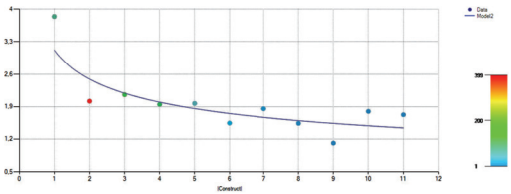
Text Segmentation
for Menzerath-Altmann Law Testing

Martina Benešová et al.



Text Segmentation for Menzerath-Altmann Law Testing

Martina Benešová et al.

Palacký University Olomouc
Faculty of Arts
2016

Reviewers:
prof. RNDr. dr hab. Jan Andres, CSc., DSc.
Mgr. Jiří Milička, Ph.D.

Authors:
Mgr. Martina Benešová, Ph.D.
Mgr. Denis Birjukov
Mgr. Jana Kovaľová
Mgr. Lenka Matoušková
Mgr. Tereza Motalová
Mgr. Denisa Schusterová
Mgr. Petra Vaculíková

This monograph could have been realized thanks to the special-purpose support for specific university research that was allotted to Palacký University in Olomouc by the Ministry of Education, Youth and Sports in 2014. The monograph is a part of the collective work within the SGS-IGA project titled Segmentation for testing the Menzerath-Altmann law and the hypotheses related to it II, no. IGA_FF_2014_083.

First Edition

ISBN 978-80-244-5131-2 (online : pdf)
ISBN 978-80-244-5112-1 (print)
DOI: 10.5507/ff.16.24451121

Contents

Foreword

The leading topic uniting the texts in this monograph is the Menzerath-Altmann law and testing its potential validity on samples in different languages (in our case, in Chinese, Japanese, and Czech), on spoken as well as written samples and on samples with newly established, pioneer unit concepts. The Menzerath-Altmann law is capable of being applied on such a variety of language aspects and is, therefore, regarded as a language universal. The monograph, this way, summarizes a follow-up research based on other pilot experiments held not only in the mentioned languages.

The validity of the Menzerath-Altmann law has been proved in many experiments, nevertheless many researches has falsified its validity and even its existence. It is, thus, necessary to continue in testing in other analyses with various samples, i.e. samples in different languages, styles etc. Exceptionality of this experiment series lies in the choice of the language of the tested samples and in satisfying the assumption to define individual units to be employed in text segmentations. The authors were obviously forced to face brand specific conditions when segmenting this text. Naturally, the so far, traditional linguistic platform has been utilized; however the authors have proved the ability to use their own abstraction and imagination. At the same time, they were able to cope with imperfections and flaws which appeared in the pilot research phase and included respective measures to remove them now.

The methodology and tools of individual experiments are depicted enough to satisfy. Mathematical and statistical tools used in experiments naturally deserve in future analyses a more detailed attention because this current research revealed many new questions in this respect; this is not but any fault of the authors' attempt and of this monograph. On the other hand, what deserves to be mentioned and noticed is a unique and manually performed segmentation of individual samples whose micro elements, language levels, language units seem to be transmittable and useful on other text types

where they can serve as methodological sets of instruments. The monograph, thus, shows, apart from others, methodologic-educational signs.

Last but not least, most of the authors' collective are young scientists. This monograph is not the first to reveal fully their promising potential for future researches and their academic potential. So let us hope that their academic careers will not hold the tendency of the Menzerath-Altmann law.

Martina Benešová

Application of the Menzerath-Altmann Law to Written Japanese – Poetic and Academic Styles

Denis Birjukov

Summary

The aim of this paper was verification of validity of the Menzerath-Altmann law on two written Japanese texts – a poem, and a seminar paper. The segmentation was conducted on a graphic principle with language units specifically designed and defined for experiments of this kind. With the poetic text, on language levels where strong graphic restrictions on maximum text length could be expected, the Menzerath-Altmann law proved valid with relatively high R^2 values. On levels where such restrictions do not apply, the parameter b turned out positive, however the values of R^2 were very low. MAL was proven invalid on two language levels of the other analysed text – a seminar paper, disputable on one, and valid on one, with surprisingly high R^2 value. At present, more experiments with identical text styles are needed.

Introduction

This paper aims to verify validity of the Menzerath-Altmann law on two written Japanese texts – a poem, and a seminar paper – by methods of quantitative-linguistic segmentation and analysis. The segmentation will be conducted on a graphic principle, and the chosen language units are: a surparagraph, a paragraph, a sentence, an intercomma, a character, a component, and a stroke. This created five language levels, labeled U_0 to U_5 from left to right: surparagraph-paragraph, paragraph-sentence, sentence-intercomma, intercomma-character, character-stroke, with the former being omitted in this experiment. Every distinctive unit received its own defini-

tion applicable only for experiments of this kind. Results received will be analysed and discussed.

