CxG in the Light of Word Vector Representations: *jia X zhen Y* for Example

Lian-Hui Tan

Department of Foreign Languages and Literatures, National Taiwan University

Computational Semantics

Prof. Shu-Kai Hsieh

January 14, 2022

CxG in the Light of Word Vector Representations: *jia X zhen Y* for Example

The construction *jia X zhen Y* has been popular on Internet forums in the recent decade, even invading daily speech as many people naturally use this construction without realizing it. This paper aims to discuss the evolution of the construction *jia X zhen Y*, from *假戲真做* to a diversity of usage, which reflects the productivity of this construction. Moreover, by taking a Construction Grammar approach (Goldberg, 1995), this paper intends to discover the possible relations between *X* and *Y* by using Facebook's fastText package (Bojanowski et al., 2017; Joulin et al., 2016a, 2016b) for word vector representation.

Literature Review

According to Liu and Luo (2020), "*zhen* and *jia* are mixed items with bi-dimensional meanings, i.e. the judge of truth-value as the descriptive meaning, and the degree of similarity/deviation between the facts and the subjective expectations as the expressive meaning." This suggests that the two meanings of *zhen* and *jia* are either logical or gradable. This paper takes both aspects of *zhen* and *jia*, since the arguments of the *jia X zhen Y* construction can be rejecting one another other (i.e. *X* and *Y* are antonymic) or they can be graded in appropriateness in its particular contexts, i.e. *X* and *Y* are unrelated.

Previous studies of word vector representations applying on Construction Grammar are often diachronic studies of change of word senses that appear in the slots (Budts & Petré, 2020; Hamilton et al., 2016), but there have not been many discussions about contemporary word vector representations in specific constructions. Desagulier (2021) compares English IL constructions (such as "in the middle/heart/center/midst of") on distributional semantic models, finding that the same methods used to interpret diachronic changes are also suitable for synchronic comparisons between different constructions.

Rambelli and colleagues (2019) proposed a new framework to combine constructions and word vector representations, along with other features including frames and events. However, this framework is too grand to be adopted in this research. Instead,

we take reference from procedures of Desagulier (2015), which collects data and uses *l'analyse collexémique covariante*, a quantitative approach of word distributions, to discuss the English construction <ADJ *as* GN>.

Method

Data Collection

We collect data sentences from Chinese Word Sketch Engine (Huang et al., 2005). We search for sentences containing *jia*, with the KWIC window sizes set at 5 tokens on both sides, and then filter those that also include *zhen*. The data sentences are collected into a table for further assessments. Some criteria of discarding data sentences are:

1. if *X* is *xi* and *Y* is *zuo*, namely a common four-word expression in sentence (1)

   (1) 假 戲 真 做
   jia xi zhen zuo
   fake drama real do
   'actually realizing an illusionary play'

2. if *jia* is used in a different sense, e.g. *jia* 'use' as in sentence (2) below,

   (2) 假 盜領 之名 真 詐財
   jia daoling zhi ming zhen zha cai
   use heist in-name-of real scam
   'using the name of a heist but actually performing a scam'

   or the noun sense *jia* 'a fake thing' as in sentence (3)

   (3) 以 假 換 真
   yi jia huan zhen
   use fake-thing swap real-thing
   'swapping the real thing with a fake thing'

3. if X and Y are the same, as in sentence (4) below

   (4) 投入 假 選票 帶出 真 選票
   touru jia zuanpiao daichu zhen xuanpiao
   throw-in fake ballot take-away real ballot
   'throw fake ballot in (the ballot box) and take away the real ballot'

4. if *jia X* and *zhen Y* are two arguments of an event, as in sentence (5) below,

