokes 1-8
一	勹 丿	丿	一
一	丿		冫
			一 一
			一 一

Character 3			
约			
Component 1	Component 2	Component 3	Component 4
纟	一	勺	丶
Strokes 1-2	Stroke 1	Strokes 1-2	Stroke 1
㇇ ㇇	一	丿 丿	丶

**TABLE 11**  
 An example of segmentation on the language level | parcelate – character  
 Source: created by authors

Parcelate 1						
翻译市场年产值约120亿元人民币。						
Character 1	Character 2	Character 3	Character 4	Character 5	Character 6	Character 7
翻	译	市	场	年	产	值
Components 1-6	Components 1-4	Components 1-2	Components 1-2	Component 1	Components 1-3	Components 1-2
番、	、	、	土	年	、	、
习	又	市	刃		厂	直
、						
Character 8	Character 9	Character 10	Character 11	Character 12	Character 13	Character 14
约	120	亿	元	人	民	币
Components 1-4	Components 1-3	Components 1-2	Components 1-2	Component 1	Component 1	Component 1
纟	1	、	一	人	民	币
、	2	乙	兀			
	0					

TABLE 12

An example of segmentation on the language level sentence – parcelate  
 Source: created by authors

Sentence 1												
Parcelate 1			Parcelate 2			Parcelate 3			Parcelate 4			
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外			来中国的学习者、投资者、工作者也逐渐增多。			仅英语学习市场年产值已超过100亿元人民币，			翻译市场年产值约120亿元人民币。			
Characters 1–20			Characters 1–3			Characters 1–19			Characters 1–14			
我	国	语	据	统	计	仅	英	语	翻	译	市	
言	教	育				学	习	市	场	年	产	
产	业	语				场	年	产	值	约	120	
言	翻	译				值	已	超	亿	元	人	
产	业	拥				过	100	亿	民	币		
有	强	大				元	人	民				
活	力					币						

TABLE 13

An example of segmentation on the language level paragraph – sentence  
 Source: created by authors

Paragraph 1	
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。我国语言教育产业、语言翻译产业拥有强大活力，据统计，仅英语学习市场年产值已超过 100 亿元人民币，翻译市场年产值约 120 亿元人民币。我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。	
<b>Sentence 1</b>	<b>Sentence 3</b>
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。	我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。
<b>Parcelates 1-5</b>	<b>Parcelates 1-6</b>
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。	我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。	我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。	我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。	我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。
我国正处在人口大流动时期，上亿农村人口向城市流动，劳动力国外输出渐成规模，学生出国留学人数与日俱增，境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。	我国信息化的发展日新月异，2012 年网民即将突破 5 亿，手机用户 9 亿，其中手机上网用户 3.5 亿；信息产品的社会普及率速度惊人，语言信息产业具有巨大的发展潜力。

After the segmentation it was necessary to quantify the sample texts, in other words the texts had to be transformed into the language of numbers in order to obtain the lengths of constructs  $x_i$ , their frequencies  $z_i$  and the lengths of constituents  $y_i$  for all language levels. Segmentation and quantification were processed by means of the Microsoft Excel program.

## 2.5 TESTING THE MODEL RELIABILITY BY MEANS OF STATISTICAL METHODS

The extracted lengths of the variables, i.e.  $x_i, z_i, y_i$  were processed in the statistical R program<sup>11</sup> by means of a statistical method, namely by means of linear regression<sup>12</sup>. On the basis of data processing, values of parameters  $A$  and  $b$  for all language levels were acquired. Owing to testing the model reliability for individual language levels it was also needed to extract coefficients of determination  $R^2$  whose values calculated by this software determine the measure of this

11 "R is a language and environment for statistical computing and graphics" (What is R?: Introduction to R, 2013).

12 "The relationship between a dependent variable and an independent variable can be graphically expressed by the regression line. The regression line is an illustration of a linear trend in data ... and it is constructed in a way which ensures that second powers of distances between all data points and the regression line are smaller than between all data points and any different line. ... Therefore during construction of the regression line we can talk about the method of the smallest squares, or more precisely about the method of the minimal sum of squares. ... Deviations of data points from the regression line are measured vertically, i.e. parallel to an axis of the dependent variable. The deviations are known as residua and it is obvious that the smaller the residua are, the more precise prediction of values the regression line provides. The tightness of data points to the regression line is known as goodness-of-fit ... If all data points lay on the regression line, there are no deviations ..." (Volin, 2007, p. 208–209; translated by authors).

The original text: „Vztah mezi závislou a nezávislou proměnnou je také možno vyjádřit graficky pomocí regresní přímky. Regresní přímka je pak znázorněním lineárního trendu v datech. ... Regresní přímka je sestrojena způsobem, který zajišťuje, aby druhé mocniny vzdáleností všech datových bodů od ní byly menší než od jakékoliv jiné přímky. ... Při sestrojování regresní přímky se proto mluví o metodě nejmenších čtverců, nebo přesněji o metodě nejmenší sumy čtverců. ... Odchyly datových bodů od regresní přímky se měří svisle, tj. paralelně s osou závislé proměnné. Říká se jim rezidua a je zřejmé, že čím menší jsou rezidua, tím přesnější predikci hodnot bude přímka poskytovat. Těsnosti, s jakou se datové body k přímce přimykají, se říká kvalita proložení. ... Budou-li všechny body ležet na regresní přímce, pak od ní nebudou mít žádné odchyly...” (Volin, 2007, p. 208–209).

reliability. “The range of the coefficient of determination is  $0 \leq R^2 \leq 1$  – the closer the values are to 1, the better the model fits. ... The values of  $R^2$  greater than or equal to 0.7 may be considered as adequate goodness-of-fit of the model in quantitative linguistics” (Andres et al., 2012, p. 15).

Observations with a low frequency (so-called extremes) which are insignificant in comparison with others were adjusted by a statistical method, namely by omitting. The values of omitted observations are cited in footnotes which are related to tables containing parameter  $A$ , parameter  $b$  and coefficient of determination  $R^2$ .

## 2.6 INTERPRETATION OF THE THE ACQUIRED DATA

An interpretation of the acquired data and their graphic visualizations was the last step of the experiment. This step is described in details in the two following parts. The first part is focused on the interpretation of the data acquired from the quantitative analysis of a scientific text; the second part is aimed at an interpretation of the results obtained from the quantitative analysis of an artistic text.



### **3. The application of the Menzerath–Altmann law to the scientific text**

The aim of this chapter is to describe the implementation of the quantitative analysis, which focused on testing the existence of language units defined by the graphical principle on respective language levels by means of the Menzerath–Altmann law applied to a sample text representing the scientific style.

Over the course of implementing certain steps, some specifics appeared in connection with analysing the scientific text. For this reason the first five steps are cited again in respective subchapters in order to maintain the sequence of the experiment's procedure. The emphasis is primarily laid on the steps with specifics. Other steps are mentioned only for the sake of comprehensiveness and their brief introduction is supplied with references to subchapters which describe them in detail. The closest attention is paid to the last step focusing on interpreting the data acquired from the analysis because it represents the core of this chapter. The data are interpreted individually within respective sections and the results are summarized in a conclusion at the end of this chapter.

#### **3.1 DETERMINATION OF THE CRITERIA FOR THE CHOICE OF THE SAMPLE TEXT AND THEIR EXPLANATION**

As mentioned above (cf. subchapter 2.1), the sample text first had to meet several requirements in order to be chosen. The fundamental step of the choice was to determine a stylistic style in which the sample text had to be written. For this reason, the analysis was aimed at the scientific style due to comparison with the data acquired from the previous experiments testing the newspaper style and the literary style and with the data obtained from a subsequent experiment which was realized by co-author Lenka Matoušková and which tested the artistic style. Next, the scientific article had to be published under the patronage of a highly regarded scientific institution in order to guarantee its scientific quality. In connection with the following requirement, i.e. the coherence, it was

necessary to leave out scientific articles of such disciplines which insert formulas, calculations, graphs, pictures and other requisites in their texts. In the case of the length it was inevitable to take attributes of the scientific style into account. Scientific articles are characterized by a greater length; therefore, it was necessary to increase the maximum of the text length from 3,500 to 5,500 Chinese characters. Last but not least, the sample text had to be written in the standard form of contemporary Chinese using the simplified set of Chinese characters and to be published after the year 2002 to ensure its contemporarity.

### 3.2 THE CHOICE OF THE SAMPLE TEXT

The determined criteria were satisfied by the scientific article titled *The Economic Aspect of Language* (*Rènshì yǔyán de jīngjìxué shǔxìng*, 认识语言的经济属性). The article was published in 2012 in the academic periodical *The Applied Linguistics*<sup>13</sup> (*Yǔyán wénzì yìngyòng*, 语言文字应用) and released in 2013 on the official websites of the *Chinese Academy of Social Sciences*<sup>14</sup> (*Zhōngguó shèhuì kēxuéyuàn*, 中国社会科学院). The author of the article is professor of *Beijing Language and Culture University* (*Běijīng yǔyán dàxué*, 北京语言大学)

- 
- 13 The Applied Linguistics represents a significant academic periodical which has been published since 1992 by the Institute of Applied Linguistics under Ministry of Education (“Yuyan wenzhi yingyong” Bianjibu (Yingyong yuyanxue yanjiu zhongxin), 2005).
- 14 The Chinese Academy of Social Sciences (CASS) is the highest academic institution and a research centre in the People’s Republic of China which is focused on research into philosophy and social sciences. CASS was established in May 1977 with the aim to support the development of philosophy and social sciences. In the present CASS has six academic divisions, namely *Literature and Philosophy* (*Wénzhé xuébù*, 文哲学部), *History Studies* (*Lìshǐ xuébù*, 历史学部), *Economics Studies* (*Jīngjì xuébù*, 经济学部), *Social, Political & Legal Studies* (*Shèhuì zhèngfǎ xuébù*, 社会政法学部), *International Studies* (*Guójiè yánjiū xuébù*, 国际研究学部) and *Marxist Studies* (*Mǎkèsīzhǔyì yánjiū xuébù*, 马克思主义研究学部), which are formed of 37 research institutes and 45 research centres conducting research into 120 scientific disciplines and employing more than 4,200 resident scholars. The current president of CASS is Chen Kuíyuan (*Chén Kuíyuán*, 陈奎元). In 2011 CASS became the foremost think tank in whole Asia; it took first place among 30 institutions. This and further information cf. (Wo yuan gaikuang, 2010), (Yuan jigou, 2011), (Zheng, 2012) a (Zhongguo shehui kexueyuan, © 2013).

Li Yuming<sup>15</sup> (Lǐ Yǔmíng, 李宇明). The article is written in the standard form of contemporary Chinese and its main coherent passage contains 5,155 simplified Chinese characters.

The selected sample text is given in the appendix of this publication (cf. Appendix 1).

### 3.3 DETERMINATION AND DEFINITION OF THE LANGUAGE UNITS AND LANGUAGE LEVELS

The language units were determined precisely in the previous experiments. Their definitions are described in detail in subchapter 2.3; therefore, their enumeration suffices in this part:

*stroke – component*<sup>16</sup> – *character – parcelate*<sup>17</sup> – *sentence – paragraph*.

A determination of the relationship between two immediately contiguous language units created four language levels, cf. Table 14.

*i* represents the four language levels: *i* = 1 for paragraph – sentence, *i* = 2 for sentence – parcelate, *i* = 3 for parcelate – character and *i* = 4 for character – component.

15 Li Yuming was born in 1955 in Miyang, Henan province. In 1981 he finished his studies in the Department of Chinese Language at Zhengzhou University and then began studying a master's degree of Modern Chinese at Central China Normal University. After his graduation in 1984 he started to teach there and in 1993 he was made a professor. Three years later he was appointed the dean of Faculty of Humanities and in 1998 the vice-chancellor. After 2001 he held various functions in institutes at Ministry of Education. In the present he carries out a secretary position in Beijing Language and Culture University. Li Yuming specializes in theory of linguistics, modern Chinese, psycholinguistics, language planning etc. He has published several books and more than 300 articles and he also became the editor in chief of the periodical *The Applied Linguistics*. This and further information cf. (Li Yuming, © 2013) a (Beiyu jiaoshou: Li Yuming jiaoshou, © 2006).

16 For the purpose of this experiment, the component represents a hypothetical language unit which was determined on the basis of the graphical criterion.

17 For the purpose of this experiment, the parcelate represents a hypothetical language unit which was determined on the basis of the graphical criterion with a minimal and inevitable consideration of the syntactic criterion.

**TABLE 14**Language levels  $L_i$ ,  $x_i$  construct,  $y_i$  constituent ( $i = 1, 2, 3, 4$ )

Source: created by authors

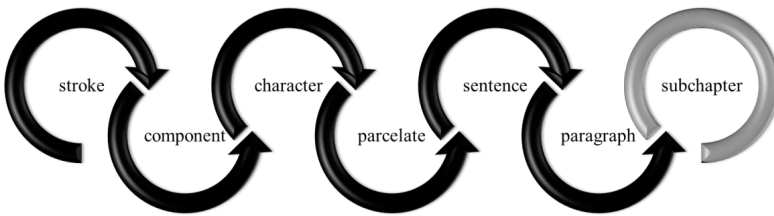
Language level L4		Language level L2	
$x_4$	character measured in components	$x_2$	sentence measured in parcelates
$y_4$	component measured in the average number of its strokes	$y_2$	parcelate measured in the average number of its characters
Language level L3		Language level L1	
$x_3$	parcelate measured in characters	$x_1$	paragraph measured in sentences
$y_3$	character measured in the average number of its components	$y_1$	sentence measured in the average number of its parcelates

### 3.4 SEGMENTATION AND QUANTIFICATION OF THE SAMPLE TEXT

After determining and defining the language units and the levels a next step followed: the segmentation of the sample text. Let us begin by defining the object at which the segmentation was aimed. The structure of the scientific articles consists of four sections. The first section includes the title of the article *The Economic Aspect of Language* (*Rènshì yǔyán de jīngjìxué shǔxìng*, 认识语言的经济属性), the name of the author Li Yuming (Lǐ Yǔmíng, 李宇明), a brief introduction of the author (*zuòzhě jiǎnjiè*, 作者简介), the date of publication and the source. A bilingual abstract (*nèiróng tíyào*, 内容提要) and keywords (*guānjiàncí*, 关键词) belong to the second section which follows. The third section begins with a main passage which is divided into four subchapters. In the end of the article the author adds explanatory notes (*zhùshì*, 注释) and a list of bibliographical references (*cānkǎo wénxiàn*, 参考文献).

Owing to the criterion for a coherent text, the quantitative analysis was performed on the main passage of the scientific article; the remaining three sections

(title, introductory information; bilingual abstract, key words; explanatory notes, bibliography) were not analysed because they represent independent coherent segments whose contents are not interconnected. As regards the division of the article into subchapters by way of subtitles, the subtitles were not taken into consideration. On the other hand, their usage in the structure of a text should not be ignored. Text segments delimited by subtitles could represent another language unit which could be called for this purpose a “subchapter”, cf. Figure 3.



**FIGURE 3**  
The language units including the “subchapter”

A connection between the “subchapter” and the immediately contiguous language unit, namely the paragraph, could create another language level: the subchapter (measured in paragraphs), which is the construct on this level, and the paragraph (measured in the average number of its sentences), which represents the constituent, cf. Table 15.

**TABLE 15**  
The alternative language level  $L_i$  including the “subchapter”,  $x_i$  construct,  $y_i$  constituent  
Source: created by authors

Language level $L_i$	
$x_i$	subchapter measured in paragraphs
$y_i$	paragraph measured in the average number of its sentences

Due to its length, the chosen sample text is not appropriate for testing the mutual relationship between these language units. Not only the number of empirically obtained observations could be insufficient, but also their frequencies could be very low. Therefore, the length of the scientific article, which was limited purposefully because of comparing all sample texts representing the different stylistic styles, does not allow conducting a quantitative analysis testing this language level. On the other hand, there is an option to realize another experiment which will be primarily focused on analysing the subchapter and the paragraph. For purpose of the experiment the length of sample texts should substantially exceed the maximum determined on 5,500 Chinese characters.

Eventually, in connection with defining the object of segmentation it is necessary to mention that the main passage of the scientific article contains Latin letters. Aside from a transcription of a Swiss economist's name into Chinese characters, the author states its original version written in the Latin alphabet: “弗朗斯瓦·格林 (François Grin)” (Li, 2013). This version given in parenthesis repeats the mentioned information; therefore, it does not represent an inseparable part of a parcelate with regard to the content. In view of this fact it was not taken into consideration because the analysis is exclusively related to Chinese characters.

Defining the object of the quantitative analysis allowed proceeding to segmentation. Before its implementation it was necessary to determine precisely which approach will be chosen in the instance of those particularities which appeared in the sample text and which are not rooted in Chinese. Let us begin by numbers written down by way of Arabic numerals.

Firstly, it was necessary to determine to which language units numbers and numerals used in the sample text belong. In contrast to Chinese characters, neither numbers nor numerals occupy the graphic fields, in other words, numerals adhere to each other. For this reason the number (both integer and rational number) was considered as one character and numerals as its components consisted of a certain number of strokes. The chosen approach is illustrated with the following example: the number 2009 is identical to one character composed of four components. It appears that the usage of Arabic numerals in Chinese

texts is probably related to economization of the language. This alternative of numerical notation represents a faster and easier way of writing in comparison with Chinese characters, not only in the case of handwriting, but also in the case of electronic records of language.

Next, let us turn to decimal numbers to which a particular attention must also be paid. As regards a decimal mark which separates the integer part from the fractional part, the Chinese character *dian* (diǎn, 点) is replaced with a decimal point. It means that the decimal point is regarded as one of the decimal number's components consisted of one stroke. This form is used probably due to the economization of the language because it represents a faster and easier way of writing. This phenomenon is illustrated with examples of differences in the number of components and strokes between the Chinese character *dian* and the decimal point, cf. Table 16. Both forms of the numerical notation are highlighted in bold type.

**TABLE 16**

Illustration of differences in the number of components and strokes between the two versions of the decimal numbers' notation  
Source: (Li, 2013); created by author

Arabic numerals			Chinese characters		
Number	Number of components	Number of strokes	Number	Number of components	Number of strokes
5.68	4	5	五 <b>点</b> 六八	12	19
6.33	4	4	六 <b>点</b> 三三	15	19
1.43	4	7	一 <b>点</b> 四三	10	14

The Chinese expressions for enumeration, such as *di-yi* (dì-yī, 第一), and *shouxian* (shǒuxiān, 首先) for “firstly”, *di-er* (dì-èr, 第二) and *qici* (qíci, 其次) for “secondly” and *di-san* (dì-sān, 第三) for “thirdly”, were replaced by Arabic numerals with dots, i.e. 1., 2., 3. etc. This form is considered as

one character, which is composed of two components – the numeral and the dot, and it could also be related to the economization of the language.

The last comment concerns percentages. The Chinese term for the percentage – *bai fenzhi* (bǎi fēnzhī, 百分之) followed by Chinese characters for numerals or by Arabic numerals – was replaced with the faster and easier form – a percent sign %. Although the percent sign does not occupy the graphic field, it is regarded as one character consisting of three components. It means that the entire expression of the percentage is made up of two characters – a number and the percent sign. The economization of the language could also exert a strong influence on this form of numerical notation.

Anomalies also appeared in the usage of punctuation. As mentioned earlier, selected punctuation marks establish the borders of a sentence (。 ? ! ). Despite this, the colon was placed at the end of two sentences and the text which follows began as a new paragraph (i.e. a paragraph which is separated by an inserted blank line and whose first line is indented), cf. below mentioned example (Li, 2013):

“人类的经济活动与语言密不可分，而且在某些领域，语言和语言知识已经成为重要的经济资源。这可以从以下几个方面来看：

首先，语言能力是劳动力的重要构成要素。”

These deviations cannot be considered to be a human error because the usage of the colons occurred twice under identical circumstances. The sentences ended by the colons were followed by new paragraphs which represented an enumeration of certain aspects and began by Chinese terms used for the enumeration (i.e. firstly, secondly; in Chinese *shǒuxiān*, 首先, *qícì*, 其次 etc.). The determination of the language units was primarily based upon the graphic principle, for this reason the segmentation respected the scientific article's division into paragraphs and the colons became secondarily the border of the sentences.

In addition, it was necessary to specify the way of the segmentation of a sentence in which the title of the document, given in title marks (《》), contained a comma (, ), cf. the below mentioned example (Li, 2013). On the one hand, the title marks were defined as such punctuation marks which do not distinguish



the borders of parcelates (cf. subchapter 2.3, Table 8). On the other hand, the comma belongs to such punctuation marks which are valid for the parcelates' borders (cf. subchapter 2.3, Table 7). Segmentation of this sentence abided by rules determined for the purpose of this experiment and thus the title marks were not taken into consideration. When the title marks are not included, the comma will disturb the syntactic structure of the sentence; therefore, it did not delimit the parcelate for this time.

“在 2010、2011 年北京市“两会”上，北京市人大代表贺宏志先生连续提出《关于发展我市语言产业的建议》和《加强语言文化建设□促进语言产业发展》的建议，语言经济的话题首次提到了地方人民代表大会的议坛上。”

After the precise determination of approaches and after the segmentation it was possible to proceed to a next step which was aimed at the quantification of the scientific article in order to extract lengths of constructs  $x_i$ , their frequencies  $z_i$  and lengths of constituents  $y_i$ , where  $i = 1, 2, 3, 4$ ; cf. subchapter 2.4.

### 3.5 TESTING THE MODEL RELIABILITY BY MEANS OF STATISTICAL METHODS

Values obtained by quantification were inserted in the statistical R program which calculated the required data output, namely parameters  $A$ ,  $b$  and a coefficient of determination  $R^2$  for all language levels (cf. subchapter 2.5). Aside from the parameters and the coefficient of determination  $R^2$ , the program created graphical displays of the acquired results for each language level.

A statistical method was also used in the instance of observations with a low frequency. The observations whose frequencies were insignificant in comparison with others were omitted. The grey background of the cells in respective tables is used to highlight these extremes and the values of parameters  $A$ ,  $b$  and the coefficient of determination  $R^2$  extracted after the omission of these extremes are stated in respective footnotes.

### 3.6 INTERPRETATION OF THE ACQUIRED DATA

The last step, an interpretation of the acquired data, represents the core of this chapter. The following four subchapters deal with the individual language levels (L4, L3, L2, L1). Every subchapter begins by stating a table containing the lengths of variables (i.e.  $x_i$ ,  $z_i$  and  $y_i$ ). Next, let us turn to a graphical visualization with comments and finally, let us discuss the obtained results. Conclusions of the interpretations and discussions are summarized at the end of this chapter.

#### 3.6.1 Language level L4

##### CHARACTER – COMPONENT

Table 17 shows the observations empirically obtained by quantification of the scientific article:  $x_4$  represents the lengths of characters (measured in components),  $z_4$  their frequencies and  $y_4$  the average lengths of components (measured in strokes). The grey background of the cells is used to highlight the omitted observations with a low frequency ( $z_4 \leq 8$ ). The values obtained by their omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 21.

**TABLE 17**

Language level L4: character (measured in components) – component (measured in the average number of its strokes)

$x_4$	$z_4$	$y_4$
1	789	4.7136
2	1,240	3.3056
3	1,285	2.4999
4	995	2.0766
5	540	1.6815
6	134	1.9701
7	163	1.3681
8	8	1.5000
9	1	1.7778

The sample text establishes 9 characters' lengths. Average lengths of components appear within the interval of  $\langle 1.37; 4.71 \rangle$ . It is evident that the longer a character (measured in components) is, the shorter its components (measured in strokes) are, hence the relationship of the inverse proportionality between variables' lengths is valid. Nevertheless, the decreasing tendency of average constituents' lengths was interfered with characters composed of 6 ( $y_4 = 1.97$ ), 8 ( $y_4 = 1.50$ ) and 9 ( $y_4 = 1.78$ ) components. In the case of 8-component and 9-component characters, the deviation could be caused by a low frequency ( $z_4 \leq 8$ ). Ranks<sup>18</sup>, assigned by the 文林 Wenlin software<sup>19</sup> on the basis of frequency, and frequency groups<sup>20</sup> reveal that these characters belong to less frequent Chinese characters (cf. Table 18). Characters with the highest frequency in the sample text consist of 3 and 2 components ( $z_4 \geq 1240$ ).

**TABLE 18**

8-component and 9-component characters and their frequency created by 文林 Wenlin Software

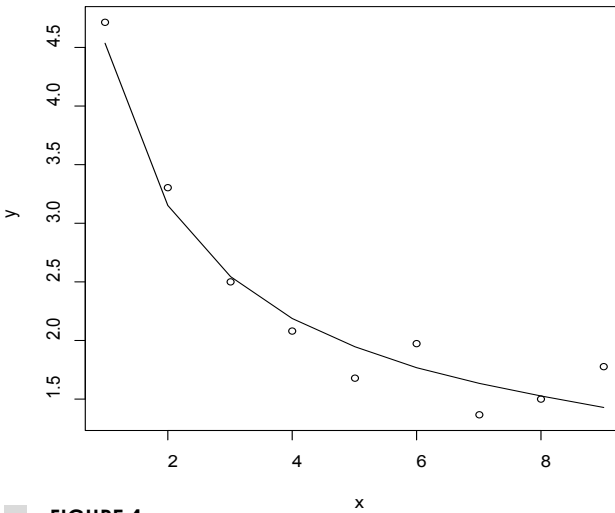
Source: 文林 Wenlin Software; created by author

Character	感	德	测	综	墨	韵	飙
Number of components	8	8	8	8	8	8	9
Rank	336	401	1,121	1,553	1,680	2,126	×
Frequency group	1 <sup>st</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	2 <sup>nd</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	×
Frequency $z_4$	1	1	2	2	1	1	1

18 The lower a value of stated ranks is, the higher a frequency of respective characters is.

19 The frequency list of Chinese characters created by the 文林 Wenlin Software for Learning Chinese, Version 4.0.2.

20 The most frequent 3,000 Chinese characters incorporated in the frequency list of Wenlin were divided into three frequency groups according to their ranks: the first group includes characters with the rank from 1 to 1,000; the second frequency group contains characters with the rank from 1,001 to 2,000 and the third frequency group embraces characters with the rank from 2,001 to 3,000. A sign of the cross was made within characters which do not have a rank due to their extremely low frequency.



**FIGURE 4**

Graphic visualization of the observations presented in Table 17

Parameter  $b$  (cf. Table 19) has a positive value which results in the down-trend and convexity of the curve visualizing the relationship between characters' and components' lengths. On the basis of the coefficient of determination (cf. Table 19) it can be concluded that the mathematical model of the MAL shows an extremely wide goodness-of-fit with empirically gained observations; its value exceeds 0,9. The sample text adheres, in an almost perfect fashion, to the assumptions of the MAL.

**TABLE 19**

Parameter  $A$ , parameter  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 17<sup>21</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
4.5406	0.5259	0.9061

21 After omitting the observations with the lowest frequency ( $z_4 \leq 8$ ): parameter  $A = 4.8117$ ; parameter  $b = 0.5974$ ; coefficient of determination  $R^2 = 0.9519$ .

On the basis of the results obtained from the scientific article, it seems that the component – a hypothetical language unit determined purposefully in accordance with the graphical principle – exists and it has proved itself as a valid language unit on this level.

## DISCUSSION

The question arises what factors could bring about this extremely wide agreement. Let us begin by a characteristic of the construct on this language level. O. Švarný takes the view that the characters represent the basic graphic units which are approximately equal in size regardless of the number of strokes composing a character. The strokes are arranged into a square or into a rectangle (whose height is not much bigger than its width); cf. (Švarný, 1967, p. 31). Maintaining the identical size of the graphic field imposes a strong pressure on the structure of the character. In other words, inside its limited space the graphic field forms a certain system of rules which regulates the arrangement of strokes and components and thereby creates a relationship of subordination between the field and characters. For this reason the graphic field, or, more precisely, its constant size could be regarded as a crucial factor.

Consequently, characters composed of one component have a potential to become more complex without losing their legibility because they can occupy a whole space of the graphic field. Owing to the absence of other components, the component of these characters can consist of a greater number of strokes (cf. Table 20) because the graphic field imposes a minimal pressure here.

In the case of characters composed of a higher number of components the potential to become more complex is limited. In contrast to one-component characters, individual components of these characters can squeeze into a much smaller space. Nevertheless, components composed of a larger number of strokes can diminish themselves only to a certain extent without losing their readability. Hence the graphic field forces to use more uncomplicated components (i.e. with a smaller number of strokes), cf. Table 20. The main conclusion drawn from this phenomenon is that economization of the graphic field's space plays a key role because demand of economization rises depending on a higher

number of components within an individual character in order to ensure its legibility.

**TABLE 20**

Simplified characters composed of different number of components and strokes

Source: created by author

Character	重	展	象	影	德	颿
Number of components	1	1	1	7	8	9
Number of strokes	9	10	11	15	15	16
Average number of strokes	9	10	11	2.14	1.88	7.78

There is also another tendency in the arrangement of characters within the graphic field. If a character consists of a large number of components, they incline to connect to one another because the graphic field does not provide them sufficient space to maintain their independence; blank space which originally separated components from each other is eliminated. It means that a number of components within a character is reduced and a component (or components) which newly emerged through the connection has (have) a higher number of strokes.

In view of this fact, traditional characters should be regulated more strictly by the above mentioned system of rules due to their higher number of strokes and components, cf. Figure 5 and 6.



**FIGURE 5**

Traditional character composed of 17 components and 32 strokes



**FIGURE 6**

Traditional character composed of 5 components and 30 strokes

The fact that these rules could be respected by traditional characters to a greater extent than in the case of simplified characters was verified by a sub-experiment applied to the scientific text transformed into the traditional set of characters by the 文林 Wenlin Software. The sub-experiment is marked according to the respective language level (i.e. 4) and the respective sample text (i.e. A). The gained results are stated below.

**SUB-EXPERIMENT 4A**

The transformation of the scientific article into the traditional form involved 2,027 characters from the total amount of 5,155. With regard to the total amount of various characters, i.e. 761, the transformation involved 282 characters. Remaining characters (3,128; 479) held their original form. Segmentation of the transformed characters abided by the selected definitions of the language units (cf. subchapter 2.3)

Table 21 presents observations empirically obtained by quantification of the transformed scientific article:  $x_{4A}$  represents the lengths of characters (measured in components),  $z_{4A}$  their frequencies and  $y_{4A}$  the average lengths of components (measured in strokes). The grey background of the cells is used to highlight the omitted observations with a low frequency ( $z_{4A} \leq 2$ ). The values obtained by their omission (parameters  $A, b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 22.

**TABLE 21**  
Sub-experiment 4A – Language level L4: character (measured in components) – component (measured in the average number of its strokes)

$x_{4A}$	$z_{4A}$	$y_{4A}$
1	739	4.7605
2	976	3.8817
3	1,011	2.9829
4	637	2.7323

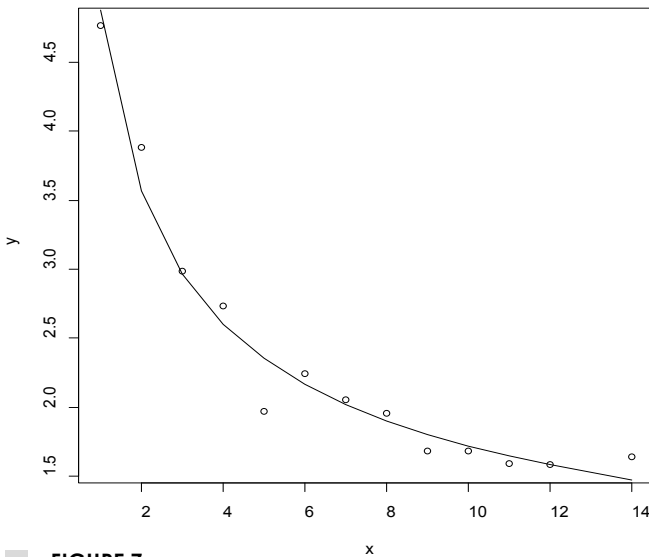
$x_{4A}$	$z_{4A}$	$y_{4A}$
5	631	1.9715
6	638	2.2422
7	128	2.0569
8	63	1.9603
9	193	1.6845
10	134	1.6866
11	2	1.5909
12	1	1.5833
14	2	1.6429

The transformed scientific article establishes 13 characters' lengths. Average lengths of its components appear in the interval of (1.58–4.76). As can be seen from Table 21, the average lengths of the components decline, whereas the lengths of the characters increase; nonetheless, two cases, i.e. characters composed of 10 ( $y_{4A} = 1.69$ ) and 14 ( $y_{4A} = 1.64$ ) components, contradict this relationship of an inverse proportionality. An exceptional case of the characters composed of 5 components ( $y_{4A} = 1.97$ ) should be also noted. Although this average length shows the downtrend, its value considerably disturbs the decrease of the following average components' lengths and thereby produces a deviation. As regards frequency, the most frequent characters consist of 3 and 2 components ( $z_{4A} \geq 976$ ). The lowest frequency was noticed in the case of characters composed of 11, 12 and 14 components ( $z_{4A} \leq 2$ ).

