Maximilian Frankowsky* and Dan Ke

Humanness and Classifiers in Mandarin Chinese

A Corpus-Based Study of Anthropocentric Classification

Abstract: Mandarin Chinese numeral classifiers receive considerable attention in linguistic research. The status of the general classifier 个 *gè* remains unresolved. Many linguists suggest that the use of 个 *gè* as a noun classifier is arbitrary. This view is challenged in the current study. Relying on the CCL-Corpus of Peking University and data from Google, we investigated which nouns for living beings are most likely classified by the general classifier 个 *gè*. The results suggest that the use of the classifier 个 *gè* is motivated by an anthropocentric continuum as described by Köpcke and Zubin in the 1990s. We tested Köpcke and Zubin's approach with Chinese native speakers. We examined 76 animal expressions to explore the semantic interdependence of numeral classifiers and the nouns. Our study shows that nouns with the semantic feature [+ animate] are more likely to be classified by 个 *gè* if their denotatum is *either* very close to *or* very far located from the anthropocentric center. In contrast animate nouns whose denotata are located at some intermediate distance from the anthropocentric center are less likely to be classified by 个 *gè*.

Keywords: anthropocentric continuum; Chinese; humanness; numeral classifiers

*Corresponding author, Maximilian Frankowsky, University of Bonn, E-mail: mfrankow@uni-bonn.de
Dan Ke, University of Münster, E-mail: d_ke0001@uni-muenster.de

1 Introduction

Many languages across the world exhibit nominal classification systems (Allan, 1977; Aikhenvald, 2000). While most Indo-European languages make

use of grammatical gender to classify nouns, almost all other languages use classifiers instead. Many speakers across the planet routinely process classifiers in their everyday communicative interactions.

Mandarin Chinese is generally regarded as a classifier language. In this language, a classifier is obligatorily used when a noun is combined with a determiner or a numeral. The choice of the correct classifier is a not a matter of rote learning, nor is it determined by a simple grammatical rule; rather, its use depends on specific semantic features of the noun such as material, shape, size, or consistency (Tai, 1994). Moreover, it is commonly assumed that the Mandarin Chinese classifier 个 *gè* may replace any specific classifier to classify/individualize a noun. However, although 个 *gè* is used as a classifier for many nouns, its degree of acceptability varies from case to case. We contend that the use of 个 *gè* is motivated by conceptual principles, i.e., classifier choice is not based on an arbitrary convention. The question whether 个 *gè* can be used as a classifier for a specific noun can be answered only by empirical research that we describe in this paper. In what follows, we first briefly characterize the classifier system of Mandarin Chinese. Next, we present our experimental studies, focusing on animate nouns and the classifiers they require. Finally, we report some relevant results from our corpus analysis and discuss three possible explanations for classifier choice in Mandarin Chinese: frequency, shape, and the anthropocentric continuum.

2 The Mandarin Chinese Classifier System

In Mandarin Chinese, a classifier is a monosyllabic morpheme that is required in noun phrases containing determiners or numerals:

(1) 三　　　本　　　字典
　　 sān　　*běn*　　*zìdiǎn*
　　 three　　CL　　　dictionary
　　 'three dictionaries'

(2) 这　　　只　　　猫
　　 zhè　　*zhī*　　*māo*
　　 this　　 CL　　　cat
　　 'this cat'

In Mandarin Chinese, it is not possible to combine the numeral 'three' and the noun 'dictionary' without the insertion of a classifier between the numeral and the noun. In (1), due to the inherent semantics of the noun 'dictionary', the classifier for books and book-like things is selected. In example (2), a classifier must be placed between the determiner 'this' and the noun 'cat'. In this case, it is one of the classifiers for animals. The two nouns in (1) and (2) differ semantically; accordingly, their classifiers are also distinct.

