Semanticity in the Chinese Graphic System

Modeling and Assessing Its Consistency

Pierre Magistry & Yoann Goudin

Abstract. Beyond the dominant representation regarding the Chinese characters (sinograms) and its supposed "ideography" among the general public, learners, and (still many) teachers, this paper addresses the semanticity of the Chinese graphic system. We tackle this assumption and investigate the consistency of semantic clues embedded in the sinograms compositions. After a brief overview of principles of the economy of the Chinese graphic system, and the typology of the different functions of components, it is reported how probabilistic and graph-based models were designed in order to assess the semantic contribution of the 214 canonical "radicals". Results show that if *some* semanticity cannot be denied, it concerns a few among not the more frequent components. Discussion then addresses the relevance of the consistency of the semanticity of the system and what is at stake with such a focus for learning and teaching sinograms.

In Memoriam Zhitang Yang-Drocourt & Georges Antoniadis.

1. Introduction

This paper follows a previous publication (Magistry, Fabre, and Goudin, 2017), where we advocate the relevance of crafting phonological-based linguistic resources for learning languages related to the Chinese script (Sinitic languages and other languages written with sinograms). Based on a litterature review in cognitive science and the question of granularity in writing systems, we then demonstrated how grapho-phonological correspondences were helpful for the machine and hopefully—beyond dominant speeches and representations—useful in class for learners of these related languages : Sinographic languages. For this paper, after phonological cues and their reliability, we want to challenge our

Pierre Magistry 0000-0002-9296-8902
ERTIM – Inalco, Paris, France. E-mail: pierre.magistry@inalco.fr
Yoann Goudin 0000-0002-0522-8261
Université Grenoble Alpes, Laboratoire LIDILEM, Bâtiment Stendhal, CS 40700, 38058 Grenoble Cedex 9, France. E-mail: yoanngoudin@yahoo.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 10.
Fluxus Editions, Brest, 2024, pp. 671–688. https://doi.org/10.36824/2022-graf-magi
ISBN: 978-2-487055-06-3, e-ISBN: 978-2-487055-07-0

previous statement by tackling the reverse question and the hypothesis of the semanticity of this graphic system. This later view is mostly shared by sinologists, language teachers and consequently their learners. Thus, it perpetuates the representation according which at least in the West and especially here in the French context, since Jesuites figurism and Leibniz, sinograms would directly note ideas and not according the way in which phonological informations are embedded in these signs among the components they are combined with and their different functions. We hence intend to seek in which extend an array of functions—discriminant component, key for indexation, phonological or semantic clues and even autonomous sinograms or just graphical sub-component—of the modern canonical 214 components referred as "radicals" in English convey—or not—semantic information. We do so by designing probabilistic models which relate form to meaning, and we rely on such models to define odd ratio of semantic contribution. Finally these odd ratio enable us to provide a graph-based analysis and visualisation of *some* semantic contribution of a few radicals, but this semanticity remains marginal compared to our previous findings on the grapho-phonological relations.

This paper is organized as follows: in Section 2, we firstly describe the context and the practices at the origin of the graphic system originated from what became China, to introduce what is at stake for learning and teaching sinograms and related lexicons of the sinogramic languages. Section 3 presents different strategies to model *radicals* reliability as semantic clues. We compare two different distributional models (Skipgram and Bert) and two different languages (Mandarin and Taiwanese). In Section 4, we apply the model to build a network of radicals able to capture *some* semantic effect among a subset of most relevant *radicals*. Follows a discussion of possible applications and perspectives.

2. Chinese Graphic System: Historical Context, Evolution and Practices

In this section, our point is to shed light on very long term and still running paradox between in one hand, the observable evolution of the sinographic system and its fundamental principles among those phonetism; and on the other hand, the cultural practice of focusing on components initially designed for discrimination, institutionalised as key of indexation and since mainly reinterpreted as reliable at least for semantic purposes if not for ideography. We finally propose a five component functions typology and discuss the relevance to consider a sixth one further in this paper.

2.1. Early Evolutive Facts: Discriminant Components

Our position is strongly grounded on works such as those—already classical—by Karlgren (1923), Boodberg (1939)—*vs* Creel (1938)—and after them among other Boltz (1994) and Baxter and Sagart (2014); such as in the French academic field, Sagart (2006) again—*vs* Vandermeersch (1994)—and in the field of teaching and learning (Yang-Drocourt, 2022); or of course Chinese scholars such as (Qiú, 1988) and (Zhèngzhāng, 2003). For our brief historical sketch here, we mainly rely on Boltz's three steps of development to whose we add a fourth one after the normalization and institutionalization of the Chinese script. After him, we vigorously state in class as in our papers that, firstly the main development principle is phonetism as early as the very first step of oracle-bones available remains, and did not change later on as documented by Li (1986) until the Song period *Guǎngyùn* 《廣韻》 rime dictionary and then until modern and contemporary times (DeFrancis, 1989). Technically, after a first *zodiographic* step during which graphs already were not pictograms and referents could not be recognized. See "elephant" and "mouth" as examples in Figure 1. Quickly, antic practitioners had to face different problems of ambiguity as soon as the sentences went longer and more complex: the *multivalence* step. Indeed some zodiographs were ambiguous because they both noted words such as "elephant" or homophonic "statue" (cf. A in Fig. 1); or were polyphonic for related concept such as the "mouth" and "to call" (cf. B in Fig. 1). In order to disambiguate these situations, practitioners started to add a set of components in order to discriminate the different meanings, the *discriminative* step. As a result, a component nowadays referred to as "man" was added to the homophonic "statue," and the conventionally called "dawn" component was added to our polyphonic representative.

