Comprehension correlates of the occurrence and deletion of "de" in Mandarin "N1 (de) N2" structures

Junyuan Zhao
Department of Chinese Language and Literature
East China Normal University
Shanghai, China
e-mail: jyzhao9802@gmail.com

Junru Wu
Department of Chinese Language and Literature
East China Normal University
Shanghai, China
e-mail: jrwu@zhwx.ecnu.edu.cn

***Abstract*—Based on corpus materials and on-line semantic judgment surveys, this paper investigates the comprehension differences related to the occurrence and deletion of "de" in the Mandarin "N1 (de) N2" structure. By applying PCA and LME modellings on a set of semantic survey data, this study provides a multi-level database of semantic measurements for a set of Chinese "N1 (de) N2" structures as well as a quantitative analysis regarding the correlation between structure-level and constituent-level semantic features. The research shows that: (1) The "de"-occurring structure is more likely to be interpreted as indefinite than the "de"-deletion structure. (2) Animacy of N1 is positively related to the grammaticality of the "de"-occurring structure, while animacy of N1 is negatively related to the grammaticality of the "de"-deletion structure. The research findings provide evidence for prototype effects in the process of language comprehension. We propose that in natural comprehension, there is a high-animacy bias for N1 regarding the "de"-occurring structure; while a low animacy interpretation for N1 is more prototypical for the "de"-deletion structure. Accordingly, the "de"-occurring structure tends to be interpreted as a possessive, while the "de"-deletion structure is more likely to be interpreted as a modifier-head structure.**

Keywords—occurrence and deletion of "de"; Chinese de structure; semantic change; quantitative study

I. INTRODUCTION

The Mandarin structure "N1 (de) N2", where the presence of the functional word "de" is optional in some conditions (e.g., 皇帝(的)女儿, EMPEROR (DE) DAUGHTER, "daughter of the emperor"), where "de" is optional) has long been under linguistic discussion. Following either a descriptive or formal paradigm, decades of efforts have been made to delineate the mechanism lying behind this phenomenon, because of its theoretical importance to the Mandarin determiner system and the Chinese word boundary puzzle. This problem is also relevant to Natural Language Processing (e.g., for entity recognition) and has also been investigated by recent corpus linguistic studies, such as by Zhang et al [1]. Zhang and his colleagues' recent study relied on formal features of the constituents (i.e., part-of-speech annotation), while whether two slightly different structures (i.e. "de"-occurring and "de"-deletion structures) stand for two distinct entities remained a question.

Among Chinese linguists, Zhu [2] was the first to systematically probe into this problem. He found that the occurrence and deletion of "de" in this structure is related to several language-internal factors, such as the fixedness of expressions, use of kinship terms and word-lengths of constituents. He also admitted that the selection of "de" is quite flexible and requires further studies.

Following Zhu, a number of researchers have looked into this problem, among which many have approached the problem from a holistic perspective and focused on the property of the whole structure. As Xu [3] suggested, constituents in the "de"-occurring structure are loosely-combined while constituents in the "de"-deletion structure are tightly-combined, which made the constituents an integrated module in the context. This theory of "module integrity" may have predictive power regarding the occurrence and deletion of "de" in such structures. However, this is an intuitive judgement and requires further empirical verification.

Other pertinent studies focused on internal semantic relations of the constituents. Yuan [4] and Wen [5] are among the first linguists to point out that the most basic semantic distinction involved is the distinction between possessive and modifier-head relations, i.e., possessives tend to select for the "de"-occurring structure while modifier-head relations (which is common cross-linguistically[6] in noun phrases) tend to select for the "de"-deletion structure.

Yuan [4] and Wen's [5] explanation accounts for a large part of materials in linguistic literatures of Mandarin. An alternative semantic explanation was provided by Si [7], who focused on the subcategories of possessives and proposed that alienable possessives must take a "de" while the case of inalienable possessives is more flexible. However, Si's explanation is subject to reproductive counter-examples (e.g. the possessive 我帽子-I hat-"my hat"[1] is alienable but is a grammatical 'de'-deletion structure).

