A quantitative investigation of English adnominal modifiers

Tsy Yih[1]
Zhejiang University
Department of Linguistics
yezi_leafy@hotmail.com

Haitao Liu✉
Zhejiang University
Department of Linguistics
lhtzju@gmail.com

Abstract

The present study employs the valency theory to approach the internal structure of noun phrases. Specifically, we adopt The Georgetown University Multilayer Corpus (GUM) to investigate the valency distribution, mean valency, and valency pattern of three subclasses of nominal heads, commons nouns, proper nouns, and pronouns. It is found that: 1) The mean valency of common nouns is significantly larger than that of proper nouns, which is far larger than that of pronouns. 2) The Mean Valency Pattern (MVP) newly proposed in this paper offers a better comparative basis for the valency distribution of different categories than the original Probabilistic Valency Pattern (PVP). 3) The three nominal classes manifest different valency patterns. Common nouns have a strong combining ability with determiners, adjectival modifiers, prepositions phrases and nominal modifiers. The combining ability of pronouns is near zero for most groups except for a few cases in certain constructions, whereas the mean valency of proper nouns are not close to zero as expected due to the improper treatment of complex proper names in annotation. The findings are expected to pave the way for further quantitative studies into the frequency structure of noun phrases.

1 Introduction

The noun is one of the most important lexical category in human language, which accounts for about 37% of word-tokens in texts (Hudson, 1994). The internal structure of English noun phrases has been a long and constant topic in theoretical linguistics, which has received much attention in the literature (Quirk et al., 1985; Rijkhoff, 2002; Keizer, 2007; Davidse and Breban, 2019). A popular theory is that the noun phrase has a layered structure (Rijkhoff, 2002) or an orbit structure (Lu, 1993). Seen from the coarsest perspective, English adnominal modifiers could be divided into prenominal lexical categories such as determiners, quantifiers, adjectives, nominal modifiers, as well as postnominal phrasal and clausal modifiers, including prepositional phrases and relative clauses (Quirk et al., 1985; Halliday, 1985). Some have studied the finer internal structure of the category, adjective (Halliday, 1985; Bache, 2000; Davidse and Breban, 2019), and even further splits might be employed to the inside world of descriptive adjectives (Dixon, 1977). Besides their own theoretical value, fine and elaborate classifications would also provide good basis for developing high-quality language resources for natural language processing, thereby being fundamental to computational linguistics.

So far, all the above-mentioned studies investigate the theoretical structure of noun phrases, with special attention paid to the preferred order when multiple modifiers occurred. For instance, we have roughly presented the above-mentioned modifiers in the order with regard to their natural closeness to the head noun (demonstrative < quantifier < adjective <nominal modifier < head noun). However, there is a lack of quantitative studies on how often these modifiers occur.

Since the adnominal modifiers are considered dependents of head nouns in the context of

[1] Tsy Yih is the transliteration of the name of the firth author in his mother tongue, Shanghai Wu Chinese. He is also known as ZI YE in Mandarin pinyin.

dependency grammar, it is possible to employ the dependency grammar and valency theory to approach this issue. There have been many previous quantitative studies on valency of verbs in English (Yan and Liu, 2021), Chinese (Liu, 2009b; Gao et al., 2014), Czech (Čech et al., 2010), Hungarian (Vincze, 2014). However, there are few studies on nominal heads. Pan and Liu (2014), which is among the few such studies to our knowledge, have conducted a quantitative survey on the distribution of adnominals in Mandarin Chinese. They made a 43-way distinction of modifiers, and found that the frequency distributions of most texts follow a Zipf's function. Nevertheless, they focus on fitting data to some mathematical functions, rather than on the linguistic functions of modifiers and explain why they have the frequency structure the way they look like .

Therefore, in the present study, we employ an English treebank with dependency annotation to investigate the frequency structure of noun phrases. Note that most of the above-mentioned studies focus on the noun phrase with a common noun as its head, while there are also two other subclasses of nominal heads, that is, proper nouns and pronouns. They are generally considered to have the ability to stand alone as noun phrases per se without modification, and this study will test if it is the case.

Two research questions to be answered are as follows:

RQ1: What are the valency distributions and mean valencies of three subclasses of nominal heads? Are there significant differences among the three groups?

RQ2: What are the valency patterns of the three subclasses of nominal heads?

2 Methods

2.1 Theoretical framework

This section elaborates on the theoretical framework, dependency grammar and valency theory, in which this study is couched.