A comparison between the gained results yields findings that the transformed text contains a higher number of characters' lengths in contrast to the original text written in simplified characters. Other lengths include characters composed of 10, 11, 12 and 14 components. In the case of constituents, although the interval, in which the average lengths  $y_{4A}$  appear, has approximately equivalent values (i.e. the interval between 1 and 5 strokes), a slight



increase is observed within the transformed variant due to the occurrence of traditional characters whose numbers of strokes are generally larger. As regards deviations, the decreasing tendency of average lengths is interfered three times in both text's variants. Despite this, the relationship of the inverse proportionality is confirmed. The most frequent characters of both text's variants are characters composed of 3 and 2 components. It should be noted that the frequency of characters which were used in the transformed text and which consist of a higher number of components (i.e. 5, 6, 8, 9 and 10 components) increases, while the frequency of characters consisting of 1, 2, 3, 4 and 7 components decreases. Characters with the greatest lengths (i.e. 8-component and 9-component characters used in the first text's variant; 11-component, 14-component and 12-component characters used in the second text's variant) belong to observations with the lowest frequency.



**FIGURE 7**

Graphical visualization of the observations presented in Table 21

It is clear from the graphical visualization (cf. Figure 7), the value of parameter  $b$  and the coefficient of determination (cf. Table 22) that the mathematical

model of the MAL shows an extremely wide goodness-of-fit with empirically obtained observations, even higher than in the previous experiment undertaken on the original text's variant. The value of its goodness-of-fit exceeds 0.95. It can be concluded that the transformed text's variant adheres, in an almost perfect fashion, to the assumptions of the MAL.

**TABLE 22**

Parameter  $A$ , parameter  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 21<sup>22</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
4.8769	0.4525	0.9597

In the case of this sub-experiment, the component – a hypothetical language unit determined purposefully in accordance with the graphical principle – has also proved itself as a valid language unit on this level.

Not only the validity of the MAL was verified in the sample text transformed into the traditional form, but also the agreement exceeding 0.95 validated that the graphic field through its inner rules for the character's arrangement plays a significant role in the formation of relationships between the language units, such as the stroke, the component and the character.

It must be emphasized that the empirically acquired observations reflect not only the mutual relationship between components and characters transformed into their traditional form, but also the relationship between components and characters preserving their original form. The transformation involved only 2,027 characters from the total amount of 5,515. Therefore, remaining characters (almost 61 %) still have a considerable influence on the final result due to their majority and do not allow exposing a mutual relationship between components and those characters which can be transformed

22 After omitting the observations with the lowest frequency ( $z_{4A} \leq 2$ ): parameter  $A = 4.9778$ ; parameter  $b = 0.4708$ ; coefficient of determination  $R^2 = 0.9580$ .

into their full forms. The increase of the coefficient of determination  $R^2$  (from 0.9061 to 0.9597) could point out that abiding by the above described rules of the graphic field is stricter within traditional forms of Chinese characters. For this reason, subsequent experiments focusing exclusively on unsimplified form will be conducted.

### 3.6.2 Language level L3

#### PARCELATE – CHARACTER

Table 23 presents observations empirically obtained by quantification of the scientific article:  $x_3$  represents the lengths of parcelates (measured in characters),  $z_3$  their frequencies and  $y_3$  the average lengths of characters (measured in components). The grey background of the cells is used to highlight the omitted observations with a low frequency ( $z_3 \leq 1$ ). The values obtained by their omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 23.

■ **TABLE 23**

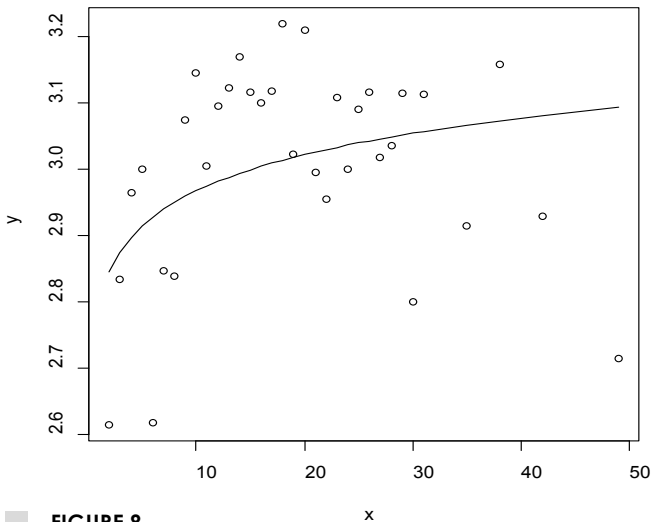
Language level L3: parcelate (measured in characters) – character (measured in the average number of its components)

$x_3$	$z_3$	$y_3$
2	13	2.6154
3	2	2.8333
4	7	2.9643
5	4	3.0000
6	17	2.6176
7	14	2.8469
8	17	2.8382
9	15	3.0741
10	20	3.1450
11	17	3.0053

$x_3$	$z_3$	$y_3$
12	35	3.0952
13	15	3.1231
14	22	3.1688
15	24	3.1167
16	22	3.0994
17	19	3.1176
18	17	3.2190
19	18	3.0234
20	15	3.2100
21	10	2.9952
22	5	2.9545
23	6	3.1087
24	6	3.0000
25	4	3.0900
26	6	3.1154
27	2	3.0185
28	2	3.0357
29	3	3.1149
30	1	2.8000
31	2	3.1129
35	1	2.9143
38	1	3.1579
42	1	2.9286
49	1	2.7143

As regards parcelates, the total number of their lengths is 34. Average lengths of their characters oscillate within the interval of  $\langle 2.62; 3.22 \rangle$ . It is

apparent that the sample text does not show the downward trend within the characters' lengths because they fluctuate around the previously cited values. The relationship of the inverse proportionality between the parcelate and the character is not proved. Concerning frequencies, they do not have high values due to a relatively high number of various parcelates' lengths. Despite this, it can be stated that the most frequent parcelates are composed of 12 characters ( $z_3 = 35$ ). The longest parcelates ( $x_3 = 30, 35, 38, 42$  and  $49$ ) appeared in the text to the smallest extent ( $z_3 = 1$ ).



**FIGURE 8**

Graphical visualization of the observations presented in Table 23

Parameter  $b$  has a negative value, hence the sample text does not adhere to the assumption of the MAL (cf. Table 24). As can be seen from the graphical visualization, the decreasing and convex tendency of the curve defined by the MAL is not observed. For this reason the mathematical model does not prove any goodness-of-fit with empirically gained observations (the coefficients of determination  $R^2$  stated in Table 24 is valid for the increasing tendency of the curve which is in direct contradiction to the assumption of the MAL).

**TABLE 24**

Parameter  $A$ , parameter  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 23<sup>23</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
2.7933	-0.0262	0.1435

It appears on the basis of the results obtained from the scientific article that the existence of the parcelate – a hypothetical language unit determined purposefully in accordance with the graphical and syntax principle – is not supported on this level. Nevertheless, its existence is not excluded because of certain factors which adversely affect the relationship between the units on this level (cf. Discussion).

## DISCUSSION

Of course, the existence of the relationship between the parcelate and the character was not proved by means of the MAL, but it is nevertheless necessary to point out that there are several factors which could exert a strong influence upon the invalidity of these language units' relationship.

A different characteristic of the units could be considered as the first factor. The parcelate measured in characters represents a variable construct. An author of a text creates its length depending on his language intelligence and expressive language skills without any restrictions. In other words, it can be said with just a little exaggeration that the parcelate can be composed of an infinite number of characters because its structure is not invariable. On the contrary, the character measured in components represents a constituent with a constant length. It means that it has an unchanging fixed structure which cannot be modified by authors. Furthermore, this unit is tied up with a certain meaning which is restricted by rules for its usage. As a consequence, in most cases this restriction

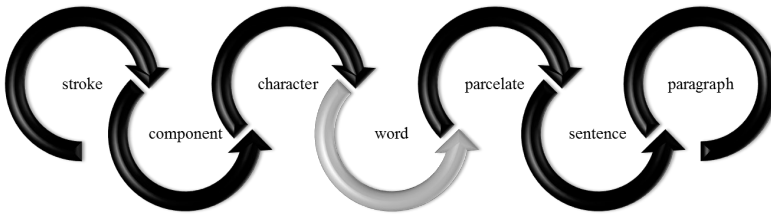
23 After omitting observations with the lowest frequency ( $z_3 \leq 1$ ): parameter  $A = 2.6651$ ; parameter  $b = -0.0484$ ; coefficient of determination  $R^2 = 0.4615$ .

does not allow purposely substitution of individual characters by alternatives composed of a lower or a higher number of components. In comparison with the parcelate, the character's length (i.e. number of components) is completely independent of the author; hence their mutual relationship could be significantly interfered with this contrast.

Not only the parcelate and the character, but also the component could exert a strong influence on the unsuitability of the mathematical model. As mentioned in subchapter 2.3, owing to the absence of a universally valid component definition, the character can be divided into components according to numerous approaches. Even the choice of a certain segmentation method does not have to be definitive because there is a possibility of its diversification into subsequent approaches, as was already demonstrated by fonts' variability (cf. Table 6). In addition, while the MAL can be valid on language level L4 on the basis of a selected component definition, nevertheless, in the case of the higher language level L3 it does not have to hold true because this definition does not have to represent an appropriately selected approach for this level. The relationship of these language units will stay an object of research and it will be analysed on the basis of different component definitions which will be verified by means of the MAL.

The invalidity of the mutual relationship on this level could also alert to an insufficient amount of language units and levels. The parcelate is composed of a certain number of characters which are regarded as independent entities. It should be noted that Chinese characters lost a logographic attribute due to the formation of polysyllabic words. Hence they started to be connected with other characters which resulted in the existence of a certain relationship between them (i.e. absent language levels). However, these experiments ignore this relationship owing to the hypothesis formulated on the basis of the graphics which, unfortunately, does not allow Chinese texts to reflect it. Linking the parcelate and the character into a language level on the one hand and omitting a language unit which is higher than character and lower than parcelate on the other hand could bring about a loss of the mutual relationship of inverse proportionality between an absent unit (the construct) and the character (the constituent) and between the parcelate (the construct) and an absent unit (the constituent).

The question is which language unit is missing between the parcelate and the character? A *word* could be the answer. Linking the word with other units could lead to the acquisition of the following language units (cf. Figure 9) and levels (cf. Table 25).



**FIGURE 9**  
The language units including the word

**TABLE 25**  
The alternative language level  $L_i$  including the word,  $x_i$  construct,  $y_i$  constituent  
Source: created by authors

Language level $L_i$		Language level $L_i$	
$x_i$	word measured in characters	$x_i$	parcelate measured in words
$y_i$	character measured in the average number of its components	$y_i$	word measured in the average number of its characters

Due to the fact that the word has been included as another language unit, the graphical principle, which was crucial for the determination of the units in these experiments, has to be left aside. Maintaining the graphical principle is not admissible because the graphics of texts written in Chinese characters does not use a space between words. Therefore, the alternative of an orthographic word cannot be taken into account, as a result of the missing function of Chinese texts to distinguish this kind of a word from each other. It is obvious that it should be necessary to choose another criterion, such as semantic, syntactic etc.



Because of the need to verify this assumption, co-author L. Matoušková performed a sub-experiment (cf. 4.6.2.) which tested the validity of a mutual relationship between a syntactic word, which represents a construct measured in characters, and the character, which is a constituent measured in components. Although the empirically gained observations showed a minimal goodness-of-fit with the mathematical model of the MAL, it is necessary to conduct other experiments which will verify the chosen definitions of a word and give precision to them.

In connection with the absence of a language unit, frequency of constructs' lengths on the lowest language level  $z_4$  should also be taken into account as one of the factors. The average lengths of characters  $y_3$  oscillate in the interval of (2.62; 3.22). It means that every parcelate is composed of characters consisting, on average, of 2 or 3 components. Regarding the number of occurrences of these characters, they comprise not only the largest proportion of the total amount of various characters used in the text, i.e. 50.07 % (which means 381 out of 761 characters), but also the majority of the total amount of all characters used in the text, i.e. 48.98 % (which means 2,525 out of 5,155 characters), cf. Table 26, Figure 10 and 11.

It seems that in connection with the high frequency of these characters and simultaneously with the absence of a language unit and level, the averaging lengths of characters in every parcelate could lead to a loss of nuances. This could be why the average lengths of characters have similar values. The addition of another language unit could bring about substantial differences between average lengths of characters and consequently it could reveal a mutual relationship between the absent language unit and the character.

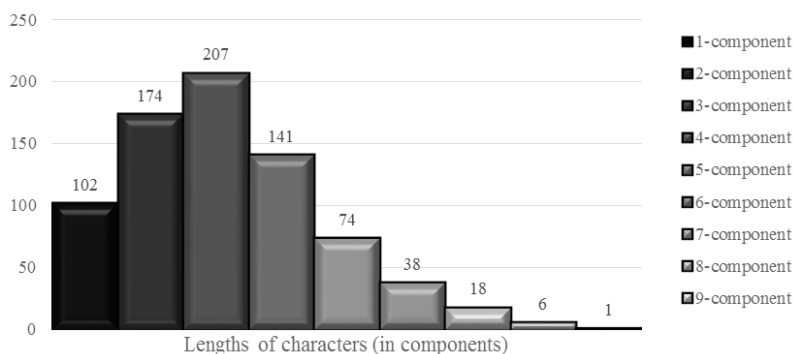
**TABLE 26**

Outline of character frequency related to the observations presented in Table 17

Length of characters (in components)	Total amount of various characters (without duplication)		Total amount of all characters (with duplication)	
	Number of characters	%	Number of characters	%
1-component	102	13.4034	789	15.3055

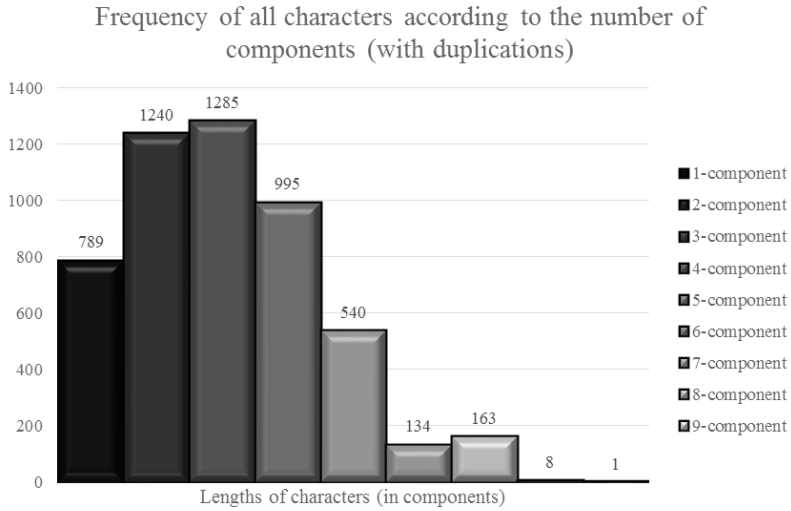
Length of characters (in components)	Total amount of various characters (without duplication)		Total amount of all characters (with duplication)	
	Number of characters	%	Number of characters	%
2-component	174	22.8647	1,240	24.0543
3-component	207	27.2011	1,285	24.9273
4-component	141	18.5283	995	19.3016
5-component	74	9.7240	540	10.4753
6-component	38	4.9934	134	2.5994
7-component	18	2.3653	163	3.1620
8-component	6	0.7884	8	0.1552
9-component	1	0.1314	1	0.0194
<b>Total</b>	<b>761</b>		<b>5,155</b>	

Frequency of various characters according to the number of components (without duplications)



**FIGURE 10**

Graphic representation of the frequency of various characters according to the number of components, related to the data presented in Table 26



**FIGURE 11**  
 Graphic representation of the frequency of all characters according to the number of components, related to the data presented in Table 26

The last factor could be the Chinese script simplification which was realized in the 1950s and 1960s. The aim of this reform was to simplify almost 3,000 traditional Chinese characters by declining the number of their strokes. As a consequence, the interventions reduced not only substantial differences in numbers of strokes within these characters, but also differences in the numbers of their components to a considerable extent. For this reason, it is necessary to conduct an analysis which will focus on the comparison of the traditional set of Chinese characters with its simplified variant. A crucial finding of the analysis will be the number of occurrences of characters consisting of 2 and 3 components. Testing the influence of the reform will be the objective of a subsequent experiment.

### 3.6.3 Language level L2

#### SENTENCE – PARCELATE

Table 27 shows the observations empirically obtained by quantification of the scientific article:  $x_2$  represents the lengths of sentences (measured

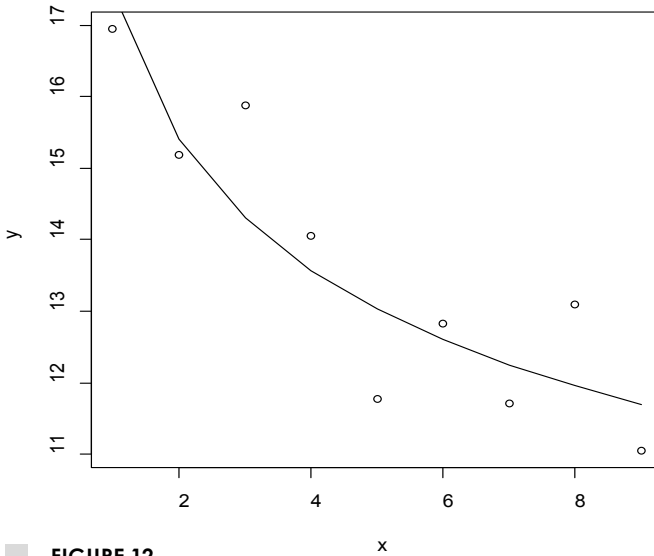
in parcelates),  $z_2$  their frequencies and  $y_2$  the average lengths of parcelates (measured in characters). The grey background of the cells is used to highlight the omitted observations with a low frequency ( $z_2 \leq 2$ ). The values obtained by their omission (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 24.

**TABLE 27**

Language level L2: sentence (measured in parcelates) – parcelate (measured in the average number of its characters)

$x_2$	$z_2$	$y_2$
1	18	16.9444
2	27	15.1852
3	30	15.8778
4	16	14.0469
5	6	11.7667
6	6	12.8333
7	2	11.7143
8	5	13.1000
9	2	11.0556

As can be seen from Table 27, 9 various lengths of the sentences occur. The average lengths of the parcelates appear in the interval of  $\langle 11.06; 16.94 \rangle$  and with the exception of three cases (i.e. sentences composed of 3 ( $y_2 = 15.88$ ), 6 ( $y_2 = 12.83$ ) and 8 parcelates ( $y_2 = 13.10$ )) they show a decreasing tendency while the sentence lengths increase. Hence the relationship of the inverse proportionality between variables is proved. As regards frequency, observations with the highest frequency ( $z_2 \geq 27$ ) are sentences composed of 2 and 3 parcelates. Sentences composed of 7 and 9 parcelates belong to the least frequent ( $z_2 = 2$ ) constructs on this level.



**FIGURE 12**  
Graphical visualization of the observations presented in Table 27

That is clear evidence that a positive value of parameter  $b$  (cf. Table 28) and a decreasing and convex curve visualizing the relationship between the language units on this level adhere to the assumptions of the MAL. It can be concluded that its mathematical model reveals a wide goodness-of-fit with the empirically obtained observations; the value of the goodness-of-fit adds up to almost 0.79 (cf. Table 28).

**TABLE 28**  
Parameter  $A$ , parameter  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 27<sup>24</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
17.4816	0.1827	0.7862

24 After omitting the observations with the lowest frequency ( $z_j = 2$ ): parameter  $A = 17.1221$ ; parameter  $b = 0.1542$ ; coefficient of determination  $R^2 = 0.7167$ .

Although the validity of the parcelate was not verified on the previous level, there is clear evidence on the basis of the results gained from the scientific article that this hypothetical language unit exists and it has proved itself as a valid language unit on this level.

## DISCUSSION

Let us return to the deviations occurring in the average parcelates' lengths. The increasing tendency of 3-parcelate sentences could be closely connected with punctuation marks and their functions. Subsequent calculations showed that 15 out of the total amount of these sentences (i.e. 30) have a higher average length of their parcelates than 2-parcelate sentences ( $y_2 > 15.19$ ). On the basis of the used punctuation marks it is possible to divide these 15 sentences into 3 groups.

1. The first group is represented by 5 sentences without any particularities.
2. The second group embraces 3 sentences using an enumeration comma. As mentioned earlier (cf. subchapter 2.3), this punctuation mark separates parts of a sentence in a coordinate relationship and for this reason it is not valid to define borders of parcelates. Hence, these parts of a sentence comprise an indivisible part of a parcelate and thereby they increase its length.
3. The third group involves sentences containing the quotation marks (“ ”) and the title marks (《 》). Contents of these punctuation marks very substantially increase the total number of characters in a parcelate and consequently its average length to a larger extent than in the previous group of sentences. These punctuation marks appeared in 7 sentences, 3 of them also include the enumeration comma or commas.

Table 29 presents 3 sentences in which a combination of the previously mentioned punctuation marks (enumeration comma, quotation marks, title marks) was used. Every sentence is divided into three parcelates which are accompanied by the numbers of their characters and the total average lengths

of parcelates within the whole sentence. The punctuation marks are highlighted in bold type.

**TABLE 29**

Illustrations of the sentences composed of 3 parcelates with their average lengths measured in characters

Source: (Li, 2013); created by author

Parcelates of the sentences	Number of characters
2012年3月2日，	6
黄少安、苏剑、张卫国三位发表的《语言经济学与中国的语言产业战略》，	28
基本上代表了我国学界在语言经济方面的认识。	20
<b>Average length</b>	<b>18</b>
国家语委全力支持山东大学、南京大学、广州大学等高校的语言经济学研究，	31
还于2008年12月29日支持商务印书馆成立了“中国语言资源开发应用中心”，	30
中心的宗旨是“致力于把语言及语言知识转化为生产力和文化商品”。	28
<b>Average length</b>	<b>29.67</b>
在2010、2011年北京市“两会”上，	10
北京市人大代表贺宏志先生连续提出《关于发展我市语言产业的建议》和《加强语言文化建设，促进语言产业发展》的建议，	49
语言经济的话题首次提到了地方人民代表大会的议坛上。	24
<b>Average length</b>	<b>27.67</b>

Regarding the remaining deviations, the increases could also be connected with the above mentioned punctuation marks simultaneously with low frequencies of the observations.

Finally, it is crucial to point out that the wide agreement between the mathematical model of the MAL and the empirically gained observations could be

caused by the length's variability of both units. The previous subchapter examining language level L3 cited that one of the reasons causing the low agreement could be the different characteristic of the language units: the parcellate represents a construct with a variable length and the character is a constituent with an invariable length (cf. 3.6.2). In the case of language level L2 the lengths of both language units are variable, in other words the Chinese language users form their length in accordance with their language intelligence and expressive language skills. Owing to the elimination of invariability, the mutual relationship is not disturbed which could result in the positive effect on the wide goodness-of-fit on this level.

### 3.6.4 Language level L1

#### PARAGRAPH – SENTENCE

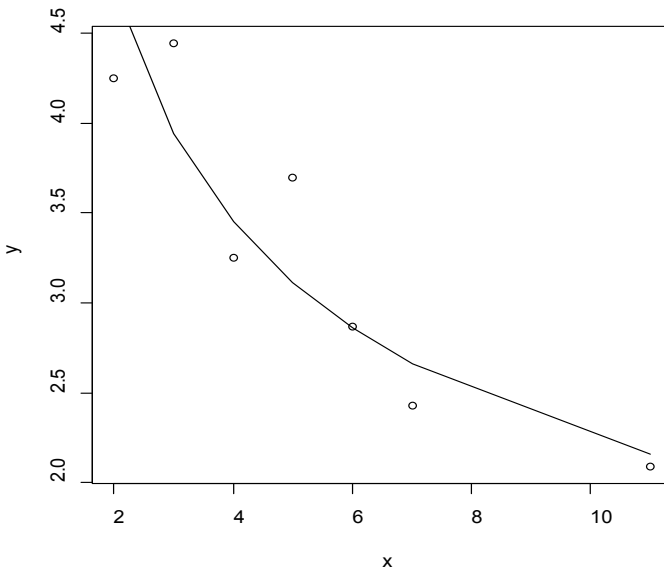
Table 30 presents the observations empirically obtained by quantification of the scientific article:  $x_i$  represents the lengths of paragraphs (measured in sentences),  $z_i$  their frequencies and  $y_i$  the average lengths of sentences (measured in parcellates).

**TABLE 30**  
Language level L1: paragraph (measured in sentences) – sentence  
(measured in the average number of its parcellates)

$x_i$	$z_i$	$y_i$
2	8	4.2500
3	3	4.4444
4	3	3.2500
5	4	3.7000
6	5	2.8667
7	2	2.4286
11	1	2.0909



The sample text establishes 7 various lengths of the paragraphs. The average lengths of the sentences appear between 2.09 and 4.44 parcelates. With the exception of 2 instances the sentence lengths show a downward trend. The opposite tendency is observed in the case of paragraphs consisting of 3 ( $y_1 = 4.44$ ) and 5 ( $y_1 = 3.70$ ) sentences. Despite this, the sentence lengths decline with the increasing paragraphs' lengths and thus the relationship of the inverse proportionality is valid. As regards frequency, the most frequent constructs are the shortest paragraphs ( $z_1 = 8$ ); on the contrary, the longest paragraph appeared in the article only once.



**FIGURE 13**

Graphical visualization of the observations presented in Table 30

As can be seen from Figure 13, a positive value of the respective parameter  $b$  (Table 31) is responsible for the decrease and the convexity of the curve. It is apparent that the relationship between the paragraph and the sentence defined by the MAL is valid. The value of a goodness-of-fit of the mathematical model with the empirically gained observations exceeded 0.85.

**TABLE 31**

Parameter  $A$ , parameter  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 30<sup>25</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
6.5604	0.4636	0.8535

## DISCUSSION

Let us consider the earlier mentioned deviations, i.e.  $y_1 = 4.44$  and  $y_1 = 3.70$ . As in the previous subchapter, the average lengths of sentences related to individual paragraphs were calculated in order to reveal which of them caused this increase. In the case of the 3-sentence paragraphs, the value  $y_1 = 4.25$  (i.e. the average length of sentences of the 2-sentence paragraphs) was exceeded by 2 sentences. Both of them contained a semicolon. The same findings were observed in the case of the second deviation. Although the value  $y_1 = 3.25$  (i.e. the average length of sentences related to the 4-sentence paragraphs) was exceeded in all cases, two of them, which has the greatest difference between the own average length and the value 3.25, also embraced this punctuation mark.

In accordance with *General rules for punctuation* the semicolon represents a type of a comma. It is used to separate those clauses which make up a part of a compound or complex sentence and which are in a coordinate relationship. The semicolon is also inserted between clauses on the first level of a multiple sentence which are not in a coordinate relationship; cf. (GB/T 15834 – 2011, 2012, p. 6). The Chinese *Baidu Baike* encyclopaedia cites the semicolon as a punctuation mark which is on the border between the comma and the full stop, cf. (Fenhao, © 2013).

The disputable function of this punctuation mark could exert a strong influence on the deviations in the average lengths of sentences  $y_1$ . Therefore a sub-experiment considering the semicolon as a punctuation mark valid for the border

25 In the case of this language level it was not appropriate to omit the empirically gained observations with a low frequency due to their small number.

of sentences was conducted. The sentence represents not only a constituent of the immediately higher language unit, but also a construct of the immediately lower language unit, hence it was inevitable to implement the sub-experiments on both language levels. The sub-experiments are marked according to the respective language level (i.e. 1 and 2) and the respective sample text (i.e. A). The gained results are stated below.

**SUB-EXPERIMENTS 1A AND 2A**

Table 32 and 33 show the observations empirically obtained by quantification of the scientific article in which the semicolon was considered to be the border of sentences.

Table 32 is concerned with the data acquired from the language level paragraph – sentence.  $x_{1A}$  represents the lengths of paragraphs (measured in sentences),  $z_{1A}$  their frequencies and  $y_{1A}$  the average lengths of sentences (measured in parcelates).

Table 33 is concerned with the data acquired from the language level sentence – parcelate.  $x_{2A}$  represents the lengths of sentences (measured in parcelates),  $z_{2A}$  their frequencies and  $y_{2A}$  the average lengths of parcelates (measured in characters).

■ **TABLE 32**  
Sub-experiment L1A –  
Language level L1: paragraph  
(measured in sentences) –  
sentence (measured  
in the average number of its  
parcelates)

■ **TABLE 33**  
Sub-experiment L2A – Language  
level L2: sentence (measured  
in parcelates) – parcelate  
(measured in the average  
number of its characters)

$x_{1A}$	$z_{1A}$	$y_{1A}$	$x_{2A}$	$z_{2A}$	$y_{2A}$
2	8	4.2500	1	20	16.8500
3	1	3.6667	2	36	14.8472
4	5	3.4000	3	36	15.0741

$x_{1A}$	$z_{1A}$	$y_{1A}$	$x_{2A}$	$z_{2A}$	$y_{2A}$
5	2	3.5000	4	20	13.4125
6	7	2.9762	5	4	13.5500
7	2	2.4286	6	5	12.8667
11	1	2.0909	8	2	12.0000
			9	2	11.0556

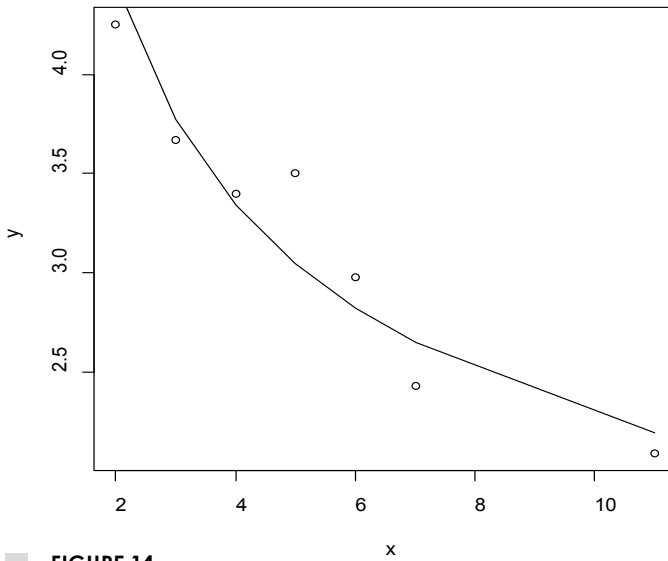
Let us begin with language level L1 (Table 32). The number of the various paragraphs' lengths remained equivalent (i.e. 7). Concerning the average lengths of sentences, the highest value of the interval, in which they appear, declined. The average length of the sentences of the 3-sentence paragraph decreased by the alternative segmentation approach from 4.44 to 3.67, hence the sentences related to the 2-sentence paragraph ( $y_{1A} = 4.25$ ) became the constituents with the highest average length and the deviation was eliminated. The lowest value of the average length remained equivalent ( $y_{1A} = 2.09$ ). The deviation occurred only in the case of average length of the sentences of the 5-sentence paragraph ( $y_{1A} = 3.50$ ) probably due to its low frequency. Apart from this sole instance, the values of the average lengths show a decreasing tendency with an increase in the constructs' lengths. It can be concluded that in the case of the alternative segmentation approach the relationship of inverse proportionality between paragraph and sentence also occurs. As regards frequency, the most frequent constructs remain the shortest paragraphs ( $z_{1A} = 8$ ) and the least frequent construct remains the longest paragraph ( $z_{1A} = 1$ ). As a consequence of the alternative segmentation, the frequency of other observations changed: the increase is observed within paragraphs composed of 4 (from 3 to 5) and 6 sentences (from 5 to 7); the decrease within paragraphs composed of 3 (from 3 to 1) and 5 sentences (from 4 to 2).