(5) 把 假 酒 裝入 真 酒瓶
ba jia jiu zhuangru zhen jiuping
BA fake wine fill-into real wine-bottle
'fill fake wine into real wine bottle'

or are referring to two entities, as in sentence (6) below

(6) 假 乘務 持 刀 搶 旅客 真 勇士 施武鬥
jia chengwu chi dao qiang lvke zhen wushi shiwudou
fake crew-member take knife rob passenger real warrior fight-back
'(a) fake crew member with knife robbed the passengers, while a true warrior fought back'

Moreover, data sentences that do not follow the *jia X zhen Y* construction but contain similar structure (e.g. *zhen Y jia X* or *X shi jia Y shi zhen*) are also included. For example, in sentence (7),

(7) 講 緩和 是 假 搞 台獨 是 真
jiang huanhe shi jia gao taidu shi zhen
talk-about peace be fake support Taiwan-independence be real
'not talking about peace but supporting Taiwanese independence'

This *X shi jia Y shi zhen* construction occurred more when *X* and *Y* are longer phrases that cannot fit into *jia X zhen Y* due to its length. Still, this kind of sentences are still recruited because their senses of *jia* and *zhen* are consistent with those of the *jia X zhen Y* construction this research focuses on.

Furthermore, the *X* and *Y* slots are not necessary preceded by *jia* and *zhen* due to the above mentioned exception structures *zhen Y jia X* and *X shi jia Y shi zhen*. Thus, manual judgements are required for each data sentence to make sure that *X* and *Y* are what the sentence is to imply.

After sentence selection and slot judgement, 961 data sentences are collected into a spreadsheet for Python to process.

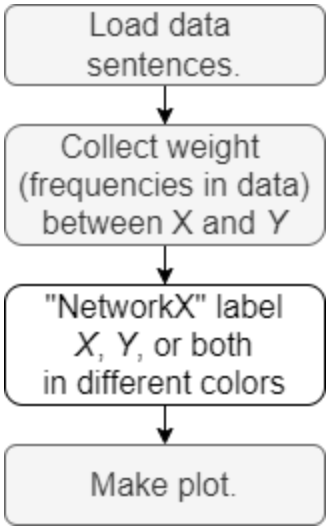
NetworkX Visualization

To better observe a potential relationship between *X* and *Y*, we use NetworkX to plot a directed graph in order to visualize the connections between *X* and *Y* nodes. The nodes are *X* and *Y* phrases, while the weights between them are the count of each *X* and *Y* pair appears in the *jia X zhen Y* construction, visualized as line widths of the edges. The

nodes are in different colors to see what slot the phrases appear in, red represents occurrence only in *X*, blue represents occurrence only in *Y*, green represents occurrence in either slots. A simplified flow chart is demonstrated in Figure 1 below.

Figure 1

Flow Chart of Making NetworkX Graph



fastText Word Vector Representation

Pre-trained Model

This model uses pre-trained model "cc.zh.300.bin" loaded from fastText (Bojanowski, 2017; Joulin, 2016a, 2016b). The words in *X* and *Y* slots are transformed into vector representations and project to two dimensions by Principal Component Analysis (PCA). The two-dimension word vectors are collected into a data frame for plotting. We build a scatter plot with the dots being phrases in *X* and *Y*, distinguished by their colors. *X* slots are in red while *Y* slots are in green. Two ellipses with the confidence level of one standard deviation are appended on the plot to better visualize the distribution of the dots.

Unmasked Model

To train a model with current data, we perform word segmentation by "CKIPtagger", while setting coerced dictionary collecting all phrases in *X* and *Y* slots to

