Acta Linguistica Petropolitana. 2021. Vol. 17.1. P. 458–477
DOI 10.30842/alp23065737171458477

Form-frequency correspondence in adjectives: A cross-linguistic corpus approach

Jingting Ye
Fudan University (Shanghai, China); Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany); jingting_ye@eva.mpg.de

Abstract. The adjective has always been a puzzle despite the long-standing discussion in the previous literature, e.g., [Chomsky 1970; Dixon 1982]. Cross-linguistically, a substantial variation can be observed regarding the syntactic behavior of adjectives. Adjectives are more noun-like in some languages, while more verb-like in other languages [Wetzer 1992, 1996]. In some languages, adjectives are marked by a copula when used as predicates, while in other languages adjectives are used as predicates without any further marking. Likewise, adjectives behave differently across languages when used as modifiers.

The aim of this study is twofold. The first aim is to explain the cross-linguistic coding pattern of adjectives with reference to the form-frequency correspondence hypothesis [Zipf 1935; Haspelmath 2008; Haspelmath et al. 2014; Haspelmath 2021]. The second aim is to test the hypothesis, using cross-linguistic corpus data from the Universal Dependencies Corpora [Nivre et al. 2017] and the BCC Mandarin Corpus [Xun et al. 2016].

According to the form-frequency correspondence hypothesis, the more frequent forms are less likely to be marked with extra markers. Within the realm of adjectives, the effect of the form-frequency correspondence hypothesis can be understood as follows. Firstly, the relative frequency of the attributive use of adjectives correlates negatively with the probability of their co-occurrence with a relativizer; secondly, the relative frequency of the predicative use of adjectives correlates negatively with the probability of their co-occurrence with a copula. These hypotheses are tested using the logistic regression model on the basis of an 84-language sample from the Universal Dependencies Corpora, which is a suitable database for the purposes of the present study because it has cross-linguistically consistent annotation for parts of speech and their syntactic contexts. In addition, I have also tested the form-frequency correspondence hypothesis based on the data from the BCC

Mandarin Chinese Corpus based on the frequency of different adjectives. These results have provided positive evidence for the form-frequency correspondence hypothesis.

Keywords: adjective, universal dependencies, frequency, typology, relativizer, copula.

Корреляция между формой и частотностью прилагательных: кросс-лингвистический корпусный анализ

Цзинтин Е

Университет Фудань (Шанхай, КНР); Институт эволюционной антропологии общества Макса Планка (Лейпциг, Германия); jingting_ye@eva.mpg.de

Аннотация. Несмотря на то, что статус прилагательного обсуждается в литературе уже достаточно давно (см., например, [Chomsky 1970; Dixon 1982]), для исследователей эта область до сих пор остается проблемной. Прилагательные характеризуются значительной типологической вариативностью с точки зрения синтаксиса. Если в одних языках они обнаруживают преимущественно именные свойства, то в других они скорее ведут себя, как глаголы [Wetzer 1992, 1996]. Так, в языках первого типа прилагательные, выступающие в качестве предикатов, требуют глагола-связки, в то время как в языках второго типа они выступают в этой позиции без какого-либо дополнительного маркирования. Аналогичным образом прилагательные, выступающие в качестве определений, ведут себя в разных языках по-разному.

В настоящем исследовании решаются две задачи. Первая из них заключается в том, чтобы объяснить формальные свойства прилагательных в разных языках с опорой на гипотезу о наличии корреляции между формой и частотностью [Zipf 1935; Haspelmath 2008; Haspelmath et al. 2014; Haspelmath 2021]. Вторая задача заключается в том, чтобы протестировать эту гипотезу на материале различных языков, используя данные корпусов Universal Dependencies Corpora [Nivre et al. 2017], а также BCC Mandarin Corpus [Xun et al. 2016].

Согласно гипотезе о наличии корреляции между формой и частотностью, более частотные формы с меньшей вероятностью присоединяют дополнительные показатели. В области прилагательных эта гипотеза может быть сформулирована следующим образом. Во-первых, предполагается, что относительная частотность атрибутивного употребления прилагательного находится в отрицательной корреляции с вероятностью маркирования этого прилагательного при помощи релятивизатора; во-вторых, предполагается, что относительная частотность предикативного употребления прилагательного находится в отрицательной корреляции с вероятностью употребления этого прилагательного с глаголом-связкой. Выдвинутые предположения проверяются методом логистической регрессии на материале выборки из 84 языков, входящих в корпуса Universal Dependencies Corpora: эта база данных представляется подходящей для целей настоящего исследования, поскольку она содержит типологически последовательную аннотацию частеречной принадлежности единиц, а также синтактических контекстов, в которых они употребляются. Кроме того, я протестировала рассматриваемую гипотезу на материале корпуса BCC Mandarin Chinese Corpus, опираясь на частотность прилагательных. Результаты исследования подтверждают гипотезу о наличии корреляции между формой и частотностью.

