*Sergey Zinin**

Corpus-driven analysis of the semantic relationship between Chinese characters and their radicals

ABSTRACT: This study analyses the semantic relationship between Chinese characters and their radicals based on unsupervised corpus methods. It discusses the origins and evolution of Chinese radicals, as well as frameworks of their understanding and usage, offered by computational linguists. It particularly addresses the attempts to re-instate radicals as native Chinese ontology. It also elaborates on such issues as multi-syllabicity of the modern Chinese language, as well as the complex, semantic nature of radical groups. Two corpora were used in this study: The Leeds Chinese Internet Corpus (hereafter abbreviated as LCIC) and the custom-made, Classical Chinese Corpus (hereafter, CTexts). The experiments were centered on applying SVD methods, Latent Semantic Analysis (LSA)/LDA topic model analysis, and cluster analysis.

The first approach targets topic model relationships between Chinese disyllabic words and single characters, as well as the relation of characters and radicals. The second approach's goal is cluster validation based on vector-space model representations.

The study discusses the experimental results and suggests other ways to analyze relationships between characters and radicals.

KEYWORDS: Classical Chinese, Modern Chinese, computational linguistics, corpus linguistics, Chinese characters, radicals, semantic analysis, Chinese classics, Warring States Project, Ctexts corpus.

* Zinin Sergey, Warring States Workshop, University of Massachusetts, Amherst, USA; E-mail: szinin@research.umass.edu

Zinin S. Corpus-driven analysis of the semantic relationship between Chinese characters and their radicals. In: Kobzev A.I. (Ed.). Scholarly Reports of the Department of China of the Institute of Oriental Studies, Russian Academy of Sciences, issue 30, The 49th Conference "Society and State in China", vol. XLIX, pt. 2, pages 492-521, Moscow, 2019.

1. Introduction

1.1. Chinese characters and radicals

Definition of radicals. The predominant type of Chinese character is the phono-semantic compound[1], consisting of a semantic graphic component ("radical") and phonetic graphic component ("phonetic")[2]. Most phono-semantic compounds' meanings are presumably related to their radicals. Radicals are graphic symbols, i.e., shared graphic components (glyphs) of characters. Many of them are also meaningful characters; others are glyphs, or their combination, that has no independent existence. Françoise Bottero defines radicals as a set of "recurrent predominantly non-phonetic constituents" that are related to a semantic classification of words[3].

Two "camps" in sinology. Several sinologists claimed that ideography (and existence of radicals) makes Chinese writing system a "universal," "language-independent" tool of communication. They are sometimes called the "radical" or "ideographic camp."[4] The opposing "phonetic camp" denies universal semantic value of Chinese characters; these scholars insist the Chinese writing system is inseparable from the Chinese spoken language. For a long time, debates between these camps have been conducted mostly in philology, concentrating on the nature of characters, i.e., whether they could be called "ideographs" or not[5].

---

[1] See Table 1 (Appendix 4) for all six types. All related materials are available publicly on the article's accompanying GitHub site, at the DOI: https://github.com/wsw-ctexts/radicals.

[2] E.g., they make up about 81% of the 7000 most frequent characters in Chinese orthography (Li and Kang, "Analysis of phonetics of the ideophonetic characters"). Considerable part of other characters is pictograms or ideograms.

[3] This relationship could be complicated. As (Bottéro and Harbsmeier, "The Shuowen jiezi Dictionary", 258) note, "the phonetic role of constituents is specified explicitly by the technical term *sheng* 聲, whereas a non-phonetic constituent X is *not* explicitly characterized as "semantic"; although we find reason to translate the technical term "*cóng* 从 X" as "has X as a semantic constituent" … the non-phonetic constituents are generally construed by Xu Shen as semantic."

[4] See, e.g., (Packard, *The Morphology of Chinese*, 309). Jerome Packard's monograph on Chinese morphology is the best up-to-date source on this subject.

[5] Packard (Packard, *The Morphology of Chinese*, 309) formulates it as a discussion on whether Chinese characters provide "direct access to meaning." He agrees that, due to the "morpheme isomorphism" phenomenon in the Chinese language, "Chinese orthography may be more likely than other orthographies to stimulate activation of the 'meaning' part of a lexical item before its 'sound' part." However, this cannot happen independently of spoken language. If the term "ideographic"

Recently, philological argumentation has been complemented by cognitive psychological studies on the relationship of characters and radicals, concentrating on the role that radicals and phonetics play in perception of Chinese written text (e.g., Chen and Ovid (eds), *Language Processing in Chinese*; Wang et al (eds), *Reading Chinese Script*, etc.).

Computational linguistic research on the semantic relationship of radicals and characters may provide important evidence for this debate. Such research could also have value for Chinese computational linguistics, especially for developing semantic ontologies of Chinese characters.

Origins of systems of radicals If radicals are supposed to be carriers of the most common semantic features, it might seem logical to assign the most common, meaningful character elements to a radicals list. However, that is not how the first known system of radicals came to be.

The practice of adding radicals to characters to differentiate meanings began in the 1st millennium BC[6]. When pictograms and "zodiographs" failed to represent speech adequately[7], scribes started using existing characters for words with similar pronunciation (paronomasia), as well as for words with different pronunciations but similar meanings (polyphony)[8]. Finally, to avoid ambiguity, they started adding a graphic to-

---

means that "the meaning of lexical item exists in a mental lexicon dissociated from and independent of the sound of that lexical item" (*ibid.*), then characters cannot be called ideographic. If the definition of "ideographic" implies there is only "possibility of relatively direct or 'early' access of the 'meaning' part" of lexical item, then "characters could indeed be considered ideographic" (*ibid.*). Another term, "logograph," is sometimes considered a weakened form of the "ideography claim;" see a recent review of these debates at (McDonald, "Getting over the Walls of Discourse"). Meanwhile, William Boltz, one of the leading authorities on the development of the Chinese writing system, uses the term logograph in its direct meaning, as a graph denoting a word (not an idea or concept).

[6] According to Boltz (Boltz, *The Origin and Early Development of the Chinese Writing System*), it is observed in the records on oracle-bones.

[7] Boltz (Boltz, *The Origin and Early Development of the Chinese Writing System*) suggests discerning in the Chinese writing pictograms and their advanced version, "zodiographs." Zodiographs are considered to be more abstract than pictograms. If pictograms stand for "things," zodiographs, remaining ideographs, stand for "words."

[8] Boltz (Boltz *The Origin and Early Development of the Chinese Writing System*, 51–55) calls this spreading of paronomastic and polyphonic practices the second stage of Chinese writing system development.

ken to the loaned character, which indicated its broad semantic, different from the loaned character[9]. This was the origin of radical semanticity.

Obviously, there were no clear-cut rules for adding radicals. Boltz (Boltz, *The Origin and Early Development of the Chinese Writing System*) assumes selection from a well-established list of the most common characters (which later became the basis for the first two-character categories on the *liushu* list). However, it does not seem that a real list of such characters ever existed before the 2nd century AC.

By the end the 1st millennium BC, Chinese scholars reflected on the written system and identified basic types of characters. The phono-semantic compound got its current name (*xíngshēng* **形聲**) in one such classification of the Chinese writing system (*liushu*)[10].

However, until the appearance of the *Shuowen Jiezi* **說文解字** dictionary (hereafter *SWJZ*), radicals were not perceived as a system.

First system of radicals. At the end of the 1st millennium BC, philology (*xiǎoxué* **小學**) was flourishing. First, dictionaries were created and organized on semantic principles.

There were at least two attempts, before the SWJZ, to create dictionaries of Chinese characters[11]. The first was an encyclopedia-like *Erya* **爾雅** (3rd century BC), structured as a description of the world. The second was a dialect dictionary *Fāngyán* **方言**, 1st century BC**)**, which was mostly concerned about character dialect pronunciations.

The concept of radicals as a system (*bù* **部**, or *bùshǒu* **部首***)* was introduced by Xu Shen **許愼** / **许慎** (ca. 58 AC–ca. 147 AC), in his dictionary *Shuowen Jiezi*[12], as a new semantic organization of the Chinese thesaurus[13]. (Boltz, "Shuo wen chieh tzu", 431) calls the invention of

---

[9] See (Boltz, "Shuo wen chieh tzu"). Actually, Boltz suggests that, for a long time, all three stages co-existed (because radicals are found as early as on the oracle bones).

[10] See detailed analysis of the lineage of classification in (Bottero, *Sémantisme et classification*).

[11] There are evidences that other dictionaries were created, but only those two survived.

