Application of Data-Driven Semantic Map Modeling in International Education of Chinese

Ying Zhang(✉)

Harbin Institute of Technology, Shenzhen, Shenzhen, China
yzhangbc@163.com

Abstract. To better reveal the patterns of negative interlingual transfers made by international learners of Chinese, large-scale cross-linguistic data are required to be compared. In presenting and analyzing big set of data from 71 languages of the world, the study employs digital technologies of data analysis and map modeling to present a Semantic Map Model (SMM) related to the notion of "universal quantification". Besides, a Communicative and Controlled Graph (CCG) based on parameters of Probability Entailment, which is also known as Weighted Map with Least Edges by frequencies of occurrences, has also verified the validity of the connective pattern of the original SMM. By comparing the detailed SMMs of Chinese and Hindi, the study instantiates how data-driven map modeling techniques shed light on predicting and correcting the possible semantic errors made by international Chinese learners.

Keywords: data analysis · SMM · CCG · Chinese education · universal quantification

1 Introduction

With the advent of web 2.0 era, digital technologies concerning data analysis and modeling have been widely used in educational practices, such as identifying [1] and evaluating [2] students' learning activities. These practices reveal correlations between different variables and shed light on course development in the field of engineering and computer science. Besides, it is demonstrated from the perspective of methodology that data analysis and map modeling enable efficient analysis of large data set, and provide helpful insight in explaining the teaching and learning issues.

However, in the discipline of international Chinese education, more challenges occur due to the great diversified language backgrounds of the international students. For instance, students from different countries make different semantic errors because of the inevitable negative interlingual transfer. To better solve the semantic errors in second language learning, deep understanding in terms of the language universals and language particularities is required. To cater for this aim, large-scale crosslinguistic data in typological research can well be used to reveal the conceptual correlations and bifurcations.

X. Yuan et al. (Eds.): ICEKIM 2023, AHCS 13, pp. 1942–1949, 2023.
https://doi.org/10.2991/978-94-6463-172-2_215

Among all the analytical tools to represent the outcome of cross-linguistic comparison, the Semantic Map Model (SMM) is one of the most powerful [3]. By employing SMM, not only can we uncover a unique semantic structure for a particular grammatical item, but also, we can have a clear picture about the universal connective pattern of these meanings that could apply to arguably all languages. Therefore, SMM has been adopted in this research.

In this study, the theoretical basis in the areas of typology and education is first discussed to explain why SMM can benefit the international education of Chinese. Then, by employing large-scale data from 71 languages in the world, we exemplify how a semantic map centered on the notion of "universal quantification" is established. In addition, how this semantic map be used to correct and even predict the semantic errors made by the international Chinese learners are demonstrated as well. Through data analysis and modeling procedures, the study emphasizes the significance of applying analytical technology to education practices of global Chinese teaching.

2 Theoretical Basis

To lay some groundwork for following instantiations, the present section sets out to discuss the theoretical issues concerning analytical tools in the field of typology and learning theories in the field of education respectively.

2.1 Typological Basis

When conducting cross-linguistic comparisons, typologists have to decide which observations are language-particular and which might lead to language universals. A tool to represent both language universals and language-specific grammatical knowledge is thus produced [3], and this approach is called the Semantic Map Model (SMM) [3].

As described by Haspelmath [3], "A semantic map is a geometrical representation of functions in 'conceptual/semantic space' which are linked by connecting lines and thus constitute a network". In this network, the more similar the functions are, the closer they are located on the map. As the semantic map is derived from cross-linguistic comparison, it is believed to represent "a universal structure of conceptual knowledge [4]"; therefore, the configuration shown by the semantic map "is claimed to be universal [3]". Apart from the universals, the semantic map could also represent language-particular facts. There is a basic working principle to set up a semantic map – the "Semantic Connectivity Hypothesis". It requires that the functions expressed by a language-particular category should occupy contiguous areas on the semantic map. This is equivalent to saying that each language-specific category could be "map[ped] onto connected regions [4]". Therefore, the language universals concerning how concepts are connected are reflected by the overall configuration of the semantic map (which is also known as "conceptual space"), whereas the particularity of forms in individual languages could also be demonstrated by the map as the functions of a form would be represented by contiguous smaller areas of the original semantic map.