Formal linguists also have been avoiding this phenomenon in their analyses because of its obscure and uncertain nature [8]. Thus far, the description and explanation regarding the occurrence and deletion of "de" in the Mandarin "N1 (de) N2" structure are still insufficient. This is partly due to complex factors involved. For instance, the findings of Feng [9] on prosodic constraints and the findings of Lu [10] on pragmatic constraints remind us that this phenomena is not solely syntactically constrained. A specific example provided by Lu was the pragmatic difference between 我的女朋友 (I DE GIRLFRIEND, "my girlfriend")

[1] A context can be: 我找不到我帽子了(I FIND NEG ARRIVE I HAT PERF, "I cannot find my hat")(NEG-negative, PERF-perfect).

and 我女朋友(I GIRLFRIEND, “my girlfriend”), with which Lu suggested that the structure with “de” can denote more than one entity while the one without “de” can only denote one, indicating the “de”-occurring structure is indefinite.

A diachronic explanation can be that the Mandarin functional word “de” as well as the structure involved have undergone complex historical changes [10]–[12], which, along with the property of Mandarin as a language lacking formal markers [13], made the semantic and syntactic interpretation of “de” obscure and fuzzy. Nevertheless, a diachronic explanation, if even exist, cannot directly help synchronic or practical predictions.

The nature of this problem obviously conflicts with the binary feature of the long-adopted descriptive paradigm. Hence, an alternative approach that may improve practical prediction of the occurrence of “de” is to assume a (possessives-to-modifier-head) semantic continuum rather than a binary contrast, similar to what has been pointed out by Downing [14] in a study on English compounds. With a probabilistic view, the problem is thus translated to a probability issue regarding grammaticality and predictive semantic features.

Taking this approach, this study focuses on the probabilistic correlations between structure-level and constituent-level semantic features, as well as the relations between semantic features and grammaticality. Inspired by previous studies, which pointed out that categories of possessives show prototype effects and can be predicted by the concreteness and animacy of constituents [15], [16], this study seeks to better model the phenomenon using these semantic features. Quantitative methods were adopted in this study, which include mixed linear logistic regression (LME) [17], Principal component analysis (PCA, with the princomp function provided in R [18]) and hypothetical tests. Target semantic features based on subjective rating were collected with an online survey.

II. Online Survey

A. Sampling of Corpus Material

Materials of the online survey were downloaded from the BCC corpus [19] using the query “../n 的 ../n”. 144560 target structures (i.e., “N1 de N2” structure with different constituents, provided along with an original context of approximately thirty characters) were downloaded under the “news” subcategory as raw materials, of which 53800 entries were downloaded time-wise and 90760 time-reversed. For better sampling results, we clustered the raw materials based on vector representations. First, all original contexts were parsed with the jieba [20] parser for Chinese in python. In order to make sure that all target structures are parsed out, we included all these structures in the parser dictionary. Next, we generated semantic representations for all target structures with a 100-dimension vector, using the word2vec package from the gensim [21] python library. The context required for word2vec was constructed by concatenating all parsed contexts of the original material. Then we performed a K-means clustering using the sklearn package [22] in python. The goodness of clusters was evaluated with silhouette scores [23] (see Table I).

TABLE I. Number of Clusters ~ Silhouette Scores

N clusters	Silhouette score
2	0.302
3	0.480
4	0.126

Table I shows that a cluster number of three yielded the best clustering result. Two thousand samples were extracted using stratified sampling, based on the clustering result.

A further filtering of samples was conducted by manual examination. Undergraduates who have received fundamental linguistic training were recruited for this task, following the criteria that a target structure should not overlap with other constituents in the original context (e.g., the structure 民族的精神 (ETHNIC GROUP DE SPIRIT, “spirits of the ethnic group”) extracted from 中华民族的精神 (THE CHINESE ETHNIC GROUP DE SPIRITS, “spirits of the Chinese ethnic group”) was rejected because the constituent 民族 overlapped with the constituent 中华民族). Eventually two hundred “de”-occurring samples were selected and two hundred “de”-deletion samples were generated accordingly by deleting “de”. The two hundred “N1 (de) N2” pairs were used as materials for the online survey.

Note that, most of the materials (“de”-occurring structures) that passed the manual examination carry an animate N1 (e.g. a person), which reveals a natural tendency of the “de”-occurring structure to select for an animate N1.

B. Survey Design

An online survey was designed on the Qualtrics [24] platform with four tasks involved: (a) grammaticality judgement, (b) concreteness rating, (c) animacy rating and (d) definiteness judgement. The grammaticality judgement and definiteness judgement targeted the structures as a whole and both “de”-occurring and “de”-deletion structures were rated (2 × 1). The concreteness rating and animacy rating targeted both the structures (“de”-occurring, “de”-deletion) and the constituents (N1, N2). Each constituent was rated three times, twice as embedded in (the two) structures and once in isolation, 3 (N1/N2/structure) × 2 (occurring/deletion) + 2 (N1/N2 in isolation).