[12] *"Shuōwén Jiězì"* **說文解字** / **说文解字**; in Wade-Giles transcription: *Shuo-wen chieh-tzu*; the title has many translation. Richard Cook suggested, "Interpreting the Ancient Pictographs, Analyzing the Semantic-Phonetic Compounds" (Cook, *Shuo Wen Jie Zi*, 1).

[13] The SWJZ reportedly contained 9,353 characters, comprising practically all characters that formed the lexicon of classic canons. Not all of them survived in the

radicals "a major conceptual innovation in the understanding of the Chinese writing system."

The *Erya* and SWJZ are similar in their intention to "describe the world" through lists of characters and could be called taxonomies. But if the *Erya* groups semantically-related characters in sections, devoted to a specific subject (e.g., "dwellings" or "utensils"), chapters in the SWJZ are divided into sections, containing characters with the same graphic component related to the section header, which is the radical.

SWJZ contains 540 such section headers (*bù)*, suggesting the universal cosmic completeness of the system[14]. Therefore, instead of optimizing radical sections, thirty-four SWJZ headers have no characters under them, while 159 headers have only one each, i.e., their role is mostly symbolic.

The order of radicals also reflects the cosmic order as known during the Han period[15]. Bottero underlines another philosophical aspect of Xu Shen's innovative classification method. In retrospect, selecting shared semantic graphs as classifiers looks natural. At that time, however, categories such as "heaven," "earth," and "man" were much more common. (Bottero, *Sémantisme et classification*, 55) suggests that the *Yi-jing*'s hexagrams, as an abstract Chinese classification system, might have influenced Xu Shen. If Xu Shen saw radicals as linguistic counterparts of hexagrams, it fits well the idea of the SWJZ's "cosmic completeness".

Radicals' system as the cause of Chinese logography. (Boltz, *The Origin and Early Development of the Chinese Writing System*) offered an original explanation of logography preservation by the Chinese writing system and the SWJZ significance in this process. Relying on analysis of recently excavated, original, pre-Han and early Han manuscripts (i.e., earlier than 2 century BC), Boltz suggests that the Chinese script at this time was very close to de-semantization of characters and creation of a real syllabary.

---

received copies of the SWJZ. (Bottero, *Sémantisme et classification*) reports that for the period from 1500 BC to 500 BC registered so far up to 2,500 characters.

[14] As (Bottero and Harbsmeier, "The Shuowen jiezi", 257) note, "It is clear that the total number of radicals was more important in Xu Shen's eyes than their functional use. ... Xu's choice of sections appears in large part to have been driven by the desire to create an unbroken, systematic sequence among the headers themselves, such that each had a natural, intuitive relationship (e.g., structural, semantic or phonetic) with the ones before and after, as well as by the desire to reflect cosmology."

[15] See, e.g., (Bottero, *Sémantisme et classification*, 164).

However, the Chinese "learned community," supporting the Confucian world-view, opposed trends "toward pure phoneticization and de-semanticization that they recognized in the script, trends in which they themselves were in all likelihood participants."[16]

Concerned scholars launched a program to increase and standardize the use of determinatives (i.e., radicals) and stop the de-semanticization process. This program successfully delivered a "normative, systematized, firmly logographic writing system that reflected the proper order of language and script, world and universe, as it should. This is what we find codified in the *Shuo wen chieh tzu* of A.D. 100, preserving the script in its now unassailable logographic integument."[17] The SWJZ radicals system marks the conservative approach's victory, which happened to be so successful that Chinese writing never again ventured to syllabic script[18].

While Galambos agrees that desemanticized character usage existed, he does not observe it as a leading trend (Galambos, *Orthography of early Chinese writing*, 24–25). Instead, he underlines the importance of the ongoing character standardization process and development of "standard (*zheng*) writing in a centralized bureaucracy" (Galambos, *Orthography of early Chinese writing*, 50) during the Han period. Without this kind of standardization, the SWJZ radical system would be impossible. It, in its turn, enforced standardization of the Chinese writing system.

Semantic and index radicals. While most radicals are independent characters, some simply represent a common element for section characters (e.g., a "dot"). As early as in SWJZ, there is a difference between real semantic radicals and simple "indicators."[19] (Bottero, *Sémantisme et classification*) notes that Xu Shen did not intend to use radicals as a character-retrieval system. They were an analytic conceptual tool. Woon (Woon, *Chinese Writing*) observes that, although most section headers happen to

---

[16] Boltz, *The Origin and Early Development of the Chinese Writing System*, 176.

[17] Boltz, *The Origin and Early Development of the Chinese Writing System*, 177.

[18] There were a few phoneticization attempts later, mostly caused by Buddhist influence. No one of them was successful (see Bottero, *Sémantisme et classification*).

[19] See (Bottero, *Sémantisme et classification*, page 8). There is a growing tendency in linguistics to use the term "key" ("index key") instead of the term "radical". As early as in 1927, (Wieger, *Chinese Characters*, 14; cf. Bottero, *Sémantisme et classification*) uses the terms "keys of the dictionary" and "the 214 keys of K'ang-hsi" for 部首 *bùshǒu*, reserving the term "radical" for any element (not just the root portion) bearing meaning. The term "clé" ("key") is used in the French tradition (e.g., Bottero 1996, page 13).

play a semantic role in the characters listed under them, radicals are not *fundamentally* semantic, but rather are "somewhat arbitrarily chosen."[20]

The radicals' arbitrariness, as semantic components, could be explained by the strong association between characters and spoken language at this time. Imre Galambos observes that the most important pattern behind characters' variability is "retention of the phonetic element. The scribes could abbreviate or leave out almost any other part of the character, could introduce new components, yet they retained the phonetic component in virtually every instance. This realization reinforces the priority of spoken language (sound) over writing (visual form), a connection easily forgotten when it comes to Chinese writing." (Galambos, *Orthography of early Chinese writing*, 3)

S*howen Jiezi* as taxonomy. A close modern counterpart of the SWJZ radicals' system would be a semantic classification, or taxonomy[21] of the WordNet type. The radicals could be compared to WordNet's base types[22], and semantic groups inside radical sections could be compared to synsets[23].

If regular English words permanently included annotation by base types or top taxonomy concepts, it would look similar to Chinese compound characters, e.g.:

[canine]_blenheim_spaniel
[canine]_silver_fox
[canine]_bitch
[person]_bitch
[feline]_lynx

The role of [canine] hypernym would be played in Chinese by the radical #94 犬 *quan*, "canine," role of [person] hypernym — the radical #9 人 *ren*, "human". The character dog *gou* 狗 consists of the shortened

[20] It is one of the earliest opinions where radicals are considered arbitrary. The idea of radical arbitrariness will be further discussed later.

[21] In this study, terms taxonomy and ontology may be synonymous despite their differences. The SWJZ is not an ontology in the sense that it is not built on a rational-logic schema of concepts.

[22] There were various numbers of these types in different versions of WordNet, e.g., (Budanitsky and Hirst, "Evaluating WordNet-based Measures", 16) mention 11 "unique beginners" concepts. This study does not look into WordNet evolutionary history; and only the very concept of top hierarchical concepts is important.

[23] E.g., the typical SWJZ gloss is of the type X Y 也, which (Bottero and Harbsmeier, "The Shuowen jiezi Dictionary", 260) translates as "X is (a kind of) Y," "X is (a way of) Y-ing,"

form of the radical #94 on the left and phonetic 句 ("sentence" *gou)* on the right side. The characters for fox *hu* 狐 consist of the shortened form of the radical #94 on the left and phonetic "瓜" *gua*, "melon" on the right:

[canine]_gou => dog

[canine]_hu => fox

The SWJZ radicals' meanings and order were supposed to be similar to the structure of the universe (like the *Erya*). Characters in sections under radicals were grouped semantically in a kind of synset.

Xu Shen's major innovation is (unlike the *Erya)* that characters in sections all share (presumably) semantic graphic component, and it happened to be one of the most effective character organization concepts in the Chinese dictionary.

The internal semantic structuring of the SWJZ drastically changed later in the *Kangxi* dictionary, where the number of radicals was reduced and characters were organized according to number of strokes[24]. Yet the radical principle itself survived.

Evolution of radical system. Over 1,500 years, there were many variations of the system of radicals, counting different numbers of radicals. Many characters had to be re-assigned to other radicals than the original SWJZ's ones. The semantic aspects of this process require further research.

In 1615, the first dictionary, using the system of 214 radicals, was published[25]. Since the *Kangxi* dictionary (hereafter, KX)[26], which accepted this system, was published in 1716, these 214 graphs have been commonly called the "*Kangxi* radicals" (rather than the "*Zihui* radicals").

Compared to the 540 section headers used in the earlier *Shuowen Jiezi*, the KX dictionary reduced the number of radicals down to 214. Some other changes happened during the reform, simplifying characters in the last half of the 20th century.