It is necessary to distinguish classifiers in the narrower sense from simple measure words (Tai, 1994; Cheng & Sybesma, 1998). The latter are used for the purpose of quantifying nouns. An example for a measure word is 斤 *jīn* 'pound'. This word is used like its equivalents in European languages. 一斤米 *yī jīn mǐ* means 'one pound rice'. The numeral refers to the unit 斤 *jīn* 'pound'. Classifiers behave differently from measure words. For example, 一条鱼 *yī tiáo yú* means 'one fish'. In this case, the numeral does not refer to a conventionalised unit of measurement but to the entity that is counted, in this case 鱼 *yú* 'fish'. European languages do not make use of classifiers and, consequently, they are not translatable. The litmus test for determining whether a word is an individual classifier or a measure word, is to put the structural particle 的 *de* between the classifier and the noun. 一斤的米 *yī jīn de mǐ* 'one pound of rice' is grammatically correct. 斤 *jīn* is therefore a measure word whereas *⋆一条的鱼 *yī tiáo de yú* 'one tiao of fish' is grammatically ill-formed. Hence, 条 *tiáo* is a classifier. However, there is no sharp dividing line between quantifying and classifying words. There are also container measures (such as 瓶 *píng* 'bottle'), group measures (such as 群 *qún* 'group'), partitive measures (such as 块 *kuài* 'piece'), and classifiers such as 片 *piàn* 'slice', which can be a partitive measure word or an individual classifier for flat objects like CDs. In the following, we use the term 'classifier' only for sortal classifiers such as 条 *tiáo*.

Mandarin Chinese has hundreds of classifiers. Almost every classifier can classify more than just one noun, and almost every noun can be classified by more than just one classifier. This means that, in contrast to noun classification in European languages where a noun usually belongs to only one noun class, Chinese speakers normally have to choose among several possible classifiers. The classifier is selected on the basis of the inherent semantics of the noun (Huang & Ahrens, 2003). The classifier 条 *tiáo* in the preceding example is chosen because of the long shape of the entity designated by the noun.

In addition to the sortal classifiers mentioned above (like 本 *běn* for books or 条 *tiáo* for fishes), there is what is often called a general classifier

(Zubin & Shimojo, 1993; Myers et al., 1999) or default classifier (Chen, 1996). It has often been claimed that the classifier 个 *gè* can replace any specific classifier to classify a noun. For this reason, 个 *gè* is considered to be the "the most commonly used classifier" (Ross & Ma, 2006: 46).

3 Possible Explanations for the Distribution of 个 *gè*

个 *gè* may indeed be used as a classifier for many nouns, but its degree of acceptance varies from noun to noun. As mentioned above, most linguists suggest that the distribution of 个 *gè* usage across nouns is arbitrary. Some researchers like Mary Erbaugh (1986) assume that, in actual language use, 个 *gè* as a classifier may not be grammatically determined. Rather, it can be sensitive to discourse-pragmatic factors. When a noun is introduced into the discourse, it is classified by an individual sortal classifier to code that a new discourse entity comes into play. When referring to the same noun referent again, speakers use 个 *gè* to classify the previously mentioned noun (ibid.: 408). According to this account, Chinese classifiers are discourse sensitive and pragmatically determined (ibid.: 44). However, a study conducted by Liu (2010) challenges Erbaugh's thesis. Liu's data show only one occurrence of 个 *gè* per 1000 concordance lines on second mention of a referent, whose first occurrence was classified by one of the most commonly used 27 classifiers. In any case, it seems to be a widely held assumption that the degree of acceptance for 个 *gè* does not depend on the semantics of the classified noun itself. To formulate this claim in the words of James Myers (2000: 196): "*Ge* is semantically vacuous". In this article, we want to challenge the view that 个 *gè* is a semantically empty classifier. On the contrary, we contend that the inherent semantics of nouns plays a crucial role in how likely the classifier 个 *gè* is to co-occur with them.

Analyses in terms of frequency have become very popular in contemporary linguistics, and proved to be a very powerful tool to account for the language structure (Haspelmath, 2006). But is it possible to explain the distribution of the Chinese classifier 个 *gè* in terms of frequency alone, whereby low-frequency nouns are classified by 个 *gè* more often than high-frequency nouns?

The second possible account would rely of the feature of shape. James Tai (1994) lists shape as an important semantic principle for the distribution

of classifiers. The classifier 条 *tiáo* is a classic example for this semantic principle. Speakers may select 条 *tiáo* as a noun classifier if the denoted object is oblong. For example, 蛇 *shé* 'snake' usually takes this classifier because of its long shape. However, sometimes nouns take 条 *tiáo* as a classifier even if not the object itself but only the object that the root morpheme denotes has a long shape.