Finally, beyond the scope of Boltz's study, we added a fourth step of development during which neology was designed on the same principle of "neography" all along the exponential development of sinograms through History and the compilation of bigger and bigger dictionaries.

2.2. Key of Indexation and Beyond?

As a result of the increasing of literary corpus over time and the evolution of the script itself for notating and archiving, after thematic ordering works such as *Ěr yā* 《爾雅》 or *Jíjiùpiān* 《急就篇》, a major game changer arose with the earliest lexicographic masterwork by Xǔ Shèn 許慎's *Shuōwén Jiězì*《說文解字》(CE 121). Despite the fact that its 1,900 years of presentation to the Emperor three years ago was notably uncommemorated, its influence is still at the chore of our representation of the Chinese writing system and the invention of the "radical". The interesting

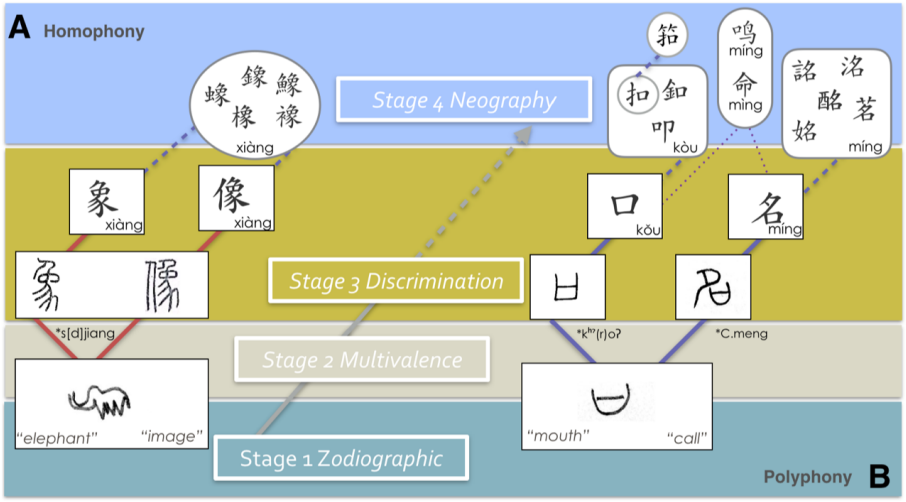

FIGURE 1. The different steps of evolution of the Chinese script after Boltz (1994, p. 64). Archaic Chinese (5th. c. BCE) reconstructions indicated with "*" by Baxter and Sagart (2014). Source: (Magistry, Fabre, and Goudin, 2017), Figure designed by Murielle Fabre.

things are that firstly, the observation between the emic designations of "radical", as it is translated in English. Initially referred as 部 *bù* "part" in the postface of *Shuowen Jiezi*, and later 部首 *bùshǒu* as the combination of 部 "part" determining the second sinogram 首 "head": "head of category". We are quite far from the Latin grammatical reference "radical," and so the changing desinence counterpart[1]. The second observation is that this proposal of indexation took centuries to become the reference system and dominant but not before the late Ming and above all Qing dynasty and the *Kangxi Dictionary* 《康熙字典》*Kāngxī Zìdiǎn* (1716) as presented by Bottéro (1996). The third is that the number for radicals did considerably vary through time, in number—but also in types—from 540 in the *Shuowen Jiezi* to 214 in the *Kangxi*, and since in contemporary times (Wèi, 2015): to 189 in *Xiàndài Hànyǔ cídiǎn* 《現代漢語詞典》first published in 1958 and 王竹溪's *Xīn bùshǒu dà zìdiǎn* 《新部首大字典》in 1988 dealing with 51,100 entries but 56 *bushou* only (Wáng, 1988). As a key of indexation, it is significant when we observe that it was chosen in the Diderot et d'Alembert's *Encyclopedia* for representing the Chinese writing system as "Clefs chinoises" in the dedicated volume of "Alphabets anciens et

1. In French, let's note that the direct reference to indexation through the use of "clé" or "clef" (cf. in Fig. 2) of an index as early as the first attempt of dictionary in the mid-18th century.

modernes" relying on the *Kangxi Dictionary* ordering and its 214 radicals. Except for those possibly used as sinograms as such, in this representation, these unautonomous components were presented as the equivalent of a character used for indexation in dedicated books as it is observable in conventionally alphabetic ordered dictionaries.