Dependency grammar (DG) is a syntactic theory based on the idea that sentences are composed of words and their part-part relations (Tesnière, 1959; Hudson, 1984; Mel'čuk, 1988; Liu, 2009a). Today it has constituted a large family of concrete formalisms and implementations, as opposed to the phrase structure grammar (PSG), which pays more attention to phrases and part-whole relations of linguistic units. DG is now more widely used in the field of computational linguistics. There is both a popular storage format .conllu and a cross-linguistic annotation scheme called Universal Dependencies (UD), which came out of a shared task (Nivre et al., 2016). 202 treebanks of 114 languages are offered in UD till version 2.8, and there are continuing endeavors to transform other existent corpora into this format.

The canonical version of DG is generally recognized to be established by Tesnière (1959), although its root might be traced back to earlier times. Within Tesnière's theory, the notion of valency plays a large part. Tesnière offers a metaphor, conceptualizing a verb as an atom in chemistry, and valency as the number of bonds. Valency, therefore, reflects the combining ability a verb has to attract arguments, or actants in his terms. Till this day, the notion of valency has undergone several changes and extensions, while it could find similar counterparts in almost every syntactic-semantic theory. The most prototypical definition of valency originally refers to the potential number of nominal complements governed by a verb. For instance, the English verbs *rain*, *cry*, *eat*, and *give* in their most common use are respectively called avalent, monovalent, divalent, and trivalent verbs in traditional terms. In other words, their valencies are zero, one, two, and three. However, there are at least six restrictions concerning this definition, and these restrictions could be removed gradually, leading to various extended definitions.

(1) 'Potential' means it is a static property of specific words, which can be found in the dictionary (Gao and Liu, 2018; 2020). When realized as dependency relations in authentic texts, the real number could be less than the potential valency. For example, *eat* in *he has eaten* has a syntactic running valency of one, although semantically there should at least be two arguments or semantic roles, EATER and EATEN. Therefore the running and static valency are alternatively called syntactic valency and semantic valency in the literature.

(2) 'Complements' refer to core, obligatory arguments rather than peripheral, optional adjuncts. For instance, *he* is a complement while *room* is not in the sentence *he sings in the room*. There have been extensions to dispel the indeterminacy between complements and adjuncts, such as generalized dependency (Liu, 2006; 2009b) or full valency (Čech et al., 2010).

(3) The original idea focuses on the nominal dependents. Despite the extensions in (2), in a

conservative form they still stick to nominal arguments rather than those tense-aspect markers or particles, which may also be annotated as dependents of the head in the treebank. Further extensions also break the latter restriction and take into account all kinds of dependents in an abstract tree, regardless of the linguistic categorization of dependency relations. This view is reflected in Yan and Liu (2021).

(4) The original idea generally considers verbs as heads. One might extend it to other lexical categories such as nouns and adjectives. However, a first-step extension only stays at those possessing nominal dependents, such as *a picture of John* or *be angry at him*. In the present study, we extend it further to consider other kinds of dependents of nouns. However, we still hold a somehow conservative view that only seven types of adnominal modifiers in UD are taken into account, rather than all the dependents of a nominal head.

(5) The original idea only concerns the number of dependents rather than its governor. The extension with regard to this aspect is reflected in Probabilistic Valency Pattern (Liu and Feng, 2007; Liu, 2009a) and the definition of dynamic valency by Lu et al. (2018), which also take into consideration the input valency. The output valency (number of dependents) and input valency (number of governers) are respectively called active and passive valency by Liu (2009b).

(6) Finally, the original idea of valency is associated with a word, while Liu (2006, 2009b, 2009b), Liu and Feng (2007), Yan and Liu (2021) have extended it to a lexical category. The present study also explores on the word-class level.

Therefore, so far the widest definition of valency equals the number of all adjacent nodes of a certain node in a syntactic tree, which is exactly Gao et al. (2014)'s dynamic valency. It equals the notion of 'degree' in graph theory. In other words, it could be operated even abstracted away from linguistic contexts.

In the present study, we investigate running, active valency of nominal heads but also restrict the scope of adnominal modifiers to certain types. Since in a real corpus, there might be errors and complex situations which are not true syntactic dependents, such as *cc* (conjunction subordinators), we hence only focus on seven types of modifiers based on the UD annotation scheme, i.e., *det* (determiners), *amod* (adjectival modifiers), *compound* (nominal modifiers), *nmod* (prepositional phrase modifiers), *nummod* (numeral modifiers), *acl* (reduced relative clauses), and *acl:relcl* (finite relative clauses).