Next, let us turn to level L2 (cf. Table 33). The alternative segmentation approach eliminated sentences composed of 7 parcelates, hence the number of various sentences' lengths is reduced from 9 to 8. The highest average length

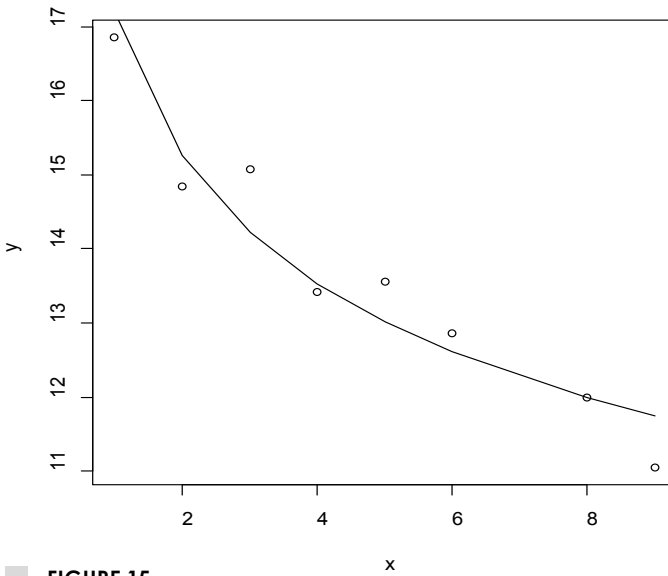
of parcelates declined from 16.94 to 16.85, the lowest remained equivalent, i.e. ( $y_{2A} = 11.06$ ). Contrary to the previous experiment, the decreasing tendency of the average lengths was interfered only with two observations. The first increase occurred within the average length of parcelates related to the 3-parcelate sentences ( $y_{2A} = 15.07$ ). As mentioned earlier, this deviation could be caused by punctuation (namely quotation marks, title marks). The second increase is observed within the average length of parcelate of the 5-parcelate sentence ( $y_{2A} = 13.55$ ) probably due to its low frequency. In spite of this, the relationship of the inverse proportionality between the sentence and the parcelate has emerged. Observations with the highest frequency remain sentences composed of 2 and 3 parcelates ( $z_{2A} = 36$ ). Sentences composed of 8 and 9 parcelates belong to the least frequent ( $z_{2A} = 2$ ) constructs on this level. Due to the alternative segmentation, the values of frequency changed almost within all observations: the increase is observed within sentences composed of 1 (from 18 to 20), 2 (from 27 to 36), 3 (from 30 to 36) and 4 parcelates (from 16 to 20); the decrease within sentences composed of 5 (from 6 to 4), 6 (from 6 to 5) and 8 parcelates (from 5 to 2).

Both segmentation approaches based on the different understandings of the semicolon proved that the relationships of the inverse proportionality between the construct and the constituent are valid on both language levels. It should be noted that the deviations from the downtrend of average constituents' lengths appeared to a lesser extent in the case of the alternative segmentation approach. The following graphical visualizations and value of a goodness-of-fit show to what extent the semicolon's function of separating sentences has influenced the results.

Positive values of both parameters  $b$  (cf. Table 34 and 35) are responsible for the decrease and the convexity of curves visualizing the mutual relationship between the variables defined by the MAL. Regarding the value of the goodness-of-fit, the coefficients of determination  $R^2$  have a higher value than in the original experiment. In the case of language level L1,  $R^2$  increased from 0.8535 to 0.9041 and in the case of language level L2  $R^2$  rose even more sharply – from 0.7862 to 0.9173.



**FIGURE 14**  
Graphical visualization of the observations presented in Table 32



**FIGURE 15**  
Graphical visualization of the observations presented in Table 33

**TABLE 34**

Parameter *A*, parameter *b* and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 32<sup>26</sup>

Parameter <i>A</i>	Parameter <i>b</i>	Coefficient of determination $R^2$
5.9661	0.4179	0.9041

**TABLE 35**

Parameter *A*, parameter *b* and coefficient of determination  $R^2$  for the mathematical model related to the observations presented in Table 33<sup>27</sup>

Parameter <i>A</i>	Parameter <i>b</i>	Coefficient of determination $R^2$
17.2152	0.1737	0.9173

Replacement of the semicolon by the full stop does not change the condition in determining borders of the pancelate. Hence this language unit remains valid.

It is evident from the results obtained by segmentation based on the semicolon's function of separating sentences that this punctuation mark inclines towards the full stop. The question arises as to whether or not the semicolon's function is disputable and as to whether or not its disputable function is influenced by the punctuation whose usage has not a long tradition in Chinese texts.

The punctuation was gradually integrated into the written Chinese language in the 20<sup>th</sup> century. Until then punctuation marks were used in the Chinese texts only sporadically and they often appeared in a form of specific symbols whose functions resembled the functions of Western punctuation (for further

26 In the case of this language level it was not appropriate to omit empirically gained observations with a low frequency due to their small number.

27 In the case of this language level it was not appropriate to omit empirically gained observations with a low frequency due to their small number.

information cf. *Biaodian fuhao*, © 2013). In the second decade of the 20<sup>th</sup> century Chinese scholars invested a great effort into the implementation of the punctuation based on the Western punctuation marks into Chinese texts. The first work using the punctuation was published in 1919 and its title is *Outline of the History of Chinese philosophy* (Zhōngguó zhéxuéshǐ dàgāng, 中国哲学史大纲) written by Chinese philosopher and writer Hu Shi (Hú Shì, 胡适), (*Biaodian fuhao*, © 2013). Finally, in 1951 the Chinese government published a document summarizing the rules for its usage. Owing to a few punctuation modifications emerging over the course of the following four decades, the *General rules for punctuation* (Biāodiǎn fúhào yòngfǎ, 标点符号用法) from 1951 were updated and published in 1990 again, for further information cf. (*Biaodian fuhao yongfa*, © 2013).

In view of this fact it is necessary to conduct subsequent experiments focused on the influence of the punctuation upon the mutual relationships between the language units in the context of its development.

Subsequent quantitative analyses, applied to sample texts whose length is greater than the length determined for the purpose of this experiment, should also be aimed at the paragraph.

### 3.6.5 Conclusion

Let us summarize the results obtained by the quantitative analysis of the scientific text by means of the MAL. As regards the lowest language level, the existence of the mutual relationship between the character and the component is valid. Not only the MAL represents an adequate and well-fitting mathematical model, but has also an extremely goodness-of-fit with the empirically gained observations ( $R^2 = 0.9061$ ). Our assumption is that on this level the sample text adheres to the assumptions of the MAL in an almost perfect fashion probably due to the identical size of the graphic field.

On the contrary, the next language level represented by the parcellate and the character does not show an agreement with the mathematical model of the MAL because the obtained results contradict the assumption of this law. There was a discussion about several causes, such as the different characteristic



of the units, or, more precisely, the variable length of the parcelate and the invariable length of the character; absence of a universally valid component definition; the insufficient amount of language units and levels, or, more precisely, an absent language unit – the word; the high frequency of constructs' lengths on the lowest language level  $z_4$ , i.e. the high frequency of characters composed of 2 and 3 components and last but not least, the Chinese script simplification.

In the instance of the sentence and the parcelate constituting the higher language level, the mutual relationship defined by the MAL is proved by the wide goodness-of-fit whose value sums up to 0.7862.

The MAL also shows itself as an adequate and well-fitting mathematical model in the case of the highest and last tested language level – the paragraph and the sentence. Not only their mutual relationship is validated, but also the goodness-of-fit has a very high value ( $R^2 = 0.8535$ ).

It can be concluded that the existence of the mutual relationships between units determined by the graphical principle was proved on three language levels: character – component, sentence – parcelate and paragraph – sentence. The sole language level whose empirically gained observations did not show the agreement with the mathematical model of the MAL was parcelate – character. The first assumption of the MAL as the adequate and well-fitting model was confirmed in the case of two levels: character – component and sentence – parcelate. The second assumption about the MAL's invalidity was confirmed only partially, i.e. in the instance of the language level parcelate – character. Against expectation, the analysis of the last language level paragraph – sentence revealed that the relationship between these units exists and thus our assumption was contradicted.

Let us turn to the hypothesis which predicts that if language units of the contemporary written Chinese are determined on the basis of the graphical principle, the mutual relationships between them exist on respective language levels and their validity is verified by means of the MAL. With regard to the component as a hypothetical language unit, it can be stated on the basis of the gained results from the scientific article that the component was verified as a valid language unit; hence the hypothesis was proved in this case. In the instance

of another hypothetical language unit – the parcelate, on the level parcelate – character its existence was not supported; on contrary, on the higher language level sentence – parcelate it was verified. According to the results it seems that the parcelate could also be considered a valid language unit because on the level parcelate – character there are several factors interfering the relationship between the units on this level.

This chapter also deals with two sub-experiments. The first sub-experiment, related to the lowest language level, i.e. character – component, was performed on the identical sample text transformed into the traditional set of Chinese characters. The sub-experiment verified both the existence of the component and existence of the relationship between the character and the component. The mathematical model of the MAL showed an extremely wide goodness-of-fit, its value is the highest, i.e.  $R^2 = 0.9597$ , within the whole experiment. The second sub-experiment tested the identical sample text regarding the semicolon as the border of sentences. It examined two language levels, namely sentence – parcelate and paragraph – sentence, and its results did not contradict the assumption of the MAL. On the contrary, it revealed that the alternative segmentation approaches showed itself as a more suitable method; the goodness-of-fit with empirically obtained observation was higher than in the case of the original segmentation approach. The value exceeded 0.9 within both language levels. The existence of the parcelate as a valid language unit was also supported.

## **4. The application of the Menzerath–Altmann law to the blog article**

This part of the work will deal with a quantitative analysis of the artistic style, which is here represented by a blog article. To verify the hypotheses, the MAL will be applied.

This part of the publication is divided into six sections. The first five subchapters present individual steps of the experiment. Firstly, a suitable sample text will be chosen according to the given criteria and there will also be given reasons for that choice. Consequently, the language units and the language levels, which arose by an arrangement of language units into mutual relations, will be briefly introduced in order to adhere to a sequence of the experiment. On every language level, there will be inserted tables containing the necessary values. These values will be used for calculating parameters and coefficients of determination and also for the construction of graphs. These steps will be carried out by means of the statistical R program. The last sixth subchapter will interpret the acquired data and it represents a gist of this chapter. The obtained results will be summarized in the conclusion.

### **4.1 DETERMINATION OF THE CRITERIA FOR THE CHOICE OF THE SAMPLE TEXT AND THEIR EXPLANATION**

In subchapter 2.1, it was said that over the course of choosing the sample texts it is necessary to comply with the criteria established for the selection of appropriate samples. Requirements on which the major emphasis is placed and from which other criteria arose are focused on the synchronic point of view in exploring Chinese language. With regard to the synchronous aspect of the research the first applied criterion is to select a sample text which is written in the standard form of contemporary written Chinese. Next criterion is to choose a sample

text which is written in simplified characters (the youngest variant of Chinese characters, i.e. *kǎi shū*, 楷书). It follows that the author has to originate from mainland China and has to use a simplified Chinese character set. The third criterion – the choice of stylistic styles – is based on the comparison of different stylistic styles, hence the artistic style which supplements the already tested stylistic styles was chosen. Following the fifth criterion, the analysed sample text has to reflect the contemporary language, therefore, the sample text should be the most current and should not be older than eleven years, i.e. should be published after 2002. The sample text has to have a distinct beginning and end and should not be interrupted by images, charts or tables. Based on the already performed experiments the sample text length has to be in the range of 2,500–3,500 Chinese characters. According to the last criterion, the author of the sample text has to be reputable and renowned.

## 4.2 THE CHOICE OF THE SAMPLE TEXT

In the course of selecting the sample, it was necessary to choose the text which is written in a certain type of the artistic style which is read by wide circles. Based on this condition, the blog article was chosen as an appropriate sample text for this experiment.

The first step was to choose a contemporary Chinese blogger. Several conditions were set before choosing the most appropriate author. Accordingly, this was not a completely random choice. The most appropriate author uses vocabulary influenced by foreign languages as little as possible. That is why it is necessary to take into account whether the author studied abroad or stayed there for a long time. Another condition was that his or her profession is not connected to journalism (if yes this profession could influence his style of writing). Last but not least, the aim was to find an article that belongs to the artistic style.

A list of Chinese bloggers on Wikipedia served as a source for a general overview (Category: Chinese bloggers, 2013). Then basic information about every mentioned blogger was found out on the Chinese *Baidu baike* encyclopedia,

and if these authors met the established criteria, their blogs were examined and the visit rates of individual blogs were found out.

Bloggers who are listed in the following table were investigated (c.f. Table 36):

**TABLE 36**  
Contemporary Chinese bloggers  
Source: created by author

Names of bloggers	Chinese characters	Blog visit rate (to the date November 11, 2013 (18:40))	Link
Han Han	韩寒	595,656,098	<a href="http://blog.sina.com.cn/s/blog_4701280b0102e7er.html">http://blog.sina.com.cn/s/blog_4701280b0102e7er.html</a>
Kong Qingdong	孔庆东	81,094,125	<a href="http://blog.sina.com.cn/u/1198367585">http://blog.sina.com.cn/u/1198367585</a>
Li Chengpeng	李承鹏	7,311,776	<a href="http://www.weibo.com/lichengpeng">http://www.weibo.com/lichengpeng</a>
Murong Xuecun	慕容雪村	3,490,721	<a href="http://blog.sina.com.cn/hawking">http://blog.sina.com.cn/hawking</a>
Muzi Mei	木子美	2,487,952	<a href="http://muzimeiriji.blog.sohu.com">http://muzimeiriji.blog.sohu.com</a>
Ran Yunfei	冉云飞	2,891,015	<a href="http://tufeilaoran.blog.163.com/">http://tufeilaoran.blog.163.com/</a>
Rao Xueman	饶雪漫	1,758,135	<a href="http://weibo.com/raoxueman">http://weibo.com/raoxueman</a>
Liu Mangyan	流氓燕	3,962,433	<a href="http://blog.tianya.cn/blogger/blog_main.asp?BlogID=19329">http://blog.tianya.cn/blogger/blog_main.asp?BlogID=19329</a>
Sima Nan	司马南	711,573	<a href="http://weibo.com/simanan">http://weibo.com/simanan</a>
Vivibear / Zhang Weiwei	张薇薇	29,066,823	<a href="http://blog.sina.cn/vikingbear333">http://blog.sina.cn/vikingbear333</a>

Names of bloggers	Chinese characters	Blog visit rate (to the date November 11, 2013 (18:40))	Link
Wang Keqin	王克勤	7,544,517	<a href="http://wangkeqin.blog.sohu.com/">http://wangkeqin.blog.sohu.com/</a>
Xu Jinglei	徐静蕾	312,455,983	<a href="http://blog.sina.com.cn/xujinglei">http://blog.sina.com.cn/xujinglei</a>
Zeng Jinyan	曾金燕	–	<a href="http://zengjinyan.wordpress.com/">http://zengjinyan.wordpress.com/</a>
Zhou Shuguang	周曙光	–	<a href="https://www.zuola.com/">https://www.zuola.com/</a> ; <a href="http://www.zuola.com/weblog/">http://www.zuola.com/weblog/</a>

From these authors, the most appropriate seemed to be blogger Han Han (Hán Hán, 韩寒), because he meets all the criteria and on top of that the most readers have visited his blog (more information about the author see below). One of the articles posted on his blog was chosen as a sample text for this experiment.

Regarding the choice of the article, it was necessary to follow criteria concerning contemporarity and length of the sample text. Hence, when selecting the appropriate article, it was proceeded from the most recently added articles which simultaneously satisfied the determined length (i.e. 2,500–3,500 characters). Besides these criteria, the number of views of the individual articles was taken into account and the article with a high number of views was chosen.

With regard to the given criteria the article *Life as I understand it* (Wǒ suǒ lǐjiě de shēnghuó, 我所理解的生活), (Han, 2012) was chosen since it meets all the criteria. The article was posted on 20<sup>th</sup> June 2012, it is written in simplified characters and contains 2,641 characters. The beginning and the end of the text are clearly distinguished and there are no embedded pictures or graphs, therefore, this article can be considered to be a coherent text. According to the information on the author's blog, 963,831 readers have already seen

this text (to 16<sup>th</sup> November 2013). The sample text is inserted into the attachment under the name Sample B (cf. Appendix 2).

### **HAN HAN (HÁN HÁN, 韩寒)**

Han Han is a popular Chinese writer, blogger, professional rally driver, musician, and also founder and chief editor of the *Party literary magazine* (*Dúchàng tuán*, 独唱团), which was first published in 2010, but it lasted only the first issue (the second issue was long overdue but finally did not occur) (Watts, 2010).

Han Han was born on 23 September 1982 in Shanghai. He began publishing his first works when he was attending junior middle school (Han Han, © 2013). In 2000, he broke through the literary scene with his bestseller *Triple door* (*Sānchóng mén*, 三重门), (Elegant, 2009). In 2005, he founded his blog, which became the most popular blog in China in 2009 (Elegant, 2010). “His caustic commentary on current events gives voice to popular outrage at official corruption and abuse of power, while avoiding direct attacks on the government that might result in censorship of his blog” (Abrahamsen, 2012). Han Han ranks among the young generation of writers who are known as the post-80s generation. He has published many prosaic collections and novels. His works usually arouse a great interest and vigorous discussion. In May 2010, he was ranked on the list of the one hundred most influential people in the world by the Time magazine. His blog has registered over 595 million hits up to now so it is the best-read blog not only in China but also in the world (Pilling, 2012).

### **LIFE AS I UNDERSTAND IT (Wǒ SUǒ LǐJiě DE SHēNGHUÓ, 我所理解的生活)**

The article *Life as I understand it* is the subject of this experiment, therefore, its structure and other features should be introduced in the first place because they can influence the results of the experiment. As mentioned above, this article is a coherent text and except the title, the author name and web link there are no interruptions such as pictures or graphs.

The article consists of 2,365 characters (these data also include punctuation marks). It is organized into 12 paragraphs which are clearly separated in the text.

Each paragraph begins with an indentation at the edge of the paper and the paragraphs are separated by an inserted blank line.

Due to the fact that the author comes from mainland China, the article is written in simplified characters. At present time, English words often occur in blogs and forums written in Chinese and authors of these words use Latin to write them down. However, in this article, Han Han does not use any foreign word and he uses Chinese character even in case when a variant form written in Latin letters is possible to use. E.g. vulgar word “pretentious bastard” (zhuāngbī, 装逼) is written down by two Chinese characters instead of a variant expression 装B (the fourth paragraph) as well as another vulgar word “fucking awesome” (niúbī, 牛逼) which can be also written as 牛B (the tenth paragraph).

As regards numbers, the author uses Chinese characters in the most cases and he uses Arabic numerals only rarely. Number 30 written in Arabic numerals represents only one exception (in the tenth paragraph). It is used in the phrase “thirty years” (sānshí nián, 30 年). In this case number 30 is formatted in the second way in which the digits adjoin together (cf. subchapter 2.3). Therefore, number 30 is regarded as one character.

The text operates only with four kinds of Chinese punctuation marks, namely, a comma, a full stop, a question mark and a colon (cf. Table 37). From the article it is apparent that the author probably tries to choose a simpler structure of the text and therefore uses only basic punctuation marks. He even does not use quotation marks in the parts where he describes a conversation with his friend. The direct speech is not marked in any way.

**TABLE 37**

Punctuation marks used in the blog article

Source: Chinese terminology from (Biaodian fuhao, © 2013); created by author

English terminology	Punctuation marks	Pinyin	Characters
Comma	,	dòuhào	逗号



English terminology	Punctuation marks	Pinyin	Characters
Full stop	.	jùhào	句号
Question mark	?	wèn hào	问号
Colon	:	mào hào	冒号

### 4.3 DETERMINATION AND DEFINITION OF THE LANGUAGE UNITS AND LANGUAGE LEVELS

A precise determination of language units is a fundamental step for any similar experiment. This experiment is built on a previous experiment in which the language units have already been precisely determined. It is important to follow the same definitions because of the possibility to compare the results of both experiments. The precise determination of the individual language units is specified in subchapter 2.3. The language units for these experiments are as follow:

*stroke – component<sup>28</sup> – character – parcelate<sup>29</sup> – sentence – paragraph.*

By linking the above mentioned language units together the four language levels, on which the validity of the MAL is examined, arise. Let  $i$  be a natural number representing four language levels, where  $i = 1$  represents the language level paragraph – sentence,  $i = 2$  language level sentence – parcelate,  $i = 3$  language level parcelele – character and  $i = 4$  language level character – component, thus the value of  $i$  can be  $i = 1, 2, 3, 4$ .

28 A hypothetical language unit.

29 A hypothetical language unit.

**TABLE 38**Language levels  $L_i$ ,  $x_i$  construct,  $y_i$  constituent, ( $i = 1, 2, 3, 4$ )

Source: created by authors

Language level L4		Language level L2	
$x_4$	character measured in components	$x_2$	sentence measured in parcelates
$y_4$	component measured in the average length of its strokes	$y_2$	parcelate measured in the average length of its characters
Language level L3		Language level L1	
$x_3$	parcelate measured in characters	$x_1$	paragraph measured in sentences
$y_3$	character measured in the average length of its components	$y_1$	sentence measured in the average length of its parcelates

#### 4.4 SEGMENTATION AND QUANTIFICATION OF THE SAMPLE TEXT

The subject of quantification of this part is the blog article *Life as I understand it*. Quantification will be carried out throughout the whole text with the exception of the title and formalities such as date of the issue, the author's name and the link. In comparison with the scientific text whose structure is more complicated, the blog article is not further subdivided.

On the basis of determined language units the selected article was segmented in the Microsoft Excel program and then the same program was used for quantifying the text. Due to this step, tables with the necessary data for  $x_i$  (length of construct)  $z_i$  (frequency)  $a_i$  (length of constituent) were obtained. Segmentation and quantification was carried out on all of the mentioned language levels L4–L1.

## 4.5 TESTING THE MODEL RELIABILITY BY MEANS OF STATISTICAL METHODS

The data from tables obtained by quantification of the article were used as an input data for the statistical R program. On account of these data it was possible to calculate parameters  $A$ ,  $b$  and coefficients of determination  $R^2$  and to construct graphs. Parameters  $A$ ,  $b$  and coefficients of determination  $R^2$  were calculated by the method of a simple linear regression in the R program (cf. subchapter 2.5, footnote 12), which was applied to a truncated mathematical formula of the MAL.

Observations which are insignificant due to their low frequency in comparison with other observations (so-called extremes) are statistically treated on every language level. Values obtained by removing them are always listed in the footnotes which relate to the table showing parameter  $A$ , parameter  $b$  and the coefficient of determination  $R^2$ .

## 4.6 INTERPRETATION OF THE THE ACQUIRED DATA

The following section of this work is focused on the interpretation of the acquired data, tables and graphic visualizations. This is a very important part of the experiment, therefore it is the largest subchapter of Chapter 4. The interpretation is proceeded from the lowest language level L4 (character – component) to the highest language level L1 (paragraph – sentence). The results and conclusion obtained from all four language levels are summarized at the end of this subchapter.

### 4.6.1 Language level L4

#### CHARACTER – COMPONENT

Table 39 presents observations empirically obtained by quantification of the blog article:  $x_4$  represents the lengths of characters (measured in components),  $z_4$  their frequencies and  $y_4$  the average lengths of components (measured in strokes). The grey background of the cells is used to highlight the omitted observation with a low frequency ( $z_4 = 1$ ). The values obtained by

its omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 30.

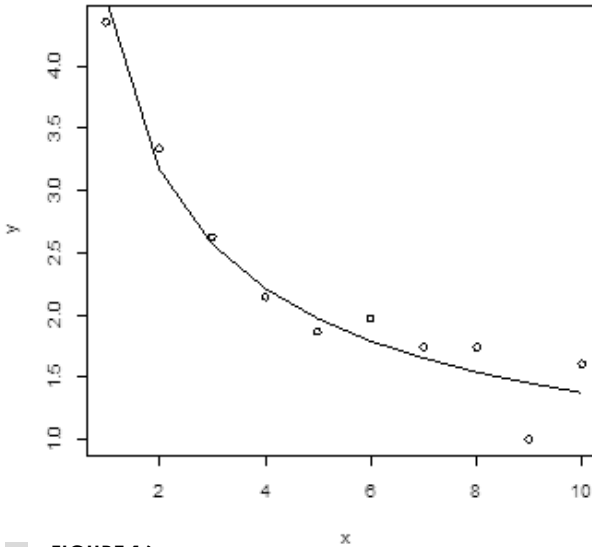
**TABLE 39**

Language level L4: character (measured in components) – component (measured in the average number of its strokes)

$x_4$	$z_4$	$y_4$
1	522	4.3544
2	670	3.3366
3	479	2.6200
4	323	2.1416
5	224	1.8652
6	58	1.9684
7	69	1.7329
8	14	1.7411
9	1	1.0000
10	5	1.6000

It is evident from Table 34 that there occur characters compounded of 1 to 10 components in the sample text. The average lengths of components appear around the values within the interval of  $\langle 1,00; 4,35 \rangle$ . It is obvious from the data that the average lengths of constituents decrease with the increasing lengths of constructs. Thus the assumption of the MAL that construct is inversely dependent on its constituents is fulfilled.

The most frequent characters are composed of two or three components. Characters consisted of four components are also relatively highly represented. As regards characters composed of five or more components, their frequencies have a decreasing tendency. Observation  $x_4 = 9$  showed the lowest frequency and it is highlighted by the grey background of the cells in Table 39.



**FIGURE 16**  
Graphic visualization of the observations presented in Table 39

It is apparent from the graphic visualization that the tendency formulated by the MAL has emerged and a mathematical model reveals an extremely wide goodness-of-fit with the empirically gained observations; the value of its goodness-of-fit exceeded 0.86. The assumption of the MAL that parameter  $b$  is a positive real number is adhered to. Its positive value is expressed by the decreasing and convex curve in the graphic visualization. The respective parameters  $A$ ,  $b$  and coefficient of determination  $R^2$  are presented in Table 40.

**TABLE 40**  
Parameter  $A$ , parameter  $b$  and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 39<sup>30</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
4.5461	0.5208	0.8698

30 After omitting the observation with the lowest frequency ( $z_4 = 1$ ) parameter  $A = 4.3123$ ; parameter  $b = 0.4554$ ; coefficient of determination  $R^2 = 0.9719$ .

Table 39 and the graphic visualization indicate that the only observation which deviates from the trend defined by the MAL is the character composed of 9 components, namely the character 洲 (zhōu). This deviation is probably caused by the low frequency of its occurrence – this character appears only once in the article and another character consisted of 9 components which should influence the average length of 9-component characters is not arisen in the text. Thus, this observation can be omitted. After omitting this sole observation, the goodness-of-fit with the empirically gained observations would increase to 0.97. In this case, the value of parameter  $A$  would be 4.3123 and the value of parameter  $b$  would be 0.4554.

From the obtained data it is obvious that the assumption of the MAL emerges on language level L4. On the basis of the results obtained from the blog article, it seems that the component – a hypothetical language unit determined purposefully in accordance with the graphical principle – exists and it has proved itself as a valid language unit on this level.

## DISCUSSION

The question arises what caused such a wide agreement with the mathematical model of the MAL and how it is possible that the structure of characters strictly complies with the tendency defined by the MAL. One possible explanation is that a graphic field in which the characters are written plays a major role in the structure of characters. As noted above (cf. subchapter 2.3), the graphic field always has the same size independent of the number of strokes or components. Thus, the graphic field affects the organization of strokes and components in the character.

If the character consists of a lower number of components, the components might consist of more strokes because they can occupy more space and their readability is not affected. One character composed of one component (i.e. the character is identical with the component) can be constituted by a complicated component because it has plenty of space and the component is recognizable. In the case where characters have a small number of components, the graphic field minimally limits the construction of the character.

Two tendencies which are connected with the number of components can be observed. The first case includes characters constituting of more components. If components are supposed to be separated from each other, they have to be composed of a smaller number of strokes due to a limited space of the graphic field and a requirement of readability. The more components the character has, the less space every components has, thereby, the number of strokes of which the component consists of decreases. Concerning the second tendency, characters consisted of fewer components tend to be more complex because they can be written into a larger space and at the same time they are easily recognizable. Consequently, these components can be composed of a higher number of strokes.

For easier reference, the examples of characters composed of higher and lower number of components are listed in Table 41:

**TABLE 41**

An example of characters composed of one, nine and ten components

Source: created by author

Character	夏	着	洲	憾
Number of components	1	1	9	10
Number of strokes	10	11	9	16
Average number of strokes	10	11	1	1.6

Tables 39 and 41 show that characters composed of one component are in most cases composed of a greater number of strokes. The average number of strokes of 1-component's characters is 4.3544. These components can appear within other characters, nevertheless, their occurrence probably decreases with the increasing number of components in the character, because it might be difficult to recognise them in the character composed of more components.

As regards characters consisting of a high number of components, the average number of strokes per one component is significantly lower in this case (cf. Table 41). Table 39 indicates that characters consisted of five or more components

have less than two strokes per one component on average. These components might be present as parts in other characters consisted of more components without major restrictions since they can be easily distinguished.

Following the data from the previous experiment, it can be concluded that these rules should be adhered to even more within traditional characters. This set of characters was simplified to the current form within the reform which took place in 1956 and 1964. The reform reduced the number of strokes within 2,236 characters, and thus, it simplified their structure. In comparison with simplified characters, traditional characters should show a wider agreement with the MAL because their original form is more complex and pressure of the graphic field should be even more intense.

Co-author Tereza Motalová decided to verify to what extent the above mentioned rules are reflected in texts written in traditional characters. The original sample text (Sample A, cf. Appendix 1) was converted into a traditional form and subsequently quantified.

Based on the data obtained from sub-experiment 4A (cf. subchapter 3.6.1), we found out that according to our expectation the value of agreement with the MAL increased. This experiment confirmed that the pressure of the graphic field within character with more complex structures is stronger.

The results obtained on this language level confirmed that a constant size of graphic field plays an important role in the case of forming the relationships between characters and components. These wide agreements with the mathematical model of the MAL can be caused by the fact that the components in more complex structures have a smaller space, therefore, components have to adhere to each other more tightly. If they touch each other, they create a new component. Thus, this component is composed of more strokes.

### 4.6.2 Language level L3

#### PARCELATE – CHARACTER

Table 42 states observations empirically obtained by quantification of the blog article:  $x_3$  represents the lengths of parcelates (measured in characters),  $z_3$  their frequencies and  $y_3$  the average lengths of characters (measured in components).

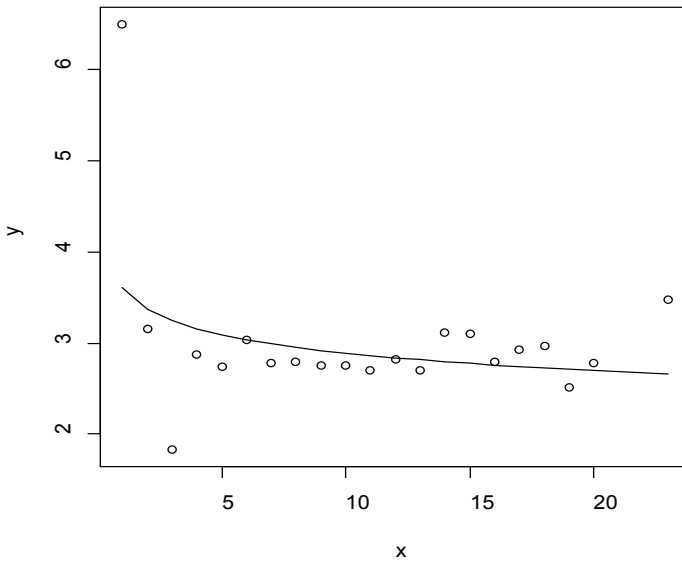


The grey background of the cells is used to highlight the omitted observations with a low frequency ( $z_3 \leq 2$ ). The values obtained by their omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 31.

**TABLE 42**  
Language level L3: parcelate (measured in characters) – character (measured in the average number of its components)

$x_3$	$z_3$	$y_3$
1	2	6.5000
2	10	3.1500
3	4	1.8333
4	35	2.8714
5	20	2.7400
6	30	3.0333
7	38	2.7744
8	24	2.7917
9	23	2.7536
10	14	2.7571
11	12	2.7045
12	14	2.8155
13	10	2.6920
14	8	3.1161
15	7	3.0952
16	6	2.7917
17	5	2.9294
18	2	2.9722
19	4	2.5132
20	6	2.7833
23	2	3.4783

It is evident from Table 42 that the examined article contains 21 parcelates with a different length. The shortest parcelate is composed of one character and the longest parcelate consists of 23 characters. The average lengths of characters on this language level obviously do not correspond to the presumption of the MAL that claims that the average lengths of characters decrease with increasing lengths of parcelates. On the contrary, the average lengths oscillate around the values within the interval of  $\langle 1,83; 6,5 \rangle$ . Concerning the parcelate's frequency, the parcelates composed of 1, 18 and 23 characters ( $x_3 = 1; 18; 23$ ) occur with the lowest frequency. These empirically obtained observations are highlighted by a grey background in the table.



**FIGURE 17**

Graphic visualization of the observations presented in Table 42

It is obvious from the graphic visualization that agreement with the mathematical model of the MAL is minimal (the value of agreement is 0.1316), but the downward trend formulated by the MAL was indicated. The low value of parameter  $b$  indicates that the curve has a constant tendency. Cf. Table 43.

