Language change in Chinese political discourse based on the relationship between sentence and clause

Renkui Hou
Guangzhou University, Guangzhou, China
hourk0917@163.com

Chu-Ren Huang
The Hong Kong Polytechnic University, Hong Kong
churen.huang@polyu.edu.hk

Kathleen Ahrens
The Hong Kong Polytechnic University, Hong Kong
Kathleen.ahrens@polyu.edu.cn

Xiaoyi Du
Central China Normal University, Huhan, China
565608551@qq.com

Abstract

The present paper explored the language change in Chinese political discourse in Mainland through the investigation of the relationship between sentence and clause based on Menzerath-Altmann law. The results showed that the relationship between sentence and clause in first and second periods (1978-1982 and 1997-2001) Chinese political discourses abide by the Menzerath-Altmann law, while they do not abide by the law in third period. The average clause length distribution in first and second periods political discourse can be fitted by the $y=ax^b$ and the fitted parameters can distinguish the different periods political speeches. The relationship between sentence and clause changed with time.

1 Introduction

Language diachronic change has been main concern of linguists from centuries in many parts of the world and is one of the topics of research for classical linguistic studies. Different from classical linguistics, which mainly concern the diachronic changes of the consecutive language elements with time, this paper considered language as a complex system and explored the change of this system from time to time. Cacoullos (1999) showed that reductive changes in grammaticalizing forms may be not only as a diachronic process but also as synchronic differences between formal and informal registers. On the contrary, it is possible that the synchronic differences were considered as the diachronic changes of language. This paper focused on one same register, political discourse, for examining the Chinese language change in order to avoid this mistake. Biber (2012) also argued strongly that reference works that describe different linguistics levels should consider register difference. Thus, we should consider register as an influencing factor for studying language change.

Quantitative linguistics consider language as a complex adaptive system and adapts to the surrounding environment all the time as like Wang (2006). The Menzerath-Altmann law, as the one of three laws of quantitative linguistics, examines the relationship between language constructs and their immediate constituents in different language domains. Based on Menzerath (1954), which proposed “the greater the whole, the smaller its parts” after he examined that dependency of syllable length on word length, Altmann generalized this hypothesis to all the language levels, formulating it as “The longer a language construct, the shorter its components” (Altmann 1980).

Altmann (1980) gave the theoretical derivation and the corresponding differential equation of the MA law, as shown in Equation (1).

$$\frac{y'}{y} = -c + \frac{b}{x} \quad \text{Equation (1)}$$

The solution to this differential equation is shown in the Formula (1):

$$y = ax^b e^{-cx} \qquad \text{Formula (1)}$$

where y is the mean size of the immediate constituents, x is the size of the construct, and parameters a, b, and c depend mainly on the levels of the units under investigation.

A large number of observations have shown that parameter c is close to zero for higher levels of language whereas lower levels lead to very small values of parameter *b*; only for intermediate levels is the full formula needed (Köhler, 2012).

The two simplified formulas were obtained when higher and lower levels were studied respectively. Formula (2) has become the most commonly used “standard form” for linguistic purposes (Grzybek, 2007).

$$y = ax^{b} \qquad \text{Formula (2)}$$

This paper explored the language change in Chinese political discourses by examining the relationship between sentence and clause based on the Menzerath-Altmann law.

1.1 Literature review

Previous studies about language change focused on sound and sound changes, word and word changes mostly. Lieberman et al. (2007) studied the regularization of English verbs over the last 1200 years and how the rate of regularization depends on the frequency of word usage. Lexicostatistics was used to calculate the evolutionary history of a set of related languages and varieties (Bakker et al. 2009, Barbancon et al. 2013). Baker (2011) focused on words that have changed their frequency and meaning in the study of change in British English over the twentieth century. Degaetano-Ortlieb and Teich (2018) have used relative entropy for detection and analysis of periods of diachronic linguistic change. Different from these previous studies, Hou et al. (2020) studied the language change in Chinese political discourse by looking at the relationship between clause and word, which showed that the fitted parameters can differentiate the different periods Chinese political discourses.