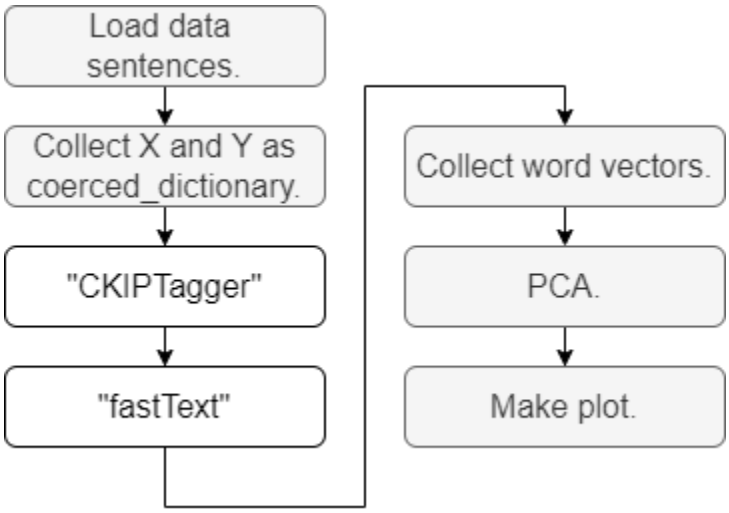
ensure that they are not out of dictionary. With a similar approach to that of the previous model, we trained a fastText model with unaltered sentences. The word vectors are consistent with those of the pre-trained model, which is in dimension 300. After using the same PCA used in pre-trained model to reduce the vector size to 2, we make a scatter plot with the dots and confidence ellipses of phrases in *X* and *Y*.

Masked Models

To make sure that the word vector representations of *X* and *Y* slots are not affected by the repetitive occurrences of *jia* and *zhen*, we replace *jia* and *zhen* with various strings. Such replacements include substituting both *jia* and *zhen* for "OOV", "*jia*", "*zhen*", or "" (empty strings). Thus 4 masked models are built by similar procedures used in either pre-trained or unmasked models. A simplified flow chart of word vector representations is demonstrated in Figure 2 below.