The other point is the—now past—extensive practice of training in Chinese philology, and nowadays sinographic languages learning: search in paper dictionaries for reading and translation purposes. Even if the last two decades relegated these artifacts to History with the successive raising of electronic devices and online resources, the fact is that the way of teaching, and consequently learning sinograms have mainly not changed yet, and still rely on the same practice of fundamental "radical" identification of any unknown sinogram, and then strokes counting and ordering, again because it is—was?—part of the main process for searching a sinogram in dictionaries. This practice is so deeply embedded that even computer resources do rely on it. Indeed, even for Unicode, sinograms are ordered by radical and complementary strokes number as we can observe it by activating the function "alphabetic ordering" in any spreadsheet for columns filled with sinograms.

2.3. Six Component Functions?

Relying on the previous two subsections, underlining the contradiction between in one hand, the principle of the development of the script—phonetism—and on the other hand, practices and representations focusing at least on the radical, we distinguish at least five different component functions and do question about the relevance of a sixth one, with examples provided by the sinograms already seen above:

- autonomous sinogram attested in the general lexicon such as 口 *kǒu* in Mandarin meaning "mouth, entrance, gate";
- phonological clue or *phonophore* after Budberg and Boltz such as in previous examples 象 in 像 or 口 in 名, respectively read *xiàng* and *míng* in Mandarin meaning "image" and "name, title, position," when they are in right side position in combination with the discriminant component 亻 "man" or 夕 "dawn" on the left;
- discriminant component such as in the different characters here above in a paleographic perspective with 亻 in 像 or 夕 in 名;
- key of indexation, *radical*, such as 口 in 台 read *tái* in Mandarin, "platform, unite" as early as the *Shuowen Jiezi* classified under the《口部》for lexicographic practices. It overlaps with the previous category as for 亻 in 像 but not systematically, such as in 名 for which the conventionalized radical is 口, and so as early as in the *Shuowen Jiezi*;
- subpart of a component with just graphic relevance as 口 in the phonophore 台 such as in 颱 *tái* "typhoon" or in 始 *shǐ* "start";

FIGURE 2. Chart of "Clefs chinoises" (Chinese radicals) in the second volume "Recueil de planches," "Alphabets anciens et modernes" part of *L'Encylopédie ou dictionnaire raisonné des sciences, des arts et des métiers* (1751-1772) p. 193.

To these five functions, we question the existence and the relevance of a sixth one:

– semantic clue? as we propose to tackle in this paper and the following sections.

In this section, through this historical sketch, we just saw how the evolution for more than two millennia, both embedded and competing different practices designed our representations and speeches about sinograms from classical exegesis to contemporary academic speeches, in classrooms then and nowadays, and even in the way sinograms are ordered in Unicode. If once, we proposed a modelisation of phonetic clues (Magistry, Fabre, and Goudin, 2017) for teaching and learning purposes, it is now time to investigate the consistency of the semantic clues.

3. Modeling Semantic Clues

3.1. Related Works

In this section, we present our attempt to provide computational models able to capture the semantic contribution of so-called radicals with a purely data-driven approach.

Some solutions have already been proposed in the literature. Among the most noticeable we can distinguish two different approaches to the question of the *semanticity* of the radicals. The first one addresses it in a lexicographic fashion, drawing from ontologies or the Generative Lexicon, esp. Hantology and HanziNet (Chou and Huang, 2006; Hsieh and Huang, 2006). The second one adopts a more task-based NLP strategy to show that the inclusion of information on radical can be helpful in semantic-related tasks such as text classification (Haralambous, 2013) or subjectivity classification (Xu and Huang, 2014). All these works make the hypothesis that there is some kind of *semanticity* in the radicals, which can be either described or relied on in practical applications. The main specificity in our present work is that we aim at detecting and assessing the reliability of potential *semantic clues*. Like the task-based approaches, we take into account all the sinograms and radicals found in our corpora, contrasting with publications on ontological approaches, which tend to cherry-pick most reliable examples for a more fine-grained description. On the other hand, we aim at producing a graphematic resource to describe the contribution of the radicals (complementing our previous work on phonetic clues). It will be more coarse-grained than ontologies but it contrasts with task-based approaches which can only show a global effect.

In our own previous work (Magistry, 2015), we proposed a model closer to the present paper. Both are based on a model of (coarse) mean-

ing similarity, which enables us to compute semantic clue reliability indices. Our first attempt was to rely on synonymy networks handcrafted by lexicographers and random walks. They have been shown to be reliable models of meaning similarity for psycholinguistics experiments and various NLP tasks. Unfortunately, such models rely on synonymy resources that are not easily available for most languages. Here we replace synonymy networks with distributional semantic and language models.

3.2. A More *Data-Driven* Approach

In this paper, we turn to distributional models which are trained from raw texts in an unsupervised fashion. We compare *Transformers*-based (Devlin, Chang, Lee, and Toutanova, 2019) models with more simple Skipgram models (Mikolov et al., 2013).