Radical-and-stroke system. Having reduced the number of radicals, the *Kangxi* dictionary also introduced the "radical-and-stroke sorting" principle of arranging characters under a radical according to the number

---

[24] Some sinologists call for re-naming "radicals" in KX system as "index keys", or "classifiers", for not being properly semantic categories.

[25] It was introduced in Mei Yingzuo's 梅膺祚 dictionary *Zihui* ("*Character treasure*") from 1615 AC. The 14-chapter *juan* 卷 ("scrolls") dictionary contained a total of 33,179 characters. It also introduced the radical-stroke system, see (Bottero, *Sémantisme et classification*).

[26] It is the largest traditional dictionary, containing 47,035 characters.

of residual strokes[27]. The *Kangxi* system of re-ordering characters and re-assigning many characters to radicals is different from the original SWJZ order. It destroyed the idea of semantically grouping characters within a radical section. Therefore, radicals in this system are "index keys" or, in Norman's terminology, "classifiers" (Norman, *Chinese*, 69)[28].

Unicode radicals' system. This study uses the Unicode (Unihan[29]) system of radical assignation. The Unihan system uses 214 radicals in the KX version, i.e., "each ideograph is assigned one of 214 radicals." In most cases, this assignment is semantic, "in the rest, the radical is arbitrary, based on the character's structure." Also, the way of ordering characters within a given radical-stroke group has changed comparing to KX; the character frequency replaced five types of strokes[30].

Arbitrariness of radicals' system. The radicals' system evolution needs an explanation of possibility re-assigning of radicals while preserving its semantic role. How could one character belong to one radical group in SWJZ dictionary, and to another in the KX dictionary, with the claim that both radicals still relate to character semantics, be valid?

On one side, not all characters changed their radical group; a few key radicals have huge lists of characters that never needed to change, e.g., "water."[31] On the other side, radical meanings could be so generic that characters, by their nature, could relate to many radicals. Boltz (Boltz, *The Origin and Early Development of the Chinese Writing System*) supports the idea of multiple semantic determinatives in one character. According to this concept, assigning a determinative graph (i.e., radical) happened a few times until characters acquired their modern form[32].

---

[27] Residual strokes are the number of strokes required to write everything in the character except the radical.

[28] Boltz prefers the term "semantic determinatives" (Boltz, *The Origin and Early Development of the Chinese Writing System*, 67).

[29] See DOI: http://unicode.org/reports/tr38/tr38–5.html#N101E4

[30] In Unicode 4.0.1, in case there are many characters with same radical and number of strokes, the order is based on frequency: the most common ones come first and the less common ones later. This study omits discussion of the consequences of character simplification reform.

[31] See table of LCIC radical statistics in Appendix 3.

[32] (Boltz, *The Origin and Early Development of the Chinese Writing Syste*, 70), "But whence the hundreds of modern characters with three, four, five, even occasionally, six constituent elements? The answer is that the "add determinative" operation was recursive."

Theoretically, layering radicals ("recycling") could go on indeterminately (but Boltz thinks that even six is very rare in practice).

As mentioned previously, analyzing character (synchronous) variability shows the most important component was phonetic, not semantic (Galambos, *Orthography of early Chinese writing*, 3). At the very early stage, a broad, semantic definition of character was less important than its phonetic attribution.

This phenomenon could be responsible for keeping semantic relationships in radical clusters, even if some characters change clusters–as far as these changes are man-made and meaningful. Wong (Wong, "Fighting Arbitrariness in WordNet-like Lexical Databases") points to the arbitrary nature of any man-made ontology, and it is well-known that the WordNet also passed through several stages of re-ordering[33].

Among other things, it means there is no need to conduct a study on the original SWJZ radicals system, and this justifies accepting the KX system in this study. Any working, man-crafted radicals' classification is good for the aims of this study.

Modern ontological interpretations of radicals. The first Chinese versions of WordNet (e.g., *HowNet*) were developed as knowledge databases and based on "sememes,"[34] or other independently developed classifications, instead of radicals or WordNet's base types. Very soon, researchers like Shun Wong and Karel Pala (Wong and Pala, "Chinese Radicals and Top Ontology in WordNet", Wong and Pala, "Chinese Characters and Top Ontology in EuroWordNet") noticed and investigated the radicals system's similarity to top concept systems of ontologies. Having compared the Chinese radicals and Top Ontology Entities (*EuroWordNet*), Wong and Pala reported very "interesting relations can be found between Chinese radicals and First Order Entities and partly also Second Order Entities" (Wong and Pala, "Chinese Characters and Top Ontology in EuroWordNet"), but no direct correspondence between radicals and Third Order Entities of the SUMO. However, as (Anderson et al., "Base Concepts in the African Languages") indicated, there is no need for base concepts to be mapped to Third Order Entities; these re-

[33] The WordNet's system of base types also was re-hauled a few times; pruning and balancing of branches is an ongoing process.

[34] Sememes (their numbers varied from 700 to 2000, see (Cai et al., "HowNet Based Chinese Question Classification")) were originally selected from 6000 Chinese characters (not polysyllabic words) in a multi-phase process. They could be related to EuroWordNet through SUMO (Alvez et al., "Consistent annotation of EuroWordNet").

searchers already consider Chinese radicals to be "basic concepts" for Chinese WordNet[35].

More researchers tried to re-introduce radicals as a valid framework for building computational linguistics ontologies, as well as to map HowNet (Chinese WordNet) onto the radicals system. (Wong, "Base Concepts in the African Languages", 236) argued that, "unlike most natural languages, the Chinese language displays a considerable amount of semantic information even at the character level." (Chou and Huang, "Hanzi Grid", 8) made a claim that "radicals, the semantic symbols, do form a robust and well-accepted conceptual system." (Wong, "Fighting Arbitrariness in WordNet-like Lexical Databases") considers the radical system a more solid foundation for building a concept system for ontology[36]. Other ongoing projects are the Hantology (Chang, *Gender Roles Reflected in Chinese Botanical Fixed Expressions*)[37], Hanzi Genes (Hsieh, *Hanzi, Concept and Computation*)[38], Hanzi Grid (Chou et al., "Hanzi Grid"), etc.[39]

Radicals' systems, WordNet base types, and informational ontologies all attempt to represent the world's most basic structures. While radicals originated from an ancient nature-philosophical world-view, ontologies tend to reflect modern conceptual hierarchies. How successful mapping the latter is to the former remains to be seen[40].

---

[35] Anderson et al., "Base Concepts in the African Languages", 3761.

[36] (Wong, "Fighting Arbitrariness in WordNet-like Lexical Databases", 237), "while lexical databases often rely on subjective and even ad hoc judgment on concept classification, the semantic relatedness displayed by such clusters of Chinese characters provides a means to concept classification which is more objective, more explicit and, hence, easier to capture."

[37] Hantology is supposed to be a "Prototypical Cross-cultural Knowledge Platform." (Chou and Huang, "Hantology") find that the "Chinese writing system can be treated as a linguistic ontology since it represents and classifies lexical units according to semantic classes." To meet the need of computer applications, as well as the Chinese philological studies, Chou and Huang propose a language resource called Hantology (Hanzi Ontology).

[38] (Hsieh, "*Hanzi, Concept and Computation*") promotes "a new theoretical framework called Hanzi Genes Theory …This theory is based on the discovery of the interpretation of the conceptual dimension of Chinese characters."

[39] (Chou and Huang, "Hanzi Grid") linked the Chinese radicals to the Suggested Upper Merged Ontology (SUMO).

[40] The author is grateful to Gerald Penn, who suggested the subject for this research, to Bruce Brooks for ongoing support of Warring States Workshop's Ctexts project, to Sergei Sharoff for providing the LCIC research corpus, and to Radim Řehůřek for support in reining in GenSim.

2. Words, characters and radicals in the written Chinese

2.1. Chinese word problem

Until the 20th century, most Chinese texts were written in Classical Chinese, a predominantly monosyllabic language[41]. Radicals were invented in this written environment[42]. It seems more beneficial to research the character-radical semantic relationship on a corpus of pre–20th century texts. Any corpus-driven study of the modern Chinese should address an important issue: at what degree could single characters be considered meaningful carriers in a corpus? In addition to being words, are they responsible for creating meaning in text?

Unsupervised research on modern text is complicated, not only by difficulties with word segmentation of modern Chinese texts;[43] there is an ongoing discussion on the nature of the Chinese polysyllabic word itself. It is assumed that most words in the modern written Chinese language are disyllabic words, i.e., written by two characters (Hsieh, *Hanzi, Concept and Computation*) [44].