As one of the best-known laws of quantitative linguistics, the MA law establishes the interrelations between successive hierarchical levels of language, providing evidence that language is a self-organizing and self-regulating system. Previous research has validated the MA law at different language levels. For example, Köhler (1982) conducted the first empirical test of the MA law at the sentence level as far as we know, analyzing German and English short stories and philosophical texts. Teupenhayn & Altmann (1984) studied the relationship between the length of sentences and its clauses measured in words, which showed that their empirical data meets Menzerath’s law. Buk and Rovenchak (2008) studied the dependence of clause length in terms of words and syllables on the sentence length in terms of clauses in Ukrainian. They propose the formula (2) can fit the clause length counted in words on the sentence counted in clauses. Formula (3) represents one of the generalizations of MA law (Wimmer & Altmann 2005).

$$F(x)=Ax^{b}e^{-c/x} \qquad \text{formula (3)}$$

Motalová et al. (2014) and Ščigulinská and Schusterová (2014) verified the validity of the MA law applied to contemporary written and spoken Chinese respectively. Benešová (2016) tested the potential validity of the MA law on samples in different languages and attempted to test the concept of this language universal. Hou et al. (2017) studied that relationship between sentence and clause in different Chinese registers. And they concluded that the relationship between sentence and clause abide by the Menzerath-Altmann law in written formal Chinese registers, while the relationship did not abide by the Menzerath-Altmann law in informal Chinese register. Xu and He (2020) studied the relationship between English sentence and clause based on the MA law in academic spoken and written registers and showed that the fitted parameters can differentiate these two registers. Jiang and Ma (2020) explored the relationship between sentence and clause based on the MA law and showed that this law held true for both original and translated texts, and the fitted parameters could differentiate the translational language from the original language.

Except for the application of the MA law to different language levels, there are some researches studying the theory and formula per se. Köhler (1984) interpreted the parameters by using Formula (2). He assumed that *a* represents a quantity depending on the language and language levels, *b* might represent a shortening tendency and might describe the range of structural information that has to be stored for each language construct.

This paper selected Formula (2) to study the relationship between sentence and clause in different periods Chinese political discourse as like Hou et al. (2017) and examined whether the relationship between sentence and clause in different Chinese political discourses abide by the Menzerath-Altmann law, and, if so, fitted parameters can differentiate the different periods Chinese political discourses.

1.2 Data and methodology

The Report texts on Works of Government of China in three different periods in last 40 years were selected to establish the corpus. These three different periods are 1978-1982, 1997-2001 and 2016-2020 respectively. The first five years, 1978-1982, are the initial stage of the reforming and opening up in China. Hong Kong was returned to China in 1997. The last five years, 2016-2020, is the 13th five-year plan was initiated and finished. We chose three periods with a difference of 20 years.

The texts of Report were segmented and Parts of Speech tagged using the Chinese Lexical Analysis System created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS).

This paper examined the relationship between sentence and clause in Chinese political discourse. According to the Menzerath-Altmann law, the average clause length decreases with the increases of the sentence length. Thus, the first question should be resolved is that the definitions of Chinese sentence and clause. Then the sentence length and clause length should be defined secondly.

The sentence is considered to be a basic linguistic unit in all languages. Different from Indo-European languages, a sentence in Chinese texts is not easily delineated for lack of a reliable convention to mark end-of-sentence, and because of frequent omission of sentential components, such as subjects and predicates (Huang & Shi 2016). Consequently, Chinese sentences are often defined in terms of characteristics of speech, rather than texts (Lu 1993; Huang & Shi 2016). Chao (1968) and Zhu (1982) defined a sentence as an utterance on pauses and intonation changes at the boundaries of sentences.

Considering the complexity of Chinese sentence and the aim of this paper, we mark the Chinese texts separated by periods, question marks and exclamation marks as sentences. Generally speaking, the length of language construct should be counted as its immediate component(s). We assumed that the immediate component of sentence is clause. The sentence length is measured in the number of constituents, clauses. In Köhler (1982), the number of clauses is determined by counting the number of finite verbs in a sentence. Different from that the number of Chinese clauses is determined by the number of commas or semicolons in a sentence. The length of clause is measured as the number of the words included. Words is considered to be the segments delineated by blank spaces in the texts segmented by a Chinese lexical analysis system.