An example noun phrase with all seven types realized simultaneously is presented in Figure 1. In real texts, such example hardly exists. Note that certain dependency relations might be realized more than once, like the two *amod* relations in the examples.

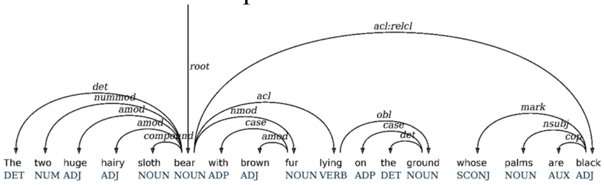
acl:relcl
root
det
nummod
amod
amod
compound
acl
nmod
case
amod
obl
case
det
mark
nsubj
cop
The two huge hairy sloth bear with brown fur lying on the ground whose palms are black
DET NUM ADJ ADJ NOUN NOUN ADP ADJ NOUN VERB ADP DET NOUN SCONJ NOUN AUX ADJ

Figure 1. The UD annotation of an example sentence

2.2 Treebank

A treebank is a corpus with syntactic annotation (Abeillé, 2003), which would facilitate the current study. The dependency treebank we use in the present study is the UD version of The Georgetown University Multilayer Corpus (GUM). The reason we employ the UD version rather than other syntax-oriented frameworks such as Surface-oriented Universal Dependencies (SUD) is that UD regards content words as head. In that case, it is easier for us to identify all the adnominal modifiers in a noun phrase. Moreover, as stated in Yan & Liu (forthcoming), the annotation scheme of UD is more delicate and specific which might be more suitable for detailed investigation based on dependency relations.

GUM is designed to be a balanced corpus including both spoken and written data with a totality of 7,409 sentences and 116,748 orthographical words (excluding punctuations). Table 1 shows its genre composition.

Table 1 The genre composition of GUM corpus

Genres	Sentences	Tokens
Academic Writings	575	12,970
Biographies	776	15,297
Conversations	684	4,555
Fictions	1,029	13,855
Interviews	1,070	15,883
News	645	12,456
Speeches	202	4,375
Textbooks	270	4,732
Vlogs	292	4,653
Travel Guides	771	13,159
How-to guides	1,095	14,813
Sum	7,409	116,748

2.3 Methods and measures

To answer the first research question, one measure employed in the present study is Mean Valency (MV). It is a property with regards to a word class. The formula of *MV* is given in (1), following Yan and Liu (2021):

$$MV = \frac{1}{N}\sum_{i=1}^{N} Val_i \tag{1}$$

where *N* is the size of the category, i.e., the common noun, proper noun and pronoun in the present study, and Val_i is the *i*-th item's valency.

As for the second question, the method applied in the present study is Mean Valency Pattern. It is a variant based on the Probabilistic Valency Pattern (PVP), proposed by Liu and Feng (2007) and Liu (2009a).

The original PVP describes the probability distribution of dependencies relations between the dependents and their heads within a dependency syntactic tree. The probability of the dependency type between the head and its *i*-th dependent is represented as w_i, which is computed as in (2):

$$w_i = \frac{f_i}{\sum_{i=1}^{n} f_i} \tag{2}$$

where f_i is the frequency of certain dependency types, and *n* is the number of all dependency types. In general, the pattern is given in a descending order with respect to w_i. Apparently, the sum of all such probabilities equals 1. For a certain node, there are either dependencies governing that node or governed by it (Liu and Feng, 2007), while we only focus on the latter ones here.

However, PVP has a problem that it cannot be compared across different head classes. Therefore, we put forwards the Mean Valency Pattern (MVP), which is composed of measures called the frequency of dependencies per head (FDPH) as defined below in (3):

$$FDPH_i = \frac{f_i}{\sum_{i=1}^{N} f_i} \tag{3}$$

where f_i is the frequency of certain dependency types, and *N* is the size of the category.