In practice, the data to be processed and the functional nodes to be presented might be quite large. Therefore, to make the data-processing computationally tractable, a second-generation of SMM has been proposed. This method is called Multi-Dimensional Scaling

(MDS). The MDS differ from SMM in the way they deal with cross-linguistic data. The first difference is that the "MDS analysis can be performed with different numbers of dimensions [5]". However, the classical SMM presentation normally has two dimensions. The second prominent difference is that there are no links in the representation of the MDS, and the similarities between functions are reflected by their closeness in the space. That is to say, the representation of the MDS is only clusters of functions. However, in the classical SMM approach, similarities are reflected by links between different functions whereas the length of the link and the spatial orientation play no role at all.

To bridge the advantages of SMM and MDS, the third approach of data analysis and algorithm has been proposed [6, 7] as CCG (Communicative and Controlled Graph). In this approach, it retains the advantage of the classical SMM which provide links to show the implicational universals. Besides, it uses the dynamic generative algorithm to capture the frequency of occurrence. Thus, in the third type of map, it provides links with numbers to show the closeness of two conceptual connections. As the data can be processed computationally in CCG, it retains the advantage of MDS in dealing with large-scale cross-linguistic data.

As SMM and CCG can provide more information about implicational universals, these two approaches will be instantiated later in Sect. 3.

2.2 Educational Basis

The emergence of digital age of web 2.0 has witnessed the advent of a new social structure, in which knowledge and the learning process are viewed significantly different from previous time. The innovation prompts the emergence of a new digital learning theory-Connectivism. Presented by Siemen and elaborated by Downes [8, 9], connectivism has become one of the most influential learning theories in the digital era. It differs from previous learning theories in the way knowledge is viewed. Connectivism regard knowledge as a dynamic emerging network rather than a static closed loop. The learning process, therefore, is dynamic as well. Personal knowledge comes from a network, and it turns out to nurture the dynamic network by feeding back into the network [10].

The learning theory-Connectivism is well in line with language education as it accords with the feature of language such as "emergent" and "interactive". "Grammar is not a static, built-in system, but a complex dynamic system with constant progress of development and evolution. The grammatical rules emerge from the actual usage among language users. [10]" Therefore, we have the grounds to apply connectivism to language learning process as it will enlighten the learners in a more natural way.

The way we capture the correlations among different languages and apply the correlations to international education of Chinese is just one of the practices of connectivism. The conceptual space based on cross-linguistic data is highly networked, and it explains how errors occur when the interlanguage system of the second language learners evolves. Besides, the digital technology involved in the data processing and map modelling procedures also align with the learning theory-connectivism.

3 SMM Related to "Universal Quantification"

This section instantiates how a semantic map model related to the notion of "universal quantification" is established, and how the same set of data is presented via the CCG approach computationally.

3.1 Cross-Linguistic Data

To obtain first-hand data, more than 40 interviewees about their mother tongues are interviewed by the author at UCLA and BLCU; moreover, second-hand data from research on languages of the world are also included to serve as supplement. As there are overlapping languages between the first-hand data and the second-hand data, the total data sources add up to 71 languages. Table 1 offers a breakdown of all the language data with the first-hand sources in bold and second-hand sources in italics. The overlapping ones are in bold and italics.

For the details of the investigation procedures and example sentences that have been used, one can refer to Zhang [11].