We used the Formula (2) to fit the relationship between sentence and clause (i.e., average clause length distribution in the sentence with certain lengths. From Formula (2), there should be a nonlinear relationship between sentence length and clause length. The initial estimates of parameters, *a* and *b*, will influence the nonlinear regression result. We transformed the nonlinear relationship in Formula (2) into linear relationship in order to avoid the influencing of initial estimates of parameters, as following.

$$y = ax^{b}$$

log(y)=b*log(x)+log(a) taking the logarithm on both the sides

Y=log(y), X=log(x)

Then: Y=b*X +log(a) Formula (1a-1)

The linear regression can be used to fit the link between logarithm of average clause length in terms of words and the logarithm of sentence length in terms of clauses. Multiple R-squared was selected to validate the fitted results. It indicates how much changes in the data can be explained by the model (Conway & White, 2013 P150). The open source programming language and environment R was used to realize the fitting procedure for regression analysis.

If the relationship between sentence and clause in these three periods Chinese political discourse was fitted by Formula (2), the texts can be represented by the fitted parameters, *a* and *b*, and were displayed in the two-dimensional space. Thus, the relationship between different periods Chinese political speeches can visualized and we can determine whether there are differences between them.

2 Frequency distribution of sentence length

The relative occurrences frequencies of sentence with certain length were calculated in three periods Chinese political discourses to establish the distributions of sentence length in terms of clause, as shown in Figure 1.

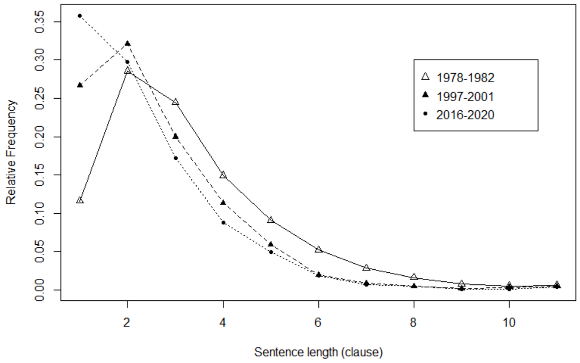

Figure 1: Sentence length distribution (clause) in different periods Chinese political discourse

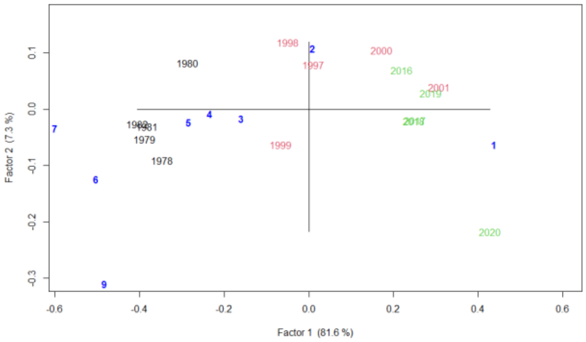

Figure 2: The result of correspondence analysis of different periods political discourses represented by sentence length distribution

From Figure 1, we can see that the relative occurrence frequency of sentence decreases with the increasing of sentence length in third period political discourse. The relative occurrence frequencies of sentence increase firstly and then decrease with sentence length in first and second periods Chinese political discourses. The short sentences occurred frequently and most of sentences in Chinese political discourse are composed of the few clauses. The relative occurrence frequencies of sentence including 1-7 clause are 96.62%, 98.66% and 98.96% from first to third period respectively. This means that we can observe the average clause length distribution in these sentences.

We used the correspondence analysis to analyze the different periods political discourses represented by sentence length distribution. The result of correspondence analysis is shown in Figure 2, from which we can see that the long sentences occurred more frequently in first period compared to the other two periods.

3 Average clause length distribution in sentences

We aim to study the relationship between sentence and clause in different periods Chinese political discourses, so it is unnecessary to determine the length of the individual clause. The average clause length in sentences with certain length was calculated as the number of words in the given sentences divided by the number of clauses included.

