http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

Towards Chinese Learner's Dictionaries for Foreigners Living in China: Some Problems Related to Lemma Selection

Mei Xue, *Department of Foreign Language, China University of Mining and Technology (Beijing), China; and Centre for Lexicography, University of Aarhus, Denmark (meix99@yahoo.com)*
and
Sven Tarp, *Sino-Danish Sindberg Centre of Lexicography, Translation and Business Communication, Guangdong University of Finance, China; International Centre for Lexicography, Universidad de Valladolid, Spain; Department of Afrikaans and Dutch, University of Stellenbosch, South Africa; and Centre for Lexicography, University of Aarhus, Denmark (st@cc.au.dk)*

Abstract: During the past decades, various dictionaries for foreign learners of Chinese have seen the light. Except for one picture dictionary which is almost completely ignored in the academic literature, none of these dictionaries has taken into account the special needs which foreigners living in China and learning Chinese may have. This contribution will discuss these needs with special focus on lemma selection. We argue that foreigners living in China, in order to meet their lexicographical needs, require additional words typically occurring in social contexts in which they often find themselves, whether or not these words have a high corpus-frequency. As a solution we therefore recommend a set of selection criteria that combines corpus frequency and context relevance. Finally, we discuss how logfiles reflecting user behaviour can be used as a new and very reliable empirical source for lemma selection for an online Chinese learner's dictionary.

Keywords: CHINESE LEARNER'S DICTIONARIES, LEMMA SELECTION, SOCIAL CONTEXTS, CORPUS FREQUENCY, CONTEXT RELEVANCE

Opsomming: Op weg na Chinese aanleerderswoordeboeke vir buitelanders wat in China woon: Enkele probleme verwant aan lemmaseleksie. Gedurende die afgelope dekades het verskeie woordeboeke vir vreemdetaalleerders van Chinees verskyn. Buiten een prentewoordeboek wat byna heeltemal in die akademiese literatuur geïgnoreer is, het geeneen van hierdie woordeboeke die spesiale behoeftes wat buitelanders wat in China woon en Chinees aanleer, mag hê, in ag geneem nie. In hierdie artikel word hierdie behoeftes, met spesiale fokus op lemmaseleksie, bespreek. Ons argumenteer dat buitelanders wat in China woon, addisionele woorde benodig wat tipies voorkom in sosiale kontekste waarin hulle hulself dikwels bevind, ongeag of hierdie woorde 'n hoë korpusfrekwensie het of nie. As oplossing hiervoor beveel

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

ons 'n stel seleksiekriteria aan wat korpusfrekwensie en konteksrelevansie kombineer. Laastens bespreek ons hoe loglêers wat gebruikersgedrag weerspieël, as 'n nuwe en baie betroubare empiriese bron vir lemmaseleksie vir 'n aanlyn Chinese aanleerderswoordeboek gebruik kan word.

Sleutelwoorde: CHINESE AANLEERDERSWOORDEBOEKE, LEMMASELEKSIE, SOSIALE KONTEKSTE, KORPUSFREKWENSIE, KONTEKSRELEVANSIE

1. Introduction

The last three decades have witnessed a rapidly increasing worldwide interest in learning Chinese. According to the recent statistics issued by Confucius Institute Headquarters (Hanban), as many as 100 million people around the world were learning Chinese as a foreign or second language by 2015 (Yang and Zhang 2017). This growing population of non-native learners of Chinese cannot be seen isolated from China's increased role and projection in the world. But apart from that, there may be many specific reasons why people decide to learn Chinese. They may have Chinese ancestors and aspire to re-establish the relations with their roots. They may live next to a Chinese speaking community in their own country and need Chinese as a means of communication. They may want to study Chinese because they plan to visit China and are interested in its rich history and culture. In 2014, about 26 million foreigners visited China (Liu 2016). Finally, some foreigners may already, for various reasons such as business, study or family, live in China for a shorter or longer period. In fact, Song (2013) estimates that in 2013 there were several million foreigners who were either registered as foreign residents or were staying in China, a number which will probably grow in the nearby future. Most of these people may wish to learn Chinese in order to manage in their daily and professional life.