(a) The grammaticality judgement asked the participants to decide whether the target structure is acceptable. (b) The concreteness rating asked the participants to rate the concreteness of the target constituents on a 5-point scale, using the instructions translated from previous English studies on semantic concreteness [25]; target constituents were marked red. (c) The animacy rating required the participants to rate the animacy of the target constituents on a 5-point scale. The definiteness judgement was based on the result of the acceptability judgement of the Chinese existential sentence 有…在这儿 (HAVE … AT HERE, “there be … here”), following a previous study [26] on the definiteness of Chinese N-de-N structures.

Twenty-four native Standard Mandarin speakers participated in the on-line survey and were paid. We balanced the two hundred materials across twenty-four participants in a way such that a single N-(de)-N pair is to be rated by three different participants on all target semantic features. To avoid

potential lexical learning effects[2], we set a time limit of five seconds per structure to tasks that involve grammaticality/acceptability judgements.

III. Data Analysis

We received twenty-four responses for all twenty-four surveys distributed and data of 199 structures were valid out of the two hundred samples. There were 14328 (199 N1-N2 pairs × 3 participants × 24 measurements) data points in total. We carried out the following statistical analysis in R (version 3.5.3).

A. Probalistic Distribution of Concreteness

First we looked at the distribution of concreteness of constituents:

On the contrary, the concreteness distribution of N1 does not fit the canonical distribution of concreteness as found by previous researches, which suggests that the structures (with and without "de") have semantic constraints on the N1 constituent, as compared with the distribution in a natural context where target words are separately examined.

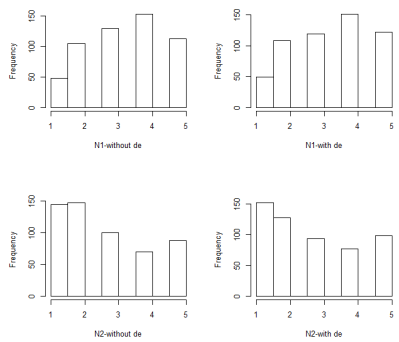

Figure 1. Concreteness distribution of N1 and N2

As shown in Figure 1, the concreteness distributions of N1 and N2 differs in both structures with and without "de". The concreteness distributions of N1 in both structures are bimodal, while the concreteness distributions of N2 in both structures resembled a normal distribution. The distribution of N2 aligned with the results of previous studies [25], [27] on the concreteness of unrestricted English words based on a large scale survey.

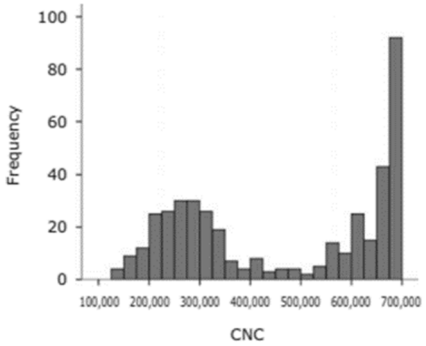

Figure 2. Concreteness distribution of unrestricted English words, from Della Rosa (2010) [24] (CNC-concreteness)

B. Semantic Comprehension of Constituents within Structures versus in Isolation

Next, we further investigated the constraints that "de"-occurring and "de"-deletion structures exert on the constituents. A linear regression was performed, using the individual concreteness of N1 as the independent variable and the concreteness of N1 in the structures as the dependent variable. The regressions for both structures were performed. The results were plotted as is shown in Fig.3, using a black line (y = x) for reference.