Figure 2

Flow Chart of Making Word Vector Representation Scatter Plot



Results

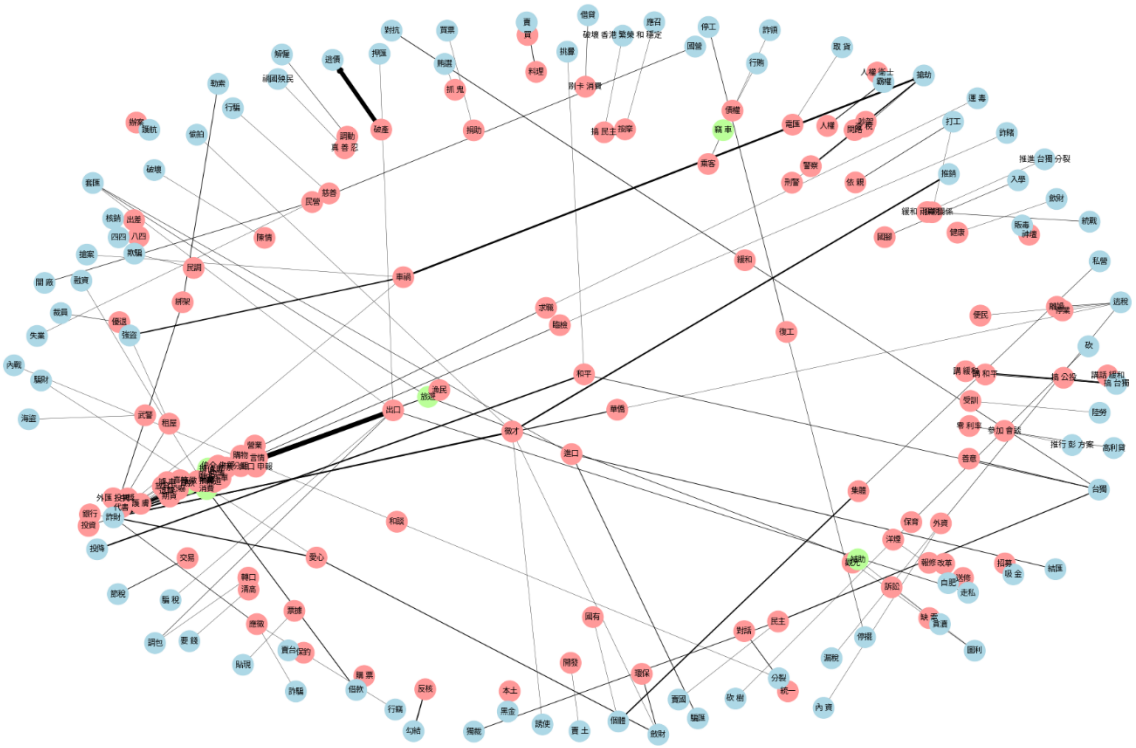
NetworkX Visualization

Figure 3 is the result of the directed graph of *X* and *Y*. At first glance, we can see that most *X* and *Y* are islands. This is because Chinese Word Sketch Engine is a corpus of

newspaper articles, and most *jia X zhen Y* constructions are used to emphasize on crimes or political opinions, of which neither is repetitive.

Figure 3

Directed Graph of jia X zhen Y (X and Y as nodes)



Moreover, there is a cluster of nodes in the bottom-left corner of Figure 3. The clustering may represent a potential topic. However, topic modelling is not the primary focus of this paper and is awaiting future research.

Looking closer into the counts of nodes, we find that there are more unique phrases in *X* slot (330) than in *Y* slot (279). Though the difference is not drastic, the nature of the corpus can account for this count difference. Since the *Y* slots contain crimes that the news is reporting about, the types of crimes are limited, such as *逃稅* 'tax evasion', *走私* 'smuggling', etc. On the other hand, *X* slots often contain the *modi operandi* or disguises of crime, which are more diverse, such as *外資* 'foreign direct investment', *合資* 'joint venture', etc.

Finally, the direction vectors between *X* and *Y* are shifts of sentiment polarity, such as from positive to negative, from neutral to negative, from negative to positive, etc. However, most instances in the collected data sentences belong to the previous two types. This is also due to the news article nature of the CWS corpus, causing biases that *Y* will be negative when *zhen Y* is a crime or malicious intentions disguised as a neutral or positive *X*. For example, in sentence (8), *X qiuzhi* is neutral and *Y zhacai* is negative.

(8) 假 求職 真 詐財
jia qiuzhi zhen zhacai
fake looking for jobs real scamming
'not looking for jobs but scamming'

fastText Word Vector Representation

Figure 4 contain scatter plots of word vector representations of the models mentioned above. From Figure 4 below, we can see that the pre-trained model is quite different from other models that are trained by data sentences, while not much differences occur among masked and unmasked models. This is normal since pre-trained model is trained based on a much bigger corpus, so the word vector representations are much denser than what we have trained on a selective corpus