Ключевые слова: прилагательное, универсальные зависимости, частотность, типология, релятивизатор, глагол-связка.

1. Introduction

The adjective has been recognized as a mixed category in the previous literature, e.g., [Chomsky 1970; Wetzer 1996], and there is a great cross-linguistic variation regarding the behavior of adjectives in the attributive and predicative position. Chomsky [1970] argued that the adjective has a mixed feature of [+N] and [+V], which is also reflected in the cross-linguistic coding pattern of adjectives. Wetzer [1992, 1996] identified two types of adjectives: nouny and verby. In some languages, adjectives pattern more like nouns in that they are used with a copula in the predicative position whereas in other languages, adjectives pattern more like verbs because they are used with a relativizer in the attributive position. This phenomenon is illustrated by the examples from English and Mandarin Chinese, as shown in (1)–(2). In English, a copula is used with adjectives in the predicative position but there is no extra marker in the attributive position. On the other hand, Mandarin Chinese has a relativizer in the attributive position but does not use any marker in the predicative position.

English

(1) a. *red flower*

b. *The flower* ***is*** *red.*

Mandarin Chinese

(2) a. *hóng* ***de*** *huā*
red REL flower
'red flower'

b. *Huā hóng*
flower red
'The flower is red'.

This paper attempts to explain this type of coding asymmetry using the form-frequency correspondence hypothesis (FFCH). According to FFCH, more frequent forms are more predictable and, therefore, they are less likely to be marked with an extra marker [Zipf 1935; Greenberg 1966; Haspelmath 2008; Haspelmath et al. 2014]. More specifically, the claim of this article concerning the cross-linguistic coding pattern of adjectives is generalized as follows:

> Cross-linguistically, a higher relative frequency of attributive use correlates with a lower probability of adjectives occurring with a relativizer in the attributive position. Similarly, a higher frequency of predicative use correlates with a lower probability of adjectives occurring with a copula in the predicative position.

This hypothesis will be tested using frequency data based on a sample of 84 languages from the Universal Dependencies Corpora. In the rest of the article, I will first introduce the form-frequency correspondence hypothesis and address the data in *Section 2*. The results are discussed in *Section 3*. Finally, I will draw some conclusions in *Section 4*.

2. Theoretical background and data

2.1. The form-frequency correspondence hypothesis

The form-frequency correspondence hypothesis (FFCH) was first introduced in Zipf's seminal work [Zipf 1935]. In this book, he used statistical evidence to persuasively show that languages demonstrate a negative correlation between the length of the words and their relative frequency. In particular, Zipf claimed that words with higher relative frequency tend to have shorter forms. This insight has proven to be true in many subsequent studies. For instance, the Leipzig Corpora Collection [Quasthoff et al. 2014; Leipzig Corpora Collection 2018] provides copious evidence from more than 100 languages to support Zipf's law.

Zipf's law has been extended to grammatical markers and been used to explain the cross-linguistic coding patterns in many previous typological studies. For example, Greenberg [1966] pointed out that the unmarked features that are coded with shorter forms (or at least with forms of the same length) are more frequent than the corresponding marked features. More recently, Hawkins [2004] also suggested that frequency factor can be used as an important tool to predict the cross-linguistic variation of the complexity of forms. Haspelmath showed that frequency can be used to explain a number of cross-linguistic phenomena concerning coding asymmetries [Haspelmath 2008; Haspelmath et al. 2014; Haspelmath 2021]. I will follow the form-frequency correspondence hypothesis as generalized by Haspelmath:

> "When two grammatical construction types that differ minimally (i.e. that form a semantic opposition) occur with significantly different frequencies, the less frequent construction tends to be overtly coded (or coded with more segments), while the more frequent construction tends to be zero-coded (or coded with fewer segments), if the coding is asymmetric" [Haspelmath 2021: 2].

With respect to the coding patterns of adjectives, the attributive and predicative uses are considered to be a minimal pair. By comparing this minimal pair cross-linguistically, four possible coding patterns can be identified, as shown in *Table 1*. These coding patterns are also illustrated in examples (5)–(8).