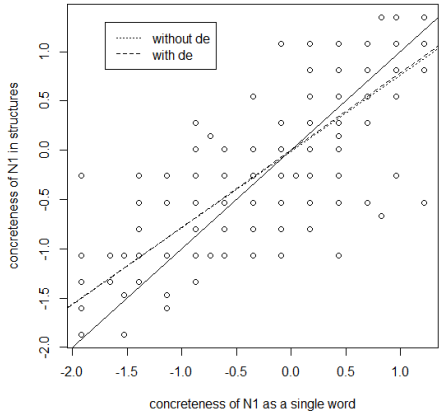

Figure 3. Concreteness of bare N1 v.s. Concreteness of N1 in structures

Figure 3, along with the following t-tests, shows that different structures yield different interpretations of concreteness regarding N1. For the "de"-deletion structure, $t(197) = 26.54$, $p < 0.001$, $R^2 = 0.78$; for the "de"-occurring structure, $t(197) = 24.29$, $p < 0.001$, $R^2=0.75$. Both distributions showed a pattern of "neutralization", which means that the concreteness of N1 in both structures covers a smaller scale than the original bare noun. This effect applies to both structures, suggesting that "N1 (de) N2" structures alters the semantic interpretations of their constituents. We also found a similar difference regarding constituent animacy in bare noun form versus in structures, see Figure 4.

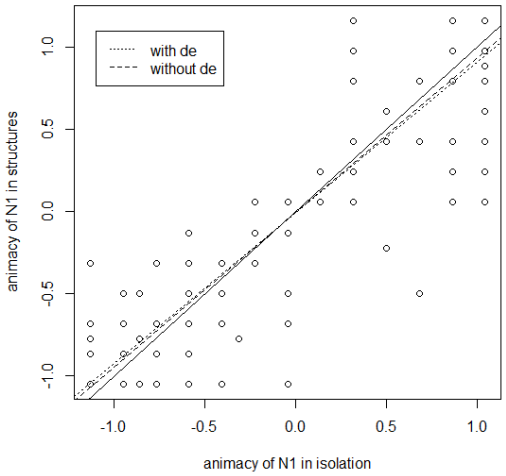

Figure 4. Concreteness of N1 in isolation v.s. Concreteness of N1 in structures

[2] In a pilot survey, we noticed that without time limits, many apparently-unacceptable constructions were marked as acceptable. This is probably due to the participants' integration of the novel constructions into their mental lexicon in a relatively short period of time (i.e. lexicalization on an individual level).

We further investigated the relations between semantic features of constituents and the structure as a whole with LME modelling [17]. Taking the concreteness/animacy of the whole structure as the dependent variable and semantic features of the constituents as the independent variables, models for both structures showed that the semantics of the whole structure correlates with the semantic features of N2 ($t = 26.45$, $p < 10^{-5}$), but not with N1. These findings provided quantitative evidence for the headedness of the Chinese modifier-head structure.

C. Definiteness and "de" Occurence

To investigate the definiteness of different structures, we performed a paired t-test on definiteness scores. Since the raw definiteness data were binary regarding the task involved (i.e., 1-definite, 0-indefinite), by-pair means were calculated beforehand. It turned out that the definiteness of the two structures are significantly different, $t(197) = 2.65$, $p = 0.009 < 0.01$. The "de"-occurring structure is more likely to be interpreted as indefinite and the "de"-deletion structure is more likely to be interpreted as definite. This finding aligned with previous pragmatic studies [10].

D. Principal Component Analysis for Semantic Features

Based on the matrix of correlation, we found apparent multi-collinearity across the semantic measurements (i.e., a high correlation between different semantic features, such as between the concreteness of N1 in both structures). Therefore, we performed a PCA on these predictors before investigating the structure of the semantic predictors that may account for the grammaticality of "N1 (de) N2" structures. As shown in Figure 5, a minimum of four principal components can explain most of the variance within the semantic predictors.

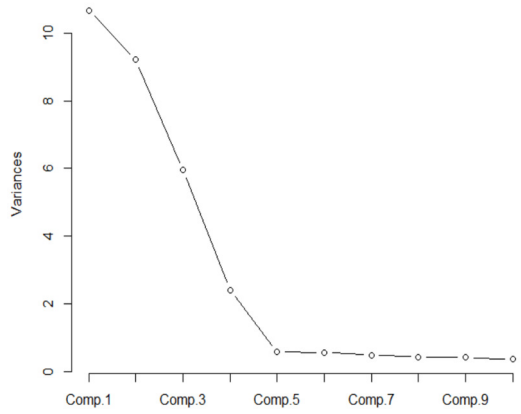

Figure 5. Variance of components of PCA

As is shown in Table II, the four principal components explained relatively separately the four theory-relevant semantic features (i.e., the concreteness and animacy of both N1 and N2), which did not conflict with our further analysis.