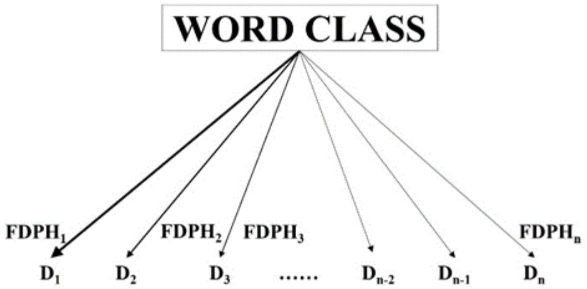

Figure 2. Graphic representation of MVP

Figure 2 demonstrates the graphic representation of the mean valency pattern of a certain word class. The linguistic significance of FDPH lies in that it reflects how many dependents will appear on average given each occurrence of a head of certain category. In addition, it could also be used to measure the degree of obligatoriness for certain dependency relations or types of dependents. There are several points worth mentioning. Note first that the sum of FDPHs does not amount to 1. Therefore MVP is not a probability distribution. Yet all the FDPHs add up to the MV of that category. Second, the sum of FDPH, or even each FDPH could exceed the number of one in theory. This means that for each occurrence of the head, there might appear more than one dependent. The reason is that there could be more than one repeatable dependency relation such as *amod* (adjectival modifiers). Third, as can be seen from (2) and (3), the numerators are the same in two formulae. Hence the internal pattern of PVP, or the proportion of different dependency relations is kept in MVP. Yet MVP prevails in that it offers the comparative ability across categories.

3 Results and discussion

3.1 Valency Distribution of Modifiers for three types of nominal heads

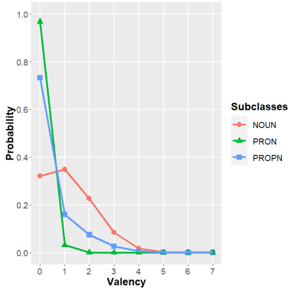

Figure 3. The valency distribution of three nominal head in the GUM English treebank

Figure 3 shows the valency distribution of common nouns (NOUN), proper nouns (PROPN) and pronouns (PRON). It is indicated that for common nouns, the most frequent valency is 1, whose frequency is a little higher than that of val = 0, while for proper nouns and pronouns, the highest frequency occurs at val = 0. More specifically, pronouns are almost completely avalent (having zero modifiers), whereas some of the proper nouns are monovalent, divalent, or trivalent (i.e., having one, two and three modifiers respectively).

As for the mean valency of each nominal head, the results are demonstrated in Table 2.

Nominal heads	Mean Valency (MV)
common nouns	1.136
proper nouns	0.417
pronouns	0.033

Table 2 The mean valencies of three nominal heads

Note that the mean valency of nouns in English is smaller than that of verbs, which is 1.801, compared with the finding by Yan and Liu (2021).

The one-tailed Wilcoxon test reveals that the valency median of common nouns is significantly higher than that of proper nouns (W = 152767815, p = 2.2e-16 < 0.05), while the valency median of proper nouns is then significantly higher than that of pronouns (W = 56680962, p = 2.2e-16 < 0.05). The results indicate commons nouns absorb the most modifiers, and next follows the proper nouns. The first part of the result can be explained by the finding in theoretical linguistics that determiners are generally taken to be an obligatory category. Common noun denoting types of objects have to go through a number of processes to be realized into noun phrases in real texts (Langacker, 1987; Croft, 1990). Yet why proper nouns have a certain mean value valency needs explanation. We hypothesize that both proper nouns and pronouns are supposed to be roughly avalent since they can stand alone as noun phrases. In theory, proper nouns can only be non-restrictively modified but these cases are expected to be rare in texts. What these types of modifiers are will be revealed in the next section.

3.2 The valency patterns of the three subclasses of nominal heads in English

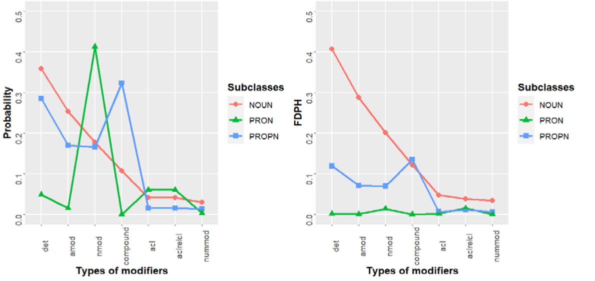

Figure 4. The distribution of PVP and MVP of three subclasses of nominal heads (Left panel: original PVP; right panel: FDPH)