Table 1. Four coding patterns in the attributive and predicative use of adjectives

Coding Types	Attributive use	Predicative use	Example Language
Zero coding	unmarked	unmarked	Koyra Chiini
Equipollent coding	marked	marked	Japanese uninflected adjectives
Attributive coding	marked	unmarked	Lango
Predicative coding	unmarked	marked	English, Jarawara

Koyra Chiini (Songhay, Africa)

(3) a. *ni* *beer*
2SG big
'You were big'. [Heath 1999: 73]

b. *har* *beer* *di*
man big DEF
'the big man' [Ibid.: 73]

Japanese (Japonic, Eurasia)

(4) a. *rippa* ***na*** *setubi*
impressive **REL** facilities
'impressive facilities' [Backhouse 2004: 59]

b. *setubi* *rippa* ***da***
facilities impressive **COP**
'Facilities are impressive'. [Ibid.: 57]

Lango (Nilotic, Africa)

(5) a. *rwòt* ***à*** *ràc*
king **REL** bad
'a bad king' [Noonan 1992: 104]

b. *rwòt* *ràc*
king bad
'The king is bad'. [Ibid.: 106]

Jarawara (Arawan, South America)

(6) a. *jifari* *tati*
banana unripe
'an unripe banana' [Dixon 2004: 339]

b. *jifari* *tati* ***amake***
banana unripe **COP**
'The banana is unripe'. [Ibid.]

From the examples quoted above, it is clear that the latter two of these four coding patterns are asymmetric. I argue therefore that they might be explained by their relative frequency. In other words, the attributive coding pattern (i.e. the occurrence of relativizers with adjectives) correlates negatively with the relative frequency of the attributive use, and the predicative coding pattern (i.e. the occurrence of copulas with adjectives) correlates negatively with the relative frequency of the predicative use. This hypothesis will be tested using the frequency data from the Universal Dependencies Corpora and the Mandarin Chinese Corpus.

2.2. The Universal Dependencies Corpora

The Universal Dependencies (UD) is a cross-linguistic project that uses a consistent annotation of grammar. The current version of UD consists of 150 treebanks in 90 languages [Nivre et al. 2017; Croft et al. 2017]. In the last decade, some typological studies have attempted to explain the cross-linguistic pattern of the word order employing the data from UD [Liu 2010; Naranjo, Becker 2018; Levshina 2019]. As they have shown, UD proves to be a valuable source for cross-linguistic studies.

UD is a suitable database to test the hypotheses discussed above, primarily because a part of speech and a function is attributed to every word. In the corpora, various labels are used to tag the different functions and parts of speech; the labels that are relevant for the present discussion are presented in *Table 2* (p. 465).

For adjectives in each language, I have included all the occurrences of amod and nmod as the frequency of the attributive use, all

Table 2. The labels in the UD corpora

Function	Labels	Definition in UD
Attribution	amod	An adjectival modifier of a noun (or pronoun) is any adjectival phrase that serves to modify the meaning of the noun (or pronoun).
	nmod	The nmod relation is used for nominal modifiers.
Predication	root	The root of a sentence is the predicate of the main clause. This may be a verb, a predicate adjective, or a nominal in a copular construction.
	xcomp	An open clausal complement (xcomp) of a verb or an adjective is a predicative or clausal complement without its own subject.
Reference	nsubj	A nominal subject (nsubj) is a nominal which is the syntactic subject and the proto-agent of a clause.
	obj	The object of a verb is the second most core argument of a verb after the subject.
	iobj	The indirect object of a verb is any nominal phrase that is a core argument of the verb but is not its subject or (direct) object.
	obl	The obl relation is used for a nominal (noun, pronoun, noun phrase) functioning as a non-core (oblique) argument or adjunct. This means that it functionally corresponds to an adverbial attaching to a verb, adjective or other adverb.

the occurrences of root and xcomp as the frequency of the predicative use, and all the occurrences of nsubj, obj, iobj, and obl as the frequency of the referential use. In this way, I was able to collect the frequency data for adjectives in attributive, predicative, and referential use. Since in some languages the frequency of adjectives is too low, I have excluded six languages.[1] The final sample consists of 84 languages from 16 language families, and the areal distribution of the sample is presented in *Fig. 1*, which shows that the sample covers four macro-areas:

[1] The languages that I excluded are Skolt Sami, Komi Permyak, Assyrian, Wolof, Coptic, and Warlpiri.

Eurasia, Africa, Papunesia, and South America. It is also shown that the sample is not a balanced one. There are clearly more languages in Eurasia than in other areas.