Table 1. Data sources by language families and branches

Indo-European	Baltic	**Lettish**
	Balto-Slavic	***Russian***, *Czech*
	Creole	**Antiguan Creole**
	Germanic	**German**, *English*
	Hellenic	**Greek**
	Indo-Iranian	**Bengali**, **Hindi**, **Persian**, **Tajik**, **Urdu**
	Italic	**French**, **Italian**, **Portuguese**, ***Spanish***
Sino-Tibetan	Chinese	**Jin (Fangshan)**, **Min (Taipei)**, **Min (Yun'ao)**, **Southwestern Mandarin** (**Chongqing**, ***Wuhan***, ***Changde***, ***Liuzhou***), ***Wu (Shanghai)***, ***Cantonese (Hongkong)***, *Mandarin*
	Tibeto-Burman	***Tibetan***, ***Nuosu*** (***Yi language***), *Lahu*, *Zauzou*, *Tujia*, *Phola*, *Jingpho*, *Qiang*, *Naxi*, *Jino*, *Derung*, *Rgyalrong*, *Hani*
Niger-Congo	Atlantic-Congo	**Ghomala**, **Kinyarwanda**, **Pular**, **Setswana**, **Shona**, **Swahili**
Austronesian	Malayo-Polynesian	**Indonesian**, **Malay**, **Tagalog**
Turkic	Karluk	**Uzbek**
	Kipchak	**Kazakh**
	Oghuz	**Turkish**
Tai-Kadai	Tai	**Thai**, ***Zhuang***, *Tai Lü*
	Kam-Sui	*Dong*, *Mulam*, *Maonan*, *Mak*, *Sui*
	Hlai	*Hlai*
	Kra	*Lachi*, *Gelao*
Hmong-Mien	Hmongic	*Hmong*, *Kiong Nai*, *Yuno*
Austroasiatic	Mon-Khmer	***Vietnamese***, *Wa*, *Jing*
Afro-Asiatic	Semitic	**Arabic**
Koreanic		***Korean***
Japonic		**Japanese**

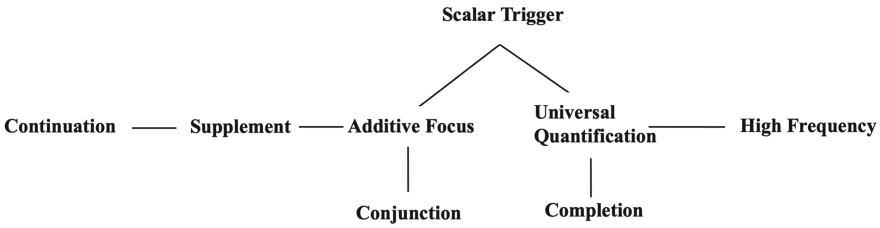

Fig. 1. SMM related to "universal quantification"

3.2 Semantic Map Model Related to "Universal Quantification"

When examining the language form in Chinese that can express the notion of "universal quantification", one may find that it may express other notions as well, like "additive focus" or "scalar trigger". However, for language forms in German or Bengali, they can also use the additive focus particle to express the notion of "scalar trigger". Different languages use one form to express similar group of notions is by no means an accident; instead, it demonstrates universal connections among these notions. Semantic Map model is just a useful tool to reflect these conceptual correlations.

Based on the data mentioned in 3.1, it is found that 18 languages use the same form to represent both "additive focus" and "scalar trigger" which demonstrates the conceptual correlations between these two notions. 10 languages employ the same form to represent both "additive focus" and "supplement" which suggests these two notions should be adjacent as well. Following this way, a semantic map model based on cross-linguistic data is established as shown in Fig. 1.

In the above SMM approach, the comparison is completely by hand, and the semantic map is built gradually by putting more nodes and links into the existing pattern under the requirement of the Semantic Connectivity Hypothesis. To better verify the validity of the map, all the above cross-linguistic data are prepared into the CCG algorithm, in which the data are automatically compared and weighted[1]. Besides, the similarities between the pairs of notions are processed by a dissimilarity algorithm according to "the number of forms that share the notions compared. Then we get a similar but more detailed graph as follows.

Comparing Fig. 1 with Fig. 2, it can be noticed that the connective pattern remains the same whereas the latter one provides more information about the occurrence of frequency. The consistency between SMM and CCG proves the validity of the connective pattern.

