ROCZNIK ORIENTALISTYCZNY, T. LXV, Z. 1, 2012, (s. 107–119)

NORBERT KORDEK

Segmentotactics of Mandarin Chinese

Abstract

The aim of this paper is to propose an extension of Bańczerowski's phonotactic theory and introduce a research project on Mandarin Chinese segmental linguistics in the proposed theoretical framework. The proposal includes a brief summary of Bańczerowski's original framework and a more detailed description of its extension to various levels of linguistic analysis. The application of the consequent extended framework to the analysis of the phonetics, phonology (phonemic and syllabic levels), morphology and writing system of Mandarin Chinese is proposed in the main part of the paper.

Keywords: linguistics, Mandarin Chinese, segmentotactics

1. Bańczerowski's framework

In his only published paper to date on his phonotactological theory, Bańczerowski shows its application to a fragment of Mandarin Chinese (MC) phonotactics, using the example of the *pinyin* transliteration.[1] In the introduction to his research he provides a non-formal account of the axiomatic phonotactological theory. Such an introduction is also necessary for our purposes, but is should be stressed that the theory in question is in fact a complex axiomatic system. More importantly, in this paper we treat the phonotactological theory as a departure point for a more flexible segmentotactological framework which is capable of multi-level analysis. In this approach, phonotactology is a member of the class of segmentotactological theories.

[1] Bańczerowski 2009.

Conformity with Bańczerowski's work requires us to distinguish between:

– segmentotactology (a class of linguistic theories)
– segmentotactics (the subject matter of segmentotactological theories).

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

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

(i) it should be sufficiently representative of the vocabulary of a given language;
(ii) the entries should be solely words (not including syntagms composed of more than one word);
(iii) it should be accessible in an electronic form;
(iv) the word-entries should be given in phonetic transcription.[2]

Confronted with reality of Chinese electronic dictionaries, such conditions turn out to be rather demanding and present the most challenging task in conducting the research. For example the *MDGB English-Chinese Dictionary* (CC-CEDICT) that was used by Bańczerowski contains numerous syntagm-entries, which means that it does not satisfy condition (iv).

1.1. Terminology

The uniqueness of Bańczerowski's phonotactological theory is reflected in the terms he uses. That is one reason why the terminology needs to be explained here. Another reason is that the framework expanded to any other level of linguistic analysis will utilize analogons of the term coined for the need of phonotactics. The definitions are quoted after Bańczerowski.

An ***utterance*** is "a spatio-temporal physical object, individual and concrete, produced *hinc et nunc* by a definite speaker in a definite time and space… In a certain sense an utterance is a linear object consisting of phonical substance, having its beginning, duration and termination in time, and immediately preceded and succeeded by pauses."[3]

A ***vocabulon*** (*actual word*) is a "maximal unit of linear, that is, sequential, ordering of an utterance. Putting it differently, the linear structure of an utterance may be imagined as a sequence consisting of vocabulons as always linearly continuous and relatively easy distinguishable units within utterances."[4]

A ***phonaton*** is "any subvocabulonic part or segment of various size, provided it is linguistically relevant. Each phonaton is also as individual and concrete as its corresponding vocabulon and it is always a linearly continuous unit. Needless to say, every vocabulon

[2] Ibid., p. 9.

[3] Ibid., p. 10.

[4] Ibid.

will be treated as a particular kind of phonaton. Thus, every vocabulon is also its own subphonaton."[5] Further Bańczerowski distinguishes two kinds of phonatons:

– proper,
– virtual,

the latter being an asubstantial object (zero segment, a pause, a moment of silence).[6]

A ***phonon*** is a minimal phonaton; this term is preferred to ***sound*** or ***actual phone*** for technical reasons.[7]

A ***phone*** is a set of homophonous (auditory indistinguishable) phonons.

A ***vocable*** is a set of homophonous and homosignificant vocabulons. The term ***word*** would be ambiguous in this terminological setting.

To the notions of a phonon and a phone there correspond two possible linear representations of vocabulons and vocables:

– ***phononotacton***,
– ***phonotacteme***,

the former being the linear representation of vocabulons in terms of phonons, the latter the linear representation of vocables in terms of phones.[8] In research practice the notion of phonotacteme will be utilized much more frequently.