All non-native learners of Chinese may demand specially designed dictionaries to assist the learning process, but their needs and expectations may not be exactly the same when they are living inside versus outside China. In this article, we will argue that this is a distinction to which too little attention has been paid, especially in practical dictionary making. We will therefore discuss Chinese learner's dictionaries with special focus on the needs which foreigners living in China may experience in terms of the required lemmata. As a conclusion, we will present some principles that can guide the selection of lemmata in an *online Chinese learner's dictionary* that takes advantage of the available technology. However, in order to put the discussion into perspective we will start with a brief excursion into the Western tradition of learner's dictionaries.

2. The western tradition

The term *learner's lexicography* was coined in Britain as a direct result of the

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

pioneering work of H. Palmer, M. West and A.S. Hornby and the publication of the first dictionaries specifically designed to assist foreign learners of English, cf. Cowie (1999). With the gradual development of English as a lingua franca in a large part of the world in the years following the Second World War, the monolingual English learner's dictionaries almost obtained a cult status and strongly influenced the making of similar dictionaries elsewhere in the world. In this period, learner's dictionaries saw the light in countries like Germany, France and Spain; cf. Hernández (1989), Zöfgen (1994), Wiegand (1998), Welker (2008), among others. The languages spoken in these three countries are all big languages in terms of the number of native speakers as well as the hundreds of thousands, if not millions, of learners interested in studying them not only inside but also outside the geographical areas where they are traditionally spoken. Most foreign learners of English, for instance, are studying this second language in their native countries. This situation differs dramatically from the situation in other European countries, like the Scandinavian countries, with a relatively small number of native speakers. In these countries, only a limited number of foreign learners have shown interest in learning the respective languages beyond their national borders, probably due to their limited communicative value at an international level. This also influenced the lexicographical terminology used in these countries. In Denmark, for instance, during many years there was no Danish equivalent to the English *learner's dictionary* until it was coined by Tarp (1999) and even today this new term (*lørnerordbog*) has not been used in any dictionary title. However, with the massive immigration of foreign workers starting in the 1960s, and later the arrival of many refugees escaping endless wars and natural disasters, all of them in need of learning the language used in their new country, a new type of dictionary began to see the light. This development resulted in a new and much more successful term being spontaneously coined, namely *immigrant's dictionary*, cf. Pálfi and Tarp (2009).

The immigrant's dictionary can be defined as a variant or subtype of learner's dictionaries specially adapted to the needs of immigrants and refugees living inside the geographical area where their new second language is spoken. It differs in various ways from the British *Big Five*, i.e. the prestigious learner's dictionaries published by Oxford, Collins Cobuild, Macmillan, Longman and Cambridge. Most immigrant's dictionaries are bilingual, either monoscopal or biscopal. The most emblematic of these dictionary projects is undoubtedly the Swedish *Lexins Svenska Lexicon* which is available both on paper (in a series of bilingual dictionaries) and online where it currently can be accessed from 20 different languages representing the biggest foreign language communities in Sweden, cf. Gellerstam (1999) and Hult (2016). The number of L2 lemmata in immigrant's dictionaries varies considerably but is generally much smaller than the ones treated in the monolingual English dictionaries mentioned above. These lemmata have frequently been selected according to criteria taking into account the very specific needs of the immigrants in their new life. In this regard, an immigrant's dictionary published in Spain in 2011

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

with the title *Bienvenidos* (Welcome) describes itself as the immigrants' and refugees' "first Spanish dictionary" and writes the following in its Introduction:

> It contains about 3 000 frequently used words that have been selected from a set of communicative situations that intend to cover the needs of daily life and to assist the development of the new speakers' linguistic competence (understanding and expressing themselves) and, in this way, to facilitate their full integration into social, work and family life (Martín 2011: ix).

The Spanish immigrant's dictionary is monolingual which, of course, limits its value, but it is worth noting that half of the dictionary consists of thematic tables illustrating the communicative situations mentioned, whereas the other half is a traditional alphabetically structured wordlist with definitions of each word. In this way it can be accessed both through the wordlist and the thematic tables. There is little doubt that this design makes it highly useful to most learners of Spanish at the very beginner's level, especially if it is used in combination with a Spanish language course. However, the limited vocabulary (as well as the title) suggests that its usefulness will be reduced proportionally with a growing proficiency level, and that it, after a few months, will have to be replaced by another type of learner's dictionary, preferable a bilingual one as argued by Lew and Adamska-Sałaciak (2015).