Bert models are considered the state of the art of distributional models, but are expensive and energy-intensive to train for a new language. Skipgram models on the other hand present the benefit of being easily and quickly reproducible. They can thus be tried on different corpora to obtain estimates for different languages or account for different situations of learners exposure (with learner corpora).

In this section, we report on experiments in which we compare Bert and Skipgram models trained on Mandarin corpora and two skipgram models trained on Mandarin and Taiwanese corpora.

3.3. Sinograms Embeddings Approaches

Both Bert and Skipgram are vector-space models and provide a mapping function from sinograms to high dimensional vectors in spaces where semantically similar sinograms are expected to be close to each other (so-called *embeddings* esp. when used in the context of neural neworks). The main difference between the two types of models is that the Skipgram model yield a single vector for each sinogram (type-wise) where Bert compute "contextual embeddings" and yield a different vector for each occurrence of each sinogram (token-wise), depending on the context in which the sinogram occurs.

Putting aside this distinction, and mathematical formulation to go from sinogram vectors to semantic clues estimate are very similar in the two cases. The main intuition is that these models are good at providing good substitution candidates in a semantically consistent way. Our hypothesis is that if a radical acts as a reliable semantic clue, then its substitutes should be more likely to have the same radical. To compute

our reliability score, we first turn the sinogram vector space into probability distributions (of a sinogram being replaced by other sinograms by the model) and then compute the following odd-ratio to measure the effect of the radical:

$$R_r = \frac{P(s = r \mid o = r)}{P(s = r)},$$

where R_r is the ratio computed for a radical r, $s = r$ stands for substituted radical being r, and $o = r$ stands for original radical being r. In other words, we compare the probability of selecting a sinogram with a radical r to replace a sinogram with the same radical with the probability of selecting such a sinogram in the general case. A high R means that the radical tends to be preserved accross the substitutions, which is expected if there is some kind of semantic field associated with the radical. On the other hand, a low value (esp. 1 or less), correspond to cases where the radical has no observable semantic effect.

3.3.1. *Corpora*

To compute the odd-ratio described in the previous section, we rely on two text corpora. For Mandarin, we used the *Academia Sinica Balanced Corpus of Modern Chinese* (ASBC, 中央研究院漢語平衡語料庫)[2]. For Taiwanese, we used the *Digital Archive Database for Written Taiwanese* (DADWT, 台語文數位典藏資料庫)[3] (Iunn, 2007).

3.3.2. *Assessing Semantic Clue Reliability With Bert*

Due to the high cost of training a Bert model, we rely on a pre-trained model readily available on HuggingFace[4]. This model was trained on Wikipedia data in Mandarin as a *masked language model*, which learnt to guess "masked" sinograms in a given context (a sentence). We use the same procedure and apply a *softmax* function on its output to obtain the probabilities of possible substitutions of every sinograms in the ASBC corpus (we exclude the substitutions of a sinogram by itself).

We replace every sinogram with its radical according to the information provided in the Unihan database[5] from Unicode.

We then aggregate the probabilities corresponding to all the occurrences of each radical to obtain a single probability distribution of radical substitutions for each radical. With these values, we can apply the odd-ratio and obtain an R_r value for each radical r.

2. http://asbc.iis.sinica.edu.tw/
3. https://github.com/Taiwanese-Corpus/nmtl_2006_dadwt
4. https://huggingface.co/google-bert/bert-base-chinese
5. https://www.unicode.org/charts/unihan.html

3.3.3. *Assessing Semantic Clue Reliability With Skipgram*

A Skipgram model is trained by trying to predict co-occurrences of sinograms. It produces a single vector per sinogram, based on all the contexts of its occurrences in the corpus. According to the distributional hypothesis, this vector space can be used as a model of semantic similarity.

We use *gensim* (Řehůřek and Sojka, 2010) to train a model on each corpus (ASBC for Mandarin and DADWT for Taiwanese), we obtain a vector-space for each language in which we can compute distances between sinograms.

We then compute substitution probabilities by looking at the closest neighbors of each sinograms, using the "cosmul" similarity from (Levy and Goldberg, 2014) as a the weight to be normalized. The main difference with the Bert approach is that we obtain a probability distribution for each sinogram in the vocabulary rather than one distribution for each occurrence in the corpus. This slightly changes the aggregation but it is also possible to perform the same replacement of each sinogram by its radical to compute a ratio R_r similar to the one described in the previous sections.

To distinguish between the different models to produce various probability ratios, we adopt the following notation:

$$R_{<l,m,r>},$$

where:

- l is the language (*mdn* for Mandarin or *tw* for Taiwanese),
- m is the model (*tf* for Bert Transformer and *sk* for Skipgram), and
- r is one of the radicals.

In the following experiments, we compare $R_{<mdn,tf,r>}$ with $R_{<mdn,sk,r>}$ and $R_{<mdn,sk,r>}$ with $R_{<tw,sk,r>}$.

3.4. Results

The results obtained with the described methods are shared as three CSV files corresponding to the three models $R_{<mdn,tf,r>}$, $R_{<mdn,sk,r>}$ and $R_{<tw,sk,r>}$.