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

1.2. Tactophoneme

The vocables consist of sequences of phones; a different way of putting it is to say that certain sets of phones sequentialize (tactify) in the vocables. A ***tactophoneme*** will be conceived as a set of phones that tactify in a phonotacteme. We shall avail ourselves of the example of Polish tactophoneme {*A, K, T*}[9] which is a set of three phones; out of all possible permutations eight result in phonotactemes representing the corresponding Polish vocabulons: *AKT, TAK, KAT, TKA, AKTA, KATA, TAKA, ATAK*.

The properties of a tactophoneme may be described in terms of:

(i) ***phonicity***: the number of phones which are its elements;
(ii) ***phonotactemic range***: the set of all phonotactemes generated out of it;
(iii) ***phonotactemicity*** (phonotactemic load): the number of all phonotactemes generated out of it.[10]

[5] Ibid.

[6] Ibid.

[7] Ibid.

[8] Ibid., p. 11.

[9] For reasons of illustration - at this point Polish serves as a better example than Chinese, as will be explained later.

[10] Bańczerowski 2009, p. 13.

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

(i) phonicity: 3

(ii) phonotactemic range: {*AKT, TAK, KAT, TKA, AKTA, KATA, TAKA, ATAK*}

(iii) phonotactemicity: 8.

Other important properties of phones are described by their:

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

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

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

(iv) ***phonotactemic dispersion number*** – the number of all phonotactemes in which a given phone occurs.[11]

The other important property of tactophonemes is described by their ***phonotactemic efficiency*** – the ratio between the phonotactemicity and the phonicity of a given tactophoneme.[12] The phonotactemic efficiency of the exemplary tactophoneme {*A, K, T*} equals 2.6 (its phonotactemicity is 8, and its phonicity 3). The notion of phonotactemic efficiency may be extended to the whole family of tactophonemes. In the extended interpretation the most obvious understanding, but not the only one possible, involves treating it as the ratio between the number of all phonotactemes and the number of all tactophonemes.[13]

The above terms are sufficient for our main purpose, i.e. introducing the concept of extending the framework to different levels of linguistic signs structure.

1.2. Phonotactics of Mandarin Chinese

The detailed results of phonotactemic analysis of MC and Polish are presented in Bańczerowski 2009. Due to the unavailability of electronic word databases (dictionaries) with phonetic or phonematic transcriptions, the results of Bańczerowski's research in fact pertain to the domain of orthographic systems of Polish and *pinyin* transliteration of Chinese. However it seems justified to assume that the results in Bańczerowski's paper are a fairly close approximation of the MC and Polish word tactics in terms of the phonematic structure, or at least it gives us some insight into the tactical structure of words.

The lack of appropriate databases is less problematic for MC than it is for Polish. Due to the syllable structure and the small number of syllables it is possible to create an algorithm which automatically converts syllables represented in *pinyin* transliteration into the phonetic or phonematic transcriptions.

[11] Ibid., p. 14.

[12] Ibid., p. 15.

[13] Ibid.

The restrictions on MC syllable structure make the possible linear alignments of phones or phonemes rather predictable. The small size of the MC syllabary, which consists of 404 syllables not taking tones into account, and the lack of influence of morphology on the syllable structure , also make the task much easier than in the case of Polish.

The MC phone inventory is a controversial issue.[14] In Dyczkowski et al. we proposed an inventory for MC consisting of 76 phones,[15] but in the most updated version we have significantly reduced the inventory to 52 phones:

[pʰ p̩ʰ b b̩ b̥ b̥̩ tʰ t̩ʰ d d̩ d̥ d̥̩ kʰ g g̊ f s ʃ ʐ ɕ x ʦʰ dz dz̊ ʨʰ ʥ ʥ̊ ʧʰ ʤ ʤ̊ m m̩ n n̩ ŋ ŋ̩ ɲ l j w ɥ i y u e ə ɤ æ a o ɿ ʅ]. Conducting the computational analysis on the transcribed syllabary, and consequently applied to the lexical entries in an electronic dictionary, instead of on the transcribed texts, probably requires further reduction of the inventory, but this issue is beyond the scope of this paper.

For the sake of terminological clarity it should be noted that the terms *tactophoneme* and *phonotacteme* both pertain to tactical analysis in terms of phones, which in the former case might be misleading in suggesting a phonemic analysis. It seems justified to adjust the terminology by applying the terms ***tactophone*** and ***phonotacteme*** for the *phonotactical* level of analysis, while for the *phonemotactical* level the terms ***tactophoneme*** and ***phonemotacteme*** should be used.