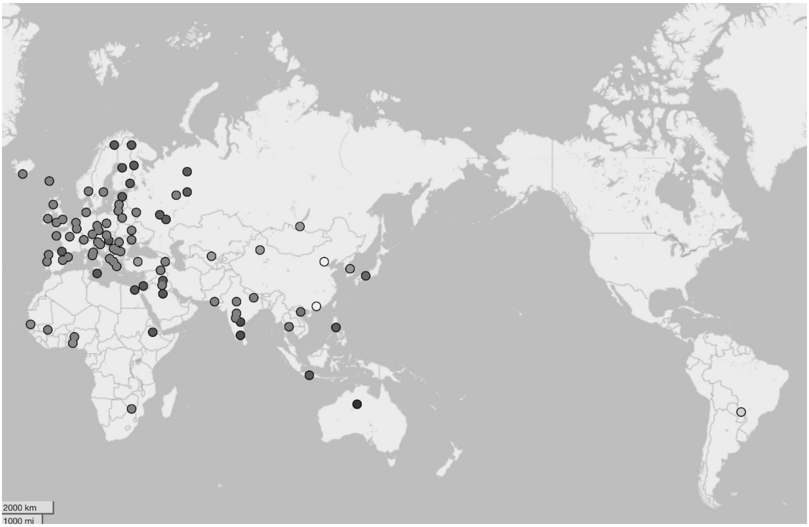

Fig. 1. The areal distribution of the sample

3. Results

In this section, I will present the main findings of my study. Firstly, the relative frequencies of adjectives in the attributive, predicative, and referential use are presented in *Fig. 2*. As it shows, the relative frequency of the attributive use is always the highest in comparison with the relative frequency of the referential and predicative use. In the previous literature, it has always been taken for granted that the primary function of adjectives is modification (or attribution) [Bhat 1994; Baker 2003; Lehmann 2013]. However, until now, this claim has never been supported with evidence from the corpus. The data in *Fig. 2* demonstrates that the most frequent function of adjectives is indeed attribution. This result is not surprising but is still worth testing.

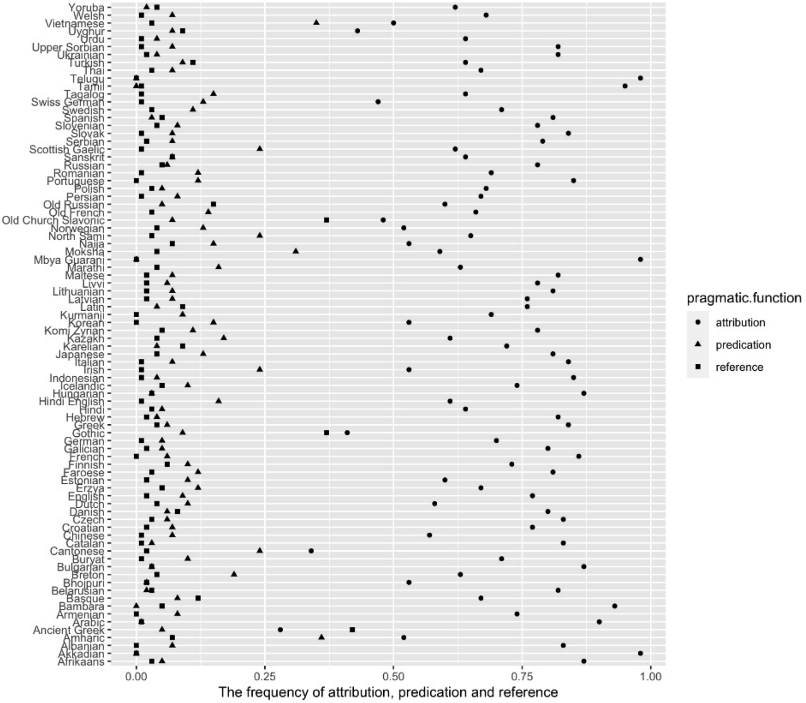

Fig. 2. The relative frequency of various pragmatic functions

The next step is to test whether frequencies correlate with coding patterns. For this purpose, I have manually collected data based on grammatical descriptions of all 84 languages from my sample regarding whether a copula or a relativizer is used with adjectives. In total, there are 73 languages that feature a copula in the predicative position, and four languages that use a relativizer with adjectives in the attributive position.

Since the relative frequency is a continuous variable and the occurrence of copula or relativizer is a binary variable, the logistic regression model is suitable for testing whether there is a correlation between these two variables. I have calculated the correlation between the relative frequency and the occurrence of copula or relativizer using the Fitting Generalized Linear Models (i.e. the “glm”) in R [R Core Team 2013]. The results are shown in *Fig. 3* and *4*. These results support the FFCH.