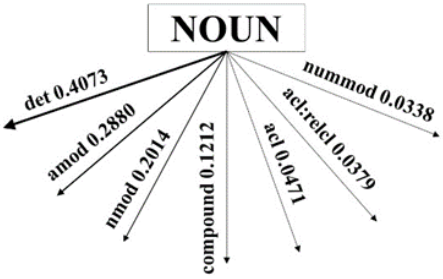

Figure 5. The PVP diagram of nouns

The left panel of Figure 4 demonstrates the original PVP distribution, whereas the right panel shows the MVP distribution. Figure 5 presents the MVP diagram of nouns, which can be transformed into the red line in the right panel of Figure 4. If we look at the left panel, at first sight, the proportion of compounds of proper nouns are much higher than that of common nouns. In addition, the number of the first three categories in two groups are close as well. Similarly, the proportion of *nmod* of pronouns also seems to be high, and its *acl* and *acl:relcl* relations are more frequent than those of common nouns. However, they are indeed not comparable since the sizes of those categories are different. Rather, the right panel offers a better comparative basis. It is shown that proper nouns do attract more compounds than common nouns, but it only exceeds a little. Moreover, the first three groups are much lower. This means that common nouns attract more determiners, adjectival modifiers and prepositional phrases than the other two subclasses.

Now let us look at the modifiers from the one with the highest frequencies to that with the lowest. It is not surprising that *det* (determiners) has the highest value since it is reported to be an obligatory modifier in English noun phrases. However, we should ask, on the contrary, why it only possesses a proportion of less than 50%. In other words, why is it not truly obligatory as we see from the data? One reason could be that of bare plurals. For plural nouns with specific references, either articles or other determiners are not necessary. Other cases include the generic reference. It is generally acknowledged that English has at least three, though neither dedicated, formal strategies to encode a concept of type, i.e., indefinite singular, definite singular and bare plurals (Du Bois, 1980). The existence of these situations contribute to the pattern we see now. The next largest group is *amod* (adjectival modifiers). This might be due to the fact that the English adjective is a mixed category to a large extent. There are several marginal subclasses such as identifying adjectives, associative adjectives, noun-intensifiers, focus markers, and metadesignatives (Davidse and Breban, 2019), or maybe even a separate class of evaluative adjectives. These subcategories might be classified into other lexical categories in other languages, but all fall within the scope of adjective in English grammar, which leads to the expansion of the *amod* group in English. The difference in the next two groups might be due to their scope of functions. The third relation, *nmod* (prepositional phrases), either plays the same role as so-called associative adjectives and nominal modifiers, or expresses the relation of the head noun to another noun with the individual referent, while the function of the fourth relation, *compound* (nominal modifiers), is more restricted, expressing only the subcategorization of types in general. Finally, the dependency type *nummod* (numeral modifiers) in UD partly reflects the function category of quantifiers to some extent. It is held by some linguists that quantifiers have a floating property, rather than lying at a certain layer (Halliday, 1985). The reason why there is only a small number of *nummod* might be that certain quantifiers do not fall under the category of the numeral in UD. Only cardinal numbers fall within this group. Yet, ordinals and fuzzy quantifiers like *many* are adjectives, while *every* is labelled determiner.

After explaining the possible functional causes of common nouns' valency pattern, the next question is why the valency of proper nouns is not close to zero. A closer look at the data reveals that a large number of examples are such as *the United Kingdom* or *University of Cambridge*. Most modifiers are indeed the internal components of a complex proper noun. This problem should be attributed to the annotation of UD as dependency treebanks are generally word-based. It is left for future improvement for the ease of better functional analysis.

Finally, as for pronouns, for most of the cases the frequencies are indeed close to zero. The only existent examples are prepositional phrases and relative clauses. Examples include *those who …*. Pronouns can be modified, though unexpectedly, in the cases where they are sometimes so less informative that one has to add an obligatory restrictive modifier to supplement information (Hopper and Thompson, 1984).

In sum, this section provides functional explanations for the valency patterns of three nominal heads. The results show that the calculation of FDPH and MVP does reflect the combining power of heads to certain dependency relations as expected.