The terminological issues, and the fact that Bańczerowski had to inquire into orthographical systems rather than phonotactic ones, paradoxically prove the versatility of the framework.

This paper aims to explore this versatility by expanding Bańczerowski's proposal beyond phonotactics.

2. Beyond phonotactics

Bańczerowski was well aware of the fact that he was exemplifying his phonotactical framework with an inquiry into a different level of language structure, which he calls graphotactic.[16]

It is a common and prevailing practice in Chinese linguistics to reduce the phonetic and phonological studies of MC to a subfield of psycholinguistics, cognitive linguistics or cognitive psychology; Myers and Tsay, for example, place the phonetic studies within the field of experimental psycholinguistics – phonetic studies "test hypotheses about how phonological knowledge (competence) is actually used (performance), and as such they can be highly relevant to phonological theory and to our understanding of how the mind works."[17] One of the consequences of this approach is the assumption that positing

[14] The issue of the phone inventory controversy was addressed in Dyczkowski et al. 2009.

[15] Ibid., p. 90.

[16] Bańczerowski 2009, p. 17. For this level we use a different term, for reasons to be explained in the following sections.

[17] Myers, Tsay 2003, p. 30.

segmental phonetic units is not necessary in MC. Segmentotactics explicitly presumes the opposite – that the segmental phonetic units as a representation of the phonological system are a valid and interesting subject of linguistic inquiry.

2.1. Orthotactics

Orthotactics is used as a replacement term for the *graphotactics* mentioned above. The reason for this becomes apparent when we are confronted with the diversity of the writing systems of world languages. It is probably justified to assume that in the case of languages using alphabetical writing systems the two terms could be used synonymously, since it seems difficult to associate them with two different levels of tactical analysis in those languages – there is no relevant graphical level of the writing system other than orthography. However, the same cannot be said of languages with non-alphabetical writing systems, such as Chinese. The graphic aspect is inherently associated with the Chinese script; on the other hand it is not immediately obvious what *orthography* means in reference to MC. The units that tactify into the written representations of vocables in the two types of writing systems are of a very different nature. The letter type units in alphabetical systems more or less directly reflect the phonetic or phonemic properties of a vocable, while in the case of the Chinese logographic script the internal structure of individual characters is not restricted by such properties of vocables. In other words, if we accept this terminological distinction, orthotactics would pertain to writing systems dependent on the phonetic and phonological properties of a given language, while ***graphemotactics*** would refer to systems independent of phonetics and phonology.

The orthotactics of non-alphabetical writing systems, including Chinese script, is not then a direct inquiry into the writing system, but instead into its alphabetical transliteration. The proposed terminology is analogous to the phonotactical case. The ***orthotacteme*** will be the linear representation of vocables in terms of letters. The ***tactorthoneme*** will be conceived as a set of letters that tactify in a orthotacteme. The following terms relate to a tactorthoneme:

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

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

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

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

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

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

As was already mentioned, in the case of analysis of the Chinese script, the orthotactic analysis is an inquiry into the transliteration system. The results based on the *pinyin* transliteration presented by Bańczerowski (2009) reflect the relevant properties of

MC. For example the orthotactemic efficiency is expected to be lower than in Polish. The reason for this is the syllable and word structure of MC and the related issue of the syllable-morpheme-word correspondence.[18] The typical word in MC consists of two syllables. Every syllable is subject to rigorous restrictions on its linear structure. Typically only one permutation of the elements of a tactophoneme is allowed (the same is true for a tactorthoneme). For example the tactorthoneme {L, O, N, G} tactifies into one orthotacteme only: {LONG}. The only theoretical possibility of increasing the orthotactemic efficiency of most MC tactorthonemes is the existence of a vocable consisting of a duplicated syllable – {LONGLONG} in this particular example. In the case of tactorthonemes that can tactify into bisyllabic vocables, for example {S, H, I, H, E}, the typical efficiency equals one, with the exception of cases where there exist orthotactemes representing the vocables with reversed syllabic linear order. In the above example the orthotactemic efficiency equals 2, since both orthotactemes SHIHE and HESHI represent vocables of MC. The restrictions on the linear order of syllables and the related small number of syllables in MC are the main factors which reduce phonotactemic and orthotactemic efficiency. On the other hand the possibility of syllable duplication and permutations in the syllabic linear order – a phenomenon non-existent in Polish – increase the efficiency. In extreme cases the efficiency may increase to values not seen in Polish:

{N, A, I}: {NAI, NIAN, NAINAI, NIANNIAN, NINA, NANI, NAINA, NA'NAI, NAN'AI, AINAN, NANAI, NAINAN, NANNAI, NINIAN, NIANNI, AINAI, NAIAI, AINA, NAAI, AINIAN, NIANAI, NI'AN'AI}. Intuitively, out of the properties having an opposite effect on efficiency, the number of syllables and the restrictions on linear order within the syllable are expected to dominate the tactical properties of MC vocables. This intuition is confirmed by the results obtained by Bańczerowski. The orthophonemic efficiency of Polish is 1.36 while that of MC is only 1.11.

2.2. Phonemotactics

The orthotactic analysis of *pinyin* transliteration was conducted mostly for the reason of its accessibility in terms of the existing databases. The linguistic importance of an inquiry into the transliteration system is perhaps questionable, but the research output can be at least considered as an approximation of the phonemotactical one. Phonemotactics is of course understood as a tactical analysis of vocables in terms of phonemes.

The phoneme inventory is much less controversial in MC than the repertoire of phones. For the purpose of the computational tactical analysis, the inventory or even the complete phonemic transcription of all MC syllables in the referential work of Duanmu[19] can be adapted.

[18] These properties have significance for every type of tactical analysis of MC, not only orthotactical.

[19] Duanmu 2007.

The proposed terminology is analogous to the phonotactical and orthotactical cases. The ***phonemotacteme*** will be the linear representation of vocables in terms of phonemes. The ***tactophoneme*** will be conceived as a set of phonemes that tactify in a phonemotacteme. The following terms are related to a tactophoneme:

(i) ***phonemicity***: the number of phonemes which are its elements;
(ii) ***phonemotactemic range***: the set of all phonemotactemes generated out of it;
(iii) ***phonemotactemicity*** (phonemotactemic load): the number of all phonemotactemes generated out of it;
(iv) ***phonemotactemic dispersion*** – the set of all tactophonemes to which a given phoneme belongs;
(v) ***phonemotactemic dispersion number*** – the number of all tactophonemes to which a given phoneme belongs;
(vi) ***phonemotactemic efficiency*** – the ratio between the phonemotactemicity and the phonemicity of a given tactophoneme.

The numbers themselves show that phonemotactic analysis should render different results than the orthotactical one. In the *pinyin* system there are 29 symbols, while there are 24 phonemes in the MC inventory.[20] Straightforward logic would tempt one to speculate that in MC the phonemotactemic efficiency will be higher than the orthotactemic efficiency, but the issue is probably more complicated than that.

2.3. Syllabotactics

One of the most characteristic typological features of MC is its syllable prominence. Due to the one-to-one correspondence of syllables and morphemes,[21] the syllabic structures also shape and determine the morphological and lexical structures.

The following terminology is proposed: the ***syllabotacteme*** will be the linear representation of vocables in terms of syllables, and the ***tactosyllable*** will be conceived as a set of syllables that tactify in a syllabotacteme. The following terms relate to a tactosyllable:

(i) ***syllabocity***: the number of syllables which are its elements;[22]
(ii) ***syllabotactemic range***: the set of all syllabotactemes generated out of it;
(iii) ***syllabotactemicity*** (syllabotactemic load): the number of all syllabotactemes generated out of it;
(iv) ***syllabotactemic dispersion*** – the set of all tactosyllables to which a given syllable belongs;
(v) ***syllabotactemic dispersion number*** – the number of all tactosyllables to which a given syllable belongs;

[20] Ibid. The number of phonemes differs according to different accounts.

[21] Also the Chinese characters are in direct correspondence to morphemes, and thus are in the same kind of correspondence with syllables.

[22] *Syllabicity*, which is a more obvious terminological choice, is already in use in phonology.

(vi) ***syllabotactemic efficiency*** – the ratio between the syllabotactemicity and the syllabocity of a given tactosyllable.

A detailed description of the syllabic system of MC is beyond the scope of this paper; some very basic properties must, however, be addressed. It was already mentioned that the syllabary of MC consists only of slightly more than 400 syllables, if the tones are ignored. The average ratio of syllables per word is also very low, since most Chinese words are bisyllabic. In theory this suggests the following:

- the low syllabocity of tactosyllables and their relatively high efficiency;
- the high efficiency of the family of all tactosyllabons.