Fig. 3 shows that when the relative frequency of the predicative use increases, the probability of an adjective occurring with a copula decreases. The actual occurrences of copulas across languages are represented by black dots distributed along either ‘1’ or ‘0’, each dot representing a language. In order to identify whether there is a coefficient between the relative frequency and the probability of adjectives occurring with a copula, I have calculated the coefficient using the Fitting Generalized Linear Models in R. The result presented below in *Fig. 3* shows that the estimated coefficient between the relative frequency and the probability of adjectives occurring with a copula is –7.385, and the p-value is significant (i.e. below 0.05).

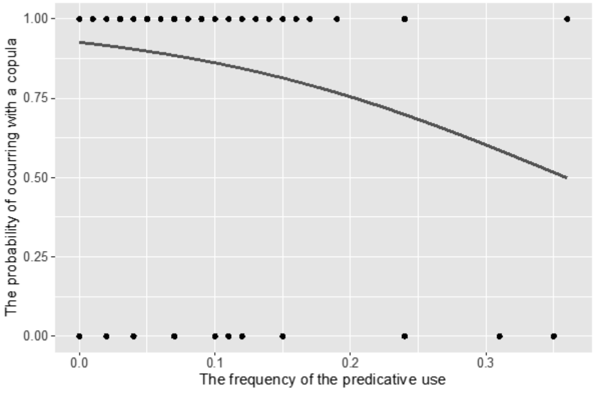

Fig. 3. The correlation between the predicative frequency and the copulas

The Coefficient of the predicative frequency and the occurrence of a copula

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	2.115	0.610	3.467	0.000527 ***
predication	−7.385	3.716	−1.987	0.046874 *
reference 15.014	14.777	1.016		0.309630

The correlation between the relative frequency of the attributive use and the probability of using a relativizer with adjectives is shown in *Fig. 4*. The general trend is that the probability of using a relativizer decreases when the relative frequency of the attributive use increases. The coefficient between these two variables is –10.414, and the p-value is 0.0219, which is also significant. However, it is noticeable that the p-value of the intercept is 0.0952, which is slightly higher than the general threshold of 0.05. This may lie in the fact that languages in UD are not balanced, and only a small number of languages use relativizers with adjectives. However, it is extremely hard to acquire relevant frequency information from other sources of cross-linguistic corpora. In addition, the effect of frequency on languages may sometimes be very hard to detect. For this reason, though the p-value of intercept is less significant, the data still suggests that there is a weak correlation between these two variables.

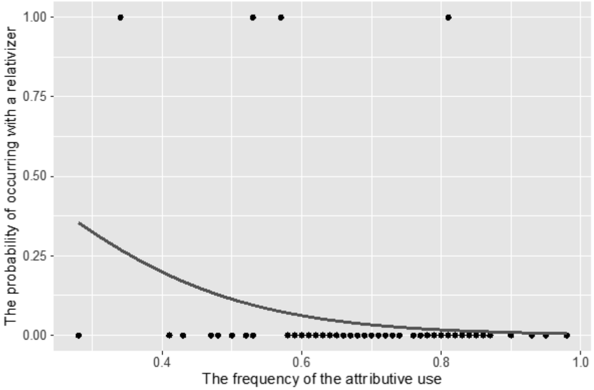

Fig. 4. The correlation between the attributive frequency and the relativizers

The Coefficient of the attributive frequency and the occurrence of a relativiser

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	4.836	2.898	1.669	0.0952 .
attribution	−10.414	4.543	−2.292	0.0219 *
reference	−43.384	35.609	−1.218	0.2231

In addition, I have also delved into the corpus data of Mandarin Chinese regarding various adjectives. In Mandarin Chinese, the occurrence of the relativizer *de* is optional, and various adjectives differ in their probability of occurring with it. Dixon [1982] has proposed seven typical semantic types that are very often coded as adjectives in various languages. Based on these semantic types, I have selected 28 adjectives from Mandarin Chinese and extracted the relative frequency information from the BCC corpus[2] [BCC; Xun et al. 2016]. The frequency information is presented in *Table 3* (next page), where "adj N" represents the token frequency of adjectives that occur without a relativizer; "adj de N" represents the token frequency of adjectives that occur with a relativizer; "adj N rel.freq" represents the relative frequency (i.e. the proportion) of adjectives that occur without a relativizer; "adj de N rel.freq" represents the relative frequency (i.e. the proportion) of adjectives that occur with a relativizer; "attr.token" represents the token frequency of the attributive use; "pred.token" represents the token frequency of the predicative use; "attr.rel.freq" represents the relative frequency (i.e. the proportion) of the attributive use; and "pred.rel.freq" represents the relative frequency (i.e. the proportion) of the predicative use.