E. Linear Mixed Effect Logistic Regression

In order to further delineate the possible different semantic interpretations associated with "de"-occurring versus "de"-deletion structures, we performed a linear mixed effect logistic regression, using the original semantic measurements and the PCA components separately.

TABLE II. LOADINGS OF PRINCIPAL COMPONENTS

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Concreteness of N1, with "de"	0.230		0.178	0.501
Concreteness of N2, with "de"	0.351	-0.164	0.138	-0.247
Concreteness of structure, with	0.332	-0.154	0.163	-0.116
Concreteness of N1, without "de"	0.205		0.194	0.493
Concreteness of N2, without "de"	0.334	-0.162	0.159	-0.191
Concreteness of structure, without	0.348	-0.152	0.164	-0.129
Concreteness of single N1	0.201		0.188	0.494
Concreteness of single N2	0.345	-0.161	0.111	-0.261
Animacy of N1, without "de"	0.191	0.529		-0.119
Animacy of N2, without "de"	0.190		-0.393	
Animacy of structure, without	0.190		-0.397	
Animacy of N1, with "de"	0.207	0.532		
Animacy of N2, with "de"	0.185		-0.391	
Animacy of structure, with	0.190		-0.397	
Animacy of single N1	0.182	0.536		-0.135
Animacy of single N2	0.158		-0.401	

We modelled the phenomenon using occurrence/deletion of "de" (a binary variable) as the dependent variable. The principal components which explained most parts of the concreteness and animacy of constituents (Comp.1 through Comp.4), were taken as the independent variable; participant and the pair ID were taken as random factors. Based on the fact that this study targeted at the semantics of originally grammatical "de"-occurring structures, we omitted the data points for which "de"-occurring structures were labelled as ungrammatical. Several different models using other original semantic features (e.g., concreteness of N1 in "de"-occurring structures) of constituents were also fit. The models were evaluated with ANOVA (analysis of variance) and the best-fit model was selected and reported in TABLE III ($X^2 = 0.153$, $p < 2 \times 10^{-16}$).

TABLE III. MODEL SUMMARY

Formula:validity_0 ~ conc_left_comp + anm_left_comp + (1 | subject) + (1 | pair_number)				
Fixed effects				
Predictor	Estimate	Std. Error	Z value	p value
Principal comp. of concreteness N1	-0.155	0.100	-1.561	0.118
Principal comp. of animacy N1	-0.170	0.056	-3.018	0.002**
Random effects				
Predictor	Variance	Std. Deviation		
pair number	2.746	1.657		
participant	0.467	0.684		

[a] validity_0: grammaticality of the "de"-deletion structure; conc_left_comp: concreteness of N1 (after PCA); anm_left_comp: animacy of N1 (after PCA). *p<0.05, **p<0.01, ***p<0.001

As is shown by model estimates in TABLE III, there is a significant negative correlation between animacy and grammaticality of the "de"-deletion structure, which indicates that the less animate N1 is, the more acceptable the "de"-deletion structure is. Inferring from the fact that the survey elicited participants' on-line interpretations, we suggest that this statistical tendency implies an interpretational tendency. The principal component for the concreteness of N1 showed a similar effect, although the effect did not reach significance.

IV. Discussion and Conclusion

This study shows different semantic and pragmatic interpretations of the Mandarin "N1 (de) N2" structure with and without "de" using a quantitative approach. We also provided a multi-level database of constituent semantic measurements.

Regarding the pragmatics of the structure, we found that the "de"-occurring structure tend to be less definite than the "de"-deletion structure. This finding aligns with Lu [10] on the definite-indefinite pragmatic differences of the two structures on a paragraph level. Moreover, the statistical analysis also indicates that the definite/indefinite distinction is not binary nor deterministic.

Regarding the semantics of these structures, the results suggest that the grammaticality of the "de"-deletion structure is negatively related with the animacy of N1, which indicates that the "de"-deletion structure tend to have a low animacy interpretation for N1, since the measurements are participants' on-line interpretations. Based on the linear nature of the statistic model used, we infer that the "de"-deletion structures that were considered unacceptable (i.e. structures that are only acceptable when "de" is present) tend to have a high animacy interpretation for N1. This quantitative evidence supports that the semantic differences between the two structures associated with structural differences, which is aligned with the very fundamental notion of grammatical constructions[28] and in the meantime rejected a possible theoretical alternative of viewing this problem as a case of ellipsis, as ellipsis hardly alters the meaning of a phrase or a sentence.