4 Conclusion

The present study will serve as a first attempt to the frequency structure of noun phrases. We found that: Firstly, the mean valency of common nouns is significantly larger than that of proper nouns, which is further larger than that of pronouns. Second, the Mean Valency Pattern (MVP) newly proposed in this paper offers a better comparative basis for the valency distribution of different categories than the original Probabilistic Valency Pattern (PVP). Third, the three nominal classes manifest different valency patterns. Common nouns have a strong combining ability with determiners, adjectival modifiers, prepositions phrases and nominal modifiers, which helps to turn a nominal head denoting coarse types of objects into a noun phrase denoting individuals from a more specific type. Yet for proper nouns, the reason for their valencies not to be zero as expected is that most of the adnominals reflect the internal structure of complex proper names rather than external restrictive modifiers. As for the pronoun, its combining ability is near zero for most groups, except for a few cases of relative clauses in certain constructions to supplement information.

This study, however, does not take into account the genre differences. It is worth investigating whether the above-mentioned properties hold across various genres. In the next step, one might continue to investigate the effect of size, genre, and language on the valency distribution. A more concrete functional annotation and cross-linguistic identification are also called for.

Acknowledgments

We thank Jianwei Yan for discussing the ideas and helping us plotting the graphs nicely.

References

Abeillé, A. (Ed.). (2003). *Treebanks: Building and using parsed corpora* (Vol. 20). Springer Science & Business Media.

Bache, C. (2000). *Essentials of mastering English*. De Gruyter Mouton.

Davidse, K., & Breban, T. (2019). A cognitive-functional approach to the order of adjectives in the English noun phrase. *Linguistics*, *57*(2), 327-371.

Dixon, R. M. (1977). Where have all the adjectives gone?. *Studies in Language*, *1*(1), 19-80.

Čech, R., Pajas, P., & Mačutek, J. (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of quantitative linguistics*, *17*(4), 291-302.

Croft, W. (1990). A conceptual framework for grammatical categories (or: a taxonomy of Propositional Acts). *Journal of Semantics*, *7*(3), 245-279.

Du Bois, J. W. (1980). Beyond definiteness: The trace of identity in discourse. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*, 3, 203-274.

Gao, J., & Liu, H. (2019). Valency and English learners' thesauri. *International Journal of Lexicography*, *32*(3), 326-361.

Gao, J., & Liu, H. (2020). Valency Dictionaries and Chinese Vocabulary Acquisition for Foreign Learners. *Lexikos*, *30*, 1-32.

Gao, S., Zhang, H., & Liu, H. (2014). Synergetic properties of Chinese verb valency. *Journal of Quantitative Linguistics*, *21*(1), 1-21.

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. London: Edward.

Hopper, P. J., & Thompson, S. A. (1984). The discourse basis for lexical categories in universal grammar. *Language*, *60*(4), 703-752.

Hudson, R. A. (1984). Word grammar. Oxford: Blackwell.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language*, *70*(2), 331-339.

Keizer, E. (2007). *The English noun phrase: The nature of linguistic categorization*. Cambridge University Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: descriptive application. (Vol. 2)*. Stanford university press.

Liu, H. (2006). Syntactic parsing based on dependency relations. *Grkg/Humankybernetik*, *47*(3), 124-135.

Liu, H. (2009a). *Dependency grammar: From theory to practice*. Science Press.

Liu, H. (2009b). Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics*, *16*(3), 256-273.

Liu, H., & Feng, Z. (2007). Probabilistic Valency Pattern Theory for Natural Language Processing. *Linguistic Sciences*, *6*(3), 32–41.

Lu, B. (1993). *The head-oriented grammar*. Shanghai Educational Publishing House.

Lu, Q., Lin, Y., & Liu, H. (2018). Dynamic Valency and Dependency Distance. *Quantitative Analysis of Dependency Structures*, 72, 145.

Mel'cuk, I. A. (1988). *Dependency syntax: theory and practice*. SUNY press.

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Zeman, D. (2016, May). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666).

Pan, X., & Liu, H. (2014). Adnominal Constructions in Modern Chinese and their Distribution Properties. *Glottometrics*, *29*, 1-30.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive English grammar*. London and New York: Longman.

Rijkhoff, J. (2002). *The noun phrase*. Oxford University Press.

Tesnière, L. (1959). *Eléments de la syntaxe structurale*. Paris: Klincksieck.

Vincze, V. (2014). Valency frames in a Hungarian corpus. *Journal of quantitative linguistics*, *21*(2), 153-176.

Yan, J. & Liu, H. (2021). Quantitative Analysis of Chinese and English Verb Valencies Based on the Probabilistic Valency Pattern Theory. *CLSW2021*.

Yan, J. & Liu, H. (forthcoming) Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures. *Studia Linguistica*.