Figure 4

Scatter Plot of Word Vector Representations in fastText Models

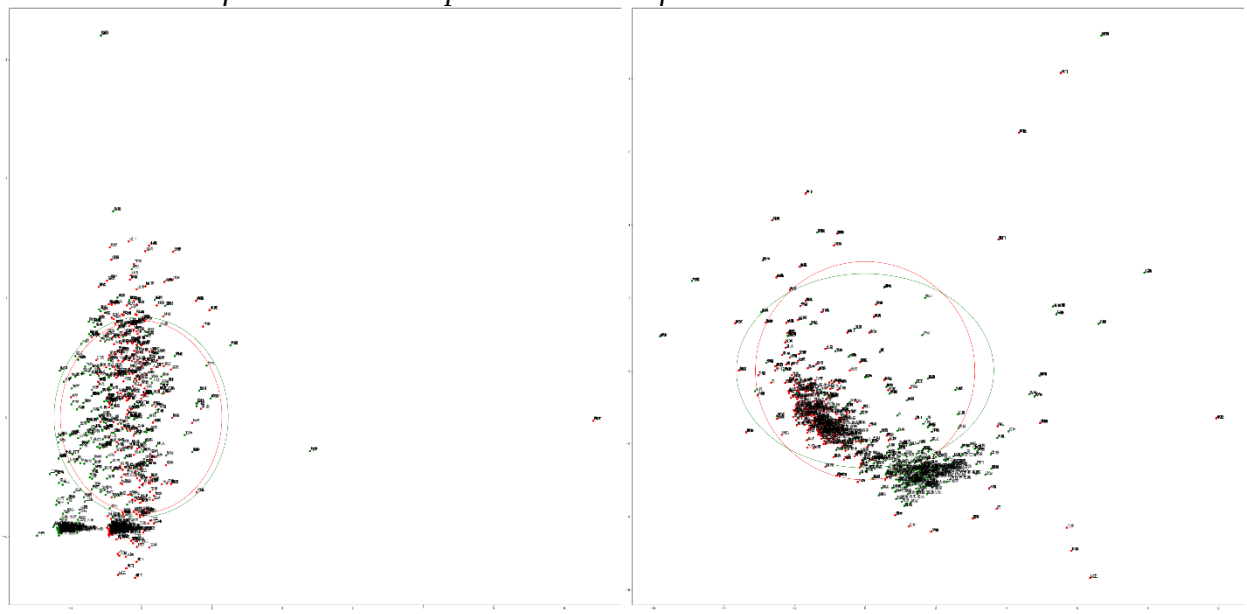
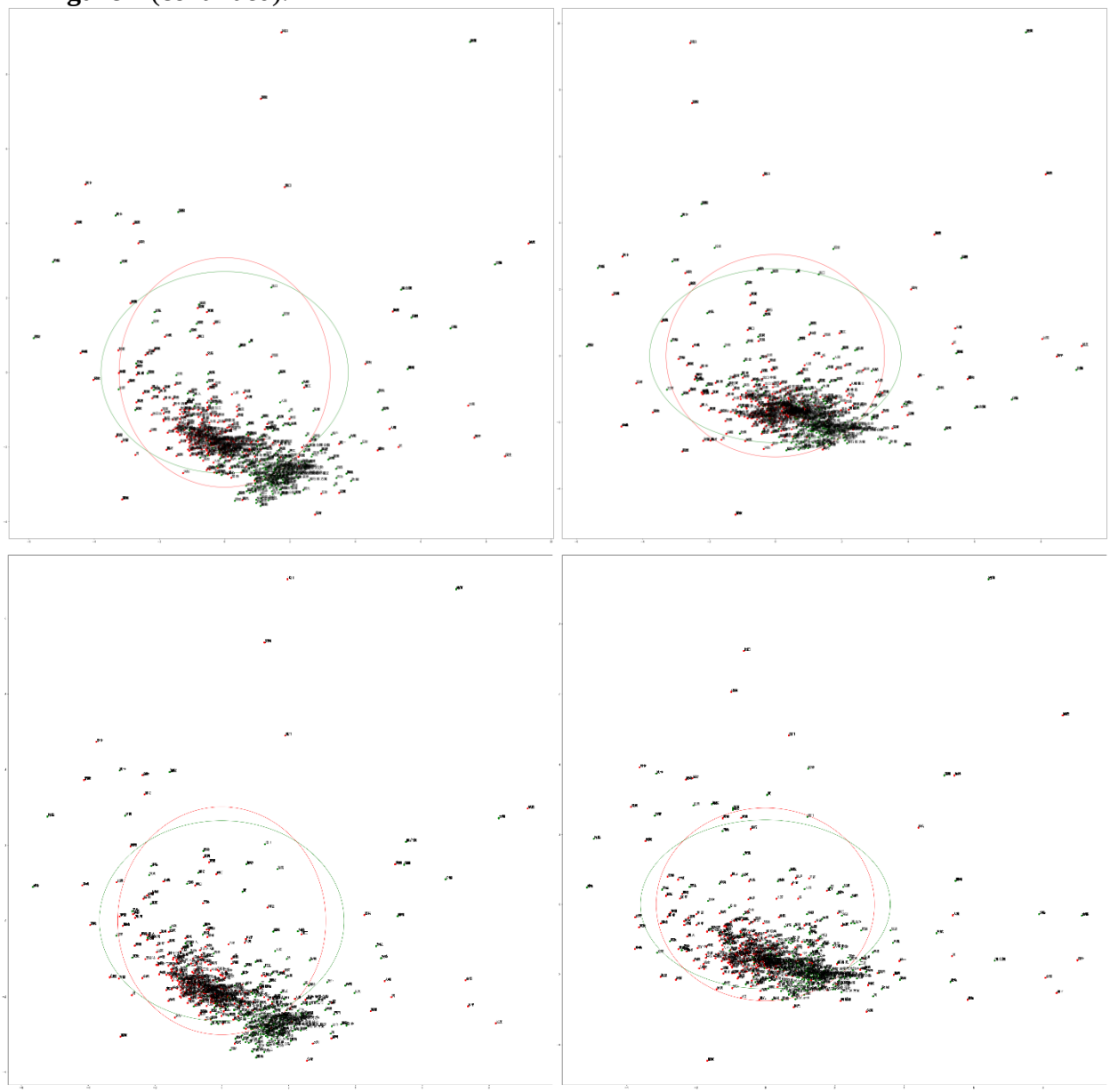