---

[41] It is not clear, though, if Chinese spoken language had mostly monosyllabic words at the moment when most characters were created. There are clear indications that spoken language was predominantly disyllabic after the 3rd century CE. According to Boltz, the Chinese language was truly monosyllabic only between 1200 to 800 BC (Boltz, *The Origin and Early Development of the Chinese Writing System*, 171). Boltz agrees with George Kennedy's concept that "the writing system, as represented by texts transmitted from the Han dynasty, and especially as registered in dictionaries, effectively camouflages the bisyllabic nature of innumerable words" (*ibid*.) Meanwhile, monosyllabic characters of the archaic Chinese language played a role in de-motivation to invent a writing system, where the phonetic aspect of a syllable would be divorced from the semantic (*ibid*.). Classical Chinese per se may be viewed differently from its later versions, *wenyanwen* or Medieval Classical Chinese, but these distinctions are not significant for this study.

[42] And, according to (Boltz, *The Origin and Early Development of the Chinese Writing System*, 177) they helped to preserve it.

[43] Unlike modern European texts (e.g., English), modern Chinese texts are still written without spaces between words (although, there are some punctuation and sentence borders). Unsupervised sentence segmentation into "word-chunks" is a very complex issue in Chinese computational linguistics. Even supervised segmentation is much more complex and ambiguous than the English one due to Chinese morphology specifics.

[44] Packard (Packard, *The Morphology of Chinese*, 313), writing about the relationship of polysyllabic words and morphemes in Chinese spoken language, concludes "the basic unit of lexical retrieval from the mental lexicon in Chinese natural

If the semantic carrier in modern written Chinese is predominantly the disyllabic word, could the radical-character relationship be studied with no supervision in a modern Chinese corpus? Semantic-wise, a disyllabic word could contain two characters with different radicals, and its semantics would be different from the semantics of both radicals. One way to approach the issue is the topic model approach; also, applications of Latent Semantic Analysis (LSA) techniques to a corpus that is represented in different ways (i.e., consisting of words, single characters and radical classes) and comparison of topics could be useful[45].

2.2. Character and radical semantics

If radicals are related semantically to characters, then characters with shared radicals should be, to some degree, semantically similar or related. Corpus-driven analysis of the radical-character semantic relationship needs to address the nature of semantic similarity of characters with shared radicals.

Important conceptual distinctions have been proposed recently between concepts of *semantic similarity, semantic relatedness, distributional similarity, and distributional relatedness* in a seminal study of (Budanitsky and Hirst, "Evaluating WordNet-based Measures"). Later developments on the subject are summarized by Peter Kolb (Kolb, "Experiments on the difference between semantic similarity and relatedness").

Similarity could be defined by the lexical relations of synonymy and hyponymy, and relatedness by "any kind of lexical or functional association" (Kolb, "Experiments on the difference between semantic similarity and relatedness"). The regular notions of similarity and relatedness are

---

speech production and comprehension is the word, and that individual morpheme access for complex words in Chinese natural speech processing is unlikely." More directly, "Chinese characters are virtually irrelevant to lexical retrieval in Chinese speech production and comprehension" (*ibid.*). This relates to the spoken word, but should have implications for modern Chinese texts, too. Packard's position is very balanced, but leaves a lot of issues to be resolved in future, as one of his reviewers' notes (San, "Review of "The Morphology of Chinese"").

[45] Some manual research on this topic suggests an optimistic outlook. (Wong, "Fighting Arbitrariness in WordNet-like Lexical Databases", 237): "A study on the composite meaning of over 3,400 randomly selected Chinese words has been performed. This study revealed that the underlying meaning of over 99% of them correlates with the meaning of their component characters." On the other side, (Wong, "Fighting Arbitrariness in WordNet-like Lexical Databases", 238) admits that disyllabicity of Chinese words is responsible for the fact that "the Chinese data also display the nature of multiple inheritance in concept formation."

presumably working for concepts. The notions of *distributional* similarity and *distributional* relatedness for words in corpus were introduced as "proxies" for conceptual notions. (Budanitsky and Hirst, 2006) emphasized the difference between semantic and distributional similarity[46].

A study of semantic relationship between radicals and characters should take into account the suggested type of relationship.

However, it seems impossible to postulate a uniform semantic relationship between radicals and characters. Both types of semantic relationships should be present in the SWJZ and in KX's radical-headed sections.

The complex process of characters' evolution was described above in section 1.1. It seems there were no standard (semantic) criteria for choosing radicals to discern meanings.

It seems, however, that the main factor in the relationship between radicals and characters should be semantic relatedness, but similarity should be observed on a regular basis also. Moreover, there is no unique relationship between some characters and a radical. As Boltz demonstrated, in multi-glyph characters there are few glyphs which could convey the vague meaning of a modern character. Multiple determinatives (which allowed several re-shuffles of radical sections, with little semantic misappropriation) create multiple concept inheritance.

Therefore, this study does not implement semantic distance measurements that use taxonomies. The SWJZ radical system was created as taxonomy, and it unequivocally affirms semantic relationships between radicals and characters. The semantic relationships, measured in the SWJZ or KX system, would indicate closeness of characters with shared radicals by default. If semantic relationships between characters and radicals were measured by distances in Chinese HowNet, which is built on the sememe framework, all that we would get, eventually, is the discrepancy between the HowNet conceptual structure and the initial radicals' taxonomy.

A corpus-driven study of the semantic relationship of radicals and characters might reveal interesting facts that taxonomy-based methods would not. It will be necessary to analyze distributional relations.

3. Semantic analysis of radicals and characters

This preliminary, corpus-based study of the semantic relationship between radicals and characters aims at identifying areas of research, based

[46] As (Kolb, "Experiments on the difference") indicates, term-document spaces based on direct co-occurrences capture relatedness, while spaces based on indirect or second-order co-occurrences capture similarity.

on application of LSA methods, and conducting several experiments. It seems that topic model analysis and cluster analysis could be correct basic approaches to the problem.

Topic model analysis. Latent topics (concepts) of corpus documents, extracted with the SVD technique of reducing word-document space, could compare character-based and word-based document meanings[47]. There could be various types of word-document space. Along with polysyllabic words and single characters, it is possible to represent characters by their radical classes and obtain the latent topics of the pseudo-documents, consisting solely of radicals. Finally, it is possible to extract latent topics of pseudo-documents, created by substitution of characters by their English glosses from Unihan database.

Topic words and character distributions contain words and characters that should be related semantically. It is also possible to extract radicals of the topic characters and compare them to the topic.

Cluster analysis. Modern Chinese researchers investigate relationships of characters with shared radicals in WordNet-type taxonomies (e.g., Huang et al, "An Ontology of Chinese Radicals"; Chang, *Gender Roles Reflected in Chinese Botanical Fixed Expressions*, etc.). This type of analysis is interesting, but it does not require corpus analysis and, therefore, is not used in this study.

Another way to evaluate semantic closeness of characters with shared radicals is to view such groups as "radical clusters," created by (partly arbitrarily) partitioning clustering, or a stage in hierarchical clustering.

This study tries to evaluate radical clustering quality by calculating average intra-cluster and inter-cluster distances between radical clusters. The goal is to understand whether characters in a radical cluster are more similar (closer) to each other than to other clusters.

4. Description of corpora and experiment settings

Modern and Classical Chinese corpora. This project used two Chinese corpora for experiments. The main one is the Leeds Chinese Internet Corpus (LCIC, kindly provided for this study by Sergei Sharoff (Sharoff, *Creating general-purpose corpora*). For comparison, a small, custom-made corpus of Classical Chinese was used.

---

[47] (Steyvers and Griffiths, "Probabilistic topic models", 12) indicate," the set of topics derived from a corpus can be used to answer questions about the similarity of words and documents: *two words are similar to the extent that they appear in the same topics*, and two documents are similar to the extent that the same topics appear in those documents."

LCIC. The LCIC has been POS-tagged with an unsupervised parser. Only characters meaningfully tagged by a Chinese POS by the parser were accepted in experiments[48].

Two-character (or disyllabic) words make up most Chinese texts. In this experiment, every two contiguous characters with the same POS were considered a "disyllabic word," i.e., they all are not lexicon words, but bigrams. Only such words were sampled into bag-of-words for experiments with words[49].

LCIC corpus statistics[50] There are 71,135 documents in the corpus; the number of recognized characters (corpus positions) is 337,382,222; the number of unique characters (types)–6,682; and the number of recognized words (corpus positions) is 208,526,733. The average length of document in characters is 4,743, in words – 2,931; the average number of characters per word is 1.61.