The latter point may be understood in different ways, namely as:

- the ratio of all tactosyllables to all syllabotactemes;
- the average efficiency of tactosyllables;
- the permutational efficiency of tactosyllables (the ratio of existing tactosyllabons to all potential ones).

Considering just the number of syllables and the average ratio of syllables per word in MC, it is quite obvious that some other factors must be also taken into account, otherwise there would not be enough 'building material' for the lexical level. MC copes with the problem by way of an extremely high homonymy of morphemes and words. On one hand, then, the system itself is potentially highly efficient, while on the other there is a phonological-lexical mechanism which decreases the efficiency. The other important factors influencing the results of syllabotactic analysis are some of the word-formation strategies in MC, which include:

- morphological reduplication,
- inversion of linear order of morphemes[23] in bisyllabic words,
- some types of grammatical affixation,

all contributing to increasing the syllabotactemic efficiency. Some examples of tactosyllables with corresponding syllabotactemic range are:

{YI}: {YI, YIYI} (duplication)
{FENG, MI}: {MIFENG, FENGMI} (inversion)
{YI, DE}: {YIDE, DEYI, YIDEDE, DEYIDE} (affixation + inversion).

The results of the syllabotactic analysis of MC are at this point only a speculation, but it is worth noticing that simply the fact that such analysis is interesting and relevant as linguistic research makes MC (and languages that share similar features of the syllabic system) unique. For example, in the case of Polish, syllabotactic analysis in the vein proposed here does not seem to be relevant. The above proposal ignores the tones in MC, but an actual analysis should take the tones into account; interesting results should be obtained by comparing the tonal and toneless analyses.

[23] Inversion of syllables is even more common, but in the case where the morphological identity of a syllable is not retained it cannot be called a word-formation strategy.

2.4. Morphotactics

The analysis of word structure, word-formation, word semantics, etc. in research practice is almost synonymous with a morphological analysis of some kind. Tactical analysis of words in terms of morphemes is usually understood as a study of their occurrence and ordering restrictions in different phonological environments. Due to the general properties of the morphological system, the analysis that is proposed here cannot be expected to render any interesting results in most languages. In MC, however, there are reasons to believe that morphotactical analysis in the vein of Bańczerowski's framework is justified.

The following terminology is proposed: the ***morphotacteme*** will be the linear representation of vocables in terms of morphemes. The ***tactomorpheme*** will be conceived as a set of morphemes that tactify in a morphotacteme. The following terms relate to atactomorpheme:

(i) ***morphemicity***: the number of morphemes which are its elements;

(ii) ***morphotactemic range***: the set of all morphotactemes generated out of it;

(iii) ***morphotactemicity*** (morphotactemic load): the number of all morphotactemes generated out of it;

(iv) ***morphotactemic dispersion*** – the set of all tactomorphemes to which a given morpheme belongs;

(v) ***morphotactemic dispersion number*** – the number of all tactomorphemes to which a given morpheme belongs;

(vi) ***morphotactemic efficiency*** – the ratio between the morphotactemicity and the morphemicity of a given tactomorpheme.

By stating that morphotactical analysis in most cases would not render any interesting result, we mean that a large number of very inefficient tactomorphemes is the expected result. MC offers more promising possibilities. The reasons are in general the same as in the case of syllabotactics:

- morphological reduplication (for example {人 *man*}: {人 *man*，人人 *people*});
- inversion of the linear order of morphemes in bisyllabic words (for example {蜜 *mì* 'honey'，蜂 *fēng* 'bee'}: {蜜蜂 *mìfēng* 'bee'，蜂蜜 *fēngmì* 'honey'}.

The expected results are relatively high morphotactemicity and morphotactemic efficiency of tactomorphemes.

2.5. Graphemotactics

Probably the most unique tactical analysis in MC refers to one of its most unique features – the script. Even an introductory description of Chinese script is beyond the limitations of this paper, so we will only mention the most relevant features which directly determine the proposed framework:

- discreteness – the characters are composed of recurring elements, except for a small number of one-stroke characters which are not compositional;

– lack of obligatory information – neither phonetic nor semantic information is required as a component part of a character;

– the spatial arrangement (including linear sequence) of components is relevant for distinguishing the characters.