Based on the relative frequency of adjectives that occur with or without the relativizer *de* and the relative frequency of adjectives that occur in attributive and predicative positions, I have calculated the correlation between the relative frequency of the attributive use and the relative frequency of the forms that omit the relativizer *de* in the attributive position. I have used three types of correlation tests that might help us to evaluate the correlation between the two continuous variables: the Pearson's product-moment correlation; the Kendall's rank correlation tau;

[2] The BCC corpora were created by the Beijing Language and Culture University and represent a balanced collection of annotated corpora containing around 15 billion Chinese characters. It covers newspapers and journals (about 2 billion Chinese characters), literature (about 3 billion Chinese characters), scientific books (about 3 billion Chinese characters), non-fiction books (about 1 billion characters), blog and weibo entries (about 3 billion Chinese characters), as well as classical Chinese (about 2 billion Chinese characters).

Table 3. The relative frequency of various adjectives in Mandarin Chinese

Chinese	**Meaning**	**adj N**	**adj *de* N**	**adj N rel. freq**	**adj *de* N rel.freq**	**attr.token**	**pred. token**	**attr.rel. freq**	**pred.rel. freq**
lao	old.animate	305 847	19 433	94.03 %	5.97 %	325 280	1 625 826	16.67 %	83.33 %
jiu	old.inanimate	68 649	10 829	86.37 %	13.63 %	79 478	58 597	57.56 %	42.44 %
nianqing	young	34 846	20 497	62.96 %	37.04 %	55 343	110 781	33.31 %	66.69 %
xin	new	766 391	2 130 059	26.46 %	73.54 %	2 896 450	22 654 314	11.34 %	88.66 %
da	big	1 481 229	353 750	80.72 %	19.28 %	1 834 979	4 807 689	27.62 %	72.38 %
xiao	small	847 534	70 554	92.32 %	7.68 %	918 088	2 532 387	26.61 %	73.39 %
chang	long	330 602	101 877	76.44 %	23.56 %	432 479	870 746	33.19 %	66.81 %
duan	short	88 863	13 481	86.83 %	13.17 %	102 344	235 906	30.26 %	69.74 %
bai	white	167 292	31 733	84.06 %	15.94 %	199 025	543 614	26.80 %	73.20 %
hei	black	96 156	13 247	87.89 %	12.11 %	109 403	427 897	20.36 %	79.64 %
lv	green	36 594	5 936	86.04 %	13.96 %	42 530	188 293	18.43 %	81.57 %
hong	red	122 733	20 261	85.83 %	14.17 %	142 994	457 511	23.81 %	76.19 %
hao	good	734 690	280 222	72.39 %	27.61 %	1 014 912	4 114 507	19.79 %	80.21 %
huai	bad	54 994	11 310	82.94 %	17.06 %	66 304	106 779	38.31 %	61.69 %

Chinese	**Meaning**	**adj N**	**adj *de* N**	**adj N rel. freq**	**adj *de* N rel.freq**	**attr.token**	**pred. token**	**attr.rel. freq**	**pred.rel. freq**
meili	beautiful	11 560	32 190	26.42 %	73.58 %	43 750	62 403	41.21 %	58.79 %
chou	ugly	6 238	1 999	75.73 %	24.27 %	8 237	51 483	13.79 %	86.21 %
ruan	soft	26 611	10 277	72.14 %	27.86 %	36 888	192 729	16.07 %	83.93 %
re	hot	75 459	20 099	78.97 %	21.03 %	95 558	540 045	15.03 %	84.97 %
leng	cold	45 843	26 693	63.20 %	36.80 %	72 536	474 016	13.27 %	86.73 %
gan	dry	96 110	16 385	85.43 %	14.57 %	112 495	951 283	10.58 %	89.42 %
zhong	heavy	209 951	74 912	73.70 %	26.30 %	284 863	1 840 078	13.41 %	86.59 %
kongju	afraid	2 727	2 487	52.30 %	47.70 %	5 214	36 303	12.56 %	87.44 %
feng	crazy	5 114	1 549	76.75 %	23.25 %	6 663	137 800	4.61 %	95.39 %
kaixin	happy	11 047	9 934	52.65 %	47.35 %	20 981	179 883	10.45 %	89.55 %
shangxin	sad	2 935	3 343	46.75 %	53.25 %	6 278	38 293	14.09 %	85.91 %
yuchun	stupid	640	3 826	14.33 %	85.67 %	4 466	9 168	32.76 %	67.24 %
congming	clever	6 639	8 138	44.93 %	55.07 %	14 777	39 533	27.21 %	72.79 %
shengqi	angry	1 126	3 550	24.08 %	75.92 %	4 676	56 905	7.59 %	92.41 %