This paper draws from two main sources, especially concerning applied methodics – previous analysis of a Japanese short story[1], and a thesis focused on more text styles with some additional experiments[2].

Hypotheses

U_1 – paragraph-sentence

Different results according to the analysed text style can be expected on the paragraph-sentence level measured in average intercomma lengths due to the fact that paragraph structures differ in different texts, and thus changes in segmentation were necessary. It is assumed that in the case of the seminar paper, MAL will not be valid, which also happened during similar experiments with Chinese texts (Matoušková 2014), and with the previous experiment with a Japanese short story. A different analysis result can possibly be expected in the case of the poem due to a slightly different approach to the definition of a paragraph. In this case one whole strophe was considered a paragraph, i.e. not termination of a line and starting a new line (with possible indentation) as is true for other types of text.[3] (This will not be a segmentation based strictly on the outlined definition of a paragraph, but even this kind of segmentation still sticks to a graphic principle. However, if termination of a line (verse) led to termination of a paragraph, then the paragraph would often terminate before a sentence or even an intercomma, and hierarchical continuity of language units would be violated.) The reason for possible validity of MAL in this case is precisely this limited (restricted) graphical length of a paragraph which means that the more this restricted paragraph contains sentences, the less should said sentences be able to contain intercommas in order not to exceed the limited maximum length of the paragraph. If this hypothesis proves to be true, it might signify that graphical text limitations can significantly influence validity of MAL

[1] Benešová, 2015.

[2] Birjukov, 2016.

[3] It can be viewed not as termination of a row at the end of a verse, but as line wrapping because of spatial restrictions, and termination of a paragraph as a "meaningful whole" will happen only with termination of a whole strophe.

(in case of graphic segmentation) – the stronger the limitations, the more pronounced MAL becomes.

U_2 – sentence-intercomma

MAL might prove itself valid on this language level. As for the seminar paper, it can be said that the text was created without strong graphic restrictions, and the author could relatively freely adjust lengths of sentences and intercommas, and to naturally shape the text into their own desired form.

As for the poem, it can be assumed that the results might be similar to the U_1 level, i.e. that the MAL will be proven valid but for somewhat different reasons. Maximum length of a sentence is to some extent influenced by the maximum length of a higher language unit – the paragraph – thus it cannot be unlimited. It should be true then that the more sentences contain intercommas, the shorter the intercommas will be when measured in the average amount of characters.

U_3 – intercomma-character

After previous experiments with Chinese texts and the Japanese short story it would not be surprising if MAL did not prove itself valid on this language level. There are several reasons for this assumption: one of them might be the usage of the intercomma unit itself which is based strictly on a graphic principle; another is the fact that there seems to be a large gap between the language units of an intercomma, and a character. (However, creation of an in-between unit is complicated by absence of an orthographically defined or at least recognisable word in Japanese.)

Motalová et al. (2013) describes other possible reasons of MAL invalidity on similar language levels in Chinese which might be relevant to Japanese texts as well and thus negative results are expected in this experiment. First, it's a fact that the intercomma is a unit with fluctuating length, while its constituents – characters – are units with fixed numbers of components (author is not able to change the number of components in any one given character). Second, it is small variability of characters in terms of included components (Motalová and Spáčilová 2013, p. 115).

U_4 – character-component

There are two contradicting problems on this language level:

On the one hand, all Japanese characters (according to the definition of a character in this paper) are to be written inside strictly defined square fields which limit the number of character components that can fit inside. Precisely this argument might support MAL manifestation on this language level. On the other hand, *kanji* characters and *kana* characters (and perhaps others, e.g. Roman script, punctuation marks etc.) might be inequivalent units, i.e. they should be viewed as hierarchically different or perhaps parallel language units. Also, while we usually refer to *kanji* characters in terms of concepts like components, strokes etc. (no matter what their definition is), it is unusual to refer to these concepts when addressing the characters of *kana* and others. Thus, while limited in the same way as *kanji* characters by fixed graphic fields (at least in the sense that one *kana* character, or a fixed amount of other characters, should fit into one field), it is possible that at this point it might be adequate to stop treating all the characters in the same way, and to further segment them in some different way. The economy principle which seems to apply for *kanji* or *hanzi* characters inside graphic fields may prove to be invalid for other types of characters due to their different evolution.

Finally, it is a question of which of these two arguments will have greater influence on the manifestation of MAL. Based on arguments mentioned above, and on the results of the previous experiment with a Japanese short story, it can be assumed that on this language level MAL might prove valid but the determination coefficient would be relatively low, and MAL will not manifest itself to the same extent as with texts written in Chinese.