The raw values are available for download on Zenodo[6] along with the code to generate the figures discussed in the present section. Visualisation of the following figures with tunable parameters is also possible online through an R-Shiny interface[7].

6. https://doi.org/10.5281/zenodo.11223724

7. https://analytics.huma-num.fr/Pierre.Magistry/Grafematik2022/

TABLE 1. Spearman rank correlation coefficients between R values from different models. It shows a very strong correlation between the two skipgram models and a strong correlation between skipgrams and transformer models.

	$R_{<mdn,tf,r>}$	$R_{<mdn,sk,r>}$	$R_{<tw,sk,r>}$
$R_{<mdn,tf,mdn>}$	1	0.609	0.606
$R_{<mdn,sk,r>}$	0.609	1	0.754
$R_{<tw,sk,r>}$	0.606	0.754	1

3.4.1. *Spearman Correlations*

A first question is whether the ranking we obtain in the three cases are consistent. To address it, we compute Spearman rank coefficients and present the results on Table 1. As we can see, there are *strong* (around 0.61) or *very strong* (0.75) correlation between the rankings obtains by the different models. Interestingly, the difference between Skipgrams and Transformer models seems larger than the difference between the two languages. This allows us to argue that the Chinese graphic system shows properties that are language independent (Mandarin and Taiwanese are far from being mutually intelligible as spoken languages).

3.4.2. *Choosing Skipgrams Models Over the Transformers*

Considering the correlations from Table 1, and more importantly, the computational costs in time and energy, we decide to focus our analysis on the Skipgram models. The cost of transformer based models makes it difficult or impossible to address new languages beside Mandarin or to conduct future experiments based on training corpus selection (especially to reproduce our results on learner corpora). Even the decoding step to compute the substitution probabilities from the pre-trained model was significantly slower compared to the training of a Skipgram model from scratch[8].

3.4.3. *Semantic Clues Ranking*

We can now confidently study the rankings obtained by our method.

The decomposition from the Unihan database follows the convention of the *Kangxi Dictionary* to divide the sinograms between 214 radicals.

In our corpora, only 205 radical are attested. In Mandarin, 74 are never realised in substitution ($R_r = P(s = r) = 0$) and 128 have $R_r > 1$. In Taiwanese the respective figures are 84 for $R_r = 0$ and 108 for $R_r > 1$.

8. Our code could easily be optimized, so exact figures are not very relevant but it took 2 days with a RTX 3090 GPU, when a skipgram model takes less than an hour to train on high end CPU.

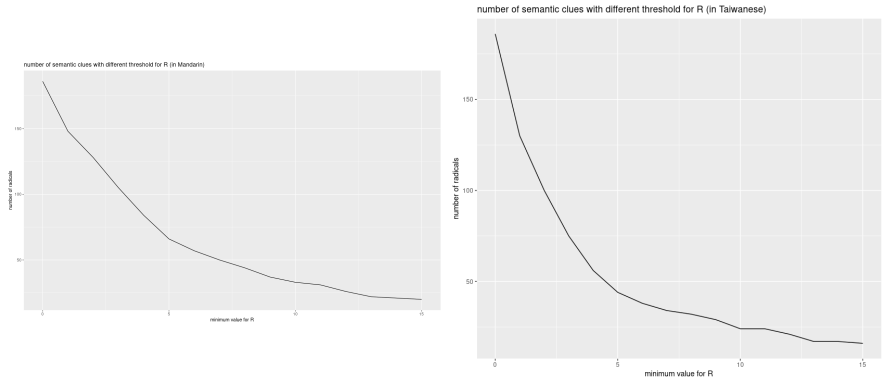

FIGURE 3. Number of r with $R_r > x$ in Mandarin on the left and Taiwanese on the right

It is not easy to define a clear cut on the R_r value to distinguish between reliable and unreliable clues. In Figure 3 we plot the number of radicals r with $R_{<l,s,r>} > x$. The graphic on the left is based on Mandarin (ASBC) data and the one on the left is based on the Taiwanese (DADWT) data. As we restrict the list to more and more reliable R, we can see that the number of radicals more likely to be kept quickly drops under 50.

3.4.4. *Comparing Taiwanese and Mandarin*

We can visualize the correlation between $R_{<mdn,s,r>}$ and $R_{<tw,s,r>}$ on Figure 4. The radicals that appears on the top right corner are the most reliable as a semantic clue.

4. Graphs of Semantic Clues

One limitation of our R metric when used as described in the previous section and especially when compared to the ontology-related approaches, is that we take the supposed meaning of each radical in isolation. We only consider substitution of a radical by itself (vs. by any radical) and our computation ignores the possibility for two different radicals to have related meanings. This is not ideal as related radicals are easily noticeable and discussed in the literature such has body parts or animals. For example, (Huang, Yang, and Chen, 2008) chose to discuss four different radicals related to "Four Hoofed-Mammals" (羊 *bovid*, 鹿 *deer*, 牛 *cattle* and 馬 *horse*).