Considering that a typical possessor should be animate and concrete, we propose that the "de"-occurring structure is influenced by the prototype effect of possessives in the psychological process of language comprehension, while the "de"-deletion structures tend to be interpreted as normal modifier-head compounds. This finding is in-line with previous cognitive studies which showed that semantic possessives have a tendency to select for a more animate N1 [15]. As mentioned in section II.A, we also noticed that the filtered materials from the BCC corpus were mostly possessives with an animate N1, which provide evidence for a preference for animate N1s in the "de"-occurring structure, in natural language use.

Although this study provided a quantitative insight for this long-discussed problem, the following points require further investigation. First, since we were not able to search for "de"-deletion structures on BCC corpus, only the "de"-occurring candidate structures were natural, and the "de"-deletion candidate structures were generated based on these "de"-occurring structures. This may have introduced some research bias. Since "de"-deletion structures are quasi-compounds according to some Chinese linguists, a further research based on natural "de"-deletion materials could benefit the investigation of their role in lexicalization. Another sampling bias may come from the fact that all materials were downloaded under the "news" subcategory, which might not reflect the real-world distribution. Also, the survey targeted at only college students whose age/education-specific linguistic background and therefore might influence the research results.

Second, the tendency as is shown by the LME model (in section III.E) have a few alternative explanations. One explanation can be made referring to the prototype effect as proposed above. However, whether this prototype effect stemmed from syntactic influence or merely from a preference regarding the use of a certain surface structure remained to be discussed. An alternative explanation can be made from a psycholinguistic perspective: since "de"-deletion structures are indistinguishable from compounds by its surface form, and it has been shown that constituents in compounds take less cognitive resources than constituents that are morphologically-free in a sentence or a structure [29], it could also be that the comprehension bias regarding "de"-deletion structures are rooted in the prototypical comprehension mechanisms of compounds. Admittedly, this interpretation requires further experimental investigation.

To conclude, this study, with a quantitative approach, revealed pragmatic and comprehension differences of the Mandarin "N1 (de) N2" structures. The results suggested possible prototype effects of the two structures, which serves as an evidence for viewing "N1 (de) N2" structures as grammatical constructions.

References

[1] C. Zhang, Z. Zhang, G. Rao, and E. Xun, "Research on extraction of simple modifier-head chunk based on Corpus," presented at the Chinese Lexical Semantic Workshop, Beijing, China, 2019.

[2] D. Zhu, Collected Works of Zhu Dexi, vol. 4. The Commercial Press, 1999.(in Chinese) (朱德熙, 朱德熙文集, vol. 4. 商務印书館, 1999.)

[3] Y. Xu, "The occurrence and deletion of "de" when using personal pronouns as modifiers revisited", Studies of the Chinese Language, no. 01, pp. 21-27+95, 2008.(in Chinese) (徐阳春, "也谈人称代词做定语时'的'字的隐现," 中国语文, no. 01, pp. 21-27+95, 2008.)

[4] Y. Yuan, "Implying predicate and its syntactic influence——on the denoting rules of "de" structures and the grammatical and semantic function of "de", Studies of the Chinese Language, no. 04, pp. 241–255, 1995.(in Chinese) (袁毓林, "谓词隐含及其句法后果——'的'字结构的称代规则和'的'的语法、语义功能," 中国语文, no. 04, pp. 241–255, 1995).

[5] Z. Wen, "Identifying the semantic relation between N1 and N2 in the 'N1 de N2' modifier-head structure", Linguistic researches, no. 03, pp. 22–27, 1999.(in Chinese) (文贞惠, "'N_1(的)N_2'偏正结构中N_1与N_2之间语义关系的鉴定," 语文研究, no. 03, pp. 22–27, 1999.)

[6] G. Fanselow, *The Parametrization of Universal Grammar*. John Benjamins Publishing, 1993.

[7] F. Si, "'Zhao Benshan's grandfather' and 'Zhao Benshan's hat'——on two kinds of possessives in Chinese", Language teaching and studies, no. 02, pp. 43–51, 2014. (in Chinese) (司富珍, "'赵本山的爷爷'和'赵本山的帽子'——漫谈汉语中的两种领属结构," 语言教学与研究, no. 02, pp. 43–51, 2014.)

[8] L. L.-S. Cheng and R. Sybesma, "De as an underspecified classifier: first explorations," *Yuyánxué lùncóng*, vol. 39, pp. 123–156, 2009.