The third possible explanation that we consider is the notion of an anthropocentric continuum. This animacy scale was used for analytical purposes by Klaus-Michael Köpcke and David Zubin in the 1990s (Köpcke & Zubin, 1996; Köpcke, 1995). It is a simple linear scale on which all living beings that populate the world can be arranged. The anthropocentric continuum reflects a folk model; in other words, it describes how ordinary people rather than experts classify animate beings. Humans are located at one end of the scale, while entities such as corals or leeches are situated at the other end. The idea behind this continuum is an old one, possibly stemming from Greek philosophy, and is also influenced by Bühler's 'I-here-now-Origo' and animacy hierarchies that have been postulated by Lakoff and Turner (1989), Silverstein (1976), and others. Köpcke and Zubin used their concept of the anthropocentric continuum to analyze the German gender system and found that animals that are close to humans are most likely to be of masculine gender. In contrast, feminine gender nouns tend to denote referents that are more distant than the denotata of masculine nouns from the human pole on the continuum. Köpcke and Zubin (1996: 484) distinguish eight classes of living beings: first are humans, followed by monkeys/predators, mammals, birds/fishes, reptiles, snakes, insects, and mollusks. We undertook to test this scale by employing an online survey and argue that this folk taxonomy has some influence on the classification of animate nouns of Mandarin Chinese.

4 Methods

4.1 Survey on the Anthropocentric Continuum

The online survey was designed to investigate the validity of Köpcke and Zubin's (1996) animacy scale. In particular, we wanted to discover whether Chinese native speakers conceptualize animals' distance from humans the

same way Köpcke and Zubin claim that westerners do, and if this mental construct correlates with grammatical choices of classifiers.

4.1.1 Participants and Task

Seventy-five monolingually raised native speakers of Mandarin Chinese (36 female, mean age 27 years, range 19–54) participated in the survey as non-paid volunteers. All participants were born and raised in China, 9 were residing as students in Germany at the time of examination, 1 in France.

The participants were asked to value 76 animals in terms of their distance to humans. We deliberately explained vaguely what 'distance to humans' could be, so that the participants would not follow any prescribed rules. The introductory text at the beginning of the survey merely stated that behavior, appearance, or the possibility to identify with the animal could be the criteria determining 'distance' and that their decision should be a gut decision. The items were presented in random order and the questionnaire consisted of 40 pages. The participants scored the items on an 11-point semantic differential scale (0 = totally different from humans and 10 = has similar characteristics with humans; see Figure 1). The questionnaire task took approximately 15 minutes to complete.
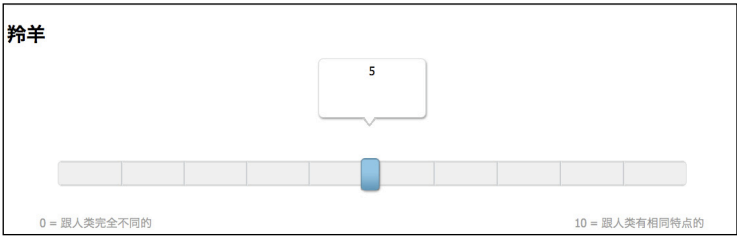

Figure 1: Example of a trial sequence used in the questionnaire

4.1.2 Results

In the vast majority of cases, participants had very similar attitudes towards the humanness of the signified animals. Thus, we got valid humanness values for every single item. The highest rated animal is the orangutan with a

score of 8.56. Except for monkeys, no animal was rated over 6. This shows that animals in general are not considered by our test group to be that close to humans.

The results mainly support the assumptions of Köpcke and Zubin. However, there are a few members of the six groups that are noteworthy. The parrot for example was scored higher than all the other birds, presumably because it is regarded as an intelligent specimen of the class. But the average score of the six animal groups corresponds to the scale of the anthropocentric continuum (see Table 1).

Table 1: Animal classes and their average score in terms of distance to humans on a scale between 0 and 10

monkeys/predators	**6.41**
mammals	4.51
birds/fishes	3.66
reptiles/snakes/amphibians	3.04
insects	2.14
mollusks	1.74

4.2 Corpus Analysis of Mandarin Chinese Animal Terms

Two sources of information were used to investigate the distribution of 个 *gè* among animal terms: data from the CCL-Corpus and data from on-line Google searches.

4.2.1 Our Corpora

The CCL-Corpus, compiled by the Center for Chinese Linguistics at Peking University, consists of about 370 million words drawn from both written and spoken contemporary Mandarin Chinese. Other corpora, such as the Lancaster Corpus for Mandarin Chinese or the Academia Sinica Balanced Corpus of Modern Chinese, consist only of a few million words and were rejected for our study as being too small. The CCL-Corpus consists of many different kinds of texts e.g. newspapers, magazines, novels, and transcripts of television shows.