Analysed texts

Text No. 1 – poem ("western" style) – poetic style

One of the analysed texts was The Raven (Ōgarasu 大鴉 in Japanese), a poem by the American poet Edgar Allan Poe in Japanese translation from 2013 which was published digitally on the iTunes Store.

The original poem is from 1845, and the – in the eyes of a modern reader – peculiar language style of the original has been kept to some extent in the Japanese translation as well. Especially obvious is quite frequent usage of *kanji* characters not included in the standardised list of characters – the *Jōyō kanji hyō*. Length of the text is 2721 characters.

Text No. 2 – seminar paper – academic style

The second analysed text is a short seminar paper by a Japanese university student in the field of literature. Length of the paper is 5594 characters. One of the criteria for selection was, among else, that a paper must had been finished and actually handed in for (successful) evaluation.

Segmentation

For this experiment, the graphic principle of segmentation was chosen because of the specifics of the Japanese language writing system and layout of the Japanese texts in comparison to often analysed European languages (in the geographical sense), as well as for continuation of similar experiments with Chinese texts and the previous analysis of a short story. Moreover, according to Claudia Prün (1994, p. 149), graphically segmented *kanji* characters do show characteristics of the Menzerath-Altmann law.

The segmentation into language units was to some extent set up in a way that after a basic briefing even a person who cannot read and write Japanese could – with the same tools – segment a Japanese text with minimal error rate. In other words, the segmentation must be to a greatest possible extent based on graphic principles.

The texts chosen for this experiment were segmented into seven language units, and consequently quantified and transformed into a format usable in the MA Studio computation software. Chosen language units are as follows:

1.	**Surparagraph**
2.	**Paragraph**
3.	**Sentence**
4.	**Intercomma**
5.	**Character**
6.	**Component**
7.	**Stroke**

This created five language levels:

U_0:	**Surparagraphs measured in paragraph length – paragraphs measured in average length of sentences**
	⇧⇩
U_1:	**Paragraphs measured in sentence length – sentences measured in average length of intercommas**
	⇧⇩
U_3:	**Sentences measured in intercomma length – intercommas measured in average length of characters**
	⇧⇩
U_4:	**Intercommas measured in character length – characters measured in average length of components**
	⇧⇩
U_5:	**Characters measured in component length – components measured in average length of strokes**

The Japanese language allows for two possible ways (directions) of text input: Left-to-right and top-to-bottom (i.e. "horizontally", when one browses through pages the same way as in English, the so-called *yokogaki*), or top-to-bottom and right-to-left (i.e. "vertically", with browsing starting from "the end" of the book compared to English texts, the so-called *tategaki*). The Japanese script have long been formed for vertical writing, so its differentiating features lay on the vertical axis – on left and right sides of the axis, but in the top-down direction. It thus can be said that if written horizontally, it loses its natural continuity of differentiating features. (Kraemerová, 2000, p. 34) However, modern texts are often written horizontally, especially academic papers and others.

MA Studio software (ver. 2.11.0.0) was used for computations, i.e. for parameters A and b, determination coefficient R^2 and other statistical data, as well as for charts.

As the frequency of occurence of individual x values often differed from each other (often thousands versus single digits in a single language level), all the empirically acquired values received a weight value according to their frequency of occurence in order not to ignore this fact during computations.

As for language units and computations, every single occurence of every language unit was accounted for. I.e. in case of repeated occurence of a structurally identical unit (e.g. the same character) in a given text, the

character was not ignored the second time around. On the contrary, it was treated like a new character with equal numerical values. The reason for this is the fact that if repeatedly occuring entries were to be ignored, then e.g. in case of the 2nd analysed text, 81.9 % of all components actually present in the text would be ignored on the language level U_4. The same with different loss rate would happen on all other language levels.

Chosen language units used for segmentation in this paper were created specifically for usage in these experiments, and thus might differ from familiar units of the same name. This is particularly true for the *character* and *component* units, which differ significantly from concepts of the normally used Japanese units of the same name. The definitions below must be understood in context of this experiment aimed at graphic segmentation, and they apply only within boundaries of this experiment. They should not be confused with usually used units or definitions.

The mathematical formula used in this experiment is $y = Ax^{-b}$.[4]

Surparagraph

The surparagraph is a unit which separates thematic pieces of text more distinctively than normal paragraphs. It is graphically represented by offsetting a piece of text by one or two empty rows and starting a new (sur) paragraph. It is not found very frequently in Japanese texts – many books contain no surparagraphs at all. It is not a commonly cited unit and it was chosen solely on the basis of its graphic distinctiveness.