and the Spearman's rank correlation rho. These three types of correlation tests lead to a result that ranges from 0 to 1 to indicate the correlation, and a greater number suggests a stronger correlation. Based on the data presented in *Table 4*, all the three correlation tests show strong correlation between the two variables. In particular, the Pearson's product-moment correlation is 0.8092 (p-value=1.863e-07); the Kendall's rank correlation tau is 0.6667 (p-value = 5.82e-08); and the Spearman's rank correlation rho is 0.7756 (p-value =3.376e-06). The result is also presented in *Fig. 5*. As shown in *Fig. 5*, the higher the frequency of an adjective used in the attributive position, the more likely it is to occur without the relativizer *de*.

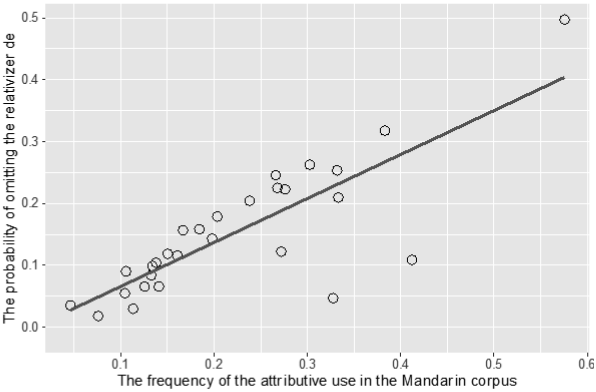

Fig. 5. The correlation based on the corpus data from Mandarin Chinese

To sum it up, although the cross-linguistic data provides but the weak evidence for the correlation between the relative frequency of the attributive use and the occurrence of a relativizer, the result based on the data from Mandarin Chinese serves as strong evidence for the correlation between these two variables.

4. Conclusion

In this paper, I have provided evidence for the form-frequency correspondence hypothesis using the data from the Universal Dependencies Corpora and the BCC Mandarin Chinese Corpus. In particular, I argue that the probability of using a copula with adjectives in the predicative position correlates negatively with the relative frequency of their predicative use; likewise, the probability of using a relativizer with adjectives in the attributive position correlates negatively with the relative frequency of their attributive use. In order to test these claims, I have studied the data of 84 languages from the Universal Dependencies Corpora in relation to the relative frequency of adjectives in the attributive, predicative, and referential use. In addition, I have also checked whether a relativizer or a copula is used with the adjectives in these 84 languages, or whether it is missing.

The data of these 84 languages has been tested using the logistic regression model, which is suitable for calculating the correlation between a continuous variable and a binary variable. The results have proven my theory: indeed, the relative frequency of the predicative use of adjectives correlates negatively with the probability of adjectives occurring with a copula; and the relative frequency of the attributive use correlates negatively with the probability of using a relativizer. The former correlation is relatively significant, while the latter is less significant. For this reason, I have also included the corpus data from Mandarin Chinese. This analysis has shown that there is indeed a strong correlation between the relative frequency of the attributive use and the probability of using a relativizer. The results of this study can be taken as evidence supporting the FFCH.

There are still some unsolved problems. This paper has provided evidence for the correlation between frequency and form. However, correlation does not imply a causal relation. Ultimately, it would be ideal to prove a causal relation between frequency and form in future research.

Abbreviations

2 — 2nd person; COP — copula; DEF — definite; FFCH — form-frequency correspondence hypothesis; REL — relativiser; SG — singular; UD — universal dependencies.

References

Backhouse 2004 — A. E. Backhouse. Inflected and uninflected adjectives in Japanese. R. Dixon, A. Aikhenvald (eds.). *Adjective Classes: A Cross- Linguistic Typology*. Oxford: Oxford University Press, 2004. P. 50–73.

Baker 2003 — M. Baker. *Lexical Categories: Verbs, Nouns and Adjectives*. Cambridge: Cambridge University Press, 2003.

Bhat 1994 — D. Bhat. *The Adjectival Category: Criteria for Differentiation and Identification*. Amsterdam: John Benjamins Publishing, 1994.