4 Application to International Education of Chinese

In the previous section, we have discussed the universal pattern of connections related to the notion "universal quantification". In this part of the discussion, we turn from the universal perspective to a language-specific perspective to see how the graph is divided into different contiguous sub-maps by different clusters of grams in a particular language.

[1] CCG can be generated via www.newlinguistics.org, designed by Chen, Z.Y., Chen, Z. N. [6]

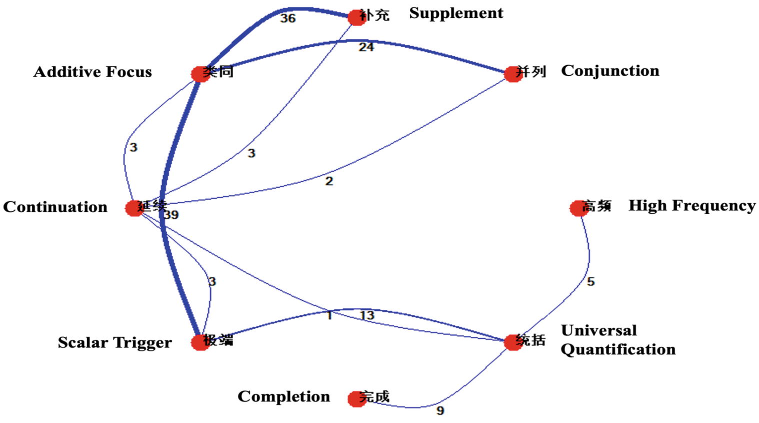

Fig. 2. CCG related to "universal quantification"

By comparing the sub-maps of different languages, we will identify and predict in what way international students will make interlanguage negative transfer, and will adjust teaching activities beforehand to avoid the occurrence of such semantic errors.

For instance, we may compare the Chinese sub-map with that of Hindi as follows:

Based on the contrast of the green frame lines in Fig. 3 and Fig. 4, we may deduce that when international students whose mother tongue is Hindi learn to use Chinese particle

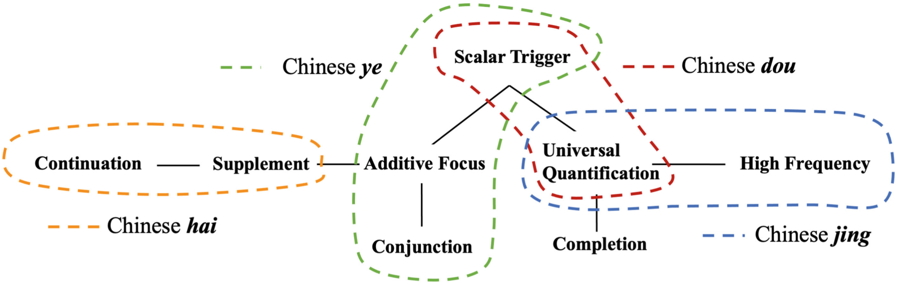

Fig. 3. Chinese SMM related to "universal quantification"

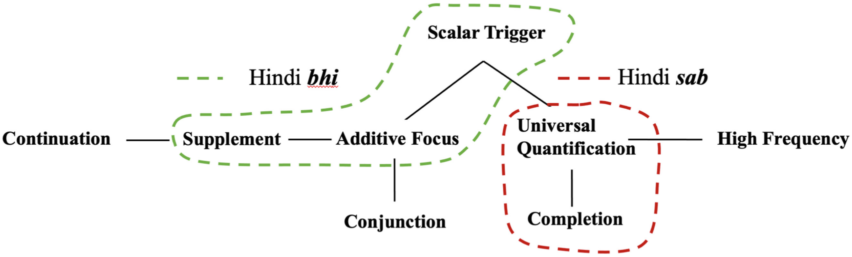

Fig. 4. Hindi SMM related to "universal quantification"

ye, (s)he will possibly extend the functions of *ye* to “supplement” due to the negative transfer of Hindi *bhi*. Similarly, when learning Chinese *dou*, (s)he might extend the use of *dou* to “completion” for the same reason of negative transfer. For instance, we search *ye* sentence written by Hindi students in HSK Dynamic Composition Corpus (http://hsk.blcu.edu.cn), and find the error did happen in between the notion of “supplement” and “additive focus”, for instance:

(1) Yǒu rén shuō fùmǔ shì háizi de dì yī rèn lǎoshī,yě shuō lǎoshī shì háizi de dì èr rèn fùmǔ. (Composition No. 200505109539102030).