3. The discussion on learner's dictionaries in China

With a very few exceptions, the Chinese tradition of making dictionaries for foreign learners of Chinese started in the late 1970s. These dictionaries were, as a rule, based upon independent reflections by Chinese lexicographers and scholars, and they were only to a limited degree influenced by the traditions in other countries. If the increasing Chinese learning population all around the world is taken into consideration, it could be argued that the number of Chinese learner's dictionaries designed to serve foreign learners is rather limited. Wei, Geng and Wang (2014) have examined the dictionaries published in China from 1978 to 2008 and conclude that there are 21 Chinese learner's dictionaries for foreign learners. However, a study conducted by Wei and An (2014: 71-72) shows that about 45 Chinese dictionaries for foreign learners have been produced since 1980. Among these dictionaries can be mentioned the one helping foreigners with the Chinese Proficiency Test (HSK) (Liu 2000), the one illustrating the use of Chinese function words (Lü 1980), and the ones helping intermediate foreign learners learn Chinese (Lu and Lü 2006a; Shi and Wang 2011; Zheng 2009). These dictionaries are either monolingual or bilingual/ bilingualized, most of the latter with English as the auxiliary language (Zhang 2010: 33). Despite their respective focus and characteristics, these dictionaries are all aimed at assisting foreigners in learning Chinese.

In this light, it is rather awkward that a number of empirical studies reveal that foreign learners inside and outside China are generally not aware of the

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

existence of many of the Chinese learner's dictionaries published in China and, hence, seldom use them (Liu 2014; Hao and Wang 2013; Xie and Li 2012; Yang 2015). Many of these dictionaries gather dust in libraries and are mainly used for research purposes (Jin 2015; Liu 2014). A study carried out by Yang (2015) shows that only about 9 percent of the foreign learners of Chinese, even the ones studying in China, use Chinese learner's dictionaries published in China. On the other hand, nearly 95 percent of the Chinese learners of English are using one of the *Big Five* British learner's dictionaries (Liu 2014). Facing such an unfortunate status with regard to learner's dictionaries, Lu and Lü (2006b) criticize that many of the so-called Chinese learner's dictionaries are nothing but the reduced versions of distinguished dictionaries designed for native Chinese, like the *Xinhua Dictionary* and *Modern Chinese Dictionary*. Such criticism has been echoed by many Chinese lexicographers (Cai 2011; Li 2013; Liu 2014; Wang 2009; Yang 2016).

The sharp contrast between the status of Chinese and English learner's dictionaries has spurred wide discussion among Chinese lexicographers on the concept, design and principles to be used in the compilation of learner's dictionaries that are specifically aimed at foreign learners of Chinese (Jin 2015; Wang 2009; Yang 2016). In an overview of the studies conducted into Chinese learner's dictionaries between 1984 and 2013, Jin (2015: 35) concludes that during the past 30 years the research has mainly focused on describing and evaluating dictionary articles from the point of view of linguistics, although there is a certain tendency to shift the focus to their usefulness in terms of the target users' needs. There is, however, a manifest lack of research on the actual needs of foreign learners as dictionary users and on the integration of modern information technology into the conception and compilation of Chinese learner's dictionaries (Jin 2015: 36). In a situation where there is an increased focus on dictionary users and where the English learner's dictionaries are rapidly moving from the printed to the digital media (Rundell 2015), Chinese learner's dictionaries to a great extent still stay in a comfort zone.

A few studies have touched upon the issue of foreign learners' actual needs in the process of planning and compiling learner's dictionaries, but without providing further specifications. Zheng (2004: 92-93) points out that it is necessary to make a distinction between the foreign learners of Chinese living in China and those who study this language in other geographic areas of the world. However, he does not explicitly elaborate on the different lexicographical needs which these two groups of foreign learners may experience, or how dictionaries should respond to such needs. Yang (2016) proposes four principles to guide the compilation of a Chinese learner's dictionary for foreign learners, namely intelligibility, utility, comprehensiveness, and explicitness. These principles, which are formulated at a very high level of abstraction, involve all the data selected for the planned dictionary and require reasoning and step-wise specifications in theory and practice. It is always easy to state that foreign learners' needs should be attended to in academic research, but

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

fragmented suggestions and ideas that are too abstract are not sufficient to plan a modern high-quality learner's dictionary in the real sense of the word. Efforts should rather be made to develop a *coherent framework* which can guarantee the production of a learner's dictionary that responds to the actual needs and expectations of the foreseen dictionary users. It is, hence, imperative to have a clear understanding of the concept of a learner's dictionary in terms of foreign language learning.