The constant tendency is mainly brought about by the observations  $x_3 = 1$  and  $x_3 = 2$ . However, these observations occur the least, therefore, these deviations can be omitted. In the case  $x_3 = 1$ , i.e. two parcelates consisted of one character, the deviation is specific because of the fact that these parcelates are composed only of one character 操 (cāo) and 滾 (gǔn) and these characters are comprised of five and eight components. In comparison with longer parcelates, whose characters are composed of two or three components on average, these two characters are consisted of a higher number of components, and therefore, they correspond to the assumption of the MAL: the fewer characters the parcelate has, the more components the character has.

However, there are only two observations, thus we cannot claim that parcelates consisting of one character tend to have more complicated characters in general. For this reason, it is necessary to implement more experiments.

**TABLE 43**

Parameter *A*, parameter *b* and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 42<sup>31</sup>

Parameter <i>A</i>	Parameter <i>b</i>	Coefficient of determination $R^2$
3.6055	0.0965	0.1316

On the basis of the results obtained from the blog article, it seems that the existence of the parcelate – a hypothetical language unit determined purposefully in accordance with the graphical and syntactic principle – is not verified on this level. Nonetheless, its existence is not excluded because of the several factors which adversely affect the relationship between the units on this level (cf. discussion).

31 After omitting the observations with the lowest frequency ( $x_3 \leq 2$ ): parameter  $A = 2.5571$ ; parameter  $b = -0.0364$ ; coefficient of determination  $R^2 = 0.0411$ .

## DISCUSSION

The low agreement with the mathematical model and dispersion of observations could be caused by several possible reasons on this language level. One of them could be the fact that the parcelate and character are very different language units. The parcelate is a unit with a variable length and the character is a unit with an invariable length. The length of a parcelate may be formed according to the author's need, while the length of a character cannot be changed at will. A character has its fixed structure and cannot be modified. Hence, the different characteristic of the units is a possible reason why the relationship formulated by the MAL was not verified on level L3.

On the other hand, in some cases authors have an option to choose another character with a similar meaning which is composed of a higher or lower number of components. In this case, there is a possibility to change the average length of the parcelate. However, it involves only a limited number of characters. The most characters do not have exact synonyms, therefore they cannot be replaced, thus the average lengths of parcelates cannot be changed in many cases.

The degree of agreement with the mathematical model of the MAL could also be affected by various approaches in determining components. Since the component does not have an exact definition, it can be determined in several ways. Various determinations of components and subsequent characters' segmentation remain an object of further research which will examine these approaches by means of the MAL.

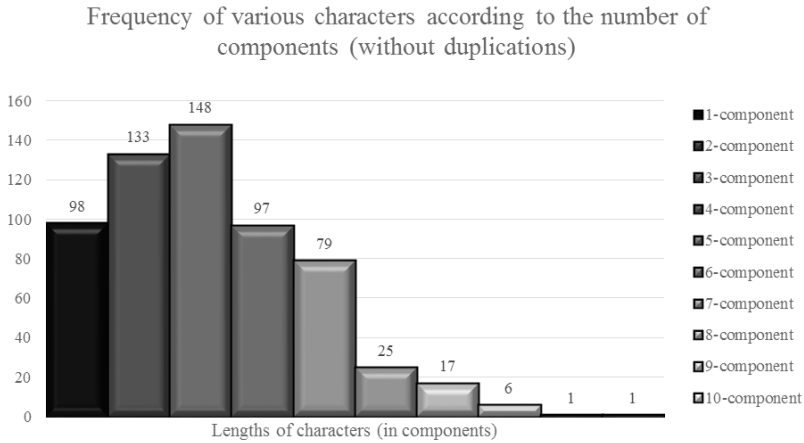
Another possible reason for the low agreement might be structural changes in characters which were caused by the centrally planned reform implemented in 1956 and 1964. The aim of the reform was to simplify Chinese characters in order to make them more accessible to the general public and to increase the literacy of Chinese people. As well as the number of strokes within 2,236 characters, the average lengths of components in characters were also reduced (cf. subchapter 3.6.1). The reform also probably brought about that the average length of majority of characters (measured in the number of components) was reduced and it seems that the lengths of the most

frequent characters settled on two or three components. The numbers of components in connection with the frequency of respective characters' lengths have probably a significant influence on this level, therefore this phenomenon will be examined below.

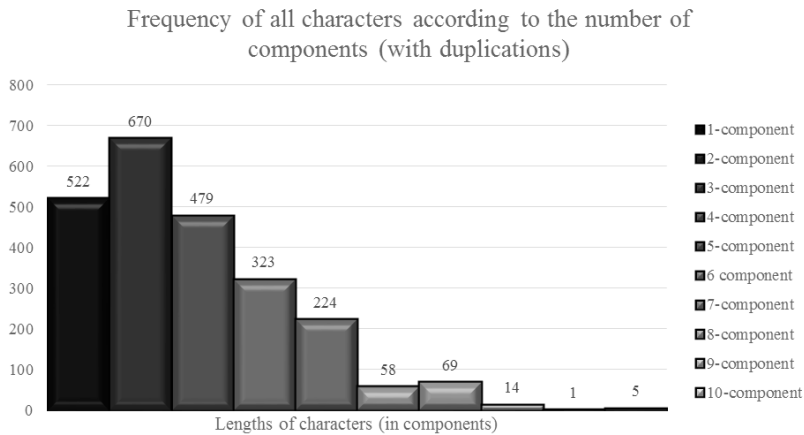
On the language level character – component, it was found out that the most frequent characters are composed of two or three components. Their high frequency might be another possible reason why the validity of the MAL was not supported. The exact data representing the frequency of characters are shown in Table 44 and they are graphically illustrated in Figures 18 and 19:

**TABLE 44**  
Outline of character frequency related to the observations presented in Table 39

Length of characters (in components)	Total amount of various characters (without duplication)		Total amount of all characters (with duplication)	
	Number of characters	%	Number of characters	%
1-component	98	16.1983	522	22.0719
2-component	133	21.9835	670	28.3298
3-component	148	24.4628	479	20.2537
4-component	97	16.0331	323	13.6575
5-component	79	13.0579	224	9.4715
6-component	25	4.1322	58	2.4524
7-component	17	2.8099	69	2.9175
8-component	6	0.9917	14	0.592
9-component	1	0.1653	1	0.0423
10-component	1	0.1653	5	0.2114
<b>Total</b>	<b>605</b>		<b>2,365</b>	



**FIGURE 18**  
 Graphic representation of the frequency of various characters according to the number of components, related to the data presented in Table 44



**FIGURE 19**  
 Graphic representation of frequency of all characters according to the number of components, related to the data presented in Table 44

Based on the data shown in Table 44 and the graphic visualization, it can be concluded that characters composed of two or three components represent over 46 % of all the different characters (i.e. 281 out of 605 different characters). For all the characters used in the text (i.e. also with duplications) it is more than 48 % (1,149 of all 2,365 characters). Other highly frequent characters are composed of one or four components.

As the 2-component and the 3-component characters constitute almost a half of all characters, thus they influence the construction of a parcelate from all the characters most significantly. It is possible that due to the small variability of characters (as regards their number of components) only a low agreement occurs. If parcelates largely consist of characters with average lengths of one to five components, then it does not depend on how long the parcelate is, the average length of characters remains constant, i.e. about two or three components. The low frequency of characters composed of more components (six or more) causes that the reciprocal proportion between the parcelate (construct) and the character (constituent) is interfered on this level.

Characters composed of two or three components occur with a high frequency not only in the case of this article, but also in the previous experiment. The following analysis will try to examine whether these characters belong to the frequent characters in the frequency list created by the Wenlin program<sup>32</sup>.

### **FREQUENCY ANALYSIS**

The rank of all the characters used in this article within the frequency list created by the Wenlin program is examined. The lower the value of the rank is, the higher the frequency of the characters is. Based on the value of ranks, three frequency groups are created. The first frequency group is represented by characters with a rank of 1–1,000; the second frequency group by characters with a rank of 1,001–2,000 and the third group by characters with a rank of 2,001–3,000.

---

32 Source: the frequency list created by the 文林 Wenlin Software 4.0.2

The crucial finding will be what proportion of characters constituted of two and three components is within each frequency group.

Table 45 shows the representation of all the characters which occur in the blog article after the removal of the duplicate characters, i.e. 605 characters in total. Characters are categorized according to their value of rank into three frequency groups. Characters without a rank form a special group which is outside these frequency groups.

**TABLE 45**

Representation of all the characters after the removal of the duplications within the frequency groups  
Source: The Frequency list created by the Wenlin Software for Learning Chinese Version 4.0.2; created by author

	<b>Total amount of various characters (without duplication)</b>	
<b>Frequency group</b>	<b>Number of characters</b>	<b>%</b>
The first frequency group (1–1000)	472	78.02
The second frequency group (1001–2000)	89	14.71
The third frequency group (2001–3000)	32	5.29
Characters outside frequency groups	12	1.98
<b>Total</b>	<b>605</b>	<b>100</b>

The majority of characters, i.e. 78 % of the characters (without duplications), which occur in the blog article belong to the first frequency group, i.e. between 1,000 of the most frequent characters.

On the basis of the obtained results, it is also necessary to find out how big a part is made up by the characters consisting of 2 and 3 components within each frequency group.



**TABLE 46**  
 Representation of the various characters according to the number of components within frequency groups and outside them  
 Source: The Frequency list created by the Wenlin Software for Learning Chinese Version 4.0.2; created by authors

Number of componentets	1 <sup>st</sup> frequency group		2 <sup>nd</sup> frequency group		3 <sup>rd</sup> frequency group		Characters outside the frequency groups	
	Number of characters	%	Number of characters	%	Number of characters	%	Number of characters	%
1-component	92	19.49	6	6.74	-	-	-	-
2-component	116	24.58	14	15.73	2	6.25	1	8.33
3-component	105	22.25	30	33.71	10	31.25	3	25.00
4-component	80	16.95	10	11.24	4	12.50	3	25.00
5-component	52	11.02	17	19.10	6	18.75	4	33.33
6-component	9	1.91	10	11.24	5	15.63	1	8.33
7-component	14	2.97	-	-	3	9.38	-	-
8-component	3	0.64	2	2.25	1	3.13	-	-
9-component	1	0.21	-	-	-	-	-	-
10-component	-	-	-	-	1	3.13	-	-

The major representation within the first frequency group (characters with the rank from 1 to 1,000) consists of characters composed of two components, this means 24.58 % (i.e. 116 out of 472 characters), and subsequently characters constituted of three components, this means 22.25 % (i.e. 105 out of 472 characters). Characters composed of three components predominate in the second frequency group (characters with the rank from 1,001 to 2,000), this means 33.71 % (i.e. 30 out of 89 characters). A similar phenomenon occurs in the third frequency group (characters with the rank from 2,001 to 3,000) where characters consisted of three components represent 31.25 % (i.e. 10 out of 32 characters). Characters which do not belong to groups of three thousand most frequent characters are mostly constituted of five components, this means 33.33 % (i.e. from 4 to 12 characters). However, in the case of these characters, only a few observations were analysed, thus it is necessary to conduct further experiments. The most frequent characters within each frequency group are highlighted by a grey background in Table 46. As regards this article, let us summarize the results: the number of components in a character increases with the increasing number of the rank. The characters composed of two or three components represent a major part in each frequency group. It is very probable that the results obtained on this language level are not rare, a similar situation should arise in the case of other sample texts.

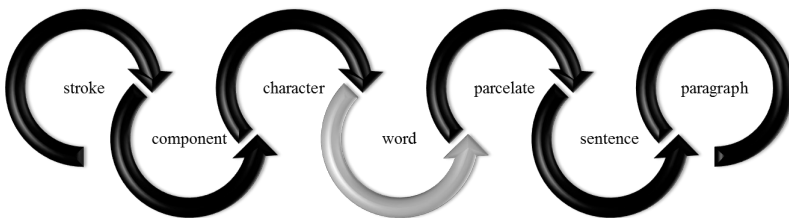
Subsequent research will focus on the exploration of 3,000 most frequent characters whose number of components will be counted. Then groups of characters with the same length (in components) will be created, and characters' percentages will be examined.

The question arises why the characters consisting of two and three components have such a high representation? One possible reason might be the fact that Chinese characters were influenced by the economization during their development; it can be said that there are two opposing forces during the course of the development of Chinese characters: the first one is the requirement for simplicity and the second one is requirement for understandability and clarity. Hence, two and three components represent a penetration of a minimal necessary requirement in terms of diversity and economy. This natural

phenomenon resulted in the reform of the previous century which officially intervened in the structure of characters, i.e. it simplified the Chinese script. And for that reason, characters composed of two and three components are pushed to a forefront and they also should occupy the foremost place in the frequency list. This presumption will be verified in the subsequent research.

The last reason which might cause such a low agreement with the MAL on this level is the absence of another language unit which would probably disperse the constant tendency. This hypothetical language unit has to be higher in hierarchy than the character and lower than the parcelate at the same time. Most likely this unit could be a word. However, when determining this unit, the graphic principle cannot be taken into consideration because texts written in Chinese characters do not operate with a word based on the orthographic principle. Chinese characters adjoin each other and they are not separated by space. From the graphical point of view, borders of words cannot be determined. Therefore, it is necessary to choose another method of determining this unit. Since the syntactic criterion was taken into account in the previous experiment when the language unit “parcelate” was determined, hence the most suitable is to keep the secondary criterion of segmentation, i.e. the syntactic criterion.

In the case of including a new language unit, the hierarchy of language units is as follows:



■ **FIGURE 20**  
The language units including the word

If the mentioned language unit is included into the hierarchy, other language levels will be acquired: word – character and parcelate – word. On the first

mentioned language level the word would represent a construct and the characters would be its constituents. On the higher language level (mentioned in the second place) the parcelate would represent the construct and the word would be a constituent of the parcelate.

**TABLE 47**

The alternative language level  $L_i$  including the word,  $x_i$  construct,  $y_i$  constituent  
Source: created by authors

Language level $L_i$		Language level $L_i$	
$x_i$	word measured in characters	$x_i$	parcelate measured in words
$y_i$	character measured in the average number of its components	$y_i$	word measured in the average number of its characters

### SUB-EXPERIMENT 3B

To verify the assumption that another language unit, namely the word, is missing in the hierarchy of linguistic units, let us perform an experiment. The sub-experiment is marked according to the respective level (i.e. 3) and the sample text (i.e. B).

The input data for sub-experiment 3B are also obtained from the blog article. The segmentation of words was carried out according to the rules of syntax introduced in the book *Úvod do studia hovorové češtiny* (Švarný, 2001). The results obtained by quantification of the data are listed below.

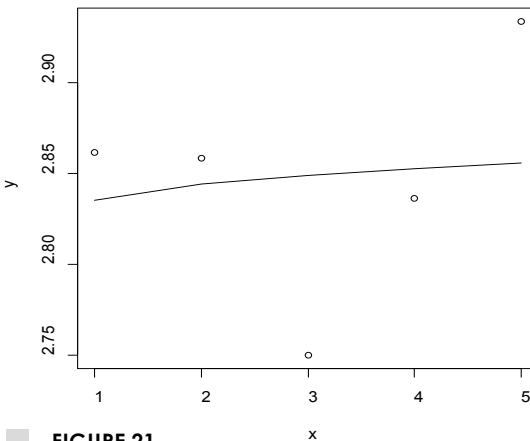
Table 48 shows observations empirically obtained by quantification of the blog article:  $x_{3B}$  represents the lengths of words (measured in characters),  $z_{3B}$  their frequencies and  $y_{3B}$  the average lengths of characters (measured in components). The grey background of the cells is used to highlight the omitted observations with a low frequency  $z_{3B} \leq 3$ . The values obtained by omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 33.

**TABLE 48**

Sub-experiment 3B – Language level L3: word (measured in characters) – character (measured in the average number of its components)

$x_{3B}$	$z_{3B}$	$y_{3B}$
1	490	2.8612
2	652	2.8582
3	144	2.7500
4	31	2.8362
5	3	2.9333

It is apparent from Table 48 that the blog article involves words compounded of one to five characters. Words consisting of two characters have the highest percent occurrence; they are followed by words consisting of one and three characters. It is evident that the average lengths of characters fluctuate around the values of two and three components independently of the length of the word, to be exact within the interval of  $(2,75; 2,93)$ . Differences between lengths of characters (measured in components) are minimal. Words consisting of five characters occurred with the lowest frequency.



**FIGURE 21**

Graphic visualization of the observations presented in Table 48

It is also evident from the graphic visualization that the downtrend of the curve illustrating the relationship between words and characters estimated by the MAL did not emerge. Assumption of the MAL, that parameter  $b$  has to be a positive real number, is not fulfilled (cf. Table 49). An agreement with the mathematical model of the MAL is only 0.0147.

**TABLE 49**

Parameter  $A$ , parameter  $b$  and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 48<sup>33</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
2.8352	-0.0044	0.0147

Based on sub-experiment 3B, it can be stated that even when another language unit “word” (measured in characters) is included into the language units’ hierarchy, the agreement with the mathematical model does not emerge due to several reasons. One of them might be an inaccurate determination of the language unit (a word), perhaps an inappropriately chosen principle of its determination, i.e. syntactic criterion. In the subsequent research, other ways of determining borders of this language unit will be used.

Another reason is connected with the lengths of characters (measured in components). As mentioned above, the majority of characters are composed of two or three components. For this reason, most words could consist of characters constituted of two or three components. Therefore, it does not matter how many characters make up a word. The average lengths of characters should always range around approximately the same values and lengths of words (measured in characters) should have a constant tendency.

33 After omitting observation with the lowest frequency ( $z_{3B} \leq 3$ ): parameter  $A = 2.8618$ ; parameter  $b = 0.0158$ ; coefficient of determination  $R^2 = 0.2612$ .

### 4.6.3 Language level L2

#### SENTENCE – PARCELATE

Table 50 shows observations empirically obtained by quantification of the blog article:  $x_2$  represents the lengths of sentences (measured in parcelate),  $z_2$  their frequencies and  $y_2$  the average lengths of parcelates (measured in characters). The grey background of the cells is used to highlight the omitted observations with a low frequency  $z_2 \leq 1$ . The values obtained by omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 34.

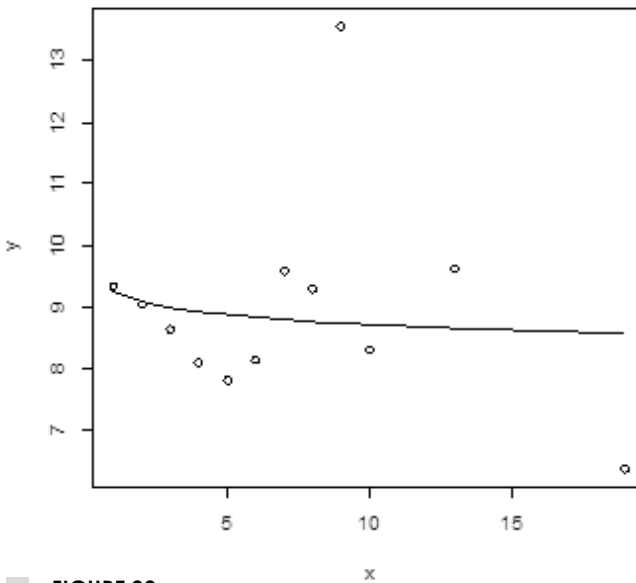
■ **TABLE 50**

Language level L2: sentence (measured in parcelates) – parcelate (measured in the average number of its characters)

$x_2$	$z_2$	$y_2$
1	9	9.3333
2	11	9.0455
3	10	8.6333
4	15	8.1000
5	5	7.8000
6	5	8.1333
7	1	9.5714
8	4	9.2813
9	1	13.5556
10	2	8.3000
13	1	9.6154
19	1	6.3684

It is evident from Table 50 that the sample text establishes 12 different lengths of sentences which are constituted within the range of 1 to 19 parcelates.

As regards all the average lengths of parcelates, they do not correspond to the assumption of the MAL that the average lengths of its constituents (parcelates) decrease with the increasing lengths of constructs (sentences). The average lengths of all parcelates fluctuate within the interval of  $(6,37; 13,56)$ . However, the downtrend can be observed in the case of values  $x_2 = 1; 2; 3; 4; 5$  which have the highest frequency. Sentences composed of seven, nine, thirteen and nineteen parcelates,  $x_2 = 7; 9; 13; 19$ , have the lowest frequency; they occurred only once in the sample text. Because their frequency is low in comparison with the other frequencies, it is possible to omit them. These observations are highlighted by a grey background of the cells in Table 50.



**FIGURE 22**

Graphic visualization of the observations presented in Table 50

Based on the graphic visualization it can be seen that the relationship of inverse proportionality between sentences and parcelates formulated by the MAL was proved but only with a minimal agreement, i.e. value of the goodness-of-fit is 0.0144.



However, when the attention is paid to the first five observations, then it can be found out that these observations reveal a downward trend of the curve. This shows that the MAL is valid in this restricted area.

**TABLE 51**

Parameter  $A$ , parameter  $b$  and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 50<sup>34</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
9.2452	0.0258	0.0144

It seems from the results that the parcelate is not confirmed as a valid language unit. However, it should be noticed that the observations with the highest frequencies have a declining tendency, hence this language unit could be valid.

## DISCUSSION

The assumption that the average lengths of parcelates (measured in characters) decline with the increasing number of sentences (measured in parcelates) was not proved with a high agreement on this level because of several possible reasons.

The first of them was already mentioned in the previous subchapter. The probable reason why the MAL did not appear is the absence of a language unit which is higher than the character and lower than the parcelate. This unit should probably be a word which should create another language level in conjunction with the parcelate: parcelate – word. The parcelate should represent a construct on this language level and words should represent its constituents. Subsequent research will be aimed at verifying this assumption and at determining the borders of the language unit “word”.

34 After omitting observations with the lowest frequency ( $z_i \leq 1$ ): parameter  $A = 9.0852$ ; parameter  $b = 0.0434$ ; coefficient of determination  $R^2 = 0.2333$ .

Another reason for the low agreement with the MAL might be an insufficient number of observations on this language level, which might have caused that mutual relations between the construct and its constituents do not reveal. Let us conduct an experiment in order to verify whether the relationship will be revealed and whether the MAL will be the adequate and well-fitting model in a longer article. The maximal length, which was set at 3,500 characters in the previous experiment, will be increased in this sub-experiment. Another blog article from author Han Han, titled *My father Han Renjun and his work* (wǒ de fùqīn Hán Rénjūn yǐjǐ tā de zuòpǐn, 我的父亲韩仁均以及他的作品) – Sample C (Appendix 3), will serve as the sample text. It meets all the criteria established for the choice of the sample texts. The article contains 8,466 characters and was posted on Han Han’s blog on 27<sup>th</sup> January 2012. The sub-experiment is marked according to the respective level (i.e. 2) and the sample text (i.e. C).

### SUB-EXPERIMENT 2C

Table 52 shows observations empirically obtained by quantification of the longer blog article:  $x_{2c}$  represents the lengths of sentences (measured in parcelate),  $z_{2c}$  their frequencies and  $y_{2c}$  the average lengths of parcelates (measured in characters). The grey background of the cells is used to highlight the omitted observations with a low frequency  $z_{2c} \leq 1$ . The values obtained by omitting (parameters  $A$ ,  $b$  and coefficient of determination  $R^2$ ) are listed in footnote no. 35.

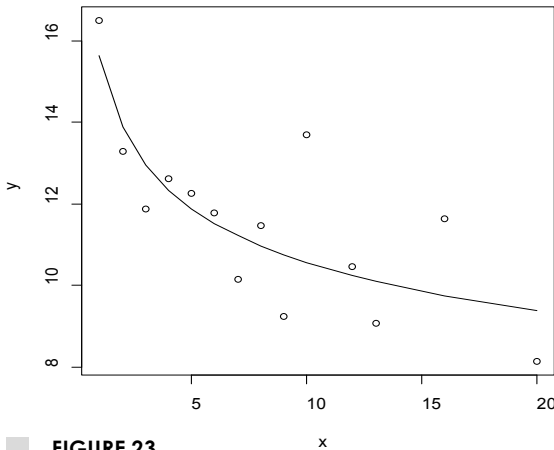
**TABLE 52**

Sub-experiment 2C – Language level L2: sentence (measured in parcelates) – parcelate (measured in the average number of its characters)

$x_{2c}$	$z_{2c}$	$y_{2c}$
1	41	16.4878
2	52	13.2885
3	40	11.8750
4	26	12.6250

$x_{2c}$	$z_{2c}$	$y_{2c}$
5	13	12.2615
6	10	11.7667
7	6	10.1429
8	6	11.4583
9	4	9.2500
10	1	13.7000
12	2	10.4583
13	1	9.0769
16	1	11.6250
20	1	8.1500

Fourteen sentences with various lengths (measured in parcelates) occurred in the blog article in total. Sentences consisting of 10, 13, 16 and 20 parcelates, i.e.  $x_{2c} = 10; 13; 16; 20$ , occur only once in the text and can therefore be omitted. These observations are highlighted by a grey background of the cells in the table. The average lengths of parcelates appear within the interval of  $\langle 8, 15; 16, 49 \rangle$ . From the table it can be seen that the average lengths of parcelates defined by the MAL have a downtrend.



**FIGURE 23**  
Graphic visualization of the observations presented in Table 52

From the visualization of relationships between the sentence and the parcelate it can be concluded that a downward trend of the curve emerged, showing the relationship between the lengths of the constructs and the lengths of constituents. The assumption that parameter  $b$  is a positive real number is satisfied. The agreement with the mathematical model of the MAL is much wider than in the case of the shorter article (cf. Table 53).

**TABLE 53**

Parameter  $A$ , parameter  $b$  and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 52<sup>35</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
15.6188	0.1699	0.5931

It is apparent from the stated data and the visualization that the assumption of the MAL emerged on language level L2 in the case of the longer article.

The assumption that the relationship formulated by the MAL will occur in the case of a larger amount of observations was correct. It is necessary to carry out more experiments which confirm or reject this assumption.

#### 4.6.4 Language level L1

##### PARAGRAPH – SENTENCE

Table 54 shows observations empirically obtained by quantification of the blog article:  $x_i$  represents the lengths of paragraphs (measured in sentences),  $z_i$  their frequencies and  $y_i$  the average lengths of sentences (measured in parcelates).

**TABLE 54**

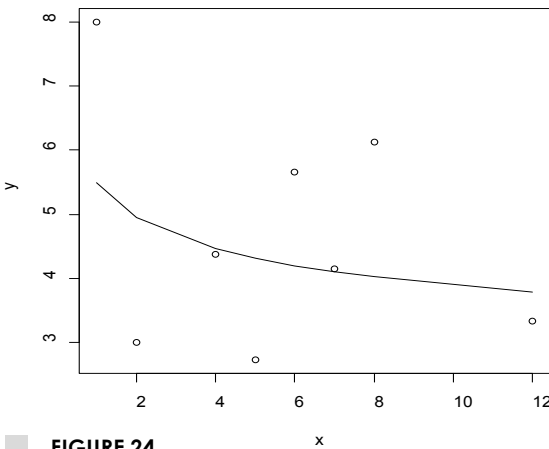
Language level L1: paragraph (measured in sentences) – sentence (measured in the average number of its parcelates)

$x_i$	$z_i$	$y_i$
1	1	8.0000

35 After omitting observations with the lowest frequency ( $z_{2c} = 1$ ): parameter  $A = 15.7993$ ; parameter  $b = 0.1897577$ ; coefficient of determination  $R^2 = 0.8081$ .

$x_1$	$z_1$	$y_1$
2	1	3.0000
4	2	4.3750
5	3	2.7333
6	2	5.6667
7	1	4.1429
8	1	6.1250
12	1	3.3333

Table 54 shows that there were 8 different lengths of paragraphs measured in sentences. The shortest paragraph is composed of one sentence and the longest of twelve sentences. Frequency of each paragraph is very low. Therefore, it is not possible to omit any observation. The paragraph consisting of five sentences occurred most often, however the frequency is also very low because there are only three occurrences. It follows from the stated data that the downtrend defined by the MAL is not observed in the case of the average lengths of sentences measured in parcelates. The average lengths rather oscillate in the range of  $\langle 2,73; 8,00 \rangle$ .



**FIGURE 24**  
Graphic visualization of the observations presented in Table 54

It is obvious from the graphic visualization that the downtrend of the curve which demonstrates the relationship between language units formulated by the MAL is only indicated. An agreement with the mathematical model of the MAL is very low (i.e. 0.1022). The assumption that parameter  $b$  is a positive real number is satisfied.

**TABLE 55**

Parameter  $A$ , parameter  $b$  and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 54<sup>36</sup>

Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
5.4888	0.1493	0.1022

## DISCUSSION

Although the frequency of observations is low on this level, the minimal agreement might also be caused by punctuation. Punctuation, which establishes the borders of sentences as well as the parcelates, which the average length of sentences is measured in, was integrated into Chinese texts relatively recently. For this reason, the usage of punctuation marks might not be stable and the determination of language units may not always follow the identical conditions. It is possible that different authors use punctuation marks in different ways. (cf. subchapter 3.6.4 and Motalová et al., 2013).

As in the case of language level L2, another reason might be the low frequency of observations which might cause that the inverse relationship between constructs and their constituents did not reveal. Whether the length of a text affects the revealing of the relationship formulated by the MAL will be verified within a sub-experiment, in which the longer text will be analysed. As in the previous sub-experiment 2C on level L2 the sample text is the article *My father Han Renjun and his work* (Wǒ de fùqin Hán Rénjūn yǐjǐ tā

36 In the case of this language unit it was not appropriate to omit empirically gained observations with a low frequency due to their small number.

de zuòpǐn, 我的父亲韩仁均以及他的作品) whose length is more than 8,400 characters. The sub-experiment is marked according to the respective level (i.e. 1) and the sample text (i.e. C).

### SUB-EXPERIMENT 1C

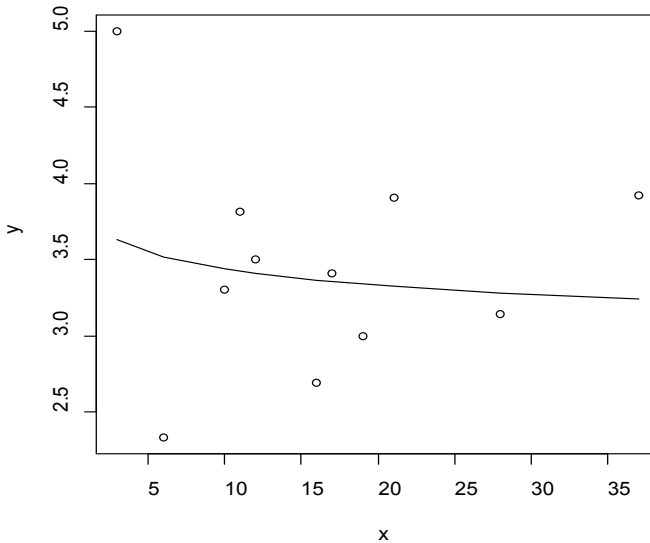
Table 56 shows observations empirically obtained by quantification of the longer blog article:  $x_{1c}$  represents the lengths of paragraphs (measured in sentences),  $z_{1c}$  their frequencies and  $y_{1c}$  the average lengths of sentences (measured in parcelates).

**TABLE 56**  
Sub-experiment 1C – Language level L1: paragraph (measured in sentences) – sentence (measured in the average number of its parcelates)

$x_{1c}$	$z_{1c}$	$y_{1c}$
3	1	5.0000
6	1	2.3333
10	1	3.3000
11	1	3.8182
12	3	3.5000
16	1	2.6875
17	1	3.4118
19	1	3.0000
21	1	3.9048
28	1	3.1429
37	1	3.9189

It can be seen from the empirically obtained data listed in Table 56 that the longer article operates with 11 different lengths of paragraphs (measured

in sentences) in total. The shortest paragraph is composed of 3 sentences and the longest of 37 sentences. However, the frequency of observations is still very low. Only the paragraph consisting of 12 sentences ( $x_{lc} = 12$ ) occurred three times. The downward trend of the average lengths of sentences is not apparent from the table and thus the assumption of the MAL that states that the average lengths of sentences (constituents) decrease with the increasing lengths of paragraphs (constructs) is probably not fulfilled even in the case of the longer article.



**FIGURE 25**

Graphic visualization of the observations presented in Table 56

It can be concluded on the basis of the graphic visualisation that the downward trend of the curve which demonstrates the relationship between lengths of paragraphs and sentences assumed by the MAL did not occur with a wide agreement even in the case of the longer article. Although the assumption of the MAL that parameter  $b$  is a positive real number is fulfilled, the agreement with the mathematical model is a mere 0.0246 (cf. Table 57). This goodness-of-fit is even lower than in the case of the shorter article (cf. Table 55).