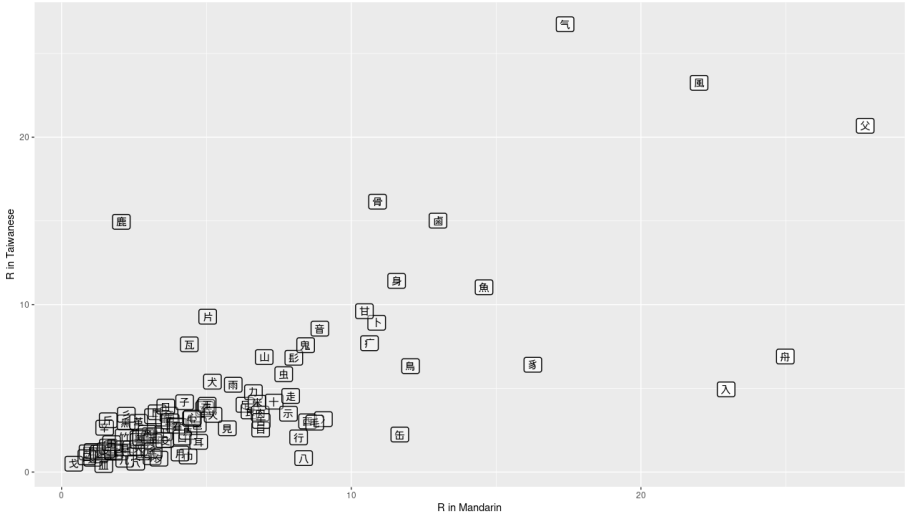

FIGURE 4. Scatter plot comparing R values in Mandarin and Taiwanese for each radical.

To explore this possibility, we propose another experiment. Firstly we compute odd ratio similar to those in the previous section, but to answer the question "what are the odds to go from radical a to radical b with a substitution following our skipgram model?" We do so for all possible pairs of radical (a, b). Secondly we build a graph in which nodes are radicals and edges are the ratio values (using the ratio as weights and cutting the edges out under a threshold). Then we apply Fruchterman-Reingold layout algorithm for spacialization (Fruchterman and Reingold, 1991) and Infomap clustering (Rosvall and Bergstrom, 2008) to color the nodes and produce Figures 6 and 7. In the printed figure, we set the R threshold to 4. It corresponds as a good balance to keep a large number of radicals while obtaining a graph with a good level of clustering modularity (hopefully creating semantic clusters). We show the modularity as a function of the R threshold on Figure 5.

The reader can experiment with different values of the threshold on our R-Shiny interface[9].

Clusters appear clearly from both Mandarin and Taiwanese datasets. These clusters indeed correspond to animals, bodypart, meteorology or food. We can thus argue that we were able to see some semanticity of the radical emerging from the data. Our measures seems consistent to the semantic classes typically discussed in the literature, with some lit-

9. https://analytics.huma-num.fr/Pierre.Magistry/Grafematik2022/

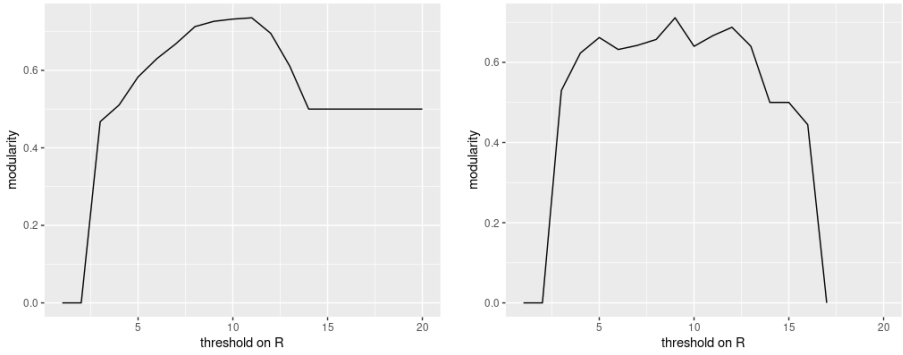

FIGURE 5. Clustering modularities as a function of the threshold on the value of R for the Mandarin (left) and Taiwanese (right) graphs