4. An unnoticed dictionary: ***My Chinese Picture Dictionary***

We will now have a brief look at a dictionary that was published in 2008, namely *My Chinese Picture Dictionary* (Wu 2008). In the scholarly discussion of the principles that should guide the design of Chinese dictionaries for foreign learners, this dictionary goes almost unnoticed. It is, for instance, not mentioned by Wei et al. (2014) who claim to offer the most comprehensive overview of dictionaries published in China between 1978 and 2008, and neither is it included in a recent overview study conducted by Wei and An (2014). This rather unnoticed existence is surprising, inasmuch as the dictionary reflects a new and different approach to Chinese learners' lexicography.

My Chinese Picture Dictionary consists of a thematic section which makes up the bulk of the dictionary as well as two indexes in English and Chinese Pinyin, respectively. The vocabulary treated is structured in 15 main themes, each further subdivided into a number of topical units. There are a total of 142 such units which are all represented in graphic tables covering various aspects of daily life such as personal information, family, school, work, shopping, dining, hospital, transportation and travel, etc. (Figure 1 provides an example of how the thematic units are represented in the dictionary). Each of the figures consists of an illustration where the words representing the different phenomena are written with Chinese characters as well as in Chinese Pinyin and English. The figures can be accessed either through the list of thematic content in the front matter of the dictionary or through one of the two indexes in the back matter where the English and Chinese Pinyin words treated in the figures are organized alphabetically.

As can be seen, in its overall structure *My Chinese Picture Dictionary* has many similarities with the Spanish immigrant's dictionary *Bienvenidos* mentioned above, as both make an endeavour to serve foreign learners' actual needs in various social situations in which they may find themselves in China or Spain. However, compared to the wordlist in its Spanish counterpart, the two wordlists, or indexes, in *My Chinese Picture Dictionary* are much more primitive in the sense that they do not offer definitions or any kind of grammatical data, not even part of speech. In addition, the overwhelming majority of selected words are nouns whereas there are few verbs and adjectives and no function words. It goes without saying that these problems, and others which will be identified in the following discussion, reduce its usefulness for foreigners

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

living in China despite its innovative approach to Chinese learners' lexicography.

5. The concept of a learner's dictionary

Foreign or second language (L2) learning is a complex process and learner's dictionaries are conceived to assist learners in different situations or contexts related to this learning process. The situations in which foreign learners turn to dictionaries are generally of a communicative or cognitive character as defined by the lexicographical *Function Theory* which will constitute the theoretical framework for the following reflections, cf. Tarp (2008). Communicative situations include *text reception* in L2, *text production* in L2, as well as *translation* from L1 into L2 or vice versa, whereas the study and assimilation of L2 *vocabulary* or *grammar* are the most relevant cognitive situations. These learning situations may vary according to the chosen didactic methods. As stated by Tarp (2004), the great challenge to learner's lexicography is to conceive and compile dictionaries that can assist learners in as many aspects of the language learning process as possible. Hence, users' needs, which may occur in the specific types of user situation, should be the starting point for learner's lexicography.

The focus on foreign learners' needs is time-honoured in learner's lexicography. Learner's lexicographic needs occur in concrete situations and are basically determined by these situations and simultaneously shaped by the learners' personal characteristics as user types. A number of variables have been identified to define the profile of foreign learners and investigate how they influence learners' lexicographic needs in concrete situations. Among these variables, the most important and relevant variables are foreign learners' proficiency in L2, native language and cultural background, age and learning circumstances (Tarp 2008). An advanced learner can in most cases resort to L2 definitions to solve his or her comprehension problems, whereas a beginner may need L1 equivalents or explanations to solve the same type of problem. A Thai learner of Chinese may have no difficulty in identifying water spinach, but a Danish speaker may wonder what it is. In short, a learner can have different lexicographic needs in different user situations and different learners in the same user situation could differ in their needs.

The purpose of a learner's dictionary is to satisfy its target users' needs. In this respect, the advent of the new information and communication technologies can help lexicography move closer than ever before to providing personalized and individualized service as claimed by Rundell (2010) and Tarp (2011), among others, a goal that can only be fully achieved in context-aware integrated information tools like e-readers and writing assistants, cf. Tarp et al. (2017). Hence, it is imperative that lexicographers planning a new dictionary project should have a coherent understanding of a homogenous group of learners' needs in specific user situations as well as their relevant characteristics as users, including their age (adult or child), first language, cultural background and L2 proficiency level. Without reference to the target users' specific

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

needs, the discussion on defining styles, examples and other type of data contained in a dictionary will be fruitless and futile in the end.