**TABLE 57**

Parameter *A*, parameter *b* and coefficients of determination  $R^2$  for the mathematical model related to the observations presented in Table 56<sup>37</sup>

Parameter <i>A</i>	Parameter <i>b</i>	Coefficient of determination $R^2$
3.8167	0.0455	0.0246

It can be summarized that the verbal formulation of the MAL, which can be expressed on this level as follows: *the longer the paragraph (measured in sentences) is, the shorter the average lengths of sentences (measured in parcelates) is*, did not reveal either in the case of the shorter blog article (composed of 2,641 characters), or in the case of the longer article (composed of 8,466 characters).

An insufficient amount of input data, i.e. an insufficient length of the article in conjunction with a low frequency of paragraphs, might cause that the MAL revealed with a minimal agreement on this language level. Although an experiment which examined an article more than three times longer was conducted (i.e. the article's length is more than 8,400 characters), only an insufficient amount of input data was obtained.

More experiments which involve analysing longer articles should be carried out. These sample texts should contain more than 5,000 characters. This length is approximately the same length as in the case of the scientific article in which the validity of the MAL was confirmed on this level.

#### 4.6.5 Conclusion

In this subchapter, the data and graphic visualisations obtained by the quantitative analysis of the blog article were interpreted. Verification of the validity (or invalidity) of the MAL was performed on four language levels L4, L3, L2 and L1. Each level consists of two language units which are immediately adjoined in the hierarchy. Let us summarize the results gained on each language

<sup>37</sup> In the case of this language unit it was not appropriate to omit empirically gained observations with a low frequency due to their small number.

level and outline possible reasons for the agreement or disagreement with the MAL.

On the lowest level L4 character – component, the goodness-of-fit with the empirically gained observations was the highest ( $R^2 = 0.8698$ )<sup>38</sup>. This wide agreement could have been caused by the constant size of the characters' graphic field. Based on this assumption, there a conclusion arises that in the case of texts written in the traditional form of characters the pressure of the graphic field has to be even more significant. Co-author Motalová conducted an experiment that examined the validity of the MAL applied to the text converted into a traditional set of Chinese characters. The results showed that the goodness-of-fit of the mathematical model has even increased.

From the resultant visualization of the relationship between the parcelate and its characters it is evident that the tendency of the curve is rather constant, i.e. the average lengths of the characters are in the range of values within two or three components. However, it should be noted that the tendency formulated by the MAL was indicated. The indication of a downtrend was caused mainly due to the observations with a low frequency which brought about deviations from the constant tendency. After omitting the least frequent observations the tendency of the curve would be more constant and the value of agreement would decrease. The minimal agreement with the MAL can be caused by several possible reasons. The first of them is the different characteristics of language units. The parcelate represents a unit with a variable length, whereas the character represents a language unit with an unchanging length. Another reason might be the structural changes in characters which were caused by the centrally planned reform implemented in the previous century. The reform reduced the number of strokes in characters and thus the number of components also changed in many characters. Along with this reason another fact emerges that almost a half of the characters from the blog article are composed of two or three components, therefore, they also influence to a large extent the construction of parcelates which are measured in characters. The last possible reason is

---

38 Without statistical adjustment.

the absence of another language unit (higher than the character and lower than the parcelate), which would probably be a “word”. Based on this assumption, the subsequent experiment was conducted. In this experiment, the additional linguistic unit “word” was determined according to the syntactical rules specified in *Spoken Chinese: Introduction to Study of Spoken Chinese* (Švarný, 2001). By linking the syntactic “word” with other units the following language levels are newly obtained: word (measured in characters) – character (measured in the average lengths of components) and parcelate (measured in words) – word (measured in the average lengths of characters). The sub-experiment was conducted on the language level word – character. The results of the research discovered that the average length of characters is almost constant with an increasing length of words, i.e. the average lengths of characters oscillate within the interval of (2,75; 2,93). Therefore, the agreement with the mathematical model of the MAL did not emerge even after the inclusion of another language unit, namely the “word”. Hence, subsequent research will be specialized on the more precise determination of this language unit.

The mutual relationship of inverse proportionality between the constructs and their constituents almost did not show on level L2 sentence – parcelate. As in the case of the previous language level L3, the minimal agreement with the mathematical model of the MAL may be caused by the absence of another language unit which has to be higher in hierarchy than the character and lower than the parcelate. Another possible reason is an insufficient number of observations. For this reason, sub-experiment 2C was conducted. In case of this experiment, the longer blog article was used as the sample text. The author of the chosen article *My father Han Renjun and his work* with the length of 8,466 characters was also Han Han. After a quantification of this article, the acquired results showed a downward trend of curve, and the agreement with the mathematical model of the MAL was relatively wide ( $R^2 = 0.5931$ ). The assumption that the insufficient length of a sample text brings about no emergence of the mutual relationship formulated by the MAL was correct.

On the last highest language level L1: paragraph – sentence, the downward trend of curves was revealed, however, the agreement with the mathematical

model of the MAL was very low. One of the possible reasons for the low agreement can be the punctuation (which determines borders of sentences). Punctuation does not have a long tradition in Chinese texts. Another reason may be an insufficient amount of observations. That is why sub-experiment 1C, which examined the longer blog article *My father Han Renjun and his work*, was conducted on this level. The declining trend of the curve visualizing the relationship between the lengths of variables (paragraph and sentence) assumed by the MAL was almost unrevealed even in the case of the longer article; the value of agreement was minimal. However, the insufficient frequency of observations of paragraphs still has a great influence on the results.

As regards the blog article, our assumption that the MAL is an adequate and well-fitting mathematical model on language levels L4 character – component and L2 sentence – parcelate was confirmed only on language level L4. On language level L2 the agreement with the mathematical model of the MAL is minimal. The assumption of the MAL's invalidity (eventually of a minimal agreement) was confirmed in the case of both language levels L3 parcelate – character and level L1 paragraph – sentence.

Regarding the hypothesis which states: if a language unit of the contemporary written Chinese is determined on the basis of the graphical criterion, the mutual relationships between them exist on a respective language level and their validity is verified by means of the MAL, in the case of the first hypothetical language unit, i.e. the component, on the basis of the results obtained from the blog article it seems that the component exists and it has proved itself as a valid language unit on level L4. Thereby, the hypothesis was confirmed in this case. However, as regards the second hypothetical language unit, i.e. the parcelate, it appears that the existence of the parcelate is not verified either on level L3 or on level L2. Nonetheless, on level L3, its existence is not excluded because of the several factors which adversely affect the relationship between the units on this level (cf. subchapter 4.6.2., Discussion). Concerning level L2, it should be noted that the observations with the highest frequencies have a declining tendency, hence, this language unit could be valid.

## 5. Comparison of the sample texts

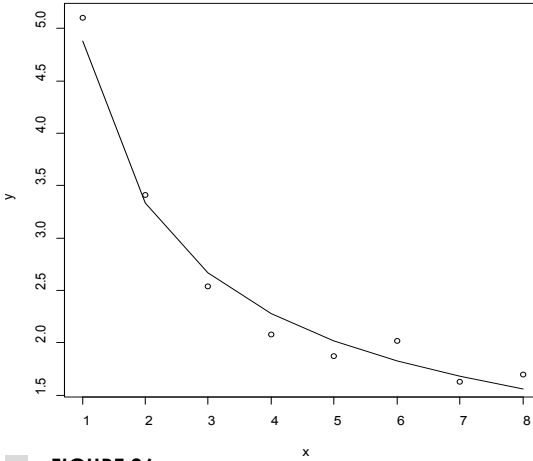
This chapter is aimed at the comparison of the results obtained by quantification of four sample texts: the newspaper article, the short story, the scientific article and the blog article.

Firstly, it should be stated that owing to the detection of the main tendencies of mutual relationships between the determined language units, the empirically gained observations of the previous experiment were statistically adjusted. It means that observations with the lowest frequency (i.e. extremes) were omitted. The data acquired by their omission are included in the article *An Application of the Menzerath–Altmann Law to Contemporary Chinese* in the *Czech and Slovak Linguistics Review*. The aim of the subsequent experiment was to increase the number of analysed sample texts written in different stylistic styles. As regards the analyses of the scientific article and the blog article, the linguistic approach embracing all observations was preferred. Therefore the results of this experiment involve observations with the lowest frequency. The data acquired by their omission are stated in respective footnotes. In order to compare all results, it was necessary to choose only one of the applied approaches. Due to examining tendencies of the stylistic styles on the individual language levels, data without omitted observations are the object of the comparison. Values of parameters  $A$ ,  $b$  and the coefficient of determination  $R^2$ , which were gained by statistical adjusting of observations obtained from the previous experiment are stated in footnotes for the sake of comparison. Owing to comprehensiveness, these data are supplied by statistically adjusted data acquired from this experiment.

This chapter is divided into four subchapters exploring respective language levels separately. Each of them begins with graphical visualizations of the mutual relationships between language units on a respective language level and with tables containing values of the parameters and the coefficients of determination. These data are followed by their interpretations. Tables including observations obtained by quantification of the newspaper article and the short story are presented in the appendix (cf. Appendix 4).

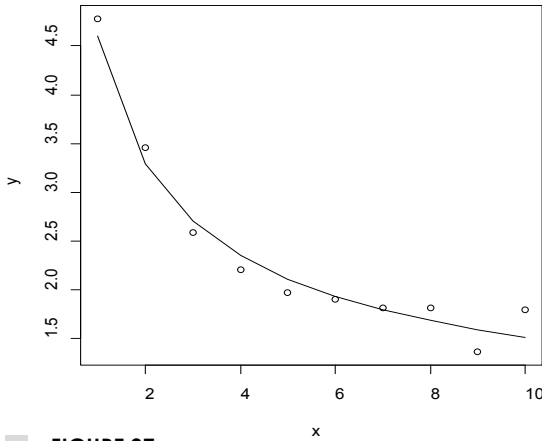
## 5.1 COMPARISON OF THE SAMPLE TEXTS ON LANGUAGE LEVEL L4

### CHARACTER – COMPONENT



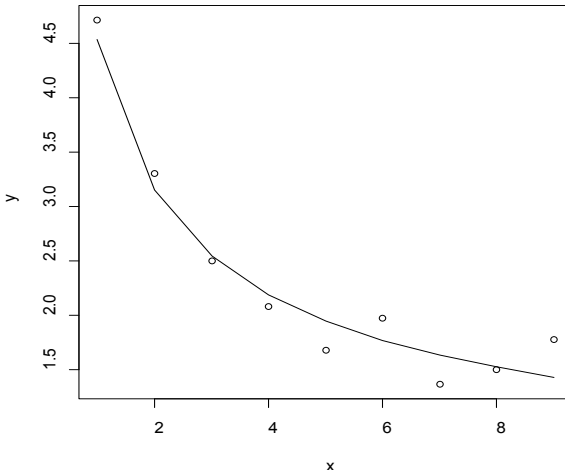
**FIGURE 26**

Graphical visualization of the relationship between the character and the component related to the newspaper article



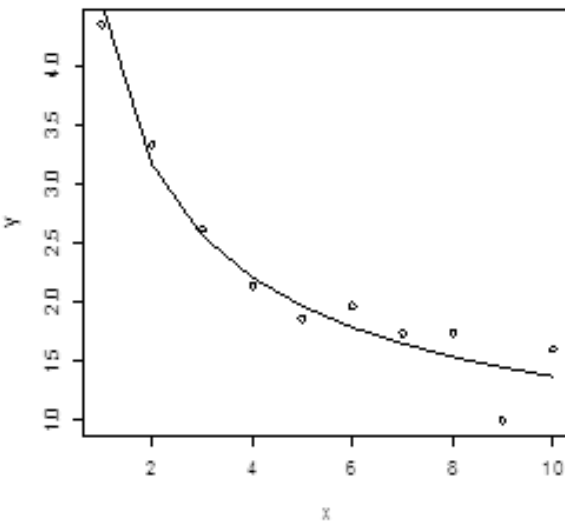
**FIGURE 27**

Graphical visualization of the relationship between the character and the component related to the short story



**FIGURE 28**

Graphical visualization of the relationship between the character and the component related to the scientific article



**FIGURE 29**

Graphical visualization of the relationship between the character and the component related to the blog article

**TABLE 58**

Parameters  $A$ ,  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the empirically obtained observations<sup>39</sup>

Sample text	Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
Newspaper article <sup>40</sup>	4.8774	0.5482	0.9658
Short story <sup>41</sup>	4.5992	0.4828	0.9389
Scientific article <sup>42</sup>	4.5406	0.5259	0.9061
Blog article <sup>43</sup>	4.5461	0.5208	0.8698

As can be seen from Table 58, all parameters  $b$  have positive values and therefore all sample texts satisfy the first assumption of the MAL. In this connection, the second assumption of the MAL is also adhered to because all curves visualizing the relationship of the inverse proportionality between the character and the component are decreasing and convex (cf. Figure 26, 27, 28 and 29). On the basis of extremely high values of the goodness-of-fit, it can be concluded that the mathematical model of the MAL proves itself as adequate and well-fitting. With the exception of the blog article (i.e.  $R^2 = 0.8698$ ), value of the agreement exceeds 0.9 in the case of the other sample texts. The highest value of the coefficient of determination is observed in the newspaper article (i.e. 0.9658).

<sup>39</sup> cf. Appendix 4.

<sup>40</sup> After omitting observations with the lowest frequency: parameter  $A = 4.9724$ ; parameter  $b = 0.5738$ ; coefficient of determination  $R^2 = 0.9717$ .

<sup>41</sup> After omitting observations with the lowest frequency: parameter  $A = 4.6546$ ; parameter  $b = 0.4941$ ; coefficient of determination  $R^2 = 0.9767$ .

<sup>42</sup> After omitting observations with the lowest frequency: parameter  $A = 4.8117$ ; parameter  $b = 0.5974$ ; coefficient of determination  $R^2 = 0.9519$ .

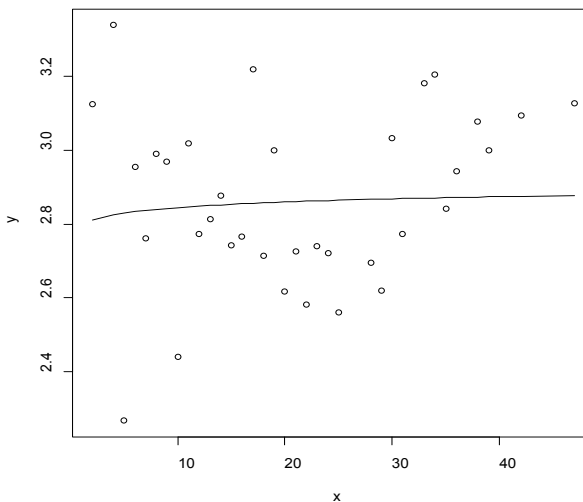
<sup>43</sup> After omitting observations with the lowest frequency: parameter  $A = 4.3123$ ; parameter  $b = 0.4554$ ; coefficient of determination  $R^2 = 0.9719$ .



The relationship between the stroke, the component and the character is unchanging. The structure of every character is formed under the influence of an identical factor, i.e. the constant size of the graphic field which causes that tendencies appearing in every character are the same. With regard to this fact it should be noted that neither stylistic styles nor texts' lengths have any influence on the results of this level. The mathematical model of the MAL shows an extremely wide goodness-of-fit with empirically obtained observation in all tested sample texts. This agreement could be probably caused by the previously mentioned tendencies (cf. subchapters 3.6.1 and 4.6.1).

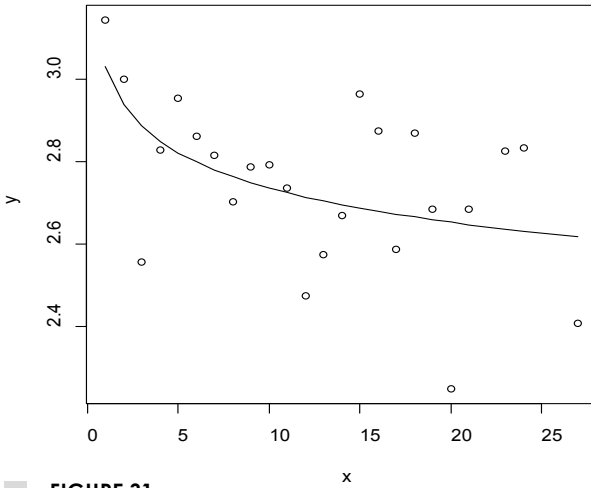
## 5.2 COMPARISON OF THE SAMPLE TEXTS ON LANGUAGE LEVEL L3

### PARCELATE – CHARACTER

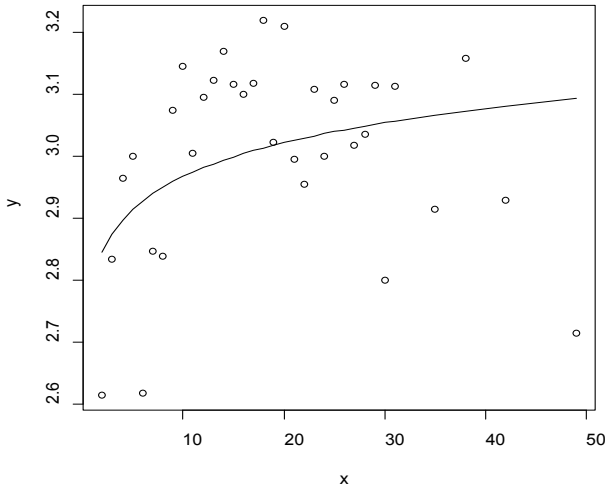


**FIGURE 30**

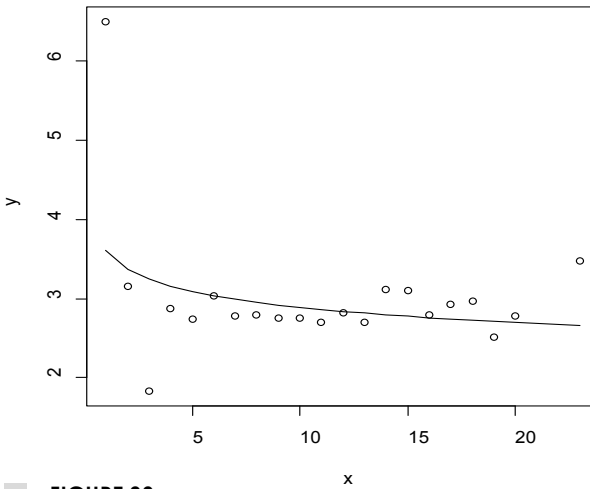
Graphical visualization of the relationship between the parcelate and the character related to the newspaper article



**FIGURE 31**  
Graphical visualization of the relationship between the parcelate and the character related to the short story



**FIGURE 32**  
Graphical visualization of the relationship between the parcelate and the character related to the scientific article



**FIGURE 33**  
Graphical visualization of the relationship between the parcelate and the character related to the blog article

**TABLE 59**  
Parameters  $A$ ,  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the empirically obtained observations<sup>44</sup>

Sample text	Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
Newspaper article <sup>45</sup>	2.7968	-0.0074	0.0043
Short story <sup>46</sup>	3.0303	0.0444	0.2470
Scientific article <sup>47</sup>	2.7933	-0.0262	0.1435

44 cf. Appendix 4

45 After omitting observations with the lowest frequency: parameter  $A = 3.0945$ ; parameter  $b = 0.0396$ ; coefficient of determination  $R^2 = 0.1034$ .

46 After omitting observations with the lowest frequency: parameter  $A = 3.0442$ ; parameter  $b = 0.0488$ ; coefficient of determination  $R^2 = 0.3497$ .

47 After omitting observations with the lowest frequency: parameter  $A = 2.6651$ ; parameter  $b = -0.0484$ ; coefficient of determination  $R^2 = 0.4615$ .

Sample text	Parameter A	Parameter b	Coefficient of determination $R^2$
Blog article <sup>48</sup>	3.6055	0.0965	0.1316

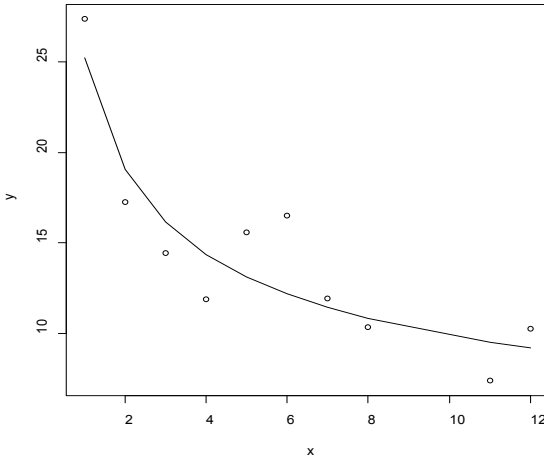
Parameters  $b$  related to the newspaper article and the scientific article have a negative value and thereby these sample texts do not adhere to the first assumption of the MAL (cf. Table 59). On the contrary, in the case of the short story and the blog article this assumption was satisfied; the parameters  $b$  have a positive value (cf. Table 59). As can be seen from the graphical visualizations, the decreasing tendency of the curves is not observed within the newspaper article and the scientific article because the empirically obtained observations oscillate in the interval of (2.4; 3.2), cf. Figures 30 and 32; the decreasing curves are noticed only in the short story and the blog article, cf. Figures 31 and 33. As regards the blog article, the downtrend is caused by the first observation with a very low frequency ( $z_3 = 2$ ). It should be noted that if this observation is statistically adjusted by omitting, the curve has a constant tendency. In fact, a decreasing tendency is observed only in the case of the short story; nevertheless, the goodness-of-fit of the mathematical model is not high (cf. Table 59). As regards all sample texts, the MAL does not prove itself as an adequate and well-fitting mathematical model on this level because of no or low value of the agreement which does not exceed 0.7 (cf. Benešová, 2011, p. 77).

On the basis of the acquired results it seems that the mutual relationship of the inverse proportionality between the parcelate and the character is disturbed. As mentioned earlier, this relationship could be impacted by several factors, such as the different characteristic of these language units, various approaches to definitions of the component, an absent language unit (alternatively an absent level) and a reduction of variability related to components' numbers of those Chinese characters which were simplified by the reform of the Chinese script realized in the 1950s and 1960s, cf. 3.6.2 a 4.6.2.

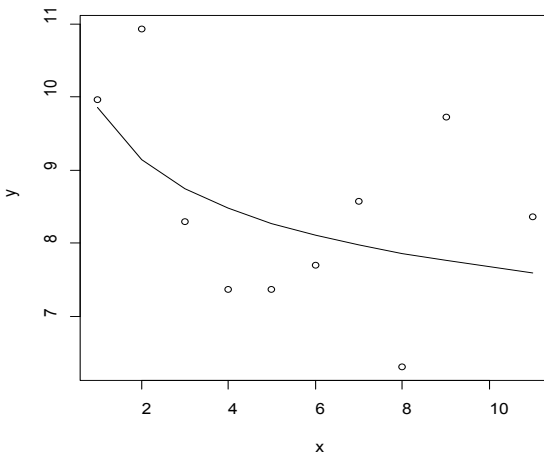
48 After omitting observations with the lowest frequency: parameter  $A = 2.5571$ ; parameter  $b = -0.0364$ ; coefficient of determination  $R^2 = 0.0410$ .

### 5.3 COMPARISON OF THE SAMPLE TEXTS ON LANGUAGE LEVEL L2

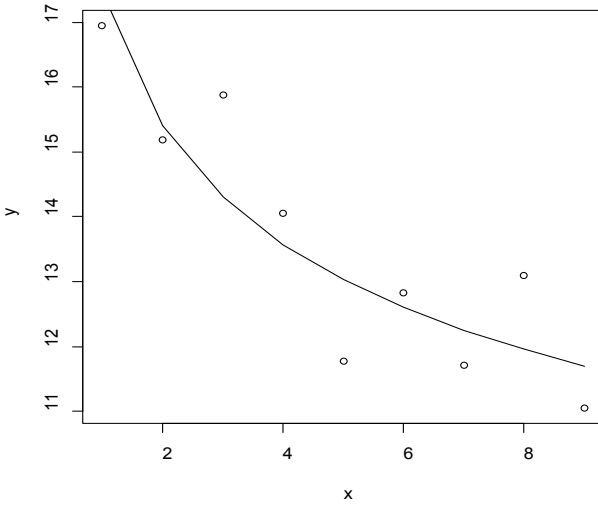
#### SENTENCE – PARCELATE



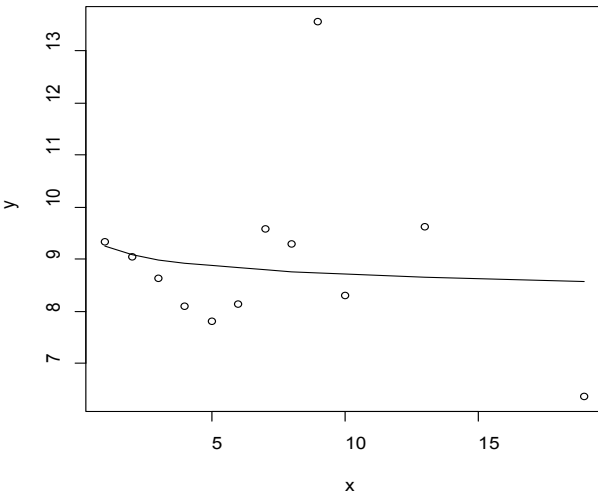
**FIGURE 34**  
Graphical visualization of the relationship between the sentence and the parcelate related to the newspaper article



**FIGURE 35**  
Graphical visualization of the relationship between the sentence and the parcelate related to the short story



**FIGURE 36**  
Graphical visualization of the relationship between the sentence and the parcelate related to the scientific article



**FIGURE 37**  
Graphical visualization of the relationship between the sentence and the parcelate related to the blog article

**TABLE 60**

Parameters  $A$ ,  $b$  and coefficient of determination  $R^2$  for the mathematical model related to the empirically obtained observations<sup>49</sup>

Sample text	Parameter $A$	Parameter $b$	Coefficient of determination $R^2$
Newspaper article <sup>50</sup>	25.2439	0.4067	0.7702
Short story <sup>51</sup>	9.8600	0.1090	0.2436
Scientific article <sup>52</sup>	17.4816	0.1827	0.7862
Blog article <sup>53</sup>	9.2452	0.0258	0.0144

It is apparent from the above stated Table 60 that the value of parameter  $b$  is a positive real number in all cases. Thus the assumption of the MAL is satisfied. The downward trend of the curve is the most evident in the case of the newspaper and scientific articles, whereas in the case of the short story it is only indicated and as regards the blog article, the tendency is almost constant (cf. Figures 34, 35, 36 and 37). It appeared that the mathematical model of the MAL proves itself as an adequate and well-fitting model in the case of the scientific and newspaper articles. Value of the agreement exceeded 0.77 (cf. Table 60). Concerning the short story and the blog article, the agreement is significantly lower, it is 0.2436 in the short story and 0.0144 in the blog article.

It implies from the previously performed experiments that the relationship between the variables formulated by the MAL is better reflected in the texts which have to follow a formal structure typical for their language style (e.g. newspaper

<sup>49</sup> cf. Appendix 4.

<sup>50</sup> After omitting the observations with the lowest frequency parameter  $A = 24.7192$ ; parameter  $b = 0.4077$ ; coefficient of determination  $R^2 = 0.8478$ .

<sup>51</sup> After omitting the observations with the lowest frequency parameter  $A = 10.6873$ ; parameter  $b = 0.2091$ ; coefficient of determination  $R^2 = 0.6870$ .

<sup>52</sup> After omitting the observations with the lowest frequency parameter  $A = 17.1221$ ; parameter  $b = 0.1542$ ; coefficient of determination  $R^2 = 0.7167$ .

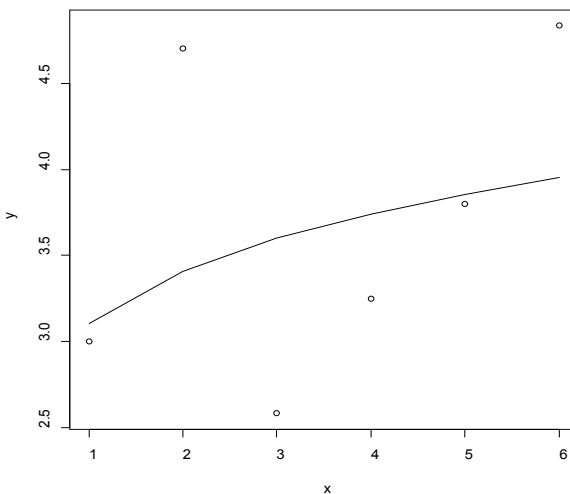
<sup>53</sup> After omitting the observations with the lowest frequency parameter  $A = 9.0852$ ; parameter  $b = 0.0434$ ; coefficient of determination  $R^2 = 0.2333$ .

and scientific articles). As regards texts with unregulated stylistic styles, it appears from the statistically adjusted data that the mathematical model shows a low agreement with them.

If we draw our attention to observations with the lowest frequency (so-called extremes), it can be found out that these observations might have a major impact on goodness-of-fit with the mathematical model of the MAL. Their chaotic dispersal conceals the downward trend which is evident within more frequent observations. Because the omission of these extremes significantly increases the value of agreement (cf. the values listed in footnote 51 and 53), we suggest to conduct research whose samples will be written in other stylistic styles.

## 5.4 COMPARISON OF THE SAMPLE TEXTS ON LANGUAGE LEVEL L1

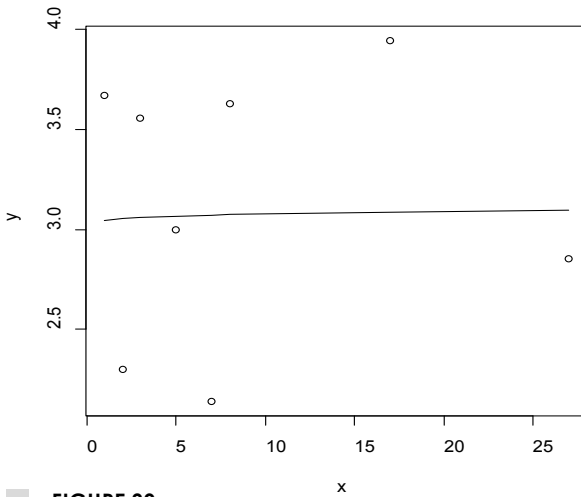
### PARAGRAPH – SENTENCE



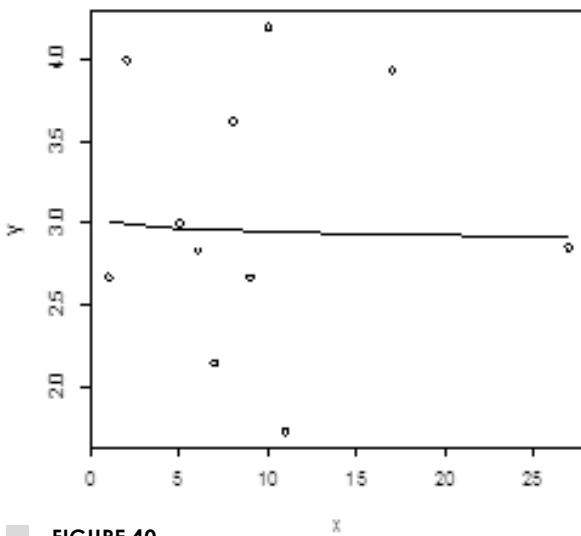
**FIGURE 38**

Graphical visualization of the relationship between the paragraph and the sentence related to the newspaper article

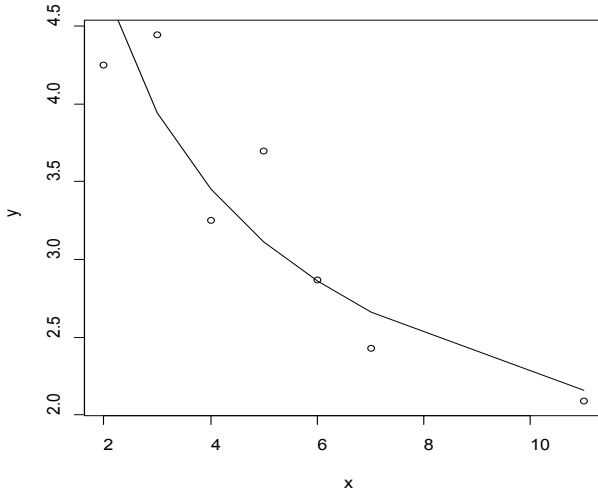




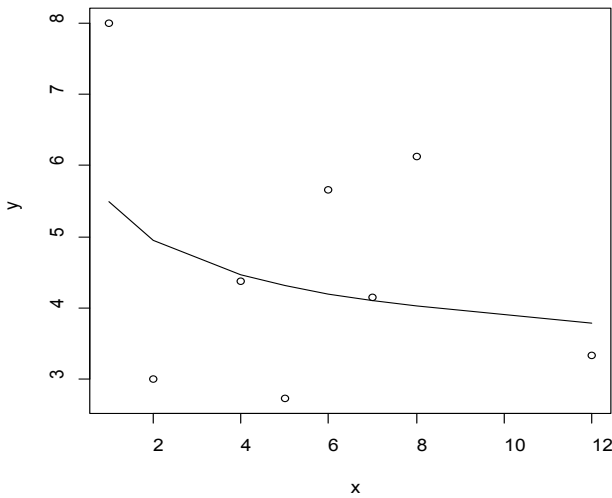
**FIGURE 39**  
Graphical visualization of the relationship between the paragraph and the sentence related to the short story – Variant 1



**FIGURE 40**  
Graphical visualization of the relationship between the paragraph and the sentence related to the short story – Variant 2



**FIGURE 41**  
Graphical visualization of the relationship between the paragraph and the sentence related to the scientific article



**FIGURE 42**  
Graphical visualization of the relationship between the paragraph and the sentence related to the blog article

**TABLE 61**

Parameters *A*, *b* and coefficient of determination  $R^2$  for the mathematical model related to the empirically obtained observations<sup>54</sup>

Sample text		Parameter <i>A</i>	Parameter <i>b</i>	Coefficient of determination $R^2$
Newspaper article		3.1068	-0.1342	0.1256
Short story <sup>55</sup>	Variant 1	3.0416	-0.0051	0.0006
	Variant 2	3.0186	0.0101	0.0011
Scientific article		6.5604	0.4636	0.8535
Blog article		5.4888	0.1493	0.1022

If parameters *b* are compared, we find out that the negative values occurred in the case of the newspaper article and the short story (sample A). Concerning other samples, the value of parameter *b* is positive (cf. Table 61). It can be seen from the graphical visualizations that the downward trend of the curves occurred in the scientific article, but in the blog article the tendency was only indicated and in the cases of other samples the tendency was not observed at all (cf. Figures 38, 39, 40, 41 and 42). Regarding the agreement of the MAL with the empirically obtained observations in the scientific article, the value is very high – 0.8535. The reason for this high agreement might be a significantly greater length of the sample, thus the observations have higher frequencies in comparison with other samples. Within other sample texts the agreements were minimal or none (cf. Table 61).