Chomsky 1970 — N. Chomsky. Remarks on nominalization. R. Jacobs, P. Rosenbaum (eds.). *Readings in English Transformational Grammar*. Waltham, MA: Blaisdell, 1970. P. 184–221.

Croft et al. 2017 — W. Croft, D. Nordquist, K. Looney, M. Regan. Linguistic typology meets universal dependencies. M. Dickinson, J. Hajič, S. Kübler, A. Przepiórkowski (eds.). *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15).* CEUR Workshop Proceedings, 2017. P. 63–75.

Dixon 1982 — R. M. W. Dixon. Where have all the adjectives gone? R. M. W. Dixon (ed.). *Where Have All the Adjectives Gone? And Other Essays in Semantics and Syntax*. Berlin: De Gruyter Mouton. 1982. P. 1–62.

Dixon 2004 — R. M. W. Dixon. *The Jarawara Language of Southern Amazonia*. Oxford: Oxford University Press, 2004.

Greenberg 1966 — J. H. Greenberg. *Language Universals: With Special Reference to Feature Hierarchies*. The Hague: Mouton, 1966.

Haspelmath 2008 — M. Haspelmath. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*. 2008. Vol. 19. Pt. 1. P. 1–33.

Haspelmath et al. 2014 — M. Haspelmath, A. Calude, M. Spagnol, H. Narrog, E. Bamyaci. Coding causal-noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics*. 2014. Vol. 50. Pt. 3. P. 587–625.

Haspelmath 2021 — M. Haspelmath. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics*. 2021. Available at: https://www.cambridge.org/core/journals/journal-of-linguistics/

article/explaining-grammatical-coding-asymmetries-formfrequency-correspondences-and-predictability/420965EC1CEA49527CCE7276B33A14D0 (accessed on 17.03.2021).

Hawkins 2004 — J. A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press, 2004.

Heath 1999 — J. Heath. *A Grammar of Koyra Chiini: The Songhay of Timbuktu*. Berlin: Mouton de Gruyter, 1999.

Lehmann 2013 — C. Lehmann. The nature of parts of speech. *STUF — Language Typology and Universals*. 2013. Vol. 66. Pt. 2. P. 141–177.

Levshina 2019. — N. Levshina. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*. 2019. Vol. 23. Pt. 3. P. 533–572.

Liu 2010 — H. Liu. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*. 2010. Vol. 120. Pt. 6. P. 1567–1578.

Naranjo, Becker 2018 — M. G. Naranjo, L. Becker. Quantitative word order typology with UD. D. Haug, S. Oepen, L. Øvrelid, M. Candito, J. Hajič (eds.). *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*. Linköping: Linköping University Electronic Press, 2018. P. 91–104.

Noonan 1992 — M. Noonan. *A Grammar of Lango*. Berlin: De Gruyter Mouton, 1992.

Quasthoff et al. 2014 — U. Quasthoff, D. Goldhahn, T. Eckart. Building large resources for text mining: The Leipzig Corpora Collection. C. Biemann, A. Mehler (eds.). *Text Mining — From Ontology Leaning to Automated Text Processing Applications.* New York: Springer, 2014. P. 3–24.

Wetzer 1992 — H. Wetzer. “Nouny” and “verby” adjectivals: A typology of predicative adjectival constructions. M. Kefer, J. van der Auwera (eds.). *Meaning and Grammar. Cross-Linguistic Perspectives.* Berlin: De Gruyter Mouton, 1992. P. 223–262.

Wetzer 1996 — H. Wetzer. *The Typology of Adjectival Predication*. Berlin: De Gruyter Mouton, 1996.

Xun et al. 2016 — E. Xun, G. Rao, X. Xiao, J. Zang. The construction of the BCC corpus in the age of big data. *Corpus Linguistics*. 2016. Vol. 3. P. 93–109.

Zipf 1935 — G. K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Mifflin, 1935.

Sources

BCC — Beijing Language and Culture University Corpus Center. Available at: http://bcc.blcu.edu.cn/lang/zh (accessed on 08.06.2020).

Leipzig Corpora Collection 2018 — *Leipzig Corpora Collection.* Available at: https://corpora.uni-leipzig.de/en?corpusId=deu_newscrawl-public_2018 (accessed on 17.03.2021).

Nivre et al. 2017 — J. Nivre, L. A. Željko Agic et al. *Universal Dependencies 2.0 — CoNLL Shared Task Development and Test Data.* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, 2017. Available at: https://universaldependencies.org (accessed on 08.06.2020).

R Core Team 2013 — *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. Available at: https://www.r-project.org (accessed on 17.03.2021).