Figure 4 (continued).



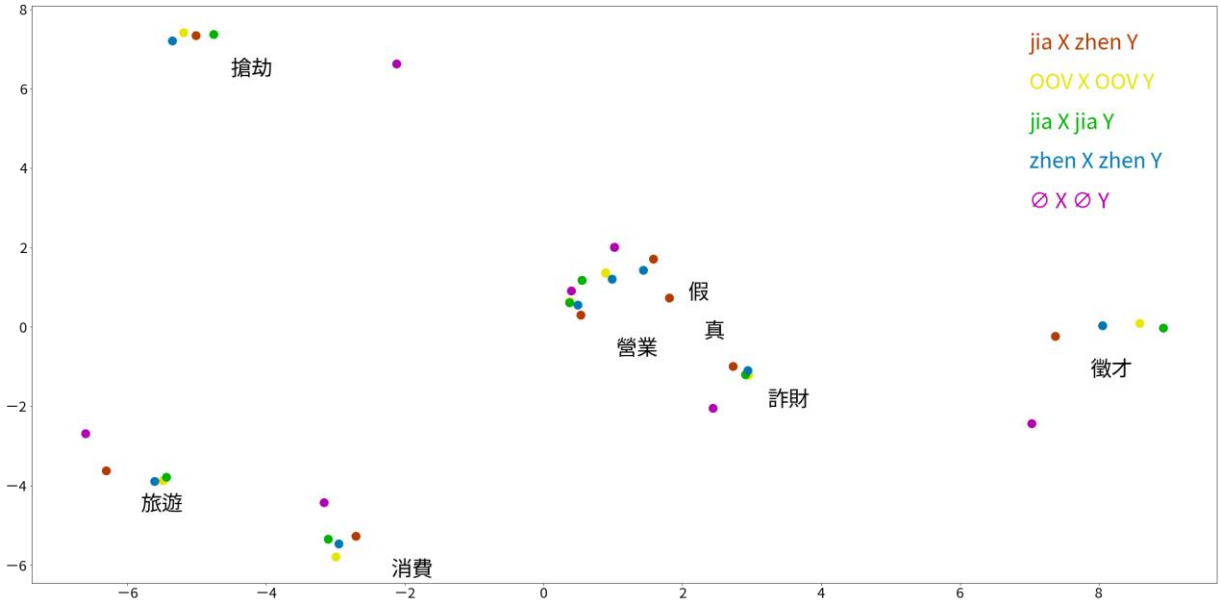
Note. From top-left to bottom-right, horizontally, are pre-trained model, *jia X zhen Y*, OOV *X* OOV *Y* model, *jia X jia Y* model, *zhen X zhen Y* model, and ∅ *X* ∅ *Y* model.