<sup>54</sup> cf. Appendix 4

In the case of this language level it was not appropriate to omit empirically gained observations with a low frequency due to their small number.

<sup>55</sup> In the case of the short story, paragraphs were not clearly graphically divided from each other. Therefore two variants of segmentation were chosen. For more information, cf. (Motalová et al., 2013).

## CONCLUSION

The experiments conducted within this work were aimed at the quantitative analyses of the contemporary written Chinese which was represented by two sample texts written in simplified characters and different stylistic styles. The analyses were concretely applied to the scientific article, which was published in the academic periodical *The Applied Linguistics* and released on the official websites of the *Chinese Academy of Social Sciences*, and the blog article, whose author is the well-known Chinese blogger Han Han.

The aim of these quantitative analyses was testing the existence of the selected language units by means of the Menzerath–Altmann law. Furthermore, the analyses were performed with the aim of comparison of the obtained results with the results acquired from the previous experiment.

As in the previous experiment, the language units were determined on the basis of the graphical criterion with a minimal and inevitable consideration of the syntactic criterion. Let us enumerate these units: stroke, component, character, parcelate, sentence and paragraph. Linking two immediately contiguous language units into mutual relationships allowed the creation of four language levels, namely character – component, parcelate – character, sentence – parcelate and paragraph – sentence. According to the hypothesis<sup>56</sup> and the results of the previous experiment it was assumed that in the case of the lowest language level (i.e. character – component) the MAL would prove itself as an adequate and well-fitting mathematical model which would show an extremely high agreement with the empirically gained observations. Another assumption concerning the next language level (i.e. parcelate – character) was that the MAL would not represent an adequate and well-fitting model, in other words its validity would not be confirmed or the value of the agreement would have a minimal value. Validity of the MAL was expected on the language level sentence – parcelate. On the contrary, as regards the last language level (i.e. paragraph – sentence) it was not presumed that the validity of the MAL would be supported.

---

56 If language units of the contemporary written Chinese are determined on the basis of the graphical principle, the mutual relationships between them exist on respective language levels and their validity is verified by means of the MAL.

Subsequently, segmentation and quantification of the sample texts were implemented according to the determined language units. In order to accept or reject the above mentioned assumptions, it was also necessary to perform a statistical analysis of the acquired data. Based on the implementation of these steps we reached the following conclusions.

Let us begin with the scientific article. Against expectations, the validity of the MAL was confirmed on three language levels. In the case of the language levels character – component and sentence – parcelate the assumption about the MAL as an adequate and well-fitting model was supported. On the contrary, the second assumption about the invalidity of the MAL was disproved because the language level paragraph – sentence yielded unexpected results. Not only was validity of the MAL verified, but also the mathematical model of the MAL showed an extremely high goodness-of-fit. In the case of the last language level parcelate – character the inadequacy of the mathematical model was confirmed as it had been expected.

Next, let us turn to the blog article the analysis of which also yielded unexpected results. The MAL proved itself as an adequate and well-fitting mathematical model only in the case of the lowest language level where the goodness-of-fit reached an extremely high value, as it had been assumed. Regarding the next language level parcelate – character the assumption about the invalidity of the MAL was also proved because the curve had a constant tendency. Contrary to expectations, the decreasing tendency visualizing the relationship between the sentence and the parcelate was only indicated and the value of the goodness-of-fit had a minimal value. Hence our assumption was rejected. Similar findings also appeared in the case of the last language level paragraph – sentence. Contrary to the previous level, this result was expected.

Lastly, let us briefly summarize the results obtained by a comparison of these sample texts with the sample texts examined in the previous experiment (the newspaper article and the short story). The language levels, namely character – component and parcelate – character, incline to similar tendencies, whereas the conclusions concerning the remaining language levels, i.e. sentence – parcelate and paragraph – sentence, differed from each other depending on the stylistic style.

The extremely wide agreement, which was showed on the level character – component within all sample texts, was probably caused by the constant size of the graphic field and its influence on the structure of characters. The acquired results not only accent its key role, but also affirm that the stylistic styles do not have any influence on the validity or invalidity of the MAL on this level.

As regards the language level parcelate – character, the validity of the MAL was not verified in the case of the newspaper article and the scientific article. Although the decreasing tendency was indicated in the short story, the agreement with the mathematical model of the MAL was minimal. In the blog article the curve had even a constant tendency. It can be concluded that the MAL does not represent an adequate and well-fitting model for this language level. Let us consider several factors which could cause this unsuitability. The first of them could be the different characteristic of the language units' lengths: the parcelate represents a unit with a variable length, whereas the character represents a unit with a constant length. The diversification of approaches determining the component on the one hand and the absence of a language unit (which is higher than the character and lower than the parcelate) related to the absence of a language level on the other hand could be considered other factors. Finally, the simplification of the Chinese traditional script could represent the last factor. This reform was implemented in the 1950s and 1960s it resulted in reducing the number of strokes and components within 2,236 characters and thereby reducing their variability.

The results of the remaining language levels differ from one another in connection with the stylistic style in which the sample texts are written. Let us begin by the language level sentence – parcelate. While a wide agreement was noticed in the case of the newspaper article and the scientific article, the short story and the blog article showed a minimal goodness-of-fit. This fact points out that the styles with a regulated structure show a higher agreement. But it should be noted that there are certain reasons which should not be left unconsidered. As on the previous language level, the absence of a language unit (higher than the character and lower than the parcelate) and the little difference in length of the sample texts could also bring about the minimal agreement of the other samples.

Next, let us turn to the last language level paragraph – sentence. A wide agreement was showed only in the case of the scientific article; in the blog article the agreement was minimal and in the remaining sample texts it is not observed at all. The different results could be caused partly by the punctuation, which was influenced by the Western tradition and which was integrated in Chinese texts in relatively recent times, partly by the low frequency of the observations, i.e. by the insufficient length of the sample texts.

The selected key factors which could have a considerable influence on the validity or invalidity of the MAL were tested by sub-experiments.

In the case of the language unit character – parcelate the MAL was applied to the scientific text transformed into the traditional set of characters. According to the assumption the constant size of the graphic field should exert a stronger influence on the traditional Chinese characters because they are characterized by a more complex structure. The extremely wide goodness-of-fit, which was supposed to be higher than in the instance of the simplified set of the characters, was confirmed by this sub-experiment, as well as the crucial role of the graphic field.

Testing the assumption about the absence of another language unit between the parcelate and the character (i.e. a word) was the subject of another sub-experiment. Over the course of the determination of the borders of the parcelate and the sentence it was inevitable to take the syntactic criterion into consideration. In order to maintain the same criteria of segmentation, the MAL verified the existence of the relationship between the syntactic word and the character. The obtained results revealed that the MAL is not an adequate and well-fitting model on this added level. On the basis of this analysis it emerged that the average lengths of characters (measured in components) oscillated independently of the word's length in the short interval between two and three components as well as on the original level parcelate – character.

Other sub-experiments were implemented on the remaining language levels, namely sentence – parcelate and paragraph – sentence. The first one was concerned with the punctuation used in the scientific article. With regard to the disputable function of the semicolon, this punctuation mark was regarded

as the border between sentences. The assumption about its inclination to a full stop was supported by the extremely wide agreement with the mathematical model on both language levels.

The second sub-experiment dealing with the level sentence – parcelate was aimed at the quantitative analysis of a sample text with a greater length. The experiment was based on the assumption that the mutual relationship between the construct and its constituent will be exposed only in the case of testing a sample which was written in the artistic style and whose text includes a larger number of characters (i.e. more than 3,500 characters). Thus the analysis was conducted on the blog article with the length of 8,466 characters. The value of the goodness-of-fit significantly increased, hence the assumption was verified.

As regards this sub-experiment realized on the level paragraph – sentence, the analysis of the longer text did not prove that the MAL is an adequate and well-fitting model on this level; the agreement was even lower than in the case of the original sample text.

The results obtained from these experiments yield a series of questions and subsequently series of suggestions to undertake other experiments. As regards the language level character – component, quantitative analyses will focus on sample texts written in traditional Chinese characters. For this reason, the analyses could examine not only traditional texts published before the simplification of the Chinese script, but also contemporary texts issued in Taiwan (or in Hong Kong and Macau). The purpose of these analyses would be comparing results with conclusions obtained from the already performed experiments and verifying the influence of the graphic field on the mutual relationship between the units on this level. The crucial findings could also reveal the influence of the traditional characters on the higher level parcelate – character.

There is also a need to undertake other subsequent experiments examining the language level parcelate – character. The mutual relationship of these units can vary depending on diverse definitions of the component, therefore, we suggest to apply different approaches to the segmentation of characters into components and to verify them by means of the MAL.



With regard to the fact that the average lengths of characters fluctuate in the range of two to three components, an analysis focusing on 3,000 most frequent Chinese characters and their numbers of components should be performed. The crucial findings would be the number of occurrences of characters consisting of two and three components. In this connection, an analysis aimed at comparison of traditional characters with their simplified forms is also required because it could reveal to what extent the reform reduced differences between the lengths of the simplified characters measured in components. The number of occurrences of the 2-component and 3-component characters within this set will also be a subject of this analysis.

The invalidity of the relationship on this level could also be related to the absence of a language unit higher than the character and lower than the parcellate. We assume that a word could be this absent language unit. For this reason it is necessary to conduct experiments focusing on the perception of this unit in the Chinese language and its precise definition.

In the case of the level sentence – parcellate there is a suggestion to undertake a greater number of experiments testing the mutual relationship of these units in connection with the length of the the sample texts. It is expected that its existence should be confirmed within texts containing a larger number of characters (more than 3,500).

Quantitative analyses of the last language level paragraph – sentence will be particularly applied to texts with a greater length in order to significantly increase the frequency of the obtained observations because a higher frequency could allow observation of the relationships between these units.

The Inadequacy of the MAL which showed almost in all sample texts could be caused by the graphical division based on punctuation marks. It should be noted that punctuation was affected by the Western style and it does not have a long tradition in Chinese texts. Hence, subsequent experiments will observe its influence on the respective units within the context of its development.

Due to testing the influence of the different stylistic styles, an exploration of the mutual relationships of the language units on the levels sentence – parcellate and paragraph – sentence will also be the subject of another experiment.

The last but not least, a quantitative analysis will be applied to sample texts with a sufficient length which could allow the inclusion of another language unit, namely a subchapter.

## Bibliographical references

### Monographies

#### **Benešová, Martina**

2011 *Quantitative Analysis of Text with Special Respect to Fractal Analysis*. Olomouc: University Palacký Olomouc (Dissertation) (in Czech).

#### **Chen, Ping**

1999 *Modern Chinese: History and Sociolinguistics*. New York: Cambridge University Press.

#### **Černý, Jiří**

1996 *History of Linguistics*. Olomouc: Votobia (in Czech).

#### **Hřebíček, Luděk**

1997 *Lectures on Text Theory*. Prague: Academy of Sciences of the Czech Republic.

#### **Hřebíček, Luděk**

2002 *Stories about Linguistic Experiments with Text*. Prague: Academia (in Czech).

#### **Kane, Daniel**

2009 *The Chinese Language: Its History and Current Usage*. Mirošovice: DesertRose (in Czech).

#### **Kučera, Ondřej et al.**

2005 *Textbook of Chinese Characters I*. Olomouc: University Palacký (in Czech).

#### **Lindqvist, Cecilia**

2010 *Empire of Chinese Characters*. Prague: Nakladatelství Lidové noviny (in Czech).

#### **Liu, Yuehua et al.**

2009 *Integrated Chinese: Textbook Simplified Characters, Level 1 Part 1*. 3rd ed. Boston: Cheng & Tsui Company.

#### **Norman, Jerry**

2012 *Chinese*. New York: Cambridge University Press.

**Qiu, Xigui**

2000 *Chinese Writing*. Berkeley: The Society for the Study of Early China.

**Qu, Lei Lei**

2005 *Chinese Calligraphy*. Brno: CP Books (in Czech).

**Švarný, Oldřich et al.**

1967 *Introduction to Spoken Chinese I*. Prague: SPN (in Czech).

**Švarný, Oldřich – Uher, David**

2001 *Spoken Chinese: Introduction to Study of Spoken Chinese*. 2nd ed. Olomouc: University Palacký (in Czech).

**Těšitelová, Marie**

1987 *Quantitative Linguistics*. Prague: SPN (in Czech).

**Vochala, Jaromír – Novák, Miroslav – Pucek, Vladimír**

1975 *Introduction to Chinese, Japanese and Korean Script I: Origin and Development*. Prague: SPN (in Czech).

**Vochala, Jaromír – Hrdličková, Věna**

1985 *Introduction to Study of Sinology: The Philological Part*. Prague: SPN (in Czech).

**Vochala, Jaromír**

1986 *Chinese Writing System: Minimal Graphic Units*. Prague: Charles University.

**Volín, Jan**

2007 *Statistical Methods in the Phonetic Research*. Prague: Epoque (in Czech).

**Wang, Li 王力**

2004 *Hanyu shigao 《汉语史稿》 A Sketch of the History of the Chinese Language*. Beijing: Zhonghua shuju. 714 p. ISBN 7-101-04199-X

**Wang, Ning 王宁**

2002 *Hanzi gouxing xue jiangzuo 《汉字构形学讲座》 Lectures on Structure of Chinese Characters*. Shanghai: Shanghai jiaoyu chubanshe.

**Wang, Xianchun**

2007 *Chinese Running Script Calligraphy for Beginners*. Beijing: Foreign Languages Press.

**Wierger, Leon**

1965 *Chinese Characters: Their Origin, Etymology, History, Classification and Signification*. 2nd ed., enl. and rev. according to the 4th French ed. New York: Paragon Book Reprint.

**Zádrapa, Lukáš – Pejčochová, Michaela**

2009 *Chinese Script*. Prague: Academia (in Czech).

**Articles****Altmann, Gabriel**

1980 "Prolegomena to Menzerath's Law". *Glottometrika* 2, pp. 124–129.

**Andres, Jan et al.**

2012 "Methodological Note on the Fractal Analysis of Texts". *Journal of Quantitative Linguistics* 19 (1), pp. 1–31.

**Havránek, Bohuslav – Horálek, Karel**

1958 "The eighth international congress of linguists in Oslo". *Slovo a slovesnost* 19 (1), pp. 47–52 (available from: <<http://sas.ujc.cas.cz/archiv.php?art=923>>) (in Czech).

**Hřebíček, Luděk**

2007 "Semantic slaps in text structures". *Slovo a slovesnost* 68, pp. 83–90 (available from: <[http://www.kb-old.upol.cz/data/soubor\\_kb\\_807.pdf](http://www.kb-old.upol.cz/data/soubor_kb_807.pdf)>) (in Czech).

**Hřebíček, Luděk**

2008 "Philology versus linguistics: Text constructs higher than sentences". *Vesmír* 87, pp. 488–490 (available from: <<http://www.vesmír.cz/clanek/filologie-versus-lingvistika>>) (in Czech).

**Kelih, Emmerich**

2010 "Parameter-Interpretation of Menzerath's Law: Evidence from Serbian". In: *Text and Language: Structures · Functions · Interrelations · Quantitative Perspectives*. Vienna: Praesens, pp. 71–78.

**LaPolla, Randy J.**

2005 "Wang Li (1900–1986)". In: *Encyclopedia of Language and Linguistics*. London: Elsevier, pp. 514–515.

**Liu, Haitao – Huang, Wei**

- 2012 "Jiliang yuyanxue de xianzhuang, lilun yu fangfa 《计量语言学的现状、理论与方法》. Quantitative linguistics: current situation, theories and methods". *Zhejiang daxue xuebao (renwen shehui kexue ban)* 《浙江大学学报(人文社会科学版)》 42 (2) (available from: <<http://www.journals.zju.edu.cn/soc/CN/abstract/abstract10497.shtml>>, cit. 10. 4. 2013).

**Motalová, Tereza et al.**

- 2013 "An Application of the Menzerath–Altmann Law to Contemporary Written Chinese". *Czech and Slovak Linguistic Review* 3 (1) (in progress).

**Uhlířová, Ludmila**

- 2005 "Quantitative linguistics in the Czech Republic". In: *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook*. Berlin: Walter de Gruyter, pp. 129–135.

## Internet articles

**Abrahamsen, Eric**

- 2012 "Han Han's U-Turn?". In: *IHT Global Opinion* (available from: <<http://latitude.blogs.nytimes.com/2012/01/26/blogger-han-han-controversy-on-democracy-in-china/>>, accessed 16. 4. 2013).

**Elegant, Simon**

- 2009 "Han Han: China's Literary Bad Boy". In: *Time Magazine* (available from: <<http://www.time.com/time/magazine/article/0,9171,1931619,00.html>>, accessed 16. 4. 2013).

**Elegant, Simon**

- 2010 "Han Han". In: *The 2010 TIME 100* (available from: <[http://www.time.com/time/specials/packages/article/0,28804,1984685\\_1984940\\_1985515,00.html](http://www.time.com/time/specials/packages/article/0,28804,1984685_1984940_1985515,00.html)>, accessed 16. 4. 2013).

**Han, Han**

- 2012 "Wo suo lijie de shenghuo 《我所理解的生活》 Life as I understand it". In: <http://blog.sina.com.cn/twocold> (available from: <[http://blog.sina.com.cn/s/blog\\_4701280b0102e7er.html](http://blog.sina.com.cn/s/blog_4701280b0102e7er.html)>, accessed 16. 4. 2013) (in Chinese).

**Han, Han**

- 2012 "Wo de fuqin Han Renjun yiji ta de zuopin 《我的父亲韩仁均以及他的作品》 My father Han Renjun and his work". In: <http://blog.sina.com.cn/twocold> (available from: <[http://blog.sina.com.cn/s/blog\\_4701280b0102e0eu.html](http://blog.sina.com.cn/s/blog_4701280b0102e0eu.html)>, accessed 16. 4. 2013) (in Chinese).

**Li, Yuming 李宇明**

- 2013 "Renshi yuyan de jingjixue shuxing 《认识语言的经济属性》 *The Economic Aspect of Language*". In: *Zhongguo shehui kexue wang: Zhongguo Shehui Kexueyuan zhuban 《中国社会科学院网: 中国社会科学院主办》* (available from: <<http://www.cssn.cn/news/677346.htm>>, accessed 1. 3. 2013) (in Chinese).

**Pilling, David**

- 2012 "Lunch with the FT: Han Han". In: *Financial Times* [online] (available from: <<http://www.ft.com/cms/s/2/3be0e84e-8896-11e1-a727-00144feab49a.html#axzz2QhVk2QHq>>, accessed 16. 4. 2013).

**Watts, Jonathan**

- 2010 "Han Han, China's most popular blogger, shuts down new magazine". In: *theguardian.com* (available from: <<http://www.theguardian.com/world/2010/dec/28/han-han-china-blogger-magazine>>, accessed 17. 11. 2013).

**Zheng, Qingting 郑青亭**

- 2012 "Zhongguo shehui kexueyuan zai yazhou zhiku paimingzhong weiju di-yi 《中国社会科学院在亚洲智库排名中位居第一》 Chinese Academy of Social Sciences is the top think tank in Asia". In: *Renmin wang: Guoji pindao 《人民网: 国际频道》* (available from: <<http://world.people.com.cn/GB/16922861.html>>, accessed 17. 3. 2013) (in Chinese).

**Websites**

- "Beiyu jiaoshou: Li Yuming jiaoshou 《北语教授: 李宇明教授. 北京语言大学》 Professors of Beijing Language and Culture University: A Professor Li Yuming" (© 2006). In: *Beijing yuyan daxue 《北京语言大学》* (available from: <[http://www.blcu.edu.cn/blcuWeb/chinese/professor\\_intro.asp?bh=130](http://www.blcu.edu.cn/blcuWeb/chinese/professor_intro.asp?bh=130)>, accessed 17. 3. 2013) (in Chinese).

- "Biaodian fuhao 《标点符号》 Punctuation marks" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/31516.htm>>, accessed 26. 3. 2013) (in Chinese).
- "Biaodian fuhao yongfa 《标点符号用法》 Basic rules for punctuation usage" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/564500.htm>>, accessed 1. 4. 2013) (in Chinese).
- "Bihua 《笔画》 Strokes of Chinese characters" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/168365.htm>>, accessed 25. 3. 2013) (in Chinese).
- "Category: Chinese bloggers" (2013). In: *Wikipedia, the free encyclopedia* (available from: <[http://en.wikipedia.org/wiki/Category:Chinese\\_bloggers](http://en.wikipedia.org/wiki/Category:Chinese_bloggers)>, accessed 22. 3. 2013).
- "Dunhao 《顿号》 Enumeration comma" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/54969.htm>>, accessed 26. 3. 2013) (in Chinese).
- "Eight principle of Yong" (2013). In: *Wikipedia, the free encyclopedia* (available from: <[http://en.wikipedia.org/wiki/Eight\\_Principles\\_of\\_Yong](http://en.wikipedia.org/wiki/Eight_Principles_of_Yong)>, accessed 31. 12. 2013).
- "Fenhao 《分号》 Semicolon" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/54972.htm>>, accessed 1. 4. 2013) (in Chinese).
- "Han Han 《韩寒》 Han Han" (© 1996–2013). In: *blog.sina.com.cn* (available from: <<http://blog.sina.com.cn/twocold>>, accessed 17. 4. 2013) (in Chinese).
- "Han Han 《韩寒》 Han Han" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/5972.htm>>, accessed 16. 4. 2013) (in Chinese).
- "Hanzi bujian 《汉字部件》 Components of Chinese characters" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/3151455.htm>>, accessed 30. 3. 2013) (in Chinese).
- "Jianghao 《间隔号》 Middle dot" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/55027.htm>>, accessed 26. 3. 2013) (in Chinese).

### **Kong, Qingdong 孔庆东**

- "Dongbo shuyuan – Kong Qingdong de boke 《东博书院——孔庆东的博客》 Classical academy Dongbo – Kong Qingdong's blog" (© 1996–2013). In: *blog.sina.com.cn* (available from: <<http://blog.sina.com.cn/u/1198367585>>, accessed 17. 4. 2013) (in Chinese).



"Kuohao 《括号》 Brackets" (© 2013). In: *Baidu baike* 《百度百科》 (available from: <<http://baike.baidu.com/view/54999.htm>>, accessed 26. 3. 2013) (in Chinese).

### **Li, Chengpeng**

"Li Chengpeng 《李承鹏》 Li Chengpeng" (© 1996–2013). In: *weibo.com* (available from: <<http://www.weibo.com/lichengpeng>>, accessed 17. 4. 2013) (in Chinese).

"Li Yuming 《李宇明》 Li Yuming" (© 2013). In: *Baidu baike* 《百度百科》 (available from: <<http://baike.baidu.com/view/682964.htm>>, accessed 17. 3. 2013) (in Chinese).

"Lianjiehao 《连接号》 Dashes" (© 2013). In: *Baidu baike* 《百度百科》 (available from: <<http://baike.baidu.com/view/55019.htm>>, accessed 26. 3. 2013) (in Chinese).

### **Liu, Mangyan**

"Liu Mangyan yuanchuang boke 《流氓燕原创博客》 Original blog of Liu Mangyan" (© 1999–2013). In: *blog.tianya.cn* (available from: <[http://blog.tianya.cn/blogger/blog\\_main.asp?BlogID=19329](http://blog.tianya.cn/blogger/blog_main.asp?BlogID=19329)>, accessed 17. 4. 2013) (in Chinese).

### **Murong, Xuecong**

"Hulu hulu 《葫芦葫芦》 Hulu hulu" (© 1996–2013). In: *blog.sina.com.cn* (available from: <<http://blog.sina.com.cn/hawking>>, accessed 17. 4. 2013) (in Chinese).

### **Muzi, Mei**

"Muzi Mei 《木子美》 Muzi Mei" (© 2013). In: *sohu.com* (available from: <<http://muzimeiriji.blog.sohu.com/>>, accessed 17. 4. 2013) (in Chinese).

"Pozhehao 《破折号》 Dash" (© 2013). In: *Baidu baike* 《百度百科》 (available from: <<http://baike.baidu.com/view/55016.htm>>, accessed 26. 3. 2013) (in Chinese).

### **Ran, Yunfei**

"Ran Yunfei boke: feihua lianpian 《冉云飞博客：匪话连篇》 Ran Yunfei's blog: Thiefs' slang" (© 1997–2013). In: *www.163.com* (available from: <<http://tufeilaoran.blog.163.com/>>, accessed 17. 4. 2013) (in Chinese).

### **Rao, Xueman**

"Rao Xueman 《饶雪漫》 Rao Xueman" (© 1996–2013). In: *weibo.com* (available from: <<http://weibo.com/raoxueman>>, accessed 17. 4. 2013) (in Chinese).

"Shengluohao 《省略号》 Ellipsis" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/55004.htm>>, accessed 26. 3. 2013) (in Chinese).

"Shuminghao 《书名号》 The titles marks" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/55021.htm>>, accessed 26. 3. 2013) (in Chinese).

### **Sima, Nan**

"Sima Nan 《司马南》 Sima Nan" (© 1996–2013). In: *weibo.com* (available from: <<http://weibo.com/simanan>>, accessed 17. 4. 2013) (in Chinese).

### **Wang, Keqin**

"Wang Keqin 《王克勤》 Wang Keqin" (© 2013). In: *sohu.com* (available from: <<http://wangkeqin.blog.sohu.com/>>, accessed 17. 4. 2013) (in Chinese).

"What is R?: Introduction to R" (2013). In: *The R Project for Statistical Computing* (available from: <<http://www.r-project.org/>>, accessed 21. 3. 2013).

"Wo yuan gaikuang 《我院概况》 About our academy" (2010). In: *Zhongguo Shehui Kexue wang: Zhongguo Shehui Kexueyuan zhuban 《中国社会科学网: 中国社会科学院主办》* (available from: <<http://www.ccsn.cn/news/140195.htm>>, accessed 17. 3. 2013) (in Chinese).

"Xiang de shufa 《象的书法》 Calligraphy of the character for 'elephant'" (© 2004–2012). In: *andian shufa 《汉典书法》* (available from: <<http://sf.zdic.net/shufa/1123/5537f8fb44fcabc5ac9a353d5b5be03.html#ks>>, accessed 21. 3. 2013) (in Chinese).

"'Xiang' zi de jiben xinxi 《'象'字的基本信息》 Basic information about the character for 'elephant'" (© 2004–2012). In: *Handian 《汉典》* (available from: <<http://www.zdic.net/zd/zi/ZdicE8ZdicB1ZdicA1.htm>>, accessed 21. 3. 2013) (in Chinese).

"Xin de shufa 《心的书法》 Calligraphy of the character for 'heart'" (© 2004–2012). In: *Handianshufa 《汉典书法》* (available from: <<http://sf.zdic.net/shufa/0815/3419eb9cbb79b0bdd53bce2b45d4e1f0.html#ks>>, accessed 21. 3. 2013) (in Chinese).

"'Xin' zi de jiben xinxi 《'心'字的基本信息》 Basic information about the character for 'heart'" (© 2004–2012). In: *Handian 《汉典》* (available from: <<http://www.zdic.net/zd/zi/ZdicE5ZdicBFZdic83.htm>>, accessed 21. 3. 2013) (in Chinese).

**Xu, Jinglei**

- "Xu Jinglei 《老徐》 Xu Jinglei" (© 1996–2013). In: *blog.sina.com* (available from: <<http://blog.sina.com.cn/xujinglei>>, accessed 17. 4. 2013) (in Chinese).
- "Yinhao 《引号》 Quotation marks" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/54994.htm>>, accessed 26. 3. 2013) (in Chinese).
- "Yu de shufa 《雨的书法》 Calligraphy of the character for 'rain'" (© 2004–2012). In: *Handian shufa 《汉典书法》* (available from: <<http://sf.zdic.net/shufa/0611/7062ebdb5d8db69266b442dd9e37d89b.html#ks>>, accessed 21. 3. 2013) (in Chinese).
- "'Yu' zi de jiben xinxi 《'雨'字的基本信息》 Basic information about the character for 'rain'" (© 2004–2012). In: *Handian 《汉典》* (available from: <<http://www.zdic.net/zd/zi/ZdicE9Zdic9BZdicA8.htm>>, accessed 21. 3. 2013) (in Chinese).
- "Yuan jigou 《院机构》 Academy structure" (2011). In: *Zhongguo Shehui Kexue wang: Zhongguo Shehui Kexueyuan zhuban 《中国社会科学院网: 中国社会科学院主办》* (available from: <<http://www.cssn.cn/news/141052.htm>>, accessed 17. 3. 2013) (in Chinese).
- "'Yuyan wenzi yingyong' Bianjibu (Yingyong yuyanxue yanjiu zhongxin) 《〈语言文字应用〉编辑部(应用语言学研究中心)》 'Applied Linguistics' – Editorial Department of Applied Linguistics' research centre" (2005). In: *Jiaoyubu yuyan wenzi yingyong yanjiusuo 《教育部语言文字应用研究所》* (available from: <<http://www.yys.ac.cn/80/publish.htm>>, accessed 17. 3. 2013) (in Chinese).

**Zhang, Weiwei**

- "Vivibear de BLOG 《Vivibear 的 BLOG》 Viviber's BLOG" (© 1996–2013). In: *blog.sina.com* (available from: <<http://blog.sina.com.cn/vikibear333>>, accessed 17. 4. 2013) (in Chinese).
- "Zhongguoshehuikexueyuan 《中国社会科学院》 Chinese Academy of Social Sciences" (© 2013). In: *Baidu baike 《百度百科》* (available from: <[http://baike.baidu.com/view/33292.htm#refIndex\\_1\\_33292](http://baike.baidu.com/view/33292.htm#refIndex_1_33292)>, accessed 17. 3. 2013) (in Chinese).
- "Zhuanminghao 《专名号》 Proper name mark" (© 2013). In: *Baidu baike 《百度百科》* (available from: <<http://baike.baidu.com/view/297482.htm>>, accessed 26. 3. 2013) (in Chinese).

“Zhuozhonghao 《着重号》 Emphasizing dot” (© 2013). In: *Baidu baike* 《百度百科》 (available from: <<http://baike.baidu.com/view/55028.htm>>, accessed 26. 3. 2013) (in Chinese).

## Norm

GB/T 15834 – 2011. *Zhongguo renmin gongheguo guojia biao zhun: Biaodian fuhao yongfa* 《中华人民共和国国家标准：标点符号用法》 National Standards of People's Republic of China: General rules for punctuation (2012). Beijing: Zhongguo biao zhun chubanshe (in Chinese).

## Software

Wenlin Institute, Inc. 文林 Wenlin Software for Learning Chinese [software]. Version 4.0.2. Wenlin Institute, Inc. Copyright © 1997–2011.