In addition to the CCL-Corpus, we used on-line Google searches as a second source of information. Google has the advantage of retrieving and reflecting current language use, which may reveal approximately how people use classifiers in their daily life. Google uses simplified and traditional Chinese characters. Typing in simplified Chinese, the results may appear in both simplified and traditional Chinese. Using Google as a source of information does entail certain problems, e.g., the results are non-filtered. Since some Chinese characters are also used in Japanese, Japanese data appeared in our searches. Thus we had to filter the results ourselves to eliminate such occurrences.

4.2.2 Inquiry

We examined the usage of five classifiers that co-occur with animals: 个 *gè*, 只 *zhī*, 条 *tiáo*, 匹 *pǐ*, and 头 *tóu*, with the terms of 76 animals from different species. All of the animal expressions are disyllabic. It is also important that none of the animal terms uses 个 *gè* as a sortal classifier. The query was always of the same structure: 'classifier'+'animal term' (e.g. 个兔子). In this way we got results for numeral+classifier-combinations as well as for demonstrative pronoun+classifier-combinations. For both Google and CCL, we accepted the output only if the classifier is clearly related to the animal term. For example, the result 一个兔子洞 *yī gè tùzi dòng* 'one ge rabbit hole' does not count for our 个 *gè*-results, because the classifier refers to the hole, not the animal. These cases were eliminated from our data. In addition to that, we excluded cases in which the noun itself does not refer to the corresponding animal but to other things such as nicknames or brand names (e.g. 骆驼 *luòtuó* 'camel' for the cigarette brand).

4.2.3 Results

As for frequency, it seems that there is no correlation with the generalization of 个 *gè*. That is, grouping all the high frequency items together to compare them with the low frequency items, we find that the average rate of 个 *gè*-classification in these two groups is nearly identical (about 15%). In terms of the generalization of 个 *gè*, group 1 does not differ from group 2, although the frequency diverges widely (CCL: 967 vs. 311; Google: 7.8m vs. 1.3m). Thus we conclude that frequency does not have a large influence on the distribution

of 个 *gè*. High frequency and low frequency items are more or less equally classified by 个 *gè* (see Figure 2).
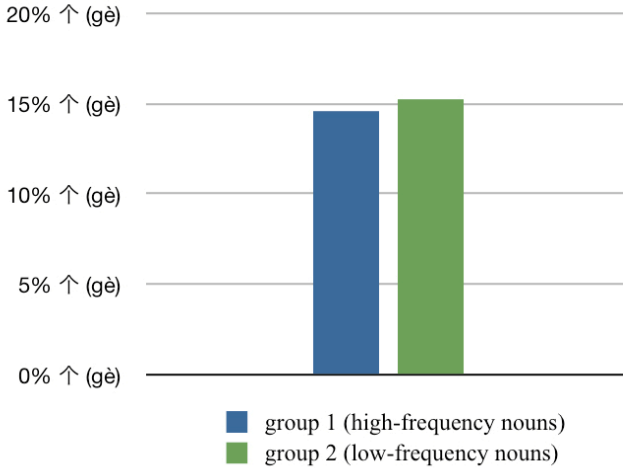
Figure 2: 个 *gè*-selection rate between high- and low-frequency animal expressions

As for shape, it seems that shape does have quite a remarkable influence on the generalization of 个 *gè* among animate nouns. As mentioned before, the classifier 条 *tiáo* is often chosen due to the long shape of the entity that the noun designates. That is why animals with a long shape are most likely to be classified by this classifier. In our data, fishes, lizards, and worm-like insects/mollusks are classified by 条 *tiáo* at a high percentage (lizards: 35%, fishes: 65%, worm-like insects: 64%). Hence, nouns that have the semantic feature of a long shape cannot be easily classified by 个 *gè*. However, this affects only 11 items in total. The classification of the remaining animal terms does not involve shape-related characteristics. Shape is therefore not a decisive feature and has no huge influence on the classification of animate nouns as a whole.