[9] S. Feng, Prosodic syntax in Chinese. Shanghai Educational Publishing House, 2000. (in Chinese) (冯胜利, 汉语韵律句法学. 上海教育出版社, 2000.)

[10] B. Lu, “On the basic and derived functions of 'de'", Chinese Teaching in the World, no. 01, pp. 14-29+2, 2003. (in Chinese) (陆丙甫, “‘的’的基本功能和派生功能——从描写性到区别性再到指称性,” 世界汉语教学, no. 01, pp. 14-29+2, 2003.)
[11] Y. Shi, “On the uniform grammatical function of 'de'", Chinese Teaching in the World, no. 01, pp. 16–27, 2000. (in Chinese) (石毓智, “论‘的’的语法功能的同一性,” 世界汉语教学, no. 01, pp. 16–27, 2000.)
[12] M. Zhu, “Grammatical functions of 地,底 in Zhu Zi Yu Lei”, Studies of the Chinese Language, no. 03, 1982. (in Chinese) (祝敏彻, 《朱子语类》 中 ‘地’‘底’ 的语法作用,” 中国语文 第 3 期, 1982.)
[13] D. Zhu, *Lectures on the Chinese Grammar*. Commercial Press, 1982.(in Chinese) (朱德熙, 语法讲义. 商务印书馆, 1982.)
[14] P. Downing, “On the Creation and Use of English Compound Nouns,” *Language*, vol. 53, no. 4, pp. 810–842, 1977.
[15] Z. Liu, Y. Ren and D. Zhong, “Constraints of animacy on possessives”, *Journal of Foreign Languages*, vol. 41, no. 04, pp. 30–43, 2018. (in Chinese) (刘正光, 任远, and 钟丹凤, “领属关系的生命度制约,” 外国语(上海外国语大学学报), vol. 41, no. 04, pp. 30–43, 2018.)
[16] N. Ge, “A review on studies of the possessive category” , *Language Education*, vol. 4, no. 04, pp. 62–68, 2016. (in Chinese) 葛娜娜, “领属范畴研究综述,” 语言教育, vol. 4, no. 04, pp. 62–68, 2016.
[17] D. Bates, M. Maechler, B. Bolker, and S. Walker, “lme4: Linear mixed-effects models using Eigen and S4,” *R package version*, vol. 1, no. 7, pp. 1–23, 2014.
[18] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019.
[19] E. Xun, G. Rao, X. Xiao and J. Zang, “Developing the BCC corpus with big data”, Corpus Linguistics, vol. 3, no. 1, pp. 93–118, 2016. (in Chinese) (荀恩东, 饶高琦, 肖晓悦, and 臧娇娇, “大数据背景下 BCC 语料库的研制,” 语料库语言学, vol. 3, no. 1, pp. 93–118, 2016.)
[20] J. Sun, *‘Jieba’Chinese word segmentation tool*. 2012.
[21] R. Řehůřek and P. Sojka, “Gensim—statistical semantics in python,” *statistical semantics; gensim; Python; LDA; SVD*, 2011.
[22] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, p. 2825−2830, Oct. 2011.
[23] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.
[24] *Qualtrics*. Provo, Utah, USA: Qualtrics, 2005.
[25] P. A. Della Rosa, E. Catricalà, G. Vigliocco, and S. F. Cappa, “Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words,” *Behavior Research Methods*, vol. 42, no. 4, pp. 1042–1048, Nov. 2010.
[26] B. H. Partee, “A note on Mandarin possessives, demonstratives, and definiteness,” in *Studies in Language Companion Series*, vol. 80, B. J. Birner and G. Ward, Eds. Amsterdam: John Benjamins Publishing Company, 2006, pp. 263–280.
[27] M. Brysbaert, A. B. Warriner, and V. Kuperman, “Concreteness ratings for 40 thousand generally known English word lemmas,” *Behav Res*, vol. 46, no. 3, pp. 904–911, Sep. 2014.
[28] A. E. GOLDBERG, “The inherent semantics of argument structure: The case of the English ditransitive construction,” *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)*, vol. 3, no. 1, pp. 37–74, 2009.
[29] D. Koester and N. O. Schiller, “Morphological priming in overt language production: Electrophysiological evidence from Dutch,” *NeuroImage*, vol. 42, no. 4, pp. 1622–1630, Oct. 2008.