The complexity of the Chinese writing system presents us with the problem of determining the most basic concepts of the tactical analysis of characters. We provisionally propose that the ***grapheme*** is a component part of a character. We have yet to determine the exact meaning of the term – it is possible to understand it as a radical-type component or as a stroke-type component – this problem will be discussed later. The remaining basic terminology is as follows: The ***graphotacteme*** will be the spatial representation of vocables in terms of graphemes. The ***tactographeme*** will be conceived as a set of graphemes that tactify in a graphotacteme. The following terms relate to a tactographeme:

(i) ***graphemicity***: the number of graphemes which are its elements;

(ii) ***graphotactemic range***: the set of all graphotactemes generated out of it;

(iii) ***graphotactemicity*** (graphotactemic load): the number of all graphotactemes generated out of it;

(iv) ***graphotactemic dispersion*** – the set of all tactographemes to which a given grapheme belongs;

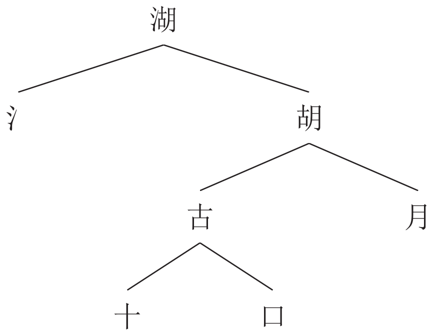
(v) ***graphotactemic dispersion number*** – the number of all tactographemes to which a given grapheme belongs;

(vi) ***graphotactemic efficiency*** – the ratio between the graphotactemicity and the graphemicity of a given tactographeme.

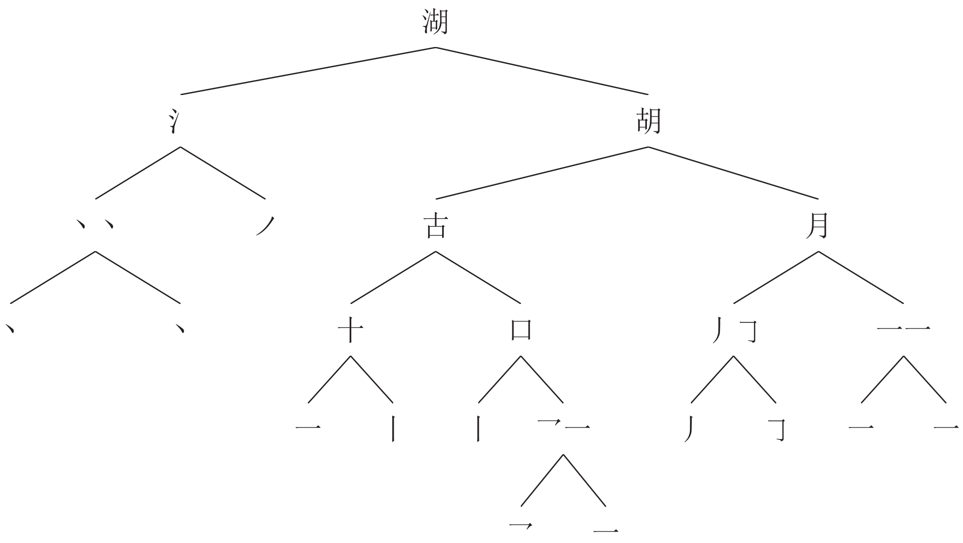
The average efficiency of the tactographemes is not expected to be high, since the majority will have efficiency equal to 1. This is due to the fact that in most cases the same set of components makes up a single character; however the character formation rules allow variations in the spatial arrangement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

{木}: {木*mù* ‘tree’，林 *lín* ‘woods’，森 *sēn* ‘forest’} (recurrence)
{一， 日}: {旦 *dàn* ‘dawn’，亘 *gèn* ‘continuous’} (recurrence)
{一,亅}: {丁 *dīng* ‘cubes’,亍 *chù* ‘footstep’} (recurrence)
{句, 多}: {够 *gòu* ‘enough’, 夠 *gòu* ‘enough’} (linear rearrangement)
{一, 大}: {天 *tiān* ‘heaven; day’, 夫 *fū* ‘man’} (spatial rearrangement).