Five types of bags-of-words. The LSI and LDA methods were applied to term-document matrices built on five types of bags-of-words, created from original documents. Beside the "original" Chinese web documents, three types of "pseudo-documents" were created to observe relations of the topics, extracted from those collections of documents, to the topicality of the original collection:

Type 1: Single characters;

Type 2: All "disyllabic words" (bigrams);

Type 3: Chinese characters, replaced by their radicals (i.e., replaced by a class representative);

Type 4: Chinese characters, replaced by a string of English words (the Unicode gloss of this character);

Type 5: Type 3 bags-of-words, where each radical was replaced by its English gloss (these study results do not represent significant interest for this study).

Classical Chinese Corpus. The early version of the Classical Chinese corpus (CTEXTS) included seven texts from the first millennium BC: *Chunqiu, Zuo-zhuan, Guliang-zhuan, Gongyang-zhuan, Shi-jing, Mao-shi*, and *Shu-jing*. There are over 190,000 characters (corpus posi-

---

[48] Foreign and unrecognized characters were discarded, as were some parts of texts where the parser failed to identify POS. Therefore, the LCIC statistics for this study could be slightly lower than the original LCIC numbers.

[49] The LCIC contains words of various lengths. Experiments including all words were conducted, and the results were less productive than with "disyllabic words."

[50] See Appendix 3 for LCIC radical statistics.

tions) and 6,562 unique types[51]. Because there are only seven texts, the corpus was split into smaller paragraphs, averaging a few tens of characters each. These paragraphs were considered "documents."[52]

This corpus is not annotated by POS; also, the text is mostly monosyllabic. Therefore, only two kinds of bag-of-words, **type 1** and **type 3**, were considered.

5. LSI and LDA topic model experiments

Experiment platform, settings and limitations The number of documents in these experiments varied from 10,000 to 20,000, depending on workstation capabilities;

Fifty topics were selected for analysis[53].

Stopwords characters and words were filtered out, unless indicated otherwise[54].

Characters with frequency lower than two were filtered out.

For each documents, five types of bag-of-words (defined above) were created.

Topics were extracted using LSI and LDA methods for sets of documents, starting from 100, then 1,000, 10,000, and 15,000 (where possible computationally). Table 3 (Appendix 4) presents the statistical data on characters and words in the experimental sets (partial corpus data).

Experiment software and limitations. All available workstations were Windows 32 bit / 3Gb RAM, with Python 2.6 as the programming language. Due to system limitations, the original Python *svd* functionality failed to process more than a couple thousand documents and the software package Gensim 0.6 was used for LDA/LSI topic extractions[55]. The *Gensim* package website (Řehůřek, *Gensim project*) describes implementation.

---

[51] In terms of character types, it is close to the LCIC, while it is more than 1,000 times smaller. Stopwords (graphs) were removed.

[52] An average Classical Chinese "document" size is about 1% of an average LCIC document, as they are paragraphs in larger documents.

[53] This may be considered a small number of topics. Mostly, 200 to 500 topics would be recommended for a corpus of this size. However, experiments with topic numbers from 25 to 200 showed too much duplication at 100 and more topics.

[54] There was a list of more than 200 characters and words. Stopword characters could make about 30% of documents in the corpus.

[55] It allows overcoming the Windows 3Gb RAM limitations (and it conveniently outputs topics). However, even this package would crash on matrices with document dimensions larger than 10K (for words; 20K for characters).

Experiments with LSI/LDA methods were limited so far by sets of documents with volume from 10,000 to 15,000 documents (40 million to 90 million characters before stopwords removal).

6. Results of topic model experiments

6.1. Chinese words topics

LSI-retrieved word topics matched broad topic classification categories that Sergei Sharoff (Sharoff, *Creating general-purpose corpora*) identified for the LCIC (e.g., natural science, applied science, social science, politics, business, life, arts, and leisure). Most prominent are *school, learning, business* and *social* and *family* life (see files in Appendix 1, Appendix 4).

Word topics showed considerable level of detail (e.g., "*Falungong*").

Experiments sampling words of varied length (mostly, one to three-syllables) and dissyllabic words showed that disyllabic sampling gives better results (more distinctive topics) than various length words (see Packard, *The Morphology of Chinese*).

Experiments with sampling all words or with a frequency of two or more (weeding out low-frequency words) showed no improvement.

LDA-retrieved word topics showed small variance and contained too many functional words[56].

Sample volume effects Topics stabilize after about 3000 documents.

6.2. Chinese character topics

Disyllabic words dominate in modern Chinese text if word and character semantics are close enough, as a manual study of 3,500 words suggests (Wong 2003).[57]

LSI-retrieved character topics The experiments showed that latent character-based topics, extracted with the LSI method, are generally close to word topics, matching the most generic word topics, indicated by (Sharoff, *Creating general-purpose corpora*): *school, learning, shopping, social, and family life*. However, character topics necessarily are not as detailed as word ones. They need to be interpreted properly.

Most often (but not always) topic characters are shared by word topics, i.e., characters related to learning, would be part of disyllabic words related to learning topics. While character topics are less detailed than the word ones, they are somewhat clearer.

---

[56] Unlike characters, removing words is more complicated in the Chinese language.

[57] (Hu et al., "Modeling Chinese documents") view characters as hidden topic-generation tools behind words, but characters used for word generation in that study.

LDA-retrieved character topics. As with words, the LDA topics, extracted by Gensim package, are not articulated well.

Sample volume effects. After about 3,000 documents, the topics stabilize.

6.3. Chinese radical topics

LSI and LDA methods. It is difficult to discover any real "topics" from texts where characters were replaced by their key representatives. At best, it is possible to deduce topics about animals or family.

It is interesting to note that radicals of topics will not match well the radicals of characters in topics that could be close to radical topics. Further, there is no consistency in radicals of characters that create a character topic; i.e., the radicals are not same (with a couple of exceptions).

Sample volume effects. After about 3,000 documents, the topics stabilize.

6.4. English words Unicode definitions

LSI and LDA methods Both LDA and LSI methods retrieve English word topics, e.g., *learning, family, social life,* etc. In general, English word topics are similar to Chinese word and character topics. English word topics even seem to be closer to character topics, rather than to disyllabic words.

Text volume effects. After about 1,000 documents, the topics stabilize.

6.5. Chinese character topics–Classical Chinese corpus

LSI-retrieved character topics

The Gensim LSI method retrieved document-related topics like "Calendar," "feudal states," and "politics," but, in general, results were dominated by numbers and seasonal characters.

LDA-retrieved character topics

The Gensim LDA method performed better (than LSI) for Classical Chinese in the topic model study than the LCIC. There are calendar topics, politics, family, etc.

6.6. Chinese radical topics–Classical Chinese corpus

Radical topics, extracted by LSI method, are richer, than those topics for the LCIC, but still not very meaningful.

7. Cluster analysis

7.1. Experiment description

Relationship of radicals and characters could be studied with a cluster analysis approach. If characters with shared radicals are semantically related (or similar), it should affect these characters' term vector positions

in term-document vector space. A group of characters, sharing radicals, could be viewed as a *cluster* of semantically similar (or related) words.

Evaluating the semantic relationship of radicals and characters in the radical clusters could be a radical clusters' validation task. The radical cluster experiment's goal is to validate the clusters' quality. There are many cluster validation criteria (see, e.g., (Bolshakova and Azuaje, "Cluster validation techniques") for an extensive review). Some of them, i.e., complete distances, centroid distances, and average to centroids distances, are not applicable in the radical cluster case, due to the clusters' variety.

In this experiment, the average linkage (average cosine distance) between all possible pairs of elements (with exclusion of self-distance) will be used to estimate inter-cluster and intra-cluster distances[58].

One is measuring cosine similarity of term vectors, using tf-idf term-document matrix. Another will measure cosine similarity of term vector pairs (rows) in the term-topic matrix, obtained by dot production of the truncated term matrix T_k and the truncated singular matrix S_k (e.g., Kontostathis and Pottenger, "Detecting Patterns", 11).

Three types of clustering were tested. The first type was character groups with the same Unihan radical. The second type was character groups with the same pinyin pronunciation token[59]. The third type was randomly selected clusters. Therefore, *pinyin* clusters were used in this study for comparison, alongside random clusters. The tones were stripped from *pinyin* tokens.

---

[58] Such internal criteria of quality, as e.g., Dunn and Davies-Bouldin indexes were calculated; however, they are not very useful for pre-defined clusters without the clustering process.