Conceiving a learner's dictionary is a complex process and involves a holistic understanding of the functions of the concerned dictionary in terms of its users and their needs in particular situations or social contexts. This article will focus on expounding the issue of lemma selection for Chinese learner's dictionaries aimed at assisting L2 (Chinese) text production. A distinction is made between foreign learners living in and outside China, as the general circumstances in which a foreign language is learned constitute an important variable that influences the learners' lexicographic needs in specific situations.

6. Lemma selection in the conception of Chinese learner's dictionaries

The issue of lemma selection has always been central in learner's lexicography and Chinese learner's dictionaries are no exception in this regard. The main questions concerning lemma selection for learner's dictionaries are summarized by Tarp (2008: 174) as follows:

1. How big should the lemma stock in learner's dictionaries be?
2. Which criteria and principles should guide lemma selection?
3. Which empirical basis should lemma selection be based on?

With these three questions in mind, in the following section we will briefly examine the practice of lemma selection in some major Chinese learner's dictionaries for foreigners. Frequent references will be made to *My Chinese Picture Dictionary*, given its unique organization of lemmata according to social contexts that foreign learners would encounter when they live in China. Based on the analysis, proposals will be made to respond to the three questions.

6.1 The present lexicographic practice with regard to lemma selection

There are two official lists of characters which are most frequently used as empirical basis for lemma selection in dictionaries for foreign learners of Chinese. The first one is *The Outline of Chinese Vocabulary and Chinese Character Level* published by the National Office for Teaching Chinese as a Foreign Language (2001) and includes 8,822 Chinese words falling into four language levels. This word list is used as vocabulary curriculum for the Chinese Proficiency Test (HSK), an international standardized test of Chinese language proficiency which assesses non-native Chinese speakers' ability to use Chinese in their daily, academic and professional lives. The other official list of characters is the *List of Frequently Used Characters in Modern Chinese* elaborated by the State Language Commission (1988). This list contains a total of 3,500 Chinese characters structured in two sections, one with the 2,500 most frequent characters and another containing the 1,000 characters that come next in frequency.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

With reference to the above-mentioned official lists of characters, *The Commercial Press Learner's Dictionary of Contemporary Chinese* (Lu and Lü 2006a), considered by Jiang (2006) to be one of the best Chinese learner's dictionaries, has a lemma stock of 2,400 Chinese characters to which should be added about 10,000 multi-character words selected as sublemmata. *A Learner's Chinese Dictionary* (Zheng 2009) includes 3,000 characters as lemmata and an additional 32,000 multi-character words and expressions presented as sublemmata. *A Dictionary of Chinese Usage* (Liu 2000) offers 8,822 single-character and multi-character words as lemmata, i.e. exactly the same amount as *The Outline of Chinese Vocabulary and Chinese Character Level* referred to above.

Wang and Liu (2014) have examined the lemma stock in eight major Chinese learner's dictionaries for foreign learners. The two authors show how these dictionaries generally claim to select lemmata with reference to the above official lists of characters but are quite divergent regarding the number of lemmata actually included. The philosophy underpinning the principles of sticking to the official teaching curriculum seems to be the belief that internalizing the knowledge of the basic vocabulary is the stepping stone for learning Chinese. This philosophy seems, to a great extent, to ignore the fact that foreign learners' needs for Chinese vocabulary may arise in authentic social situations rather than in educational contexts, especially when they are living in China.

The extensive exposure to various aspects of life in China prompts foreigners to demand a wide range of vocabulary specific to their personal situations. They need to go to local markets to buy food and vegetables, and they may also need to deal with residential issues in the local police station. Quite a number of Chinese characters relevant to realistic social situations may be ranked low-frequency in the language-teaching curriculum.