The components in the examples are the radicals, which is the natural, but not the only possible analysis. The decomposition of characters reveals a multi-layer component structure. The component structure is best represented by IC trees. The following example is the decomposition of the character ‘湖’ *a lake*:



Every single component in the tree (氵 ‘water’, 古 *gŭ* ‘ancient’, 月 *ròu* ‘flesh’, 十 *shí* ‘ten’, 口 *kŏu* ‘mouth’) is a radical or a character with a compositional function (胡 *hŭ* ‘beard’ in the above example), though at different branching levels. The first branching is the most important in the sense that it corresponds to the traditional classification of characters in terms of etymology, components type and spatial arrangement. If we continue the decomposition to the simplest elements, regardless of the complexity of a character, the branching nodes at the lowest possible levels will represent individual strokes.



The branchings of the more-than-two-stroke radicals functioning as character components (氵 (水 shuĭ) ‘water’, 古 *gŭ* ‘ancient’, 月 *ròu* ‘flesh’, 口 *kŏu* ‘mouth’ in the example) are more or less arbitrary. We propose to construct the tree in such a way that in the result the lowest nodes are arranged from left to right in an order corresponding to the stroke order in the whole character. In this way the arbitrariness is addressed by referring to another compositional property.

A few conclusions can be drawn from the above tree. The nodes in this representation do not differentiate between the component types. Not all of the nodes represent the true components of a character, for example the left branching of the true component ' 氵 ' – ('一一')[24] is a part of a true component and at the same time is composed of true ones, but is not one itself. '胡' on the other hand is a complex character on its own that is used as a true component of another character, but it is not a radical.[25] This recalls the phrasal structure of a sentence represented by X-bar syntactic trees distinguishing between the intermediary and true phrasal components. That kind of representation would be suitable for the purpose of character decomposition. In the case of complex characters (like the one in the example) the components in the highest (radicals or complex characters) and the lowest (individual strokes) nodes are a natural units of graphotactical analysis. The unaddressed problem remains the graphotactical status of the intermediate radical components (e.g. '古' '十' '月') and intermediate non-radical ones (e.g. ('一一', '丿 𠃌').[26] In the case of characters with simple structure (functioning as radicals) only the analysis in terms of strokes is available.

It is premature at this point to speculate about the results of the graphotactemic analysis of Chinese script, but it is certain that it will be both a challenging and interesting project. The most serious problem in its realization is the availability of a properly tagged database of Chinese characters.

Bibliography

Bańczerowski, J. 2009. Aspects of Chinese Phonotactics Against a Comparative Background of Polish. *Scripta Neophilologica Posnaniensia* X: 7–22.

Duanmu, S. 2007. *The Phonology of Standard Chinese*. Oxford University Press.

Dyczkowski, K.; Kordek, N.; Nowakowski, P.; Stroński, K. 2009. The Phonetic Grammar of Mandarin Chinese – a Computational Comparative Analysis. *Rocznik Orientalistyczny* LXII. 1: 80–91.

黃錦鋐 (Huang Jinhong), 張正男 (Zhang Zhengnan), 張孝裕 (Zhang Xiaoyu), 黃家定 (Huang Jiading), 葉德明 (Ye Deming). 2008. 國音學 (Mandarin Chinese Phonetics), 中正書局, 臺北.

黄普书 (著) (Huang Pushu, ed.). 2006. *汉字.字源篇学*. 林出版社, 上海.

Myers, J.; Tsay, J. 2003. Investigating The Phonetics of Mandarin Tone Sandhi. *Taiwan Journal of Linguistics* 1.1: 29–68.

Packard, J.L. 2000. *The morphology of Chinese*. *A linguistic and cognitive approach*. Cambridge: Cambridge University Press.

Rogers, H. 2005. *Writing Systems: A Linguistic Approach*. Oxford: Blackwell.

宋业瑾，贾娇燕(著) (Song Yejin, Jia Jiaoyan, ed.). 2003. *实用汉字*. 安徽教育出版社.

苏培成 (著) (Su Peicheng, ed.). 2001. *现代汉字学纲要*. 北京大学出版社.

Wierzchoń, P. 2004. *Gramatyka diakrytologiczna*. *Studium ortograficzno-kwantytatywne*. Poznań: Wydawnictwo Naukowe UAM.

[24] The shape is changed for typographical reasons.

[25] The basic function of radicals is to function as components of complex characters; not all complex characters are used in this function. Radicals may be defined as characters that can be decomposed only into strokes, while complex characters are decomposable into radicals or less complex characters.

[26] One solution to this problem is to allow ternary, quaternary, etc. branchings – the radicals, regardless of their complexity, would be decomposed directly into individual strokes. The binary trees have their advantages, though; it is difficult at the present time to decide which solution is superior.