In order to see the differences of word vector representations in masked and unmasked models, Figure 5 selects *jia*, *zhen*, and a total of six representative phrases that have appeared in only *X* slot, only *Y* slot, or both slots. Phrases that appear only in the *X* slot are *營業* 'opening for business' and *徵才* 'recruiting', while *搶劫* 'robbing' and *詐財* 'scamming' appear only in the *Y* slot. *消費* 'consuming' and *旅遊* 'travelling' have appeared in both *X* and *Y* slot. Similar to the result of Figure 4, Figure 5 indicates that

there is no significant difference between masked and unmasked models. The ∅ *X* ∅ *Y* model is more separated from other models, probably because this model has removed *jia* and *zhen*, causing adjacency with more diverse words to increase. Nevertheless, considering this is a comparison among models, this degree of deviation is acceptable.

Figure 5

Selected Word Vector Representations in fastText Models



Discussion

This paper aims at discussing *jia X zhen Y* construction by looking into the word vector representations of phrases in *X* and *Y* slots. By using node graph to see interrelationships among *X* and *Y* slots, we find that there are more diverse words in *X* slot than in *Y* slot. This incongruity in phrase counts might be influenced by the nature of the data, which are excerpted from news articles. Furthermore, by using multiple fastText model, we consistently find patterns in word vector representations, meaning that *jia* and *zhen* select words that have unique characteristics yet to be discovered. This might also an effect of news article style.

Limitations

This paper is limited by the data collection from CWS, which is both too small and too biased for news articles. It would be preferred if data sentence collections include colloquial corpora such as Ptt Forum to balance the writing styles.

Moreover, part of speech, sentiment, and phrase length vary in every data sentence, and may alter the word vector representations. These factors should be controlled.

Future Research Directions

Possible future work includes discussing the diachronic changes of the relations between *X* and *Y*, since Internet forums such as Ptt Forum have continued the use of *jia X zhen Y* construction. It would be interesting to see the semantic changes of *jia* and *zhen*. The sentiment polarities of specific phrases in the *X* and *Y* slots would be the main focus.

In addition, future work should gather more data from more corpora and compare word representations among corpora, to see if the *jia X zhen Y* construction can adapt to different writing styles. If so, what are patterns differences of the constructions in different corpora?

As for NetworkX model, it is too complicated to be interpreted easily. Future research may consider using synonymy or other methods to pre-process and group similar words into larger collective nodes.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Budts, S., & Petré, P. (2020). Putting connections centre stage in diachronic Construction Grammar. *Nodes and networks in diachronic construction grammar*, *27*, 317.

Desagulier, G. (2015). Le statut de la fréquence dans les grammaires de constructions: simple comme bonjour?. Langages, (1), 99-128.

Desagulier, G. (2021, November). Capturing horizontal links in a construction network using semantic vector spaces. In *Modelling Constructional Variation and Change: Agents, Networks and Vectors*.

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure. Chicago: Univ. of Chicago Press.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Huang C.-R., Smith S., Ma W.-Y., & ŠImon P. (2005). Academia Sinica Center for Digital Cultures. Chinese Word Sketch Engine. Retrieved December 21, 2021, from https://wordsketch.ling.sinica.edu.tw/cws/

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016a). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016b). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Liu F., Luo Q. (2020). Gradability, Subjectivity and the Semantics of the Adjectival zhen 'real' and jia 'fake' in Mandarin. In: Hong JF., Zhang Y., Liu P. (eds) Chinese Lexical Semantics. CLSW 2019. Lecture Notes in Computer Science, vol 11831. Springer, Cham. https://doi.org/10.1007/978-3-030-38189-9_17

Rambelli, G., Chersoni, E., Blache, P., Huang, C. R., & Lenci, A. (2019, August). Distributional semantics meets construction grammar. towards a unified usage-based model of grammar and meaning. In First International Workshop on Designing Meaning Representations (DMR 2019).