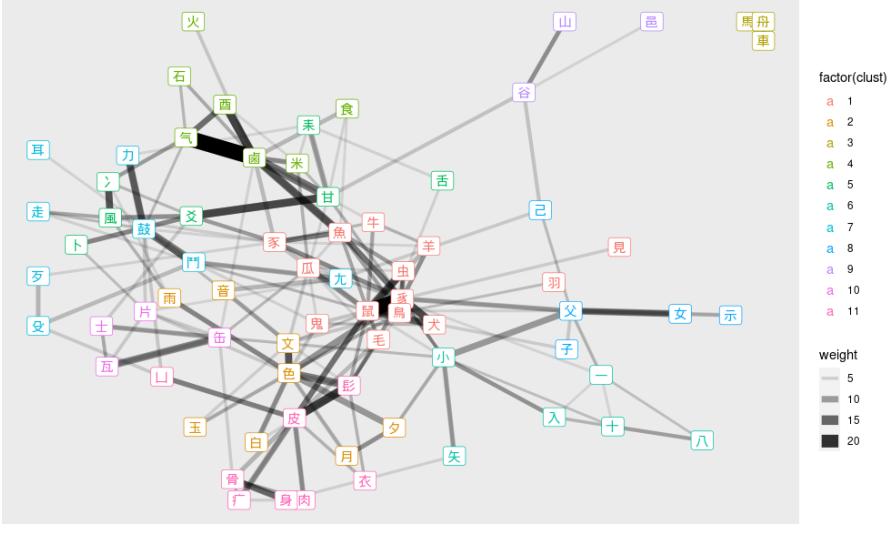

FIGURE 6. Graph of semantic clues in Mandarin with a cut at $R > 4$ with Infomap clustering and Fruchterman-Reingold layout

tle discrepancy that seems understandable as well (for example, we can observe that the radical 馬 *horses* is more related to carriages than it is to other mammals).

However, this semanticity is only observed for a few dozen of radicals. As far as our model can tell, we provide no evidence that the ontological description can be extended beyond those few components.

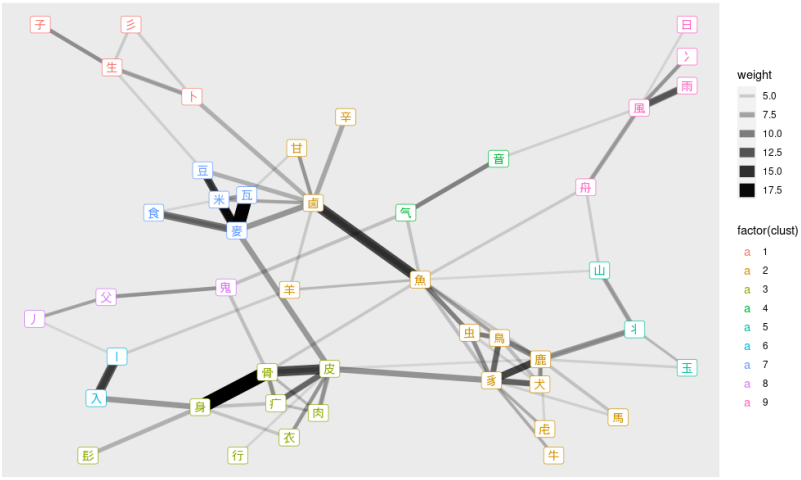

FIGURE 7. Graph of semantic clues in Taiwanese with a cut at $R > 4$ with Infomap clustering and Fruchterman-Reingold layout

5. Discussion and Future Work

Despite our initial position which is to advocate for the importance to study and teach the grapho-phonological correspondences in the Chinese Script, we wanted to provide a fair treatment of a possible semanticity effect that could emerge from the data. We were able to propose a computational model which capture some semanticity in some of the radicals. However, only a very small minority of the 214 traditional radicals can be considered as reliable *semantic clues*, much less than the *phonetic clues* we were able to detect in our previous work (Magistry, Fabre, and Goudin, 2017).

The method proposed in this work was experimented on both Mandarin and Taiwanese data. We observed a strong correlation in our rankings between the two languages. As the Chinese script is or has been used to write a variety of languages, it is interesting to observe that some properties seems to hold cross-linguistically, but it would also be relevant to describe the discrepancies. This goes beyond the scope of this paper and would require a more diverse dataset. Currently we can only compare ASBC and DADWT, but drawing robust conclusions to compare the two languages would also require some comparison between different corpora for the same language. In other words, if the correlations observed on Table 1 advocate for the robustness of our models, we can not say if the differences we can spot come from differences between the languages or simply between the two corpora. More investigations using the proposed measures are required.

For the same reason, we hope that the code and data which come with this paper[10] can foster further discussions on the Chinese script and help in teaching languages written with sinograms, but the precise rankings are to be considered as estimates and work in progress. We invite the readers to experiment on a variety of corpora of their own.

Another expected extension of this work is to run the same computation for all possible graphical component without restricting ourselves to the canonical 214 radicals from the *Kangxi Dictionary*.

6. Conclusion

In this paper, through the probabilistic models we designed, we are able to propose a graph-based analysis and visualisation of *some* semantic contribution of a few radicals, but this semanticity remains marginal compared to our previous findings on the grapho-phonological relations. It is the same for a semantic function of these graphic components. If we do not reject it, we lower it as a side function of the discriminant component. This conclusion inspires us a last question and a final statement: beyond our community of graphematicians, may this paper be relevant and used for pedagogical purposes? Indeed, our readers do not stand for themselves and do stand in order also to be read by teachers and ideally by—with—learners themselves, aiming so at to make these communities to come aware of what is at stake with general representations and dominant speech about Chinese graphic system. Thus, we firmly call for a definite shift of speeches and practices in the field of Chinese learning and teaching, from paper and radical (or alphabetical) ordered dictionaries to dynamic online open resources: a paradigmatic and radical shift.