However, the anthropocentric continuum indeed plays a decisive role in the generalization of 个 *gè* when it comes to animate nouns. Divided into three even groups in terms of humanness value (high: 10–4.13, average: 4.13–3, low: 3–0) a correlation between the humanness value and 个 *gè*-

classification becomes obvious. The more an animal differs from humans, the more likely it can be classified by 个 *gè* (see Figure 3).
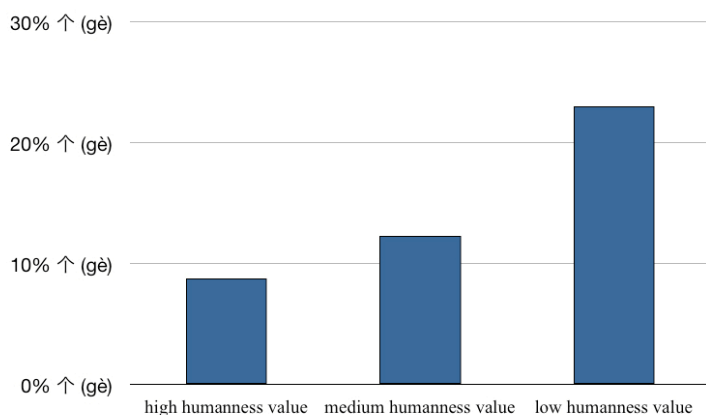

Figure 3: 个 *gè*-selection rate in relation to humanness value

The effect is weakened by the fact that 个 *gè* is one of several sortal classifiers for humans and can therefore be used easily for monkeys and other animals that are very close to human beings in terms of behavior and appearance. Divided into six groups following Köpcke and Zubin's assumptions, the 个 *gè*-classification rate forms a U-shape (see Figure 4).

One might suppose that there are two different mechanisms that could account for the U-shaped results depicted in Figure 4. On the one hand, humans are normally classified by 个 *gè*, thus a good reason to classify nouns with 个 *gè* for other living beings that are close to humans would be simple analogy. On the other hand, speakers are aware of humanness as a semantic principle for categorizing living beings in their daily lives. This principle may account for a weaker conceptual connection between a noun and its sortal classifier if the signified being is very different from human beings. 个 *gè* can therefore easily replace the individual classifier of nouns for 'very unhuman' living beings. Animate nouns are more often associated with the classifier 个 *gè* if the signified living being is very close to (e.g. 猴子 *hóuzi* 'monkey') or, on the contrary, very far away from human beings (e.g. 牡蛎 *mǔlì* 'oyster'). Nouns that signify creatures that have an average distance to humans (e.g. 羚

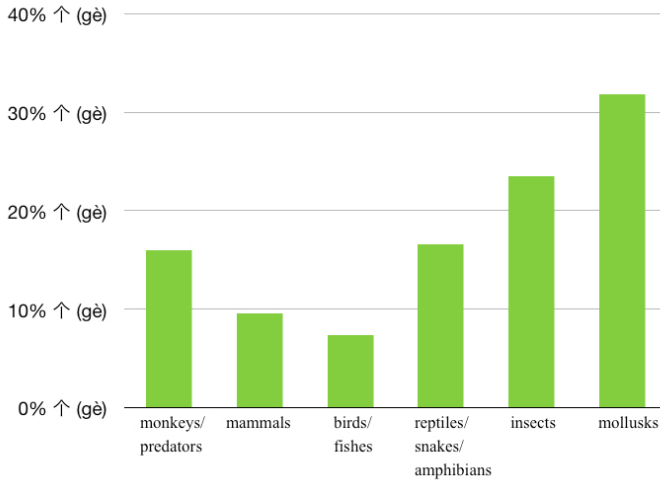羊 *língyáng* 'antelope' or 天鹅 *tiān'é* 'swan') are less likely to be classified by 个 *gè*.

Figure 4: 个 *gè*-selection rate among animal expressions in terms of animal species

5 Conclusion

The results of our research lead us to conclude that there exists a general unconscious preference for speakers of Mandarin Chinese to categorize living beings in terms of an anthropocentric continuum, such as the one developed by Köpcke and Zubin (1996). We found that whether or not a Chinese animal term is classified by 个 *gè* does not depend on the frequency of occurrence of the noun. Our account of Chinese noun classification assumes that the inherent semantics of nouns plays a crucial role. Shape, for example, has a determining influence on 个 *gè*-distribution; e.g., nouns denoting long-shaped animals are very likely to be classified by 条 *tiáo*. More importantly, our research findings indicate that the distribution of 个 *gè* is motivated to an even greater degree by speakers' conceptions of how close to or how distant from humans are other animate beings. That is to say, the more an animal differs from humans, the more likely the animal noun would be classified by 个 *gè*. The slightly higher 个 *gè*-selection rate of monkeys and predators could be due to the fact that 个 *gè* is the sortal classifier for humans. These two factors

account for the finding that terms for birds and fishes are least likely to be classified by 个 *gè*.