(2) Wǒ de fùqin shì éyǔ zhuānjiā,tóngshí yě xué le bù shǎo wàiyǔ hé wǒ guó de qí tā yǔyán,qízhōng yě xué le zhōngwén. (Composition No. 200304131539200040).

Ye in the above sentences have been mistakenly used in the function of “supplement”, which should be replaced by Chinese *hai*. The occurrence of this error is caused by semantic negative transfer, since the additive focus particle *bhi* in Hindi can simply spill over to capture the function of “supplement” whereas Chinese *ye* cannot.

5 Conclusions

The international education of Chinese entails a typological understanding about the variations across languages of the world. Only by delving into the details about the language universals and language particularities, can we find out the deep explanations of why certain semantic error occurs in students speaking certain languages.

In processing large-scale cross-linguistic data, this study demonstrates that data-driven SMM can be a useful tool in representing how the notions are connected in a universal way and how an individual language uses different language forms to split up the map into contiguous smaller subparts. The bifurcations in the way different languages split the original semantic map offer considerable insight in explaining the negative semantic transfer occurred in global Chinese education. The research instantiates how we use a semantic map to predict and to correct the semantic error made by students speaking Hindi.

The instantiation presented in this study is obviously not the only way that SMM benefits Chinese teaching. We have every reason to believe that this data-driven semantic map modeling methodology will continue to reward us with more insight to promote the Chinese education in the world.

Acknowledgement. This research is sponsored by 2022 International Chinese Education Research Project of the Ministry of Education (No. 22YH35C), Cooperative Education Project of the Ministry of Education (No. 220701094255850), Project of Educational Sciences, Guangdong (No. 2022GXJK429), 2022 Co-construction Project of Philosophy and Social Sciences, Guangdong (No. GD22XZY07) and Teaching Reform Project of HITSZ (No. 20200171).

References

1. Ww, A. , Nl, B. , Dw, B. , & Dg, C.. (2020). Exploring student information problem solving behaviour using fine-grained concept map and search tool data. *Computers & Education*, 145(5): 103731.
2. Svanstrom, M. , Sjoblom, J. , Segalas, J. , & Froling, M.. (2018). Improving engineering education for sustainable development using concept maps and multivariate data analysis. *Journal of Cleaner Production*, 198: 530-540.
3. Haspelmath, M.. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison, in M. Tomasello (eds.) *The New Psychology of Language* 2, New York: Lawrence Erlbaum Associates Publishers, 211-243.
4. Croft, W.. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
5. Croft, W., Poole, Keith T.. (2008). Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics* 34 (1): 1-37.
6. Chen, Z.Y., Chen, Z. N.. (2015). Revealing covert laws in grammar with semantic map analysis: weighted maps with least edges. *Studies of the Chinese Language* (5): 428-438.
7. Chen, Z.N., Chen, Z. Y.. (2017). Cluster and association analysis of natural languages based on inclined similarity measures. *Journal of Chinese Information Processing* (1): 205-211.
8. Siemens, G.. (2005). Connectivism: a learning theory for the digital age. *International Journal of Instructional Technology and Distance Learning* 2 (1): 3-10.
9. Downes, S.. (2012). *Connectivism and connective knowledge: essays on meaning and learning networks*. Ottawa, Ontario: National Research Council Canada.
10. Zhang, Y.. (2021). Applying digital technology to linguistic education: a connectivism-based intelligent learning system, 2021 3rd International Conference on Internet Technology and Educational Informization (ITEI), 111–115.
11. Zhang, Y.. (2019). The categorization of *Dou* in Chinese: a study from a cross-linguistic perspective. *Bulletin of Chinese Linguistics* 10 (2): 214-234.