[59] This study does not develop, at length, the subject of "phonetic radicals," as characters' phonetics are sometimes called. Some cognitive psychology research presumes that phonetics also carries some semantics, so this type of clustering should be tried, too. Packard discusses, at length, studies trying to identify whether "sound" or "meaning" are activated first in reading, and concludes there is evidence for both phenomena. He suggests that "the access of the lexicon by character orthography consists of the visual stimulus of the written character causing activation … of the lexical entry … with either the sound or the meaning potentially being activated, or coming 'on-line' first, depending on the nature of the activity" (Packard, *The Morphology of Chinese*, 305). (Boltz, *The Origin and Early Development*, 99) provides some evidence (*hsieh sheng* 諧聲 series) that phonetics could have carried some semantics since the 1st millennium BC. Also, Galambos (Galambos, *Orthography of early Chinese writing*)

7.2. Experiment settings

Due to computational limitations, only the first 9,000 documents[60] were sampled to create the term-document matrix.

The SVD was using different vector space-reducing factors — 100 and 50. Characters with per document frequency 1 and stopword characters were excluded. Both the CTEXTS and LCIC were studied.

Three types of clusters. Three types of clusters were created for the experiment. One type is regular radical clusters. Another group is pinyin clusters (character with shared pinyin Romanization). As a baseline, random numbers of groups with random numbers of characters were created:

Type 1: 214 groups of characters, having same radicals.

Type 2: Groups of characters, sharing *pinyin* reading.

Type 3: Random numbers of groups with random numbers (from 1 to 100) of not-overlapping characters, as the baseline.

Cluster quality measurement. For the intra-cluster distance, the average cosine distance between all cluster characters, except self-distance, was calculated. For the inter-cluster distances, the average cosine distance between all characters in all clusters was calculated[61].

7.3. Experiment results

The experimental results are presented in tables containing intra-cluster cosine similarity values of each group: radical clusters, pinyin clusters, and random clusters (see Appendix 3). If this distance were minimal, comparing to inter-cluster average aggregate similarity numbers[62], "hit category 2" was assigned to the group. If intra-cluster distance was one of the first ten averages, "hit category 1" was assigned to the group. Otherwise, the value was 0.

While key clusters definitely lead in numbers of "hits," they do not seem to demonstrate significant inner closeness. The hits observed mostly for groups of 2–7 (maximum at 2–3) characters, and not for majority of such groups.

---

[60] 5,760 unique tokens from 8,880 documents of total 27,571,023 characters. For tf-idf matrix, 5,595 unique tokens from 5,940 documents of total 19,276,545 corpus position were selected.

[61] Another approach would be calculating difference of intra- and inter-distances for each character. It was not used in this study.

[62] I.e., the difference between the intra-cluster average linkage and inter-cluster values was positive. An average cluster value was used for the experiment; as a variation, this value could have been calculated by element, and then averaged.

It seems that they are less random than random cluster numbers, but still not significant.

8. Discussion of results

List of appendices

In this report the list of experiment result files is reduced to necessary minimum, only most important lists are included. The appendixes could be found at the GitHub website, at https://github.com/wsw-ctexts/radicals.

RAD2019_appendix_1_LSI_characters.txt

This file contains results of GENSIM LSI topic output for 15,000 LCIC documents, sampled as characters. 50 topics were chosen, with top 10 distribution characters displayed. Characters are accompanied with their KX radicals and Unihan glosses.

RAD2019_appendix_1_LSI_words.txt

This file contains results of GENSIM LSI topic output for 10,000 LCIC documents, sampled as disyllabic words (compounds). 50 topics were chosen, with top 10 distribution words displayed.

RAD2019_appendix_1_LSI_radicals.txt

This file contains results of GENSIM LSI topic output for 10,000 LCIC documents, sampled as radical classes (of characters). 50 topics were chosen, with top 10 distribution radicals displayed.

RAD2019_appendix_1_LSI_english.txt

This file contains results of GENSIM LSI topic output for 1,000 LCIC documents, sampled as English Unihan glosses of characters. 50 topics were chosen, with top 10 distribution radicals displayed.

RAD2019_appendix_1_LSI_CTEXTS_characters_nostops.txt

This file contains results of GENSIM LSI topic output for over 5,000 CTEXTS “documents”, sampled as characters. 50 topics were chosen, with top 10 distribution characters displayed. Characters are accompanied with their KX radicals and Unihan glosses.

RAD2019_appendix_1_LDA_characters.txt

This file contains results of GENSIM LDA topic output for 15,000 LCIC documents, sampled as characters. 50 topics were chosen, with top 10 distribution characters displayed. Characters are accompanied with their KX radicals and Unihan glosses.

RAD2019_appendix_1_LDA_words.txt

This file contains results of GENSIM LDA topic output for 10,000 LCIC documents, sampled as disyllabic words (compounds). 50 topics were chosen, with top 10 distribution words displayed.

RAD2019_appendix_1_LDA_english.txt

This file contains results of GENSIM LDA topic output for 1,000 LCIC documents, sampled as English Unihan glosses of characters. 50 topics were chosen, with top 10 distribution radicals displayed.

RAD2019_appendix_1_LDA_CTEXTS_characters_nostops.txt

This file contains results of GENSIM LSI topic output for over 5,000 CTEXTS "documents", sampled as characters. 50 topics were chosen, with top 10 distribution characters displayed. Characters are accompanied with their Unihan glosses.

RAD2019_appendix_2_cluster_distances.doc

This file contains results of cluster validation analysis. Results for radical clusters, pinyin clusters, and random cluster are presented.

RAD2019_appendix_3_radical_statistics.doc

This file contains statistics on radicals for the LCIC corpus.

The table contains radical id, number of character types, having this radical, according to the Unihan system, number of documents, containing characters with this radical, total number of characters with this radical, and average number of characters with this radical per document, where these characters are present.

RAD2019_appendix_4_general_statistics.doc

This file contains general description of topical analysis. Among them, there is a table 1, containing description of six types of characters, according to the SW; table 2 with partial statistics for the LCIC, table 3 with statistics on experiments' numbers; table 4 with results of topic analysis.

Discussion

Two types of corpus experiments were conducted in this study: the LSA/LDA topic model analysis, and the radical cluster validation analysis.

In topic model analysis, LCIC and CTEXTS corpora were viewed under a variety of angles in terms of lexical units. The LCIC corpus was tested for disyllabic "words" (bigrams, contiguous two-character same POS-tagged compounds), single characters, key classes of single characters, and English glosses. The CTEXTS corpus was tested for single characters and key classes.

The most important result is that characters' topics generally fall into the same wide categories as the "words" topics; however, they are more abstract (less detailed and less diversified). It is also interesting that pseudo-texts, created from documents where characters were replaced by their Unihan glosses, demonstrated topicality, similar to characters' (but not to words) topicality.

Due to the highly abstract nature of a small set of key categories, it is impossible to extract meaningful topics from pseudo-texts created from radicals.

As a rule, topic characters' radicals (for characters' distributions) vary considerably. However, distributional characters for topics often are parts of distributional "words" for similar topics[63].

All in all, this study showed that modern Chinese texts, viewed as bag-of-words of characters, and not polysyllabic words, still hold the most important document space topics (however, they are more abstract).

Finally, the corpus-driven attempt to validate radical clusters, did not discover significant closeness of such groups. Results that may be "positive" are observed only for radical clusters of size two to four characters.

Average linkage cluster distances do not support directly claims like (Chang, *Gender Roles Reflected*, 32) "characters with horse radicals are all related to the horse in different aspects," or Chou et al. claim that "of the 444 characters containing the semantic symbol 艸, there is no doubt that they are all related to the concept 'plant'."

There is no doubt that, as (Chou et al., "Hanzi Grid") stated, "The conceptual clustering is more complex than a simple taxonomy." It seems that cluster experiments in this study confirm the thesis that "radical is more complex than a simple taxonomy" (Chang, *Gender Roles Reflected*), and large radical clusters could be a "small ontology itself" (*ibid*.) and should be broken down into smaller synsets, like in the original SWJZ thesaurus. For example, (Chou et al, "Hanzi Grid") the concepts represented by radical 馬 ("horse"), 牛 ("cow"), and 木 ("wood") also could be divided into four classes. This could explain the greater average cluster distances for smaller clusters.

According LCIC statistics, (see table in Appendix 5) most radical clusters are just a few characters, while just a few clusters include hundreds of characters. To make clusters commensurable, a new taxonomy should be built based on radicals as base types–as many Chinese computational linguists suggest[64].

Another issue to be taken into account is that similar relationships in such groups could be either semantic similarity or semantic relatedness. There is no single principle governing such groupings. Small sub-clusters, with shared radical and mixed types of relationships, should be studied with a combination of methods for similarity and relatedness.

---

[63] I.e., if the topic is "learning," characters for this topic are often characters that are parts of words for this topic.