Due to extremely low frequency of surparagraphs in chosen samples, the surparagraph-paragraph language level was labeled U_0 and omitted from the analysis, for the analysis results of this language level will probably not be very conclusive. However, in case of future analysis of longer text samples also containing surparagraphs, the existence of this language level should not be forgotten as it might after all prove important.

Paragraph

A paragraph always starts on a new row (or a column in case of vertical texts) while putting an indentation, usually represented by one empty

[4] See (Andres 2014) for usage reasons.

graphical field at its row or column start. An exception to this rule happens when a new paragraph starts with quotation marks 「 or 『 (﹁ or﹃ in case of vertical texts) or rarely a different punctuation mark, instead of an indentation. A new paragraph starts even without the indentation or a punctuation mark if the previous row or column was clearly terminated in order to create a new paragraph.

Paragraphs in Japanese texts have a relatively clearly defined usage. Although paragraphs don't have grammatical or orthographical functions, their usage is controlled by certain (though not binding) rules.[5]

Sentence

The sentence is usually terminated by a dot 。(*kuten*), the question mark ？(*gimonfu*), or the exclamation mark ！(*kantanfu*). In some cases, the sentence can also be terminated in another way (e.g. in case of an interrupted statement), and there can exist a sentence not terminated despite one of these punctuation marks being used (e.g. in case of inserted statement, quotation, highlighting of text etc.). In these instances, semantics or context have to be taken into account as well, and the sentence termination marks have to be rarely (but always in the same manner) ignored, or better said considered a mere character unit.

Intercomma

In the Japanese language, as well as in other languages, the definition of a word is relatively problematic. Several influential grammar theories exist (among others the so-called School Grammar[6] and the very similar Hashimoto Grammar[7], Tokieda Grammars[8], the Yamada Grammar[9] etc.), which vary significantly in their approach to the concept of a "word" or its equivalent. Thus, instead of a *word* unit placed in between the sentence and the character units, an easily segmentable unit of an intercomma, identically

[5] See (Bunshō no kakikata, totonoekata: 5, danraku no tsukurikata, 2014) for more information on usual usage of paragraphs in Japanese texts.

[6] *Gakkō bunpō* 学校文法

[7] *Hashimoto bunpō* 橋本文法, or *Kokugo hōyō setsu* 国語法要説 respectively

[8] *Tokieda bunpō* 時枝文法, or e.g. *Gengo katei setsu* 言語過程説 respectively

[9] *Yamada bunpō* 山田文法, or e.g. *Nihon bunpōron* 日本文法論 respectively

placed in between the sentence and the character units, was experimentally used.

Commas 、 or ， (*tōten*) in Japanese have – among other functions – an auxiliary function to help readers to read more fluently and comprehend written text more easily. Of course, some more or less applied rules of comma usage exist (usually, but not always, after certain conjunctions and particles, to connect two clauses, in listings, for differentiating between time and cause-effect consecutions, etc.) but in many cases, commas might be used while not being obligatory – their usage is not always bound by strict rules. A comma can separate clauses, delimit listings etc., but the necessity of its usage can be ambiguous and in many cases it is up to the author when and where to use a comma in questionable instances.[10]

However, one principle probably always applies: If a comma appears in a text, it provides readers with a clue as to where to make a brief pause during their reading, or at least where to slow down and take a brief look at a few following characters before continuing reading at a normal pace again. It can thus be assumed that the Japanese comma has, among other functions, the role of a graphical marker of a pause or slowdown during reading, which always applies.

This experiment is based on the graphic form of chosen sample texts. Because of certain ambiguity in the Japanese comma usage in certain instances, the intercomma segment delimited by a new sentence start or a comma on one end, and a comma or the sentence termination markers on the other end[11], was created. This segmentation method allows for easy identification of the intercomma segment on a strictly graphical basis.

[10] The following sentence (Nitsū, 2004, p. 25) makes a good example:
"あの選手はオリンピックではなばなしい活躍をした。"
As the particle で is very often followed by the particle は (hiragana grapheme は is pronounced differently when used as a particle than when used as a single syllable), it is easy to misread it the wrong way. However, in this sentence, は is not a particle and it thus should be pronounced differently than is normal during usual quick reading. To prevent confusion, one can write the same sentence (without any alteration in its meaning in either form) with a comma (1) or with the help of *kanji* (2):
1. あの選手はオリンピックで、はなばなしい活躍をした。
2. あの選手はオリンピックで華々しい活躍をした。
However, the original sentence was not erroneous as well.

[11] In this case an intercomma and a sentence are terminated simultaneously. If Arabic numerals with mathematical marks that would normally signify a border of an intercomma or a sentence appeared in a text, it would be treated as an exception and not signify a border of the intercomma or the sentence.