After calculating the average clause lengths in sentences with each length, the average clause length distributions in three periods Chinese political discourses were established as shown in Figure 3, from which we can see the link between average clause length and sentence length. From Figure 3, we can see that there are obvious decreasing tendencies of average clause length with the increasing of sentence length in most of sentences in first two periods Chinese political discourses. There is not obvious regular changing tendency of average clause length with sentence length in third period political discourse.

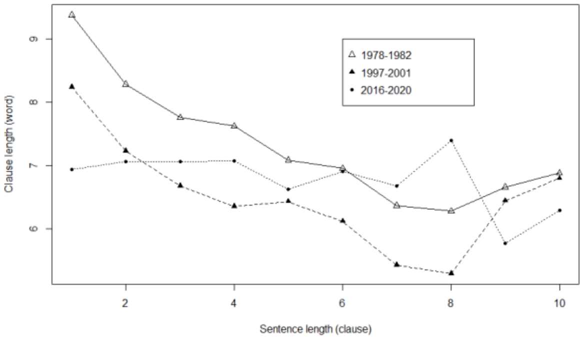

Figure 3：Average clause length distributions in sentences from different periods political discourses

The 15 political discourses from three different periods were represented by the average clause length distributions and hierarchical clustered. The Kullback-Leibler Divergence (Relative Entropy) were adopted to calculate the distance between text vectors and the sum of squares of the deviations was used to calculate the distance between two clusters in text clustering. The clustering result is shown in Figure 4. From Figure 4, we can see that third period political discourses were clustered into one cluster, the other two periods political discourses were clustered into another different cluster. The distance between these two clusters is small. Figure 3 showed that the average clause length distribution in first period political discourse is different from the other two periods. Figure 4 showed that the average clause length distribution in third period political discourse is different from the other two periods. Thus, the average clause length distribution cannot differentiate these three periods political discourse.

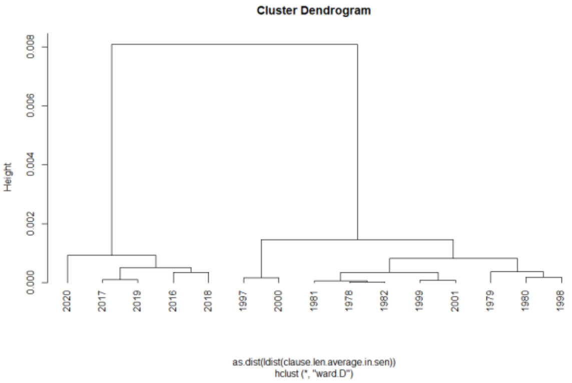

Figure 4：The text clustering result of different periods political discourse represented by average clause length

4 Fitted results of relationships between sentence and clause

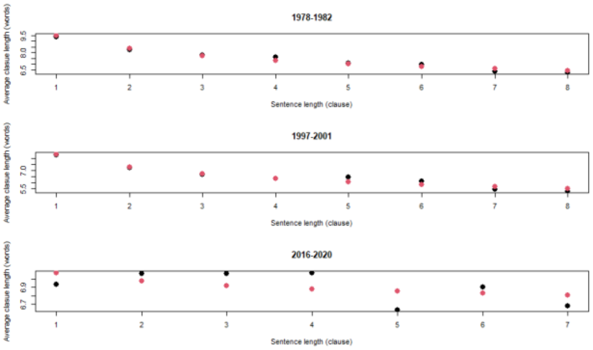

Figure 5：The fitted result of average clause lengths in different periods political discourses

The average clause length distributions, as shown in Figure 4, were fitted using the Formula (3). The fitted results are shown in Figure 5 and Table 1 respectively. From the determination coefficients in Table 1, the relationships between sentence and clause abide by the Menzerath-Altmann law in first two periods political discourses and did not in third period political discourses. Figure 5 also showed that the fitted and observed values of average clause length in third period political discourse are similar, which means the fitted result is good

Table 1: The fitted result of average clause lengths in different periods political discourses

	a	*b*	R^2	*Adjusted* R^2
1978-1982	9.498	-0.188	96.63%	96.07%
1997-2001	8.337	-0.198	93.50%	92.42%
2016-2020	7.075	-0.020	25.71%	10.85%