[64] There are more published articles supporting radicals approach, but not all of them were available (e.g., Hu and Du, "A semantic analysis of Chinese radicals").

9. Future directions

This study is a preliminary attempt to analyze the semantic relationship of characters and radicals in a corpus-driven environment. It helps clarify future studies in this direction.

First of all, it is corpus expansion. Most results were obtained for about 15%–20% of the available documents. The CTEXTS corpus size is definitely smaller than necessary for reliable results. The study results should be obtained on the full set of corpus documents, and the CTEXTS corpus also needs to be expanded at least to a few million tokens.

Second, a hierarchical taxonomy, based on radicals as "base types," should be used to study radical cluster validation. These sub-clusters should be commensurate.

Third, other cluster validation methods should be tried, e.g., methods based on co-occurrences (see, e.g., Chakraborti et al., "Acquiring Word Similarities").

Fourth, the corpus-driven environment allows running clustering experiments. There is enough evidence indicating that all known radical systems are to some degree arbitrary, and (many) characters could be assigned to different radicals. Semantic clustering could identify the most important of these relationships. The results could be compared with existing radical clusters (or sub-clusters), and the differences analyzed for better understanding of semantics of radicals in modern and classic texts.

10. Conclusions

There is growing interest in the Chinese computational linguistics in restoration of the radicals system as a native conceptual ontology for Chinese language. It is considered (Chou and Huang, "Hantology", page 8) "a robust and well-accepted conceptual system." The radical system is familiar to all literate Chinese. A considerable part of this claim is based on radicals being a semantic component of characters. Therefore, corpus-driven information on their semantic relationship to characters could help corroborate or deny the idea of building a new top concept ontology for systems similar to WordNet. This study tried to identify whether radicals, as a graphic constituent of characters, still have a considerable semantic significance.

Two types of experiments on two types of corpora were conducted to investigate this problem. Firstly, it was tested if available corpora provide any evidence that single characters still have meaning in a modern, predominantly disyllabic word environment. The topic analysis, using LSI/LDA techniques, of the modern Chinese corpus (LCIC) showed that "character topics" are similar to "word topics," even though they are– necessarily — more abstract. The results of characters' topic model

analysis on the modern Chinese corpus were close to the analytical results conducted on the (small) Classical Chinese corpus CTEXTS.

However, pure "radical topics" (analytical results of pseudo-documents consisting of radical characters) are too abstract and do not carry significant information on the corpus topic model. Further, radicals of characters included in topic character distributions showed considerable variance; i.e., characters with same radical do not cluster to create a topic.

Secondly, groups of characters sharing radicals were viewed as clusters, and the average semantic similarity of characters' vectors was evaluated through regular cluster analysis metrics. Experiments conducted by using a vector-space model and SVD did not find significant semantic cohesion in these clusters.

There are several reasons. It may be necessary to break large, radical clusters up into commensurable, related groups. However, in this study, even small clusters could not demonstrate significant similarity. Some similarity could be observed for small groups in the CTEXTS corpus; but the corpus requires a significant increase in size to corroborate this claim. Another reason is the need to apply methods which would recognize both semantic similarity and relatedness.

Finally, all radical systems are, to some degree, arbitrary. The current clusters may not be optimal, and a further clustering process could identify the most strong radical-character relationship.

Corpus-driven studies suggest a new approach to radical ontologies, relying on distributional similarity and relatedness data. They may be used for automatic retrieval of conceptual relationships between characters, and, further, evaluation of their relation to such abstract concepts that are represented by radicals. This research requires developing new, corpus-driven methods of studying semantic relations of characters and radicals.

References

Álvez Javier, Atserias Jordi, Carrera Jordi, Climent Salvator, Oliver Antoni and German Rigau. "Consistent annotation of EuroWordNet with the Top Concept Ontology." In *Proceedings of the Fourth Global WordNet Association Conference (GWC'08),* Szeged, Hungary, 2008.

Anderson Winston, Pretorius Laurette, and Albert E Kotzé. "Base Concepts in the African Languages Compared to Upper Ontologies and the WordNet Top Ontology." In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10),* Valletta, Malta, 2010,

Bolshakova Nadia F. and Azuaje Francisco. "Cluster validation techniques for genome expression data classification." *Signal Processing 83 no.1 (*2003): 825–833.

Boltz William G. "Shuo wen chieh tzu". In: Loewe, Michael (ed.). *Early Chinese Texts: A Bibliographical Guide, (Early China Special Monograph Series*

No. 2), Berkeley: Society for the Study of Early China, and the Institute of East Asian Studies, University of California, 429–442, 1993.

Boltz William G. *The Origin and Early Development of the Chinese Writing System*. American Oriental Series, vol. 78. American Oriental Society, New Haven, Connecticut, 1994.

Bottéro Françoise and Harbsmeier, Christoph. "The Shuowen jiezi Dictionary and the Human Sciences in China." *Asia Major. Third Series 21 no.1 (2008):* 249–271.

Bottéro Françoise. *Sémantisme et classification dans l'écriture chinoise, les systèmes de classement des caractères par clés du Shuowen jiezi au Kangxi zidian*. Paris, Institut des Hautes Études Chinoises, 1996.

Budanitsky Alexander and Graeme Hirst. "Evaluating WordNet-based Measures of Lexical Semantic Relatedness." *Computational Linguistics*, 32, no.1 (2006): 13–47.

Cai Dongfeng, Sun Jingguang, Zhang Guiping, Lv Dexin, Dong Yanju, Song Yan and Chao Yu. "HowNet Based Chinese Question Classification." In: *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC 20)*, 366–389, Wuhan, China, 1–3 November, 2006.

Chakraborti Sutanu, Wiratunga Nirmalie, Lothian Robert and Stuart N.K. Watt. "Acquiring Word Similarities with Higher Order Association Mining." In: *Case-Based Reasoning Research and Development, 7th International Conference on Case-Based Reasoning, ICCBR 2007*, Belfast, Northern Ireland, UK, 61–76, 2007.

Chang Chao-Huang and Cheng-Der Chen. "Automatic clustering of Chinese characters and words." In*: The Proceedings of Rocling VI Computational Linguistics Conference VI*, Taiwan, 1993, 57–78.

Chang Chu-lin. *Gender Roles Reflected in Chinese Botanical Fixed Expressions.* A Thesis for Masters of Arts, Graduate Institute of Foreign Languages & Literature, National Cheng Kung University, Taiwan, 2008.

Chen Hsuan-Chih and Ovid J.L Tzong (eds.) *Language Processing in Chinese (Advances in Psychology).* Amsterdam, Netherlands, 1992.

Chou Ya-Min and Chu-Ren Huang. "Hantology: An Ontology based on Conventionalized Conceptualization." In: Chu-Ren Huang et al. (Eds.) *Ontologies and Lexical Resources for Natural Language Processing.* Cambridge: Cambridge University Press, 2008.

Chou Ya-Min, Hsieh Shu-Kai, and Chu-Ren Huang. "Hanzi Grid: Toward a Knowledge Infrastructure for Chinese Character Based Cultures." In: Ishida T., Fussell S.R., Vossen P.T.J.M. (eds.) *Intercultural Collaboration I. Lecture Notes in Computer Science*, Springer Verlag, 2007.

Cimiano A. Hotho, and S. Staab. "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis." *Journal of Artificial Intelligence Research*, 24 (2005): 305–339.

Cook Richard S. *Shuo Wen Jie Zi - Dianzi Ban: Digital Recension of the Eastern Han Chinese Grammaticon.* PhD Dissertation. Department of Linguistics. Berkeley: University of California, 2003.

DeFrancis John. *The Chinese Language: Fact and Fantasy*, University of Hawaii Press, Honolulu, 1984.

Galambos Imre. *Orthography of early Chinese writing*. Budapest Monographs in East Asian Studies. Budapest: Eötvös Loránd University, Department of East Asian Studies, 2006.

Guan Yi, Wang Xiao-long, Kong Xiang-yong and Jian Zhao. "Quantifying Semantic Similarity of Chinese Words from HowNet." In: *Proceedings of the First International Conference on Machine Learning and Cybernetics*, Beijing, 4–5 November 2002.

Hansen Chad. "Chinese Ideographs and Western Ideas." *The Journal of Asian Studies* 52 no 2 (1993): 373–399.

Hong Jia-Fei, Sue-Jin Ker, Kathleen Ahrens and Chu-Ren Huang. "Sense Prediction Study: Two corpus-driven linguistic Approaches." In *Proceedings of the 11th Chinese Lexical Semantic Workshop (CLSW 2010).* May 21–23, Suzhou, China: SooChow University, 239–246, 2010.