## Appendix

### APPENDIX 1: THE SCIENTIFIC ARTICLE – SAMPLE A

#### 认识语言的经济学属性

李宇明

2013-2-27 15:47:01 来源:《语言文字应用》(京) 2012 年 3 期

**【作者简介】**李宇明,北京语言大学教授,主要研究领域为语言学理论、现代汉语、心理语言学和语言规划(北京100083)。

**【内容提要】**语言与经济的关系非常紧密,但在以往的社会语言意识中,语言的经济学属性并没有得到清晰的认识。在人口流动和信息化两大驱动力的推动下,语言对经济的贡献越来越显著,以至于成为不容漠视的经济现象。有些国家,语言对经济的贡献度竟然达到10%。我国语言经济学研究已经起步,而且具有光辉的发展前景。语言经济学的重要任务是,认识语言在经济活动中的作用,认识语言经济的运行规律,研究语言对社会的经济贡献度,研究语言政策的成本及其产生的经济效益,探讨促进语言经济发展的政策环境和各种举措,发展语言产业,培育语言职业,促进语言消费,使国家和个人充分赚取语言红利。

Language is closely tied to economy; however, the economic attributes of language, in the sociolinguistic awareness so far, have not been clearly understood. Population mobility and information technology have been two driving forces that make language an important contributing factor of economic growth. Language, therefore, becomes a critical economic matter. In some countries, language accounts for about 10 % in economic growth. In China, the study of economic linguistics has just started and foresees

a promising future. The important task of the economic linguistics is to understand the role of language in economic activities, and to find out the operational rules of language economy, and to study the economic contribution of language for the society as well as cost-effectiveness of language policies. This paper will explore the policies and methods of promoting the development of economic linguistics in the hope of developing language industry, nurturing language professions, and promoting language consumption to the benefit of the country and individuals alike.

**【关键词】**语言/经济 language/economy

现代语言规划十分关注的三个概念是语言意识、语言政策和语言实践。语言意识也称语言意识形态,是指社会对语言的认识和态度,是语言政策和语言实践的思想基础,有什么样的语言意识,才可能有什么样的语言政策,产生什么样的语言实践。在当今社会的语言意识中,必须意识到语言的经济属性,从而在制定语言政策时自觉进行经济学的考量,并制定出在经济活动中能够充分发挥语言作用的政策。

### 一、语言是经济活动中不可缺少的要素

人类的经济活动与语言密不可分,而且在某些领域,语言和语言知识已经成为重要的经济资源。这可以从以下几个方面来看:

首先,语言能力是劳动力的重要构成要素。语言是人的本质属性之一,是人类最为重要的交际工具和思维工具,语言能力与劳动能力总体上呈正相关。比如:1.由于各种病理原因而失去语言能力的人,如失语症患者、聋哑人等,是劳动水平相对较低、适应性较差的弱劳动力。2.文盲只有口语能力,没有书面语能力,在脑力劳动在人类的经济活动中的比例逐渐上升的时代,文盲已经成为质量较低的劳动

力,很多经济活动都无缘参与。3.人的语种能力总体上看与收入相关,具有单语能力的人,与具有双语能力、三语能力的人相比,其经济收入总体上要低。

在一些特殊的劳动阶层和经济领域,语言能力的地位会更加重要。比如工程设计、广告策划、劳动管理等,文盲等没有书面语言能力者,是无法承担的;语言艺术家、电视节目主持人、导游、导购、公司售后服务人员等,都需要较高的语言能力;同声传译、国际组织雇员等工作领域,与单语人无缘。

在以体力劳动为主、社会分工较粗的时代,对劳动者的语言能力要求不高;但是到了信息化时代,社会分工急剧加细,脑力劳动的比重急剧加大,跨地区、跨国家的经济活动急剧增多,劳动者培养的时间越来越长,语言能力在劳动力构成中的比重也急剧增大。在每年大学毕业季的用人单位的招工要求中,虽不见得出现语言能力的字眼,但是仔细分析,关于语言能力的要求其实成了用人单位的主要考虑因素。语言能力在今天已经成为劳动力的重要构成要素,教育部门和劳动培训机构对此应有清醒认识。

其次,经济活动需要通过语言来组织进行。语言是信息最为重要的载体,人类的生活须臾离不开语言,离开语言社会就将崩溃;经济活动是人类社会生活的重要部分,经济活动需要通过语言才能组织起来。正因如此,许多经济学家、语言学家都比较重视经济学术语的规范,重视经济学文本的修辞,并由此形成了最早的经济语言学。经济语言学主要关心的是经济活动中语言使用的得体与效率,当然也试图利用经济学规律来解释语言现象。

再次,语言和语言知识已经成为重要的经济资源。在一些特殊的经济活动领域,语言及其知识甚至具有“生产资料”的性质。比如语言教育活动,语言知识、语言教育方法是教师的资本,通过语言知识的传授和语言训练,学生获得语言知识,并内化出语言能力。语言教育活动还需要有一系列的保障和评估系统,例如语言教学设施、教

科书、工具书、各种录放语言的设备、衡量学生学习水平的考试活动及有关证书等，这些保障和评估系统，都涉及语言教育的经济问题。当然，语言教育有事业和产业之分，语言教育事业需要成本投入，涉及经济问题；语言教育产业则主要是经济活动。

再如语言文字艺术活动，其基本凭借是语言和文字，其成果是语言文字艺术产品。语言文字艺术可以粗分为语言艺术和文字艺术，小说、诗歌、话剧、相声、评书等是语言艺术，它们通过对语言的艺术运用形成艺术语言。艺术语言一般都是有声语言，或可以成为有声语言，并可以同其他方式结合形成戏曲、电影等综合艺术。文字艺术最为典型的是书法。除此之外，还有其他文字艺术，如将文字变形而形成的别有韵味的“文字图画”“文字雕塑”；如用文字或汉语拼音设计的商标、图案等等；如将金银珠宝镂雕成文字形状，或是将文字及其变形刻附在金银珠宝上。语言文字艺术不仅具有艺术价值，而且也往往具有较高的经济价值，形成附加值很高的语言文字艺术产业。

最值得关注的是，现代语言技术的发展所形成的现代语言经济。“语言技术”这一概念提出的时间不长，但语言技术已经有悠久的历史。文字的创制是古代最为重要的语言技术，口头语言借此打破了时间和空间的限制，可以传后达远。之后的笔墨纸砚、印刷术、打字机、电报、电话、传真、留声机、录音机、广播等等，都是至今仍用的最为重要的语言技术。现代语言技术是用计算机处理语言文字所形成的一系列技术，它使语言知识及其应用有可能成为工业标准，成为语言技术产品，形成各种专利，比如语言文字的各种规范标准，各种语言数据库，各种键盘输入法，各种处理语言文字的软件和计算机字库等等。现代语言技术使人类语言知识成为十分重要的经济资源。

## 二、语言对社会的经济贡献

但是，由于语言与人、与经济活动的关系太密切，密切到人们很少关注语言的经济意义。正如空气、阳光是人类片刻不能离开的，



然而在人类相当长的历史长河里，它们都没有成为商品；只有到了空气污染严重、阳光常被遮蔽之地，它们才可能具有商品价值，例如现在的一些楼盘推销商，会将楼盘所在地的空气质量、房间的向阳状况等包装为卖点，空气和阳光在这里具有了商品意义。

语言的经济属性较早引起人们的有意识关注，是语言能力与个人收入之间的关系。随着国际经济一体化进程的不断加快，人口流动的半径加长、规模加大、频率提高，移民成为当今社会的重要现象，并引发出一系列问题。国外很多研究发现，移民的语言能力同其就业状况、经济收入水平呈正相关。此类研究开垦了语言经济学的处女地。20 世纪末期，信息化浪潮奔涌而起，狂飙突进，语言是计算机处理的主要对象，语言信息化是信息化的基础，语言、语言知识、语言技术等等，成为高新经济的重要资源和重要增长点，许多信息产业都可以视为语言产业。

人口流动和信息化，是促进语言经济发展的两大驱动力，也是促使人们关注语言的经济属性的两个重要方面。当今世界，促进语言经济发展的这两大驱动力更加强劲，人口流动几乎成为社会常态，信息化以加速度的态势发展，可以预见，语言对社会的经济贡献将会持续加大，语言的经济属性将会不断彰显，因此，社会语言意识中必有语言经济的一席之地。

当带着语言经济的眼光来观照人类生活时，蓦然发现语言经济对人类社会是如此的重要。日内瓦大学弗朗斯瓦·格林 (François Grin) 教授的研究表明，瑞士语言的多样性，为瑞士每年创造 500 亿瑞郎的收入，约占瑞士国内生产总值的 10%。瑞士有德语、法语、意大利语、罗曼山地语四种语言，瑞士公民一般都能够掌握三种语言，综合院校的大学生需要掌握四种语言。瑞士还是联合国、欧盟以及很多国际组织的所在地，有许多语言资源可以利用。瑞士经济的发展，语言在里面占了很大的比重。10% 的说法可靠与否，由经济学家去判断，但是瑞士的经济发展跟它丰富的语言资源确有关系。

世界许多国家都开始注意语言产业的发展,并不断有相关的报告问世。据报道,美国的命名产业 1999 年的年产值就达到了150 亿美元。有人估计,全世界翻译市场年产值可达1万亿元人民币;全球英语教育市场,除大学和政府培训机构外,约有 600 亿美元的规模。

《2009 年欧盟语言行业市场规模报告》是笔者见到的当前最为全面的语言行业状况报告。该报告指出:2008 年欧盟成员国的语言市场总产值达 84 亿欧元,其中语言技术领域的产值为 5.68 亿,电影字幕和配音领域为 6.33 亿,语言教学领域为 16 亿,会议组织中的多语言服务为 1.43 亿。该报告预测,欧盟 2015 年语言行业的实际产值可达 200 亿欧元。

我国正处在人口大流动时期,上亿农村人口向城市流动,劳动力国外输出渐成规模,学生出国留学人数与日俱增,境外回内地、海外来中国的学习者、投资者、工作者也逐渐增多。我国语言教育产业、语言翻译产业拥有强大活力,据统计,仅英语学习市场年产值已超过 100 亿元人民币,翻译市场年产值约 120 亿元人民币。我国信息化的发展日新月异,2012 年网民即将突破 5 亿,手机用户 9 亿,其中手机上网用户 3.5 亿;信息产品的社会普及率速度惊人,语言信息产业具有巨大的发展潜力。

由于语言经济学刚刚兴起,人们对语言经济的业态状况还不怎么了解,语言经济的数据采集系统尚未建立,甚至缺乏有效的语言经济计算方法,因此,还无法对语言经济进行全面而科学的统计。但是仅上面的事例已经颇具魅力,由此已可窥见语言经济的巨大社会意义,感受到认识语言的经济学属性的巨大社会意义。

### 三、促进我国的语言经济学研究

与语言经济的社会实践相比,我国对语言经济的认识显然滞后。但是近些年来,学界、业界和政界已开始关注语言经济问题,呈现出良好的发展势头。

山东大学黄少安教授是我国语言经济学研究的首倡者，他率先招收语言经济学的博士研究生，他的团队连续召开中国语言经济学论坛，并成立了语言经济研究中心。2012年3月2日，黄少安、苏剑、张卫国三位发表的《语言经济学与中国的语言产业战略》，基本上代表了我国学界在语言经济方面的认识。南京大学中国语言战略研究中心徐大明教授，带领他的团队在研究语言政策、语言国情的同时，开展了语言经济学的理论与实践研究，成果引人注目。广州大学屈哨兵教授对语言服务保持着浓厚的研究兴趣，今年初还举办了“语言服务研究高级论坛”。

国家语委是专司语言事务的国家部门，近些年来在工作实践中发现语言经济问题十分重要，认识到语言不仅是国家的“软实力”，而且也是“硬实力”。国家语委全力支持山东大学、南京大学、广州大学等高校的语言经济学研究，还于2008年12月29日支持商务印书馆成立了“中国语言资源开发应用中心”，中心的宗旨是“致力于把语言及语言知识转化为生产力和文化商品”。《国家语委“十二五”科研规划2011年项目指南》，首次把“语言经济与语言产业发展战略研究”列为重要的科研方向，表明语言经济问题已经开始进入国家的语言规划。在2010、2011年北京市“两会”上，北京市人大代表贺宏志先生连续提出《关于发展我市语言产业的建议》和《加强语言文化建设，促进语言产业发展》的建议，语言经济的话题首次提到了地方人民代表大会的论坛上。

2010年9月28日，在国家语委支持下，北京市语委的研究基地——“北京语言产业研究中心”，在首都师范大学揭牌成立。这本来是一个平常的日子，但在中国语言规划史上也许会不平常，因为这是我国第一个以“语言产业”为专门研究对象的科研机构。该中心建立伊始，就着手调研北京语言产业的发展状况，建立北京语言产业数据库，思考语言产业的发展政策；据说还筹划举办我国第一届语言产业论坛，编辑出版《语言产业研究通讯》，翻译《2009年欧盟语言行业市场规模报告》，出版《北京语言产业的发展及政策研究》等。理论与实践并重，引进与创新同举，近期与长远共谋。令人欣喜的

是，在北京语言产业研究中心成立仅一年多的时间里，就撰写了我国第一部专门研究语言产业的著作《语言产业导论》。

《语言产业导论》虽不能说字字珠玑，但却处处闪光。特别是关于语言产业的分类、语言产业的要素分析、语言产业的业态梳理及其案例征引等等，读起来引人入胜，不忍释卷。这部著作给人许多新理念、新知识、新数据，而且能够引发读者许多质疑与思考，诸如：语言真能生钱吗？语言产业真的存在吗？哪些职业是语言职业？怎样统计语言对社会的经济贡献度？哪些消费属于语言消费？语言消费与语言产业、语言职业的发展有何关系？个人如何收取语言收益？国家如何赚取语言红利？……只提供知识的著作，培养的是被动消费型的读者，只能给社会以精神填充；而能够引发质疑与思考的著作，培养的是主动创造型的读者，能够促进社会的精神再生产，甚至引发社会行动。

尽管有如上的进展，但我国的语言经济学研究还处在起步阶段，研究工作需要一个案例一个案例地收集，一步一步地前进。不过宏观上看，语言经济学的研究应当完成两项基本任务：

第一，研究语言对社会的经济贡献度。这需要有一个科学的研究框架，需要广泛搜集与语言相关的经济领域的的数据，通过大数据的统计分析得出结论。当前的最大困难是，我们还没有一个有效的计算语言经济的方法，哪些产业属于语言产业？如何计算语言产业的经济效益？在非语言产业中语言是否能够产生经济效益？如果能够产生经济效益的话，如何计算非语言产业中因语言而产生的经济效益？当然，当前也无法从国家现有的统计数据中得到语言经济方面的数据，因为统计口径中没有语言经济。

第二，研究语言政策的成本及其产生的经济效益，探讨促进语言经济发展的政策环境和各种举措。过去对语言政策的评估，主要侧重于社会效益；当认识到语言的经济属性之后，还应当对语言政策进行经济学的评估。比如我国民族地区的双语教育，从语言经济学的角度看，需要多大的经济投入？这一政策将为民族地区的经济发展带

来多大效益?掌握双语的学生的经济收入将会比单语学生有多大提高?进而可以改善多少民族家庭的经济状况?再如,我国有一系列促进经济发展的政策,对这些政策应当进行语言经济学的考察,看看是否有利于促进语言经济的发展,应当做哪些必要的政策补充,或是在这些政策环境中设计出促进语言经济发展的各种举措。

#### 四、结束语

在当今社会,不包含语言的经济学属性的意识,不是与时代契合的语言意识。在语言经济可能影响到10%的经济生活的今天,社会必须树立清晰的语言经济意识,仔细观察语言经济活动,全面收集语言经济数据,认识语言经济的运行规律,发展语言产业,培育语言职业,促进语言消费,使国家和个人充分赚取语言红利。

最后我想引用林毅夫先生的一段话:“这些年来我在各种场合经常讲中国经济现象是经济学研究的一个大金矿,研究中国的经济问题有可能产生一批世界级的经济学大师。我的信念源自经济理论的作用在于解释经济现象,其贡献的大小由所解释的现象的重要性决定。……经济学成为一门独立的社会科学以来,世界级的经济学家大多先后产生于作为世界经济中心的英国和美国。中国的经济规模很有可能在21世纪30年代超过美国,中国将有可能逐渐成为一个新的领导经济学思潮的国际中心。如果我的乐观预测是正确的,中国经济学界的第一个诺贝尔经济学奖获得者,最有可能是来自于从事制度经济学研究的经济学家。”<sup>①</sup>如果林毅夫先生的推断是正确的话,我国在语言经济学的研究上也完全有可能后起居上,因为我国有十分丰厚的语言经济学研究资源。

本文原是《语言产业导论》的序言,这次发表在内容上作了一些补充。

注释:

<sup>①</sup>林毅夫《“当代制度分析前沿系列”总序》,见鲁宾斯坦(2004)。

**【参考文献】**

- [1] 薄守生. 语言规划的经济学分析[J]. 制度经济学研究, 2008, (2).
- [2] 蔡辉. 语言经济学: 发展与回顾[J]. 外语研究, 2009, (4).
- [3] 陈章太. 语言资源与语言问题[J]. 云南师范大学学报(哲学社会科学版), 2009, (4).
- [4] 贺宏志, 陈鹏. 语言产业导论[M]. 北京: 首都师范大学出版社, 2012.
- [5] 黄少安, 苏剑, 张卫国. 语言经济学与中国的语言产业战略[N]. 光明日报, 2012-03-02.
- [6] 江桂英. 中国英语教育: 语言经济学的视角[M]. 厦门: 厦门大学出版社, 2010.
- [7] 李现乐. 语言资源与语言经济研究[J]. 经济问题, 2010, (9).
- [8] 李现乐. 语言服务与服务语言——语言经济视角下的语言应用研究[D]. 南京大学博士学位论文, 2011.
- [9] 李宇明. 中国语言规划论[M]. 北京: 商务印书馆, 2010a.
- [10] 李宇明. 中国语言规划续论[M]. 北京: 商务印书馆, 2010b.
- [11] 李宇明. 语言也是“硬实力”[J]. 华中师范大学学报, 2011, (5).
- [12] 林勇, 宋金芳. 语言经济学评述[J]. 经济学动态, 2004, (3).
- [13] 刘国辉, 张卫国. 语言经济学在中国的发展: 2009(首届)中国语言经济学论坛综述[Z] <http://weiguozhang.blog.sohu.com/138597516.html>, 2010-04-07/2010-05-12.
- [14] 刘建达. 试论语言中的经济现象[J]. 山东外语教学, 1989, (3).
- [15] 鲁宾斯坦. 经济学和语言(钱勇、周翼译)[M]. 上海: 上海财经大学出版社, 2004.
- [16] 马慈君. 语言经济学视野下的大学英语教育[J]. 云南民族大学学报(哲学社会科学版), 2010, (1).
- [17] 宁继鸣. 汉语国际推广: 关于孔子学院的经济学分析与建议[D]. 山东大学博士论文, 2006.
- [18] 屈哨兵. 语言服务研究论纲[J]. 江汉大学学报, 2007, (6).
- [19] 宋金芳, 林勇. 语言经济学的政策分析及其借鉴[J]. 华南师范大学学报(社会科学版), 2004, (6).

- [20] 汪丁丁. 语言的经济学分析[J]. 社会学研究, 2001, (6).
- [21] 王世凯. 略论我国语言资源的开发与利用[J]. 云南师范大学学报(哲学社会科学版), 2010, (5).
- [22] 王振顶. 汉语国际传播的语言经济学研究[J]. 云南师范大学学报(对外汉语教学与研究版), 2009, (6).
- [23] 徐大明. 有关语言经济的七个问题[J]. 云南师范大学学报(哲学社会科学版), 2010, (5).
- [24] 许其潮. 语言经济学: 一门新兴的边缘学科[J]. 外国语, 1999, (4).
- [25] 杨依山. 语言经济学理论框架初探[J]. 山东社会科学, 2007, (10).
- [26] 张卫国, 刘国辉, 陈屹立. 语言与收入分配关系研究评述[J]. 经济学动态, 2007, (7).
- [27] 张卫国. 语言的经济学分析: 一个初步框架[D]. 山东大学博士学位论文, 2008.
- [28] 张忻. 语言的经济学与大学英语教育[J]. 中南大学学报(社会科学版), 2008, (3).
- [29] Coulmas, F. *Language and Economy*, Oxford: Blackwell Publishers Ltd, 1992.
- [30] Grin, F. The economics of language: Match or mismatch? *International Political Science Review*, 1994, 15, 27~44.
- [31] Grin, F. The economics of language: Survey, assessment, and prospects, *International Journal of the Sociology of Language*, 1996, 121, 17~44.
- [32] Grin, F. European research on the economics of language. *Homepage of Languages and the Economy: Canada in an International Perspective*, 2000, 11/13/2000.
- [33] Grin, F. Language planning and economics. *Current Issues in Language Planning*, 2003, 1, 1~66.
- [34] McManus. Labor market costs of language disparity: An interpretation of Hispanic earnings differences, *The American Economic Review*, 1985, 4, 818~827.
- [35] Mettewie, L. & L. van Mensel. Multilingualism at all costs: Language use and language needs in business

in Brussels, Sociolinguistic 23: Language Choice in European Companies, 2009.

- [36] Vaillan-court F. The economics of language and language planning. Language Problems and Language Planning, 1982, 2, 162~178.

责任编辑: 隋萌萌



## APPENDIX 2: THE BLOG ARTICLE – SAMPLE B

### 我所理解的生活

多天前参加比赛，来了一个久未见面的朋友。他现在的工作是在给明星做经纪。整个周末他就在我们车队的帐篷里。周日分别，他对我说，其实我的自我开发做的并不好，形象管理有问题，如果有职业的经纪人可以打理一下，必然远不是今天的模样。这样，回去给你一个总结的邮件。

刚才他打来了电话，说你问题太多，邮件说不清楚，比如你在比赛那天一直双反你知道么？我当时就晕了，我只知道双规和单反，双反我真不知道。弄半天才明白，所谓双反，原来是衣服穿反了，而且内外和前后都反了。我说我出门太急，真没注意，也没人提醒我，难怪一整天觉得脖子有点勒。

朋友说这个问题不大，你本来就粗心，但是容易被人取笑，但要命的是，你在车队帐篷的沙发上乱睡觉，你睡觉的时候总共有十二个人来拍过你的睡姿，五个是挂记者证的，四个是车队成员，三个是其他车手，其中有两个是故意拍丑态大头照的。我查了一下，其中五个人发微博了。有一张照片很难看，影响形象，你身边也没人拦着人家拍照，这在我们这行里是绝对不允许的。我说这我也没办法，熬夜看欧洲杯，的确睡眠不够，你教我怎么才能睡的玉树临风？

朋友继续教育我，面部表情是其次，关键是你团着身子睡，手还一直塞在你自己的裆部，这个猥琐的动作绝对是破坏形象的，照片如果上传，有些网友看见了容易反感。我说我又没把手塞在那些网友的裆部，我碰了自己的鸟，关他们鸟事。反感就拉到呗。朋友说不是这样的，你是一个公众人物，现在又是微博的时代，谁都能随手拍，越夸张的传播越快，你要确保自己的每一张照片不能影响你的形象，比如你那个手放的位置不对，很容易被一下转发数千条。

我说这我实在没办法，空调温度太低了，只要一冷，我就自动睡成捂裆派了，从小就这样。总不能我睡觉，雇几个保镖拦着不让拍照，这也太装逼了。

朋友还指出了一堆问题，比如随意让人合影，人家递过来什么都签，会留下隐患。我说不，人家如果递过来一百块人民币我就不签。朋友肯定道，不错，你还算有这个意识，我们行业里有明星在递过来的钞票上签字的，结果被网友骂死。破坏人民币肯定不好。我说不是的，是因为我不想把自己的名字和老毛放在一起。

朋友痛心疾首道，你看，你这种话又不能乱说，得罪的人太多。你在车队聊天也是这样，什么都说，而且常出脏话，你要知道，如果现场有一个不怀好意的，把你说的那些用手机记录下来，放到网上，是很大的负面新闻。你知道当时帐篷里多少人，十八个，你都认识么？我回答说有几个不认识。朋友听筒差点掉地上：有几个不认识你就那么说话？你考虑过后果么？你一睡醒就和人合影，有一撮头发翘的跟天线宝宝似的，人家还开着闪光灯，照片效果可想而知。你看你衣服的配色，是很乡土的，最关键，再不拘小节，裤子拉链还是要拉上的。总之，你太随便了，也没有一个专业点经纪人帮你，你如果不严格的对自己的形象进行系统的管理，就不能保持神秘感和名人的气质。你如果对自己有一个好的定位，有合适的人帮你运作和服务，调整一下你的社交圈子，你能赚的远比现在多很多。你告诉我，你打算怎么经营自己？你是怎么想的？

我说，整个周末只想一个问题，我和对手差了零点三秒，我该怎么追回来？我能惦记着出门要穿衣服已经不错了，哪还顾得上搭配。

挂了电话，夜深人静，回想朋友所言，有些也对。我在帐篷里口无遮拦，对所有人没有设防，要是真有人偷拍偷录或者微博直播的，的确会惹麻烦。对陌生人的提防与否取决与你的出厂原始设定，我喜欢先把人设定成好人，再从中甄别坏人，有些人反之。但所谓的甄别方式其实就是被坑一次。我相信以诚相待，也相信倒霉认栽。

至于衣着，这个夏天我就买了十件白体恤，以往冬天我也就两件黑皮衣，鞋子就盯着那么一两双穿，我是去比赛的，又不是去比美的，赛车开砸了我在乎，衣服穿难看了我真不在乎。

我所理解的生活就是除了造谣以外，去造其他一切东西。我心中的造化，就是你创造了多少文化。既然30年前，我以一挑亿，跑赢了其他所有的精子，那么我必然生来就是牛逼的，我来到这个世上，总要留下点痕迹。我承认衣着光鲜，举止优雅也是一种对美好的创造，但这方面我不太拿手。我承认这个社会，很多人觉得你只要不说脏话，说点假话，空话，套话，造谣，大肆的造谣，疯狂的造谣都不算是道德败坏，但我觉得反之，并且还要对道貌岸然的傻逼们加一个操字，是的，这会让那些道德惊诧家们浑身颤抖，严厉批判，大呼小叫，满地打滚，然后突然起立，开始审判，解决之道就是再说一次，操。我就操了，但我既不操你也不操你全家，我操这世道，这世道觉得文绉绉的诬陷没问题，这世道让那些不说粗话但最缺德的人能做道德评判家，这世道让那些话不脏但心眼脏手段脏的人当道，这世道能任意颠倒黑白混淆是非，这世道觉得公众人物或者随便谁说一个操字就不应该，那就操翻这世道。

我所理解的生活就是做着自己喜欢的事情，养活自己，养活家人。生活它不是攀爬高山，也不是深潜海沟，它只是在一张标配的床上睡出你的身形。我也不觉得留有遗憾是一种缺憾美，相比之下，干砸了倒是一种美。我喜欢的事情远不止写点东西和赛车，我还做很多事，有些做的不够好，有些做的很失败，和朋友聊天时，我直接告诉他们，这事我特喜欢，也干过，但我真的不适合，丢人了。我就最讨厌听见有人这么说，要是我去干这事，一定比某某某干的好。滚。你在台面上看见我成功一次，我在台面下就干砸十次，那又如何，我又没死，不停的干就行了，人们只会记住你成功的那一次。

我所理解的生活就是和自己喜欢的一切在一起，我曾经在快餐厅看上一个姑娘，犹豫五分钟，没敢去和人家说话，结果人家走了，我到现在都很遗憾。在那一刻，我就是白痴，我去了又如何，最坏的

结果无非就是他男朋友从厕所里出来。哪天若要死了，遗憾这事没干，那事没干，还不如自吹这事干成了，自嘲那事干砸了。我现在干的事足够多，陪伴家人爱人和孩子，每年比赛接近二十场，又开始写新的小说和游记，除了偶然进棚拍杂志，其他时间真没有精力来捋伤自己，更没心思去考虑什么形象和定位的问题，觉得我观感欠佳的，挪步就是，我只负责制造作品，不负责用户体验，也没有售后服务，更不会根据大家的口味来改进。你若喜欢，便是晴天，你若讨厌，也是晴天。谢谢这位朋友给我的忠告和精心的设计，我知道我会为我的性格和生活方式吃无数亏，吞无数恶果，但至少大到理想，小到闪念，我几乎都没有放过，所以就算我的生活里充满挫败甚至后悔，但遗憾并不多。朋友，感谢你所说的一切，世间万千种宠爱，无数种人心，得之我幸，不得我也没什么不幸。但我只认可一点，就是出门再匆忙，裤子拉链还是得拉好。

## APPENDIX 3: THE BLOG ARTICLE – SAMPLE C

### 我的父亲韩仁均以及他的作品

今天一大早，我的父亲给我电话，说写了一篇文章来说明一下，问我这样写能不能把事情说清楚了。我觉得特别的凄凉。一开始，他们说我有团队，并重金鼓励网友举证，结果千万网友中没有人能举证出身边的亲朋好友属于我的写作团队，于是他们又说金波是我少年成名的推手，结果发现金波 98 年的时候还在河南的一个罐头厂工作。他们最后的一招就是把所有的脏水泼到了我父亲身上。说我的父亲替我写了我少年时候的文章，因为我少年时候的文章特别的老成，不可能是 17 岁的学生写的。这个非常可笑，我在很多的场合说过，我小时候喜欢阅读钱钟书梁实秋和很多民国作家，因为我觉得他们文字好。在一个人刚开始写文章的时候，你阅读谁，必然模仿谁。而了显得渊博和少年老成，我还摘录了很多典故或者英语，准备随时引用在文章里显摆。而我的父亲则对民国文人兴趣不大，所以我们两人的文字非常好辨认。如果这样去加罪文学作品，我在十七岁的时候还发表过两篇写大学生活的小说，十五岁的时候还发表过写成年人生活的散文，当时我非常得意于杂志社的编辑都不知道我的真实年龄，现在想来，这些岂不都是死罪。

我只想说明，这就是一个好作家的底子。我的作品里的“我”如果一直是一个初中生形象出现，那岂不是和我喜欢的作家钱钟书梁实秋相去甚远？只要文学作品里的“我”不符合我的实际身份，我就是在造假，就是有人代笔？真是可悲，这么说的人恐怕从来没有体会到文学的乐趣。对于一个少年来说，文学的快乐就是可以把自己藏起来。不过对于一个写作者来说，还要对人解释这些，的确更可悲。

我甚至发现还有伪科学证明法，忽悠了不少人。有理科生做了一个科学统计，选取了八九本书，其中有我父亲的书为 A，我的四本书 B, C, D, E，其他作家的书 G, F。分析书中关键词出现的次数，比如三

个“的”“得”“地”，比如“因为”“所以”，结果发现，我父亲的书和我的那四本书出现这些词的总次数差不多，而另外两本书明显要数倍与ABCDE。实验得出的结论是，我和我父亲的风格雷同。这篇文章充满了我看不懂的专业术语，很多试图证明我的书是我父亲代写的朋友也像捡到了宝一样兴奋不已转发不止。甚至其中还有科普人士，他们认为最终还是科学和数据来说话了。我真的不知道该说什么好，A, B, C, D, E 都是十万字，G 是三十万字，F 则超过了五十万字。所以结果当然是ABCDE的常用词出现次数差不多，而G是他们的三倍左右，F则是五倍左右。像我这样的理科白痴也知道做这种对比的时候，抽样对象的字数得是一样的吧。我不明白这位理科生为什么这么恨我，虽然我一直对理科生特别有好感，我小时候的梦想就是科学家。但是理科更需要严谨，能在完全缺乏条件的情况下依然把实验做的津津有味并得出结论，也算有本事。而文科生们更不能“我觉得他 17 岁写的文章太成熟了，所以必然是中年男人写的”来论证问题吧

好了，下面请欣赏我的父亲@韩仁均叔叔 写的一篇文章，文章有点长，希望大家能耐心的读完，再来判断我的文章是不是我的父亲代写的。最后还附有我的父亲以前曾经发表过的作品，我的父亲为人正直，可爱，他的文笔非常的淳朴流畅，根本就不像我十六岁那样，看似老道，其实做作。我父亲念书的时候没有初中高中，根本没学过也不会英语，其中又经历文革，构陷者居然能想象出我父亲帮我写出一本讲初高中生活的校园长篇《三重门》。也欢迎大家拿我父亲的文章和我的文章来对比。很多人劝我不要再与这些人纠缠，因为他们就是要找各种茬搞臭你，你说什么都没有用。他们的逻辑就是走街上看人不爽，上前一口咬定你十年内必然杀过一个人。如果你不能自证十年内每一分钟的去处，那么我枪毙你是天经地义。关键是居然真有看客似乎觉得他们很正义，是替天行道。我知道有些人是不会错过这个好机会的，但是我相信明白人如果能够看到现在，也都能明白是怎么回事了。感谢所有支持我的朋友。今天我希望为我父亲讨一个清白，不过，这真是一个荒谬的时代，加害者像个原告一样，大摇大摆，不需要任何证据，只需要想象力，就能够损毁你和你家人的