Considering the distance between Chinese and other languages, there are several words and expressions describing common social phenomena typical for Chinese society, for instance the complex system of addressing forms. These phenomena may be absent in other cultures and language communities, but does this mean that foreign learners should ignore the corresponding vocabulary, since it is not part of their language and culture? Or should they just learn it for receptive purpose, since they may never have to use these words? In order to answer these questions, it is necessary to consider the geographic and linguistic communities where the targeted foreign learners live, when it comes to selecting lemmas for Chinese dictionaries targeting this user segment. Although part of the academic literature emphasizes the importance of considering foreign learners' daily life and study in China when selecting lemmata, no further and detailed elaborations on this challenge have yet been made (Li 2013: 36; Wang 2009: 569; Wang and Liu 2014: 73; Yang 2016: 47). Wang and Liu (2014: 73), for instance, explicitly state that the core vocabulary in the official curricula cannot be incorporated directly as lemmata in learner's dictionaries and that lemma selection requires practical experience and expertise from the lexicographers. However, the abstract selection principles of frequency, common errors, and levels of core

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

vocabulary suggested by these two authors do not solve the problem. It is therefore imperative to develop practical methods that are easier to handle.

The publication of *My Chinese Picture Dictionary* seems to be a practical response to the abstract discussion on lemma selection for Chinese learner's dictionaries. The strength of this dictionary is the presentation of 4,200 lemmata organized in 15 thematic social contexts defined in the dictionary (See figures 1 and 2 in this regard). The thematic organization of lemmata inevitably challenges the rigid levels of the wordlists designated in the official curriculum, although the preface states that the '15 thematic units are categorized according to the *International Curriculum for Chinese Language Education* published by the Office of Chinese Language Council International' (Wu 2008).

The fact that all the words included in *My Chinese Picture Dictionary* are illustrated with pictures makes it easy for non-native speakers to identify the referents and associate the vocabulary with things and phenomena in their social life. This is especially helpful to newly arriving foreigners who want to learn Chinese and become familiar with Chinese culture. Disregarding the improper translations in some cases, the bilingual dimension with its inclusion of English equivalents attached to the presented Chinese words also lowers the threshold to use this dictionary, at least for the users who are native speakers of English or have a certain proficiency level in this language. The important question of access to the words treated in the illustrations is solved by the appended English and Chinese indexes.

However, the lack of a clear definition of users and functions of *My Chinese Picture Dictionary* results in many problems, which to a certain extent reduces its value in practical use. Consequently, we will briefly discuss some of these problems because of their relevance for our vision of a Chinese learner's dictionary for foreigners living in China.

First of all, the social contexts treated in *My Chinese Picture Dictionary* do not always seem to be relevant to the envisaged user group. Some contexts like *Construction Sites* and *Martial Arts* are quite distant from most foreign learners' daily life and the words grouped under these topics are therefore not the most relevant to their actual needs. For instance, twenty-four words, both single-character and multi-character, are listed in connection with *Martial Arts*, but one may wonder in what social contexts foreign learners will have contact with the specific vocabulary describing the movements in martial arts like 二指禅 (*two-finger meditation*) 形意拳 (*intent-shaped fist*), etc. Even average Chinese people seldom have contact with these moves of martial arts in their daily life.

Secondly, the depiction of the social contexts tends to be skewed in the dictionary. For instance, people generally go to the local police station to deal with civil issues, like applying for residential permission or registration, but the words provided under *Police Station* mainly focus on crime and violence, such as 谋杀 (*to murder*), 绑架 (*to kidnap*), 手铐 (*handcuffs*), and 警棍 (*police baton*), and therefore deviate from the daily routines in China; cf. Figure 2. Moreover, the number of items listed in some thematic contexts seems to be too modest in

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

comparison with the expected user needs. This is, among others, the case with the fruit and vegetables shown under *Nutrition* as well as the food referred to in connection with *Western Restaurant*.

Thirdly, the words and terms presented in some thematic units like *Hospital* and *Sky* appear to be too specialized even for native Chinese speakers. Anatomical terms like 上腔静脉 (*superior vena cava*), 腓骨 (*fibula*) or 趾骨 (*phalanges*) are medical terms and unfamiliar to people who are laymen within this field. The same holds true for technical terms like 对流层 (*troposphere*), 同温层 (*stratosphere*), 积雨云 (*cumulonimbus cloud*) which are rather challenging for laymen and rarely appear in daily communication.

In the fourth instance, pictorial illustrations in general greatly limit the classes of the words included in the dictionary as not all words, particularly verbs, adjectives and adverbs, can be illustrated in a simple and easily understandable way. As a result, most vocabulary presented in *My Chinese Picture Dictionary* are nouns, of which multi-character words make up the majority. This may block foreign learners' mastery of the single Chinese characters, as Chinese characters are independent meaning units and very productive in combination usage.