Table 2: The fitted result of average clause length distributions in 15 political discourses

	a	*b*	R^2
1978	8.750174	-0.156	98.72%
1979	11.23413	-0.200	92.99%
1980	8.502068	-0.198	84.67%
1981	9.216761	-0.161	94.89%
1982	9.208571	-0.151	98.98%
1997	8.186424	-0.207	80.43%
1998	8.131819	-0.157	81.15%
1999	8.207432	-0.173	88.99%
2000	8.409639	-0.158	83.97%
2001	7.953944	-0.149	95.02%
2016	7.028888	0.008	1.05%
2017	6.987738	0.013	8.21%
2018	7.28553	-0.035	61.55%
2019	7.129599	-0.023	13.63%
2020	6.760901	-0.066	25.03%

The average clause length distributions in 15 Chinese political discourses from three different periods were fitted using Formula (2). The fitted results were shown in Table 2. From Table 2, the average clause length distributions in five political discourses from third period were not fitted by the Formula (2) and the relationship between sentence and clause did not abide by the Menzerath-Altmann law. This conclusion is consistent with the fitted results in Table 1.

The 10 Chinese political discourses from first and second periods were represented by the fitted parameters, *a* and *b*, of average clause length distributions and were visualized in a 2-dimensional space as shown in Figure 6. From Figure 6, we can see there is a boundary between 10 Chinese political discourses from first and second periods. The values of *a* can differentiate these two periods political discourses and they decreases from first to second period. There is not the obvious difference of *b* between these two periods.

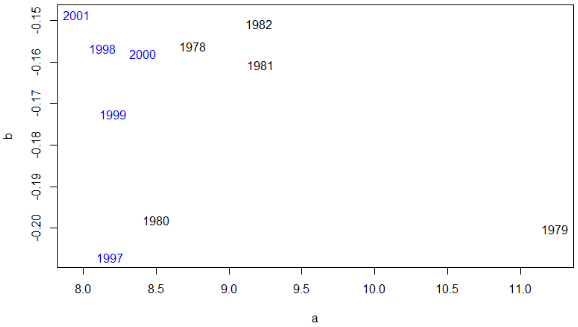

Figure 6: The relative position of 10 Chinese political discourses from first and second period

5 Conclusion

This paper examined the language change in Chinese political discourse based on the Menzerath-Altmann law. Classical linguistics mainly concerns the diachronic changes of the consecutive language elements, for example sound, with time. Different from that, this paper considered language as a complex system and explored the change of this system from time to time. The relationship between sentence and clause in three different periods Chinese discourses were fitted using Formula (2). The fitted result showed that the relationship between sentence and clause in third period political discourse did not abide by the Menzerath-Altmann law. Hou et al. (2017) showed that the relationship between sentence and clause does not abide by the Menzerath-Altmann law in informal Chinese register. Thus, it is possible to consider the political discourse in third period as informal register. The Chinese political discourse is changing to more colloquial and the link between clauses in sentence is becoming weak. In addition, the sentence length decreases from first period to third period with time.

The fitted result also showed that the relationship in first two periods political discourses abide by the Menzerath-Altmann law. The fitted parameters, *a* and *b*, can differentiate these two periods Chinese political discourses. The values of *a* in first period are larger than that in second period Chinese political discourses. This means the average clause length in the sentence with the same length in second period is smaller than that in first period. There is not obvious difference of *b* values between these two periods. The average values of *b* in first period are smaller than that in second period. The exact differences of *b* values between different periods and when political discourse is changing to colloquial should be further examined.

From the experiments, we concluded that we can explore the language change from the systematic perspective. The Menzerath-Altmann law studied the relationship between language construct and its immediate components and considered the language as a system. Thus, it is feasible to explore the language change based on the Menzerath-Altmann law.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful and helpful comments.

Funding support. Research on this paper was funded by National Social Science Fund in China (Grant No. 16BYY110), the Hong Kong Polytechnic University Grant 4-ZZFE, the Ministry of education of Humanities and Social Science project of China (Grant No. 21YJA630031).