名誉，而且什么罪名都敢往你头上加，而受害者却要像个被告一样，不停的出示着人证物证，也未必能还自己一个清白：

## 说说我自己

韩仁均

之所以想说说自己，是因为最近忽然有人发掘了我的超凡能耐，把我描述得能操控一切，既能写出《三重门》让韩寒“默写”，又能一手操控组织严密完整的全国新概念作文大赛，想买通谁就买通谁，想得什么奖就得什么奖，并且能一路代笔代思想，最起码有一个微型耳麦佩在韩寒耳边，可以随时指导这话怎么说，这事怎么做，而且能一路走红 10 多年。现在就差操控全国公务员考试的事还没来得及上提出来。当然说说我自己我也不会把自己贬得一钱不值，我只是把一个真实平凡的我告诉大家。

我生于 1957 年，韩寒之前说我生于 1958 年，那是他把她妈妈的年龄记成我了。小学中学都在村里念的，中学当时只有四年，叫做中一中二中三中四，没分初高中。读好四年中学后就在村里务农。1977 年恢复高考制度后，拼命复习，但理科外语都不懂，只得报考文科类的，后来被华东师范大学（当时好像叫上海师范大学）中文系录取。1978 年初入学。进校后，在学生名单上看到了好几个当时已经非常有名的作者，感觉以后要在这里和这么些同学一起度过几年的学习生涯很荣幸的。但事实上正经的课都没上过一节，第一个星期只是开会学习劳动之类，还有新生身体复查。结果是 GPT100 多一点，肝功能不正常（当时指标 40 以下算正常），于是就住进了师大后门那儿的肝炎隔离病房。住进去后检查的范围更大，还查出是澳抗阳性（就是现在的大三阳小三阳之类），被定性为乙型肝炎。以后每隔一些时间查一次，一直没有全部正常。最后 10 个月后，好像是 1978 年 12 月份，做了最后一次检查，同时抽了三份血样，分送华山医院等三家医院检验，三份结果完全不同，一份正常，二份不正常。于是，被认为还没痊愈，就被退学。因为大家都谈肝色变，我在病房里也很识相，不出去接触人，所以还没来得及认识一个同学就离开了学校，回到了家

乡亭新公社（乡和镇的叫法是后来的事）。因为多少算考取过大学了，所以回来就到了亭新公社的文化站工作。当时的文化站和现在的文体中心完全不是同一回事，就一间办公室，就我一个人。所做大部分的工作就是为公社机关服务，比如开会拍照，并自己冲印贴在公社门口的画廊里，布置会场等等。文化站里的上级业务指导对口单位是县文化馆和县文化局。

县里每年对各公社（乡镇）有业务考核指标，比如群众文艺创作、演出等等。为了应付这些考核，或者说是工作吧，各个公社乡镇就得组织人员进行创作。组织者自己当然也得写。这个时期，硬着头皮学写故事、表演唱、小散文等等。金山有故事创作演讲的传统，县里也经常组织培训等。后来就陆陆续续的写了些故事及散文等东西。发表在上海的解放日报市郊版，上海的故事会、故事大王及外省市的一些故事报刊上。因为我到文化站后就开始学写些小东西，但又感到自己的名字太过普通，就取了个笔名叫韩寒，当时韩寒还没有出生，但实际上没怎么用到，只在一二个小豆腐块上用了一下，后来觉得舍不得，而且本来发表东西又比较少，所以就决定把这个笔名作为我未来的儿子或者女儿的名字。所以 1982 年儿子出生后就叫韩寒了。韩寒出生那年国家又开始了高等教育自学考试，我因为受英语和理科的影响，也只能选择了文科，当时还是选择了华东师大中文专业的考试。自考每年考两次，每次最多可以报考4门课，大专在 10 门课左右，但还好工作时间比较空和自由，所以看书的时间比较多，晚上也是差不多全部时间花在读教材背题目上，当时也还不会打麻将，也没其他娱乐活动，电视机也只有一个黑白 14 英寸金星吧。我每次报的 4 门课一般都能通过 2 门，分数大都是 60 分多一点那种，只够及格，最高也只考过 70 几分吧。所以考了两年半，就完成了华东师大中文专业专科 10 门课程的学分，拿到了自学考试专科毕业证书。专科考过后，本科阶段还要考近 10 门课，觉得一些课太难了，像古代汉语、古代文学等等，所以后来就不考下去了（但韩寒多次在采访里把我说成是本科，因为他对这些专科本科的本来就没有什么概念）。但说实话，这种考试真的只为考试而考试，所学 10 门课在工作中根本用不到的，考过后也很快就忘掉了。但这个文凭也给我带来了好处，后



来 1985 年文化站转体制时派到了用场，成为我转成事业单位编制的有利条件，户口又迁出农村到城镇，户口在当时的中国决定着好多东西。后来直到 1994 年底，当时时兴办区县报，金山也要办一份金山周报。当时文化局的一位领导调到县委宣传部筹办这份报纸，他觉得我去做这个工作也合适，就把我调了过去。当时的金山报 4、5 个采编人员，采编合一。甚至划版样、校对都是自己负责，一周一期。期间也还评了个编辑还是记者的中级职称。我一直没有入党，因为我没想过要进官场混，也自知没能力在官场混。在金山报时后来要提一个副主编，部领导决定民选，大家无记名投票，我被大家暗算。后任的宣传部长对我说，可惜你不是党员，是党员的话一切都好办。甚至还想着有意“培养”我入党。我一笑置之。再后来我觉得我没义务付出那么多去负那个责任，就辞去了副主编的职务。2005 年底我们几个金山报的元老就各奔东西，我就去了我们区里的清水衙门文化局工作，做一个没有实职的主任科员，直到 2008 年底提前退休。

我到金山报工作后，韩寒还在亭林读小学。因为自己没什么社会背景，还有觉得当时县城朱泾的罗星中学教学质量什么的在全县算是比较好的，全县比较好的学生大都在这所学校里读书，里面有几个特色班，就赞助了几千元钱（相当于择校费吧），让韩寒的初中在罗星中学就读。我想等韩寒长大工作后，他的大部分同学将会是我们这个县里方方面面管事的头面人物，有这么一个人际关系基础，那对他的工作和发展会有所帮助。而我自己认识的朋友中，职位最高的也只是处级干部吧，我帮不上韩寒什么，以后只有靠他自己了，当时这么想的。记得韩寒进罗星中学后，摸底考试几门功课平均考了 91 分（满分 100 分），当时他自我感觉非常好，想这下总会名列前茅了，不料只在班级第 50 名左右，倒数前列，不禁感叹那些同学读书成绩真好。他比较牛逼的是作文，当时一篇介绍自己的作文《我》，让当时的语文老师彭老师赞不绝口。有时为炫耀，一节作文课写二篇。一开始韩寒的成绩还算比较均匀发展，但由于和教数学的班主任老师关系老是处不好，影响了他对这门课的兴趣，但中考前恶补一下后他的数学还是考得可以的，反而是语文考得不理想。他的应试作文在那种正式的考试模式中老是得不了高分。在区里的传统作文比赛里也能拿

二等奖，因为他的文章不是传统作文比赛喜欢的类型。但他还是比较喜欢看书的，尤其喜欢民国的，钱钟书和梁实秋等人的文章，家里的一本《围城》不知被他翻过多少遍，第一本翻烂后我又买了一本。因为喜欢，所以他后来在第一本书《三重门》里刻意的模仿他的偶像钱钟书《围城》的风格一点也不奇怪。

韩寒中考考了 468 分，有体育长跑比赛第一名的 8 分加分，就是 476 分，松江二中因为他长跑的成绩好，就又降低了几分特招了他。他当时是寄宿在学校的，不是每个星期都回来，那时松江到金山还得要转二次公共汽车。我发现他写《三重门》是在他差不多要写好的时候。我就对他说，要不要我拿去帮你复印一遍，不然弄丢后无法弥补。他同意了，说等写完后。写完后我就拿到对外营业的金山县图书馆复印室去自费复印了一份。我是这个时候才有了看的机会，总的感觉是非常幽默，有点像《围城》的那种笔法，文笔非常老练，而且我猜想，书中那个主人公身上，可能有他自己的影子。书中描写的生活，也从初中延续到了高中。这是一种没有生活的人想象不出来的情景。一般作者的第一本书，以自己的生活为基础，模仿自己喜欢的偶像风格写很常见。现在看到有人竟怀疑《三重门》是我写的，那我真要谢谢他们的抬举了，我要写得出《三重门》，我早不是现在的我了。这种只要有脑子，而且能正常思维的人都想得明白的。不是一代人，文笔和经历完全不一样，你能写得出那种情景那种意境那种感觉吗？现在一些人觉得他们不可能做得到，所以断定韩寒在他们这个年龄也做不到。一个 17 岁孩子的文章是他父亲代写的并且还能走红中国十多年，是在编神话吗？如果大家都可以随意恶意的假设和推测，我也可以把你臆想得什么都不是。

韩寒在松江二中读书过程中得了那种浑身奇痒的疥疮，后来学校怕感染其他同学让他回家养病一个星期。我是在那天回家看到他写的《求医》才知道他得疥疮的。那天回家时他去外面玩了，但文章在桌上。我看了以后觉得很幽默，笑坏了。因为我事先看见过新民晚报上一则上海萌芽举办新概念作文大赛的消息，所以就建议韩寒将这篇《求医》和另一篇《书店》参加新概念作文比赛。因为他觉得这次

比赛没收参赛费，应该是真正的比赛，不是那种常见的近乎骗人的活动，而且既然是新概念，就不是应试作文的那种，就参加了。后来寄出去后一直没有回音，韩寒和我都有点失望，直到那天上午，一位叫胡玮蔚的萌芽编辑将电话打到家里问为什么不去参加前一天的新概念复赛才知道那天新概念作文比赛就要揭晓颁奖了。（很多评委在初赛的时候就留意到了韩寒的两篇文章，觉得特别老练，就和现在大家的怀疑一样，所以他们委托萌芽的胡玮蔚编辑给了我们电话，一方面是爱惜人才，怕因为客观原因错过了比赛，一方面也想当面考验韩寒）。韩寒说没有接到复赛通知（当时我们一家住在 50 多平米的老公房，楼下的邮箱都是没有锁的），后来胡玮蔚去问了评委后再打电话过来说评委同意韩寒中午前赶到上海市区比赛的地方再考一遍，我带了韩寒就急急忙忙的赶到车站那里找了辆黑车去市区，到那边已经接近中午了。接下来就是评委即兴出题现场一个小时写出《杯中窥人》的事。这事居然让阴谋论者认为我是开了后门事先知道了题目写好后让韩寒背的。这真是天地良心了，我们知道韩寒其实入围了是在当天的上午，此前萌芽的编辑我一个都不认识。包括李其纲，我也是根本不知道那个出题老师叫李其纲，出题的老师叫李其纲是我后来在有关新概念作文比赛和韩寒的补考的新闻报道中才知道的。而且我根本写不出这种文章。我的文章根本不是这个风格的。再说如果新概念作文比赛可以舞弊的话，那韩寒真的不可能有这次机会，因为这个比赛很隆重，有很多的教授和著名作家作评委，真的这个比赛要走关系的话，参赛的学生里有这方面能力的家长实在太多了，能得到好处的肯定不会是我们，我相信有这种能力的家庭也不会只住在 50 多平方米的老公房里。我可以这么说，一切能够靠钱靠关系靠舞弊能获得的好处，都不会轮到我们先得到，我们甚至连号都排不上。这种污蔑直接玷污了这个严肃的比赛。韩寒作为新概念作文比赛的参赛者，我只能告诉这些我所知道的情况。

好多人都说我应该回应，其实我觉得这事真的无聊透了。韩寒诚然有很多不足之处，看他不顺眼可以直接批评或者骂他，而且一直以来也不少这么做的人，但用这种全靠自己的臆想和主观判断来污蔑和传播，我觉得就十分下作和无耻了。我对韩寒说，无论你怎么说，

他们还会无中生有找各种各样的茬来污蔑你的，因为他们就是看你看不顺眼。有一次韩寒比赛翻车了（拉力赛很容易翻车，他参加拉力赛好几十场一共翻车过两次，算是冠军车手里翻车最少的），他回头发现在新闻的留言里最多的都是咒韩寒怎么还没死的，所以他早就知道了在很多人喜欢他的时候，也有很多人不喜欢他，他几乎从来不回应那些对他的辱骂，甚至有些谣言也不回应，但我觉得这次真的太过分了。包括这篇东西，我也只是说说自己，说说当时的一些情况。其实只要是韩寒的真正的读者，就会发现韩寒的文章和书从一开始到现在，其行文风格是有一条清晰一致的成长脉络的，像幽默什么的只是一开始比较刻意，一直在掉书袋，用典故，和现在的很多专家写的一样，那是因为受到了钱钟书的影响。很多典故和生僻的书本或者英语都是他硬记下来为了炫耀而背的，你要写过文章都知道，你随便记住或者摘抄下几个很生僻的东西，想要硬放到文章里是很容易的。韩寒的引用都不算特别自然，算是明显的故作老成。后来更趋于自然和内在。一直到了近几年，他说他写文章要做到不用典而把事情说清楚。我觉得这是他的进步。一个十五岁的孩子到三十岁肯定是在不停的进步的，这也是人生中一个人改变最大的阶段，阴谋论的人不能要求韩寒永远和十五六岁的时候写的文章一个样子，否则就是有假，那样倒是要被人笑死了。我找了一下，发现韩寒学生时代写在各种作业本笔记本和各种不规则纸上的那些文章手稿都在，当时保存这些也不是为了日后打官司或者让人研究，这些不经意间保存的资料，现在看来是多么珍贵。看到这些原始资料，我不得不再一次为我的儿子骄傲，很多人，包括我，十几岁时根本做不到他那样，韩寒虽然有点虚荣，有点故作老成，但是他做到了，他还在不停的努力和进步。我到现在还是认为，韩寒的这种文风是骨子里的，不是随便可以模仿的。如果这年头连手稿都不能用来证明什么了，那大家都不要写作了，真是所谓欲加之罪，何患无辞了。韩寒现在决定要出《三重门》的手稿，你们到时还会惊羨一个十七岁的高中没毕业的少年写的字竟然有这么好这么老练，可以自己去看一下或者发奋一下，能有几个大学生研究生博士生的字能有这么好看。韩寒不是一个特别喜欢应酬和交际的人，别人可能在玩的时间，他在想东西，写作，阅读，练字，练车。他最大的娱乐就是有时候踢一次球或者周

末和朋友们打一个《使命在召唤》的真人射击类的电脑游戏，他的名气不算小，但是他平均一个月都没有一个饭局。大家之所以觉得他能做很多的事情，精力很旺盛，是因为他把别人可能用来应酬和娱乐的时间都用在工作上。你们可能不知道吧，如果没有比赛和游戏，他几乎每天晚上的八点开始写作或者看书，一直到早上六点，连续十个小时都在书房里。所以他的博客大多都是凌晨发的。虽然他口头上不承认，一直说他在玩，但这个就好像一个考试很好的学生喜欢说他在家里从来不复习一样。没有想到，他的努力反而成为了他的过错。有些人甚至拿出了韩寒小学两年级的作文要说明韩寒未来写文章不好，这真的是不厚道的，其实大部分的小学在两年级的時候还在教认字，那个年代很多的学校是三年级才开始写作文的，我不知道现在的小学生是什么样的，我记得韩寒在两年级的時候好像是主动写作文给语文老师批阅的，也就是说这不是老师布置的作业，同年级里的大部分同学都还没有开始写作文。这种推测是有点胡搅蛮缠了。你不能因为刘翔学走路的时候会摔跤就推测他将来肯定跑不快所以有猫腻。况且我觉得韩寒那篇作文写的挺好的。可能他也觉得以8岁小孩子的水平来说还不错，就自豪的贴在了博客上，要不然大家也不会看到这些。韩寒写文章很快，修改也不多，这个可以从他的手稿里看出来。韩寒回应麦田的文章，其实写了两个小时，但是4点，6点，8点，甚至10点又修改过，这说明韩寒一夜睡不着，很在意这件事情，他修改的内容是让文章更简洁一些，语气也更缓和一些，包括一些不太礼貌的气话都删除了，可以说在一个作家遭到了污蔑又没有办法自证以后，他已经做的很平和。我知道他很珍惜自己的名誉，所以一定很生气。从半夜两点修改到十点也是韩寒自己写在博客里说的，以表示他很愤怒，气得都有点不会写文章了。他不说反而没有人拿这个来说事，他就是太老实太坦诚了，自己说了，结果吃了亏。被有的博士生拿来有意曲解为韩寒一般写两千字的文章都要花十个小时，所以有代笔，他的考试文章也不可能一小时写出来。有人也质疑说韩寒这几篇文章写的文采不如以前，说明以前是有人代笔的，这几篇文章可都是被迫的回应和申明啊，而且还是他人构陷在先，结果你不去怪那个加害人，反而要百般刁难受害人。这么说的入真的是太没有意思了。

再回到我。我是从看到韩寒写的《求医》等文章后就慢慢不再写东西的，因为我觉得我已经写不过他了，有点不好意思也懒得再写了，这种感受也许一些从事文化工作方面的父亲能够体会。我以前写的也只是一些农村题材的故事和一些应景宣传用的东西，根本没涉猎过中篇小说。所以说韩寒的长篇小说是我写的很滑稽。和韩寒写的东西一比，反而是我写的实在太小儿科了。另外我也不怕浅薄，还要告诉大家，我根本不会英语，我们那个年代，从小学到四年中学到自学专科，从来就没有英语这门课。（大家好，我是韩寒，插入一下，我的父亲刚才经过回忆纠正了一下，说他们好像在四年中学里有过几节不正规的英语课，但是好像只学会了字母，他本人几乎不认识一个英语单词）我书读得比韩寒少多了，韩寒说的好多典故都是我不知道的，我甚至四大名著都没全看过（不过韩寒也没有看全过，因为他当时读书有一点炫耀的成分，要去读那些同学们都没有读过的书，才好像显得他很有学识），三国演义努力了几次都没看完第一回，我只看我喜欢的一些东西。他的书中文章中对好多事情的分析思考看法观点，让我受到过好多的启发。我是从心里佩服他的，当然从没嫉妒过他。我后来在韩寒的建议下申请了提前退休。现在的生活，就像我前几天在微博上说的，彻底告别了闹钟，一般睡到上午 10 点左右自然醒吧，先在被窝里和马桶上手机上会网，然后洗洗涮涮，弄点吃的，打扫下卫生，中午开始电脑上网，挂上QQ和MSN方便有人找和联系事情，然后看看新闻，翻翻微博，有需要在电脑上处理的事处理一下，下午出去办办事，有时去老家看看父母，遛一下金毛和萨摩耶，对了，金毛叫戛戛，萨摩耶叫闹闹，我的微博头像就是闹闹。韩寒博客头像上的金毛是几年前已经去世的木木。有时去看看小孙女。所谓含饴弄孙吧。好多朋友希望我贴小孙女照片，这个，等稍过几天小孙女的照片正式发布后再贴张我和她的合影吧，绝对小美女一个。当然我也没你们想象的老。我自然年龄 56 岁，社会年龄就是你们看上来的年龄大概要小 10 岁，心理年龄也许还要小 10 岁。没韩寒英俊，但五官还算端正。如果有麻友来约，晚上就会去打半夜麻将。回家后再上网，看会新闻，翻会微博，有时看几集美剧，二三点钟后挂起电话洗洗睡去。就这么循环，很快，一年，一年，慢慢老去。

但韩寒还只是一个30岁的青年。我希望韩寒能生活在一个正常的人与人之间具有基本信任和交流的社会环境里。你可以不同意他的观点，去和他争论，去批评他，可以看不顺眼直接骂他，但不要去用恶意的揣测去诬陷他，污蔑他。用各种谣言和臆测来扼杀一个韩寒轻而易举，而且我知道很多人想这么做，今天终于有了机会，所有一直不爽韩寒的人终于可以团结起来。不过我觉得韩寒不是那么轻易可以用谣言扼杀的。或者等他被扼杀了，你们就知道想“人造”一个韩寒是一件不可能的事情了。对不起大家，我已经很久不写长得东西了，所以写的很啰嗦，谢谢你们可以耐心的读完。也许在阴谋论的人眼里，我是故意写的这么啰嗦来和韩寒的东西很简洁的风格区分开来的。那么好吧，我找到了两篇 1999 年的时候，就是韩寒写《三重门》的时候我发表在《故事会》和《现代农村》上的小故事，大家可以和韩寒当时的文章对比一下。我知道如果要牵强附会，那么你也能找出我这个文章和韩寒的某个文章一两个用词是一样的，一两个形容词是一样的，一两个转折词是一样的，甚至可以说我发表在《故事会》等报刊上的文章就是韩寒写的甚至这篇文章就是韩寒写的。我们两个人是儿子代老子写，老子代儿子写，我们两个人太闲了。还是要说，欲加之罪，何患无辞。

2012 年 1 月 27 日

## APPENDIX 4: TABLES CONTAINING THE DATA EMPIRICALLY OBTAINED FROM QUANTIFICATION OF THE NEWSPAPER ARTICLE AND THE SHORT STORY

The tables show observations empirically obtained by quantification of the newspaper article and the short story:  $x_i$  represents the lengths of constructs (measured in constituents),  $z_i$  their frequencies and  $y_i$  the average lengths of constituents (measured in the immediately lower units). The grey background of the cells is used to highlight the omitted observation with a low frequency.

Let  $i$  be a natural number representing four language levels, where  $i = 1$  represents the language level paragraph – sentence,  $i = 2$  language level sentence – parcelate,  $i = 3$  language level parcelele – character and  $i = 4$  language level character – component, thus the value of  $i$  can be  $i = 1, 2, 3, 4$ .

**TABLE A**

Language level L4: character (measured in components) – component (measured in the average number of its strokes)

$x_4$	The newspaper article		The short story	
	$z_4$	$y_4$	$z_4$	$y_4$
1	463	5.0994	733	4.7763
2	710	3.4127	782	3.4565
3	696	2.5393	731	2.5860
4	340	2.0824	425	2.2047
5	220	1.8718	214	1.9729
6	83	2.0161	101	1.9076
7	43	1.6246	49	1.8192
8	7	1.6964	22	1.8125
9			6	1.3704
10			2	1.8000



**TABLE B**

Language level L3: parcelate (measured in characters) – character (measured in the average number of its components)

$x_3$	The newspaper article		The short story	
	$z_3$	$y_3$	$z_3$	$y_3$
1	<del> </del>	<del> </del>	7	3.1429
2	4	3.1250	11	3.0000
3	<del> </del>	<del> </del>	21	2.5556
4	14	3.3393	32	2.8281
5	6	2.2667	26	2.9538
6	22	2.9545	30	2.8611
7	6	2.7619	41	2.8153
8	14	2.9911	36	2.7014
9	11	2.9697	27	2.7860
10	5	2.4400	26	2.7923
11	10	3.0182	24	2.7348
12	7	2.7738	13	2.4744
13	12	2.8141	13	2.5740
14	14	2.8776	11	2.6688
15	7	2.7429	13	2.9641
16	4	2.7656	4	2.8750
17	4	3.2206	6	2.5882
18	8	2.7153	3	2.8704
19	3	3.0000	3	2.6842
20	3	2.6167	2	2.2500
21	8	2.7262	5	2.6857
22	5	2.5818	<del> </del>	<del> </del>

$x_3$	The newspaper article		The short story	
	$z_3$	$y_3$	$z_3$	$y_3$
23	2	2.7391	2	2.8261
24	3	2.7222	1	2.8333
25	2	2.5600		
27			1	2.4074
28	2	2.6964		
29	3	2.6207		
30	1	3.0333		
31	1	2.7742		
33	1	3.1818		
34	1	3.2059		
35	2	2.8429		
36	1	2.9444		
38	1	3.0789		
39	1	3.0000		
42	1	3.0952		
47	1	3.1277		

**TABLE C**

Language level L2: sentence (measured in parcelates) – parcelate (measured in the average number of its characters)

$x_2$	The newspaper article		The short story	
	$z_2$	$y_2$	$z_2$	$y_2$
1	8	27.3750	23	9.9565
2	14	17.2500	34	10.9265
3	7	14.4286	25	8.2933

$x_2$	The newspaper article		The short story	
	$z_2$	$y_2$	$z_2$	$y_2$
4	11	11.8636	11	7.3636
5	3	15.6000	12	7.3667
6	1	16.5000	6	7.6944
7	3	11.9048	1	8.5714
8	3	10.3333	2	6.3125
9			2	9.7222
11	1	7.3636	1	8.3636
12	1	10.2500		

**TABLE D**

Language level L1: paragraph (measured in sentences) – sentence (measured in the average number of its parcelates)

$x_1$	The newspaper article		The short story – Variant 1		The short story – Variant 2	
	$z_1$	$y_1$	$z_1$	$y_1$	$z_1$	$y_1$
1	2	3.0000	9	3.6667	3	2.6667
2	5	4.7000	10	2.3000	2	4.0000
3	4	2.5833	3	3.5556		
4	3	3.2500				
5	2	3.8000	1	3.0000	1	3.0000
6	1	4.8333			1	2.8333
7			2	2.1429	1	2.1429
8			2	3.6250	2	3.6250
9					1	2.6667
10					1	4.2000

$x_i$	The newspaper article		The short story – Variant 1		The short story – Variant 2	
	$z_i$	$y_i$	$z_i$	$y_i$	$z_i$	$y_i$
11	<del></del>	<del></del>	<del></del>	<del></del>	1	1.7273
15	<del></del>	<del></del>	<del></del>	<del></del>	<del></del>	<del></del>
17	<del></del>	<del></del>	1	3.9412	1	3.9412
27	<del></del>	<del></del>	1	2.8519	1	2.8519

# Index

## A

algebraic linguistics 15  
 Altmann, Gabriel 11, 16–17, 19

## C

coefficient of determination 47–48,  
 57–58, 60, 63, 65–67, 69–70, 76–77,  
 82, 85, 87, 91, 99–101, 105, 107,  
 116, 118–119, 121–122, 124, 126,  
 129, 133, 136, 139, 143, 147  
 component, 12–13, 28, 31–33,  
 41–42, 51–52, 54–56, 58–67, 70–75,  
 88–90, 97–100, 102–105, 107–118,  
 130–132, 134–137, 140, 148–153,  
 192–193  
 computational linguistics 15  
 constituent 16–18, 41–42, 47, 52–53,  
 59, 64, 70–73, 80, 83–85, 98, 100,  
 111, 116, 120–122, 126, 128, 131,  
 152, 192  
 construct 16–18, 41–42, 47, 52–53,  
 57, 61, 70–73, 76, 80–81, 83–85, 89,  
 98, 100, 111, 116, 120–122, 124,  
 126, 128, 131, 152,

## E

extremes 48, 57, 99, 133, 144

## F

frequency group 59, 111–114

## G

graphic field 12, 35, 54, 56, 61–62,  
 66–67, 88, 102–104, 130, 137,  
 150–152

## H

Hřebíček, Luděk 11, 26, 41

## CH

Chinese characters 9, 12–13, 22–33,  
 35, 40–44, 50–52, 54–56, 58–76,  
 78–80, 83, 88–90, 92, 96–100,  
 102–119, 121–122, 127, 129–132,  
 134–140, 148–153, 192–194  
 Chinese language 11, 21–23, 26–27,  
 50–51, 54, 80, 91, 96, 153

## J

jiaguwen viz oracle bone script

## K

kaishu viz regular script

## L

language level 12–13, 16–17, 20,  
 27–28, 31, 41–47, 49, 51–54,  
 57–58, 61, 63, 66–67, 70–73,  
 75–78, 80, 82–85, 87–91, 97–100,  
 102, 104–109, 111, 114–117, 119,  
 121–122, 124, 126–127, 129–134,  
 137, 140–141, 144, 147–153,  
 192–195

language unit 12–13, 16–17, 20,  
27–29, 31–33, 35–36, 40–42, 49,  
51–54, 56, 61, 63, 66, 70–73, 77–78,  
80, 83, 87–91, 97–98, 102, 107–108,  
115–116, 118, 121, 126, 129–133,  
140, 148–154, 192

linear regression 47, 99

## M

mathematical linguistics 11, 15

Menzerath–Altmann Law 11–13,  
16–20, 27, 49, 60, 66, 69–71, 73,  
77, 79, 81, 85, 88–91, 97, 99–102,  
104, 106–109, 115, 118, 120–126,  
128–132, 136–137, 140, 143–144,  
147–153

Menzerath, Paul 11, 16

## P

paragraph 12–13, 28, 40–43, 46,  
51–54, 56, 80–84, 88–90, 95–99,  
124–125, 127–129, 131–132,  
148–149, 151–153, 192, 195

parameter A 47–48, 60, 63, 66, 70,  
77, 82, 87, 99–101, 105, 107,  
116, 118–119, 121–122, 124, 126,  
129, 133

parameter b 17–18, 48, 57–58, 60,  
63, 65–67, 69–70, 76–77, 81–82, 85,  
87, 99–102, 105–107, 116, 118–119,  
121–122, 124, 126, 128–129, 133,  
136, 139–140, 143, 147

parcelate 12–13, 28, 35–36, 38–45,  
51–52, 54, 57, 67–73, 75–76, 78–81,  
83–85, 88–90, 97–98, 104–108,  
111, 115–116, 119–127, 129–132,  
137–141, 148–153, 192–195

punctuation 13, 28, 35–40, 56, 78–79,  
85, 87–88, 96, 126, 132, 151, 153

bracket 39

colon 36, 56, 96

comma 36, 38, 56–57, 82, 96

dashes 37–40

ellipsis 37–38

emphasizing dot 37–38

enumeration comma 36, 78

exclamation mark 36, 40

full stop 36, 40, 82, 87, 96, 152

middle dot 37–38

proper name mark 38–39, 163

question mark 36, 40, 96

quotation marks 36, 38, 78, 85, 96

semicolon 36, 82–83, 85, 87, 90, 151

title marks 56–57, 78, 85

titles marks 37

## Q

quantitative linguistics 11, 15, 48

## R

rank 59, 111–112, 114

reform of simplification 13, 23–25,  
75, 89, 104, 108, 115, 130, 140, 150,  
152–153

**S**

script 9, 11, 13, 22–26, 28, 35, 75, 89,  
115, 140, 150, 152

bronze script 22

clerical script 23

great seal script 23

oracle bone script 22–24

regular script 23–26

small seal script 23

sentence 12–13, 28, 35–37, 39–43,  
45–46, 51–53, 56, 75–76, 78–85, 87,  
89–90, 97–99, 119–129, 131–132,  
144, 148–153, 192, 194–195

simplified characters 20, 25, 64, 92,  
94, 96, 104, 148, 151, 153

stroke 12–13, 28–33, 35, 41–42,  
51–52, 54–55, 58–59, 61–66, 75,  
97–100, 102–104, 108, 130, 137,  
148, 150, 192

basic strokes 30

compound strokes 30

elementary strokes 30–31

hooked simple strokes 30–31

simple strokes 30

stylistic style 12–13, 20, 26–27, 49, 54,  
92, 133, 137, 144, 148–150, 153

artistic style 26–27, 49, 91–92, 152

literary style 26, 49

newspaper style 12, 26, 49, 150

scientific style 26–27, 49–50

**T**

traditional characters 25, 62–63, 65,  
75, 90, 104, 130, 151–153

**W**

word 16, 35, 37, 72–73, 89, 115–118,  
121, 131, 151, 153

written Chinese 11–13, 20–21, 87, 89,  
91, 132, 148

## **KATALOGIZACE V KNIZE – NÁRODNÍ KNIHOVNA ČR**

Motalová Tereza

An application of the Menzerath–Altmann law to contemporary written Chinese / Tereza Motalová, Lenka Matoušková. -- 1. vyd. -- Olomouc: Univerzita Palackého v Olomouci, 2014. -- 201 s. -- (Qfwfq)

ISBN 978-80-244-4192-4

811.581 \* 81'42 \* 81-13 \* 81'324

- Chinese language
- written texts
- segmentation (linguistics)
- quantitative linguistics
- monographs
- čínština
- psané texty
- segmentace (lingvistika)
- kvantitativní lingvistika
- monografie

495.1 – Chinese language [11]

811.58 – Čínské jazyky [11]



**An Application of the Menzerath–Altmann Law to Contemporary  
Written Chinese**

Tereza Motalová

Lenka Matoušková

8. svazek Edice Qfwfq

Výkonný redaktor: Jiří Špička

Jazyková redakce: Jana Kynclová

Odpovědná redaktorka VUP: Jana Kreiselová

Sazba: Lenka Horutová

Obálka: Martina Šviráková

Vydala a vytiskla Univerzita Palackého v Olomouci

Křížkovského 8, 771 47 Olomouc

[www.upol.cz/vup](http://www.upol.cz/vup)

e-mail: [vup@upol.cz](mailto:vup@upol.cz)

Olomouc, 2014

1. vydání, 201 stran

č. z. 2014/610

ISBN 978-80-244-4221-1

Publikace je neprodejná