Lastly, but not less important, the marking of word classes is missing in the dictionary. This leaves foreign learners almost helpless in terms of identifying part of speech of the presented lemmata. The lack of grammatical data and collocations also greatly reduces the value of *My Chinese Picture Dictionary* in communicative situations. As the dictionary claims to demonstrate different Chinese social contexts by means of pictures, it is not easy to understand why the 'new vocabulary' should be 'useful not only in Chinese culture but also in Western societies' (cf. Preface in Wu 2008). Given the limited vocabulary illustrated by pictures in typical Chinese settings, the dictionary seems to be too ambitious when it comes to 'help students to learn and use Chinese words to talk about different cultures' (cf. Preface in Wu 2008).

In summary, *My Chinese Picture Dictionary* opens new perspectives for lemma selection in Chinese dictionaries for foreign learners living in China. It seems, nonetheless, that the practical method of selecting lemmata according to specific social situations requires further reflections and refinements in order to overcome the problems identified above. The selected themes do not, to a large extent, represent the most typical social contexts in real life. The dictionary describes its target users as 'students', a generic term that tends to blur the profile of the users. The critical remarks put forward in this and the previous sections suggest that the principles applied to select lemmata are not sufficiently well-considered to achieve the desired high-quality learner's dictionary for non-native speakers living in China.

6.2 Some proposals

The analysis in the previous sections indicates that two main criteria have been

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

used to select lemmata for existing Chinese learner's dictionaries, i.e. frequency based on corpora, and something that could be called context-relevance. As none of these criteria applied separately seem to fully meet the needs of adult foreigners living in China, this article proposes that the selection of lemmata for an *online Chinese learner's dictionary* targeted at this audience should follow a combination of the two criteria mentioned, namely *frequency* and *relevance*. The application of these two criteria will be discussed in the following sections.

First, it goes without saying that the criterion of frequency provides solid empirical evidence for the occurrence of a word in actual language use. Using corpora to assist lemma selection is widely practiced in dictionary-making; cf. Rundell and Kilgarriff (2011), Hanks (2012), among many others. As shown in the previous sections, most Chinese learner's dictionaries claim to make reference to the national teaching curriculum, various corpora or frequency dictionaries containing the most frequent Chinese characters and words. Against the rapid development of corpora and the ever-increasing size of these, the two frequency dictionaries of modern Chinese words published in 1986 and 1990 tend to be outdated. The current representative corpora of modern Chinese are generally organized and constructed by national institutions or universities, like the corpus of modern Chinese (50 million characters), the CCL corpus (307 million characters), the corpus of Chinese texts by international students in Beijing (1 million characters), the interlanguage corpus of Chinese from global learners (50 million characters) and the HSK dynamic composition corpus (4 million characters) (Zhang 2015).

The availability of these corpora undoubtedly provides a huge amount of data, which offers a solid empirical basis for the frequency information about the candidate lemmata for Chinese learner's dictionaries. Statistic measures can be used to identify and select the words which remain stable in terms of their corpus coverage, their time sensitivity and diachronic classification. The fact that such words have a stable occurrence in the corpora indicates that they express vigour and versatility. Finally, the national teaching curriculum can also be considered a reliable source for candidate lemmata.

However, corpus frequency cannot be taken as the only criterion to select or exclude a lemma, despite the essential role played by the statistical significance in lemma selection. There is possible divergence between the high-frequency words in corpora and the words required in specific social situations as it has been argued by Guo et al. (2014) as well as Zhang (2015). Few foreigners will read a dictionary from one end to another in order to learn Chinese. Often, they are driven to Chinese dictionaries by practical problems in specific communicative situations and acquire the knowledge of Chinese from dictionaries incidentally. The discussion in the previous chapter indicated that the vocabulary provided under topical units such as *police station* and *hospital* may not be the most relevant to foreigners living in China whereas other words may be much more relevant to them in these contexts. The criterion of relevance should therefore be applied as an additional criterion when it comes to determining

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

whether a Chinese word should be included in a Chinese learner's dictionary.