References

Altmann, G. (1980). Prolegomena to Menzerath´s law. Glottometrika 2, 1-10.

Bakker, D., Muller, A., Velupillai, V., Wichmann, S., Brown, C.H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. and Holman, E.W. (2009). Adding typology to lexicostatistics: A combined approach to language classification. Linguistic Typology 13(1), 169–181

Baker, P. (2011). Times may change, but we will always have money: diachronic variation in recent British English. Journal of English Linguistics, 39(1), 65–88.

Barbançon, F., Evans, S., Nakhleh, L., Ringe, D. and Warnow, T. (2013). An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica 30, 143–170.

Benešová, M. (2016). Text segmentation for Menzerath-Altmann law testing. Palacký University, Faculty of Arts.

Biber, D. (2012). Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory, 8(1), 9-37.

Buk, S., & Rovenchak, A. (2008). Menzerath–Altmann law for syntactic structures in Ukrainian. Glottotheory, 1(1), 10-17.

Cacoullos, R. T. (1999). Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. Language variation and change, 11(2), 143-170.

Chao, Y. R.. (1968). A Grammar of Spoken Chinese. Berkeley and Los Angeles: University of California Press.

Conway, D., & White, J. (2012). Machine learning for hackers. Translated by Chen, K., Liu Y. & Meng, X. China Machine Press. Beijing. China.

Degaetano-Ortlieb, S. and Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 22–33.

Grzybek, P. (2007). Do we have problems with Arens' law? A new look at the sentence-word relation. In P. Grzybek and E. Stadlober. Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of His 75th Birthday, 62, 205.

Hou, R, Chu-Ren Huang, Hue San Do & Hongchao Liu (2017): A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law, Journal of Quantitative Linguistics. 24(4): 350-366.

Hou, R., C.-R. Huang & K. Ahrens. (2020). Language change in Report on the Work of the Government by Premiers of the People's Republic of China. In Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation. Pp.100-111.

Huang, Chu-Ren and Dingxu Shi. 2016. A Reference Grammar of Chinese. Cambridge : Cambridge University Press.

Jiang, Y. & R. Ma. (2020). Does Menzerath-Altmann Law Hold True for Translated Language: Evidence from Translated English Literary Texts. Journal of Quantitative Linguistics. Doi: 10.1080/09296174.2020.1766335.

Köhler, R. (1982). DasMenzerathsche Gesetz auf Satzebene. In W. Lehfeldt & U. Straus (Eds.), Glottometrika 4 (pp. 103 – 113). Bochum: Brockmeyer.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In W. Lehfeldt & U. Straus (Eds.), Glottometrika 6, 177-183. Bochum: Brockmeyer.

Köhler, R. (2012). Quantitative syntax analysis (Vol. 65). Berlin: Walter de Gruyter.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. Nature 449(7163): 713-716.

Lu J. (1993). The features of Chinese sentences. Chinese Language Learning. No.1, 1-6.

Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes (Vol. 3). F. Dümmler.

Motalová, T., Spáčilová, L., Benešová, B., Kučera, O. (2014). An application of Menzerath-Altmann law to contemporary written Chinese. Křížkovského, Olomouc: Univerzita Palackého v Olomouci.

Teupenhayn, R., & Altmann, G. (1984). Clause length and Menzerath's law. Glottometrika, 6, 127-138.

Wang W. S.-Y. 王士元. (2006). Language is a complex adaptive system 语言是一个复杂适应系统. Journal of Tsinghua University (Philosophy and Social Science). 21(6):5-13.

Wimmer, G. & Altmann, G. (2005). Unified derivation of some linguistic laws. In R. Köhler, G. Altmann & R. G. Priotrowski (Eds.), Quantitative Linguistics (pp. 791-870. Berlin – New York: de Gruyter.

Xu, L. & L. He. (2020). Is the Menzerath-Altmann Law Specific to Certain Language in Certain Registers? Journal of Quantitative Linguistics. 27:3, 187-203. Doi: 10.1080/09296174.2018.1532158.

Zhu D. (1982). Grammar handouts. Commercial Press