References

Baxter, William Hubbard and Laurent Sagart (2014). *Old Chinese: a new reconstruction*. Oxford New York: Oxford University Press.

Boltz, William G. (1994). *The Origin and Early Development of the Chinese Writing System*. American oriental series 78. New Haven, Conn: American Oriental Soc.

Boodberg, Peter (1939). "'Ideography' or 'Iconolatry'?" In: *T'oung Pao* 35.1, pp. 266–288.

Bottéro, Françoise (1996). *Sémantisme et classification dans l'écriture chinoise: les systèmes de classement des caractères par clés du Shuowen Jiezi au Kangxi Zidian*. Mémoires de l'Institut des hautes études chinoises v. 37. Paris: Collège de France, Institut des hautes études chinoises.

10. https://doi.org/10.5281/zenodo.11223724.

Chou, Ya-Min and Chu-Ren Huang (2006). "Hantology-A Linguistic Resource for Chinese Language Processing and Studying." In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Ed. by Nicoletta Calzolari et al. Genoa, Italy: European Language Resources Association (ELRA).

Creel, Herrlee Glessner (1938). "On the Ideographic Element in Ancient Chinese." In: *T'oung Pao* 34.1, pp. 265–294.

DeFrancis, John (1989). *Visible speech: the diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Devlin, Jacob et al. (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Fruchterman, Thomas M. J. and Edward M. Reingold (1991). "Graph drawing by force-directed placement." In: *Software: Practice and Experience* 21.11, pp. 1129–1164.

Haralambous, Yannis (2013). "New Perspectives in Sinographic Language Processing through the Use of Character Structure." In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 201–217.

Hsieh, Shu-Kai and Chu-Ren Huang (2006). "When Conset Meets Synset: A Preliminary Survey of an Ontological Lexical Resource Based on Chinese Characters." In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia: Association for Computational Linguistics, pp. 385–390.

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

Iunn, Un-Gian (2007). "New Manifestation of the Taiwanese vernacular literature–Introduction to Digital Archive for Written Taiwanese." In: *National Museum of Taiwanese Literature Communication* 15, pp. 42–44.

Karlgren, Bernhard (1923). *Analytic dictionary of Chinese and Sino-Japanese*. Paris: P. Geuthner.

Levy, Omer and Yoav Goldberg (2014). "Linguistic Regularities in Sparse and Explicit Word Representations." In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ed. by Roser

Morante and Scott Wen-tau Yih. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 171–180.

Li, Hsiao-ting 李孝定 (1986). 漢字的起源與演變論叢 *[The Origin and Evolution of Han Characters]*. 台北 [Taipei]: 聯經出版事業公司 [Linking].

Magistry, Pierre (2015). "Modèles computationnels des indices sémantiques et phonétiques dans l'écriture chinoise." Linglunch du Laboratoire de Linguistique Formelle, Paris Diderot.

Magistry, Pierre, Murielle Fabre, and Yoann Goudin (2017). "Indices phonologiques des sinogrammes: de l'étude de l'acquisition à la modélisation pour l'apprentissage." In: *Revue TAL* 57.3, pp. 41–65.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Qiú, Xīguī 裘錫圭 (1988). 文字学概要 *[Chinese Writing]*. 北京 [Beijing]: 商務印書館 [The Commercial Press].

Řehůřek, Radim and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora." In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pp. 45–50.

Rosvall, Martin and Carl T. Bergstrom (2008). "Maps of random walks on complex networks reveal community structure." In: *Proceedings of the National Academy of Sciences* 105.4, pp. 1118–1123.

Sagart, Laurent (2006). "L'emploi des phonétiques dans l'écriture chinoise." In: *Écriture chinoise: données, usages et représentations*. Ed. by Djamouri Redouane and Françoise Bottéro. Paris: EHESS.

Vandermeersch, Léon (1994). *Études sinologiques*. 1re éd. Orientales. Paris: Presses universitaires de France.

Wáng, Zhúxī 王竹溪 (1988). 新部首大字典 *[Great Dictionary ordered by New Radicals]*. 上海 [Shanghai]: 上海翻译出版公司 [Shanghai Translation Publishing House].

Wèi, Lì 魏励 (2015). *Understanding the Radicals of Han Characters [*汉字部首说解*]*. 北京 [Beijing]: 商務印書館 [Commercial Press].

Xu, Ge and Churen Huang (2014). "An Analysis of Radicals-based Features in Subjectivity Classification on Simplified Chinese Sentences." In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*. Phuket,Thailand: Department of Linguistics, Chulalongkorn University, pp. 495–502.

Yang-Drocourt, Zhitang (2022). *L'Ecriture chinoise: Au-delà du mythe idéographique*. Paris: Armand Colin.

Zhèngzhāng, Shàngfāng 郑张尚芳 (2003). 上古音系 *[Old Chinese Phonology]*. 上海 [Shanghai]: 上海教育出版社 [Shanghai Education Press].