The criterion of relevance is referring to the likelihood of a word occurring in one of the social situations which foreigners living in China most typically encounter in their day to day life. Although not among the most frequent words in a corpus, such a word may nonetheless be frequent and typical in the mentioned situations and therefore relevant as a lemma candidate in a learner's dictionary for this specific segment of users. As illustrated above, the likelihood of foreigners needing the vocabulary related to civilian services is much higher than their needs for the words about violent crimes in Chinese police stations. Hence, the vocabulary related to civilian services should be prioritized in the lemma selection for foreign beginner learners in China without ignoring that related to various sorts of crime. The same applies to the high-culture vocabulary like the words related to martial arts or other specialized subject fields. This does not mean that Chinese learner's dictionaries should limit the lemmata to low-culture survival words. It is to be understood that foreign language learning is a continuum and foreign learners' needs for vocabulary vary in numbers and scopes in the continuum of learning Chinese, starting from the survival needs and advancing to the needs for specialized and high-culture vocabulary.

In short, the more typical a word is in social situations in which foreigners learning Chinese in China frequently find themselves, the more often they will have contact with it although it may not display the same degree of frequency in a corpus. In order to identify words often appearing in relevant social contexts, it is first of all requisite to determine the respective contexts. The Council of Europe (2001) defines four social domains in the Common European Framework of Reference for Languages (CEFR): personal, public, educational and occupational. These domains can also be used to determine the various social situations typical of Chinese society. With reference to the CEFR framework, Tseng (2014: 27) has further specified 12 situations: personal data, work, education, housing, family and environment, daily routines, relaxation, interpersonal relationship, travelling, body and health, shopping, and food. Each of these situations or main themes can be further subdivided into a number of topical units as was the case with the thematic tables in *My Chinese Picture Dictionary*. It may be assumed that the words typically occurring in these contexts are relevant to foreigners living in China, even if they rank low in the general corpora. These words should therefore be selected as lemmata, for instance based on a collection of texts covering each of the situations in question and using the criterion of relevance.

Finally, the size of the lemma stock is subjective to the proficiency stages through which foreign-language learning develops. A learner's dictionary can be designed to assist its users in the first phase of foreign-language learning or to follow them until a more advanced proficiency level. There is therefore no absolute number of lemmata that can be recommended for a learner's dictionary. It all depends on its purpose and specific user segment. If the dictionary is

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

primarily conceived to assist learners at the beginner's level with production of Chinese text, then a reduced vocabulary may satisfy the learners' needs in this respect. However, if the dictionary is also supposed to cover the learners' needs in relation to text reception, and if the learners are living in China and exposed to Chinese every day, then a much bigger vocabulary is required even for beginners.

Among Chinese scholars there are various proposals as to the size of the lemma stock relevant to foreign learners of Chinese. Li (1999: 58), for instance, suggests that the national teaching curriculum should cover 10,000 to 12,000 words in order to meet foreigners' communicative needs in Chinese. Guo et al. (2014: 12) propose that 13,000 words could decently satisfy foreigners' needs. Zhang (2015) proposes 4,000 words for Chinese learner's dictionaries targeted at foreign beginners, an additional 6,000 for intermediate learners and a further 10,000 for advanced learners of Chinese. In total, the Chinese learner's dictionary should include about 20,000 words according to Zhang (2015: 42). As a starting point, the proposed size of 20,000 lemmata seems feasible and reasonable in a printed dictionary for foreign learners of Chinese. However, when it comes to future online dictionaries the problem may have to be approached in a different way as we will see below.

7. Perspectives

Dictionaries are human-made tools designed to assist possible users looking for information in order to solve different types of problem, as they have been defined in the *Routledge Handbook of Lexicography* by Tarp (2018). This suggests that people consult dictionaries when they have specific information needs and that the dictionaries should contain the corresponding lexicographical data, including the relevant lemmata, whereas the inclusion of superfluous data and lemmata can be regarded as a waste of time and money. In this respect, the best way to satisfy user needs in terms of lemma stock is to include the words which users actually look up. Until recently, lexicographers have generally only been able to guess the words that are relevant to their specific users. They have therefore resorted to indirect selection criteria like corpus frequency and context relevance as discussed above. At present, these selection criteria may still be recommended for dictionaries designed to assist foreign learners of Chinese living in China. However, these criteria are about to change radically in the nearby future as a much more reliable empirical basis is being developed, namely *logfiles* which trace user behaviour in dictionary consultation. Logfiles can be used either as a supplementary (see below) or as a primary source for lemma selection. When we speak about "radical change" we refer to the latter, i.e. the use of logfiles as the primary source for lemma selection which *totally replaces the corpus* as the main empirical basis for this