Hsieh Shu-Kai. Hanzi, Concept and Computation: A Preliminary Survey of Chinese Characters as a Knowledge Resource in NLP. In Philosophische Dissertation angenommen von der Neuphilologischen Fakultät der Universität Tübingen, 2006.

Hu He, Du Xiaoyong, Tian Xuan and Ruixue Bai. "A Preliminary Study on the Semantic Strength of Chinese Radicals." In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 2, 658–662, 2007.

Hu He and Xiaoyong Du. "A semantic analysis of Chinese radicals International." *Journal of Business Intelligence and Data Mining* 3: no 4 (2008): 426–436.

Hu Wei, Shimizu Nobuyuki, Nakagawa Hiroshi and Huanye Sheng. "Modeling Chinese documents with topical word-character models." In *Proceedings of the 22nd International Conference on Computational Linguistics — Volume 1* (COLING'08), Association for Computational Linguistics, Stroudsburg, PA, USA, 345–352, 2008.

Huang Chu-Ren, Ya-Jun Yang and Sheng-Yi Chen. "An Ontology of Chinese Radicals: Concept Derivation and Knowledge Representation based on the Semantic Symbols of Four Hoofed-Mammals." In: *The 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC2008),* Philippines: De La Salle University-Manila, 189–196, 2008.

Kolb Peter. "Experiments on the difference between semantic similarity and relatedness." In: Kristiina Jokinen and Eckhard Bick (Eds.) *NODALIDA 2009 Conference Proceedings,* 81–88, 2009.

Kontostathis April and William M. Pottenger. "Detecting Patterns in the LSI Term-Term Matrix." In: *IEEE ICDM02 Workshop Proceedings, The Foundation of Data Mining and Knowledge Discovery (FDM02)*, 243–248. Maebashi, Japan, 2002.

Kontostathis April and William M. Pottenger. "A framework for understanding Latent Semantic Indexing (LSI) performance." *Inf. Process. Manage* 42, no 1 (2006): 56–73.

Landauer T.K., Foltz P.W. and D. Laham. "Introduction to Latent Semantic Analysis." *Discourse Processes,* 25 (1998): 259–284.

Lee Keekok. *Warp and Weft, Chinese Language and Culture*. New York: Eloquent Books, 2008.

Li Y. and J.S. Kang. "Analysis of phonetics of the ideophonetic characters in Modern Chinese." In: Y. Chen (Ed.), *Information Analysis of Usage of Characters in Modern Chinese*, Shanghai Education Publisher, Shanghai, 84–98, 1993.

McDonald Edward. "Getting over the Walls of Discourse: 'Character Fetishization' in Chinese Studies." *The Journal of Asian Studies* 68, no. 4 (2009): 1189–1213.

Mihalcea Rada, Corley Courtney and Carlo Strapparava. "Corpus-based and Knowledge-based Measures of Text Semantic Similarity." In: *Proceedings of the American Association for Artificial Intelligence (AAAI 2006)*, Boston, July 2006, 2006.

Nenadić Goran, Spasić Irena and Sophia Ananiadou. "Term clustering using a corpus-based similarity measure." In P. Sojka et al. (Eds.): *Text, Speech and Dialogue* — TSD 2002, Springer Verlag, 151–154, 2002.

Norman Jerry. 1988. *Chinese*. Cambridge University Press, UK.

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

Penn Gerald. "Quantitative Methods for Classifying Writing Systems." In: *Workshop on Writing Systems and Linguistic Structure, Proceedings of the 18th* International Congress of Linguists (CIL–18), vol. 2 (2008): 175–176.

Řehůřek Radim. Gensim project: *http://nlp.fi.muni.cz/projekty/gensim/index.html*., 2010.

San Duanmu. "Review of "The Morphology of Chinese: A linguistic and cognitive approach" by Jerome L. Packard." *Diachronica* 19, no. 1 (2002): 188–198.

Serruys Paul L.-M. "On the System of the Pu Shou **部首** in the Shuo-wen chieh-tzu **說文解字**". Zhōngyāng yánjiùyuàn lìshǐ yǔyán yánjiùsuǒ jíkān (**中央研究院歷史語言研究所集刊**, Journal of the Institute of History and Philology, Academia Sinica), 55, no. 4 (1984): 651–754.

Sharoff Sergei**.** "Creating general-purpose corpora using automated search engine queries." In: Marco Baroni and Silvia Bernardini (eds.), *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna, 2006.

Sproat Richard. "Corpus-Based Methods in Chinese Morphology (Lecture)". In: *The 19th International Conference on Computational Linguistics (COLING 2002)* Taipei, Taiwan, 2002.

Steyvers M. and T.L. Griffiths. "Probabilistic topic models." In: T. Landauer, D. McNamara, S. Dennis and W. Kintsch (eds.). *Latent Semantic Analysis: A Road to Meaning.* Mahwah, NJ: Erlbaum, 2006.

Ralph and Inhoff Albrecht (eds.). *Reading Chinese Script: A Cognitive Analysis*. Psychology Press, 1999.

Wang Lixun. "Exploring parallel concordancing in English and Chinese." *Language Learning & Technology* 5, no. 3 (2001): 74–184.

Wieger S.J. *Chinese Characters: Their Origin, Etymology, History, Classification and Signification. A Thorough Study from Chinese Documents*. Translated from the French original ca. 1915 by L. Davrout, S.J., orig. Catholic Mission Press; reprinted in US — Dover; Taiwan — Lucky Book Co., 1927.

Wong Ping Wai and Yongsheng Yang. "A Maximum Entropy Approach to HowNet-Based Chinese Word Sense Disambiguation." In *Proceeding SEMANET'02 Proceedings of the 2002 workshop on Building and using semantic networks*, Vol. 11. Association for Computational Linguistics Stroudsburg, PA, USA, 2002.

Wong Shun Ha Sylvia. “Fighting Arbitrariness in WordNet-like Lexical Databases — A Natural Language Motivated Remedy.” In: Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): *The Proceedings of the 2nd Global WordNet Conference (GWC 2004)*, Brno, 234–241, 2003.

Wong Shun Ha Sylvia and Karel Pala. “Chinese Characters and Top Ontology in EuroWordNet.” In *Proceedings of the Global WordNet Conference'2002*. Mysore University, Mysore, India, 224–233, 2002.

Wong Shun Ha Sylvia and Karel Pala. “Chinese Radicals and Top Ontology in WordNet.” In: *Text, Speech and Dialogue—Proceedings of the Fourth International Workshop, TSD* 2001, Pilsen, 10–13 September 2001, University of West Bohemia, Springer, Berlin, 2001.

Woon Wee Lee. *Chinese Writing: Its Origin and Evolution* (in English; Chinese title: **漢字的原始和演變**). Univ. of East Asia, Macau, 1987.

Xǔ Shèn. *Shuōwén Jiezì*. In: Duàn Yùcái. **說文解字注** Shuōwén Jiezì Zhù (“Commentary on the Shuōwén Jiezì”), 1815 (reprint).

С.В. Зинин*

Корпусный анализ семантических отношений китайских иероглифов и их ключей

АННОТАЦИЯ: Статья посвящена анализу семантической связи иероглифов и их ключей, на примере семи классических китайских текстов входящих в состав корпуса “Warring States Workshop (WSW) Ctexts” и Лидского корпуса современного китайского языка. С помощью различных методов компьютерной лингвистики автор пытается выявить семантические отношения между китайскими словами, составляющими их иероглифами и ключами этих иероглифов. Статья содержит обзор современной литературы по вопросу о происхождении системы ключей, и их возможной связи со значением иероглифа. В современной компьютерной лингвистике наблюдается интерес к семантическим иерархиям, образованных ключами. Проведенные эксперименты позволяют утверждать, что, с точки зрения метода тематического анализа, тематика текстов, состоящих из двусложных слов, близка к тематике текста, состоящего из отдельных иероглифов, являющихся частями двусложных слов, в то время как ключи не дают определенного ответа на вопрос о тематике текста.

КЛЮЧЕВЫЕ СЛОВА: китайский язык, корпусная лингвистика, китайские иероглифы, ключи иероглифов, семантический анализ, китайский канон, Warring States Project, корпус Ctexts.

* Зинин Сергей Васильевич, Проект «Сражающиеся царства», Массачуссетский университет, Амхерст, США; E-mail: szinin@research.umass.edu

CONTENTS

We congratulate

History and ethnography

Source studies and canonical texts studies

Literature and arts

目录

庆祝

历史与民族志学

史料学与经学

文学与艺术

СОДЕРЖАНИЕ

Поздравляем

История и этнография

Источниковедение и каноноведение

Литература и искусство