References

Aikhenvald, A. Y. (2000). *Classifiers: A typology of noun categorization devices*. Oxford: Oxford University Press.

Allan, K. (1977). Classifiers. *Language*, *53*, 285–311.

Chen, A. (1996). *Numerative und Numerativkonstruktionen im Deutschen und im Chinesischen. Eine konstrastiv-typologische Untersuchung*. Tübingen: Niemeyer.

Cheng, L., & Sybesma, R. (1998). Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua Journal of Chinese Studies*, New Series, *28*, 385–412.

Erbaugh, M. S. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. In C. Craig (Ed.), *Noun classes and categorization* (pp. 339–436). Amsterdam: Benjamins.

Erbaugh, M. S. (2006). Chinese classifiers: Their use and acquisition. In: H. T. Li, E. Bates, & O. Tzeng (Eds.), *The handbook of East Asian psycholinguistics* (pp. 39–51). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511550751005

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of Linguistics*, *42*, 25–70. doi: 10.1017/S0022226705003683

Huang, C. R., & Ahrens, K. (2003). Individuals, kinds and events: Classifier coercion of nouns. Language Science, 25(4), 353–373. doi:10.1016/S0388-0001(02)00021-9

Köpcke, K.-M. (1995). Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache. Ein Beispiel für die Leistungsfähigkeit der Prototypentheorie. *Zeitschrift für Sprachwissenschaft*, *14*(2), 159–180.

Köpcke, K.-M., & Zubin, D. A. (1996). Prinzpien für die Genuszuweisung im Deutschen. In E. Lang & G. Zifonun (Eds.), *Deutsch – typologisch. Jahrbuch des Instituts für Deutsche Sprache 1995* (pp. 473–491). Berlin: de Gruyter.

Lakoff, G., & Turner, M. (1989). *More than cool reason: A field guide to poetic metaphor*. Chicago: The University of Chicago Press.

Liu, W. (2010). A preliminary study on the substitution of the general classifier *Ge* for specific classifiers. [Unpublished ms].

Myers, J., Gong, S. P., & Shen, Z. G. (1999). The semantic content of the general classifier in Mandarin. International Association of Chinese Linguistics Eighth Annual Conference, Melbourne, Australia.

Myers, J. (2000). Rules vs. analogy in Mandarin classifier selection. *Language and Linguistics*, *1*(2), 187–209.

Ross, C., & Ma, J. H. S. (2006). *Modern Mandarin Chinese grammar: A practical guide*. Oxon: Routledge.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. Dixon (Ed.), *Grammatical categories in Australian Languages* (pp. 112–171). Canberra: Australian Institute of Aboriginal Studies, 112–171.

Tai, J. H.-Y. (1994). Chinese classifier systems and human categorization. In M. Y. Chen & O. Tzeng (Eds.), *In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change* (pp. 479–494). Taipei: Pyramid Press.

Zubin, D. A., & Shimojo, M. (1993). How "general" are general classifiers? With special reference to *ko* and *tsu* in Japanese. *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society*, *19*(1), 490–502.

Bionotes

Maximilian Frankowsky

Maximilian Frankowsky (b. 1986) studied German Language and Literature at the University of Münster, Germany. His research interests include psycholinguistics and nominal classification systems. He has given talks at conferences in Nanjing and Prague and was an intern at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig, Germany, and at International Studies University in Xi'an, China. Currently he is working as a research fellow in the Department of German Studies at the University of Bonn. Additionally, he is involved in an EEG study in cooperation with the Institute of Psychology of the University of Münster.

Dan Ke

Dan Ke (b. 1989) studied linguistics at the University of Münster, Germany. Her research interests include psycholinguistics, language typology, and historical linguistics. She has given talks at conferences in Nanjing and Prague and held internships at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, and the Institute of Psychology in Münster. Currently she is working at the Institute of Chinese Studies at the University of Münster. Additionally, she is involved in an EEG study in cooperation with the Institute of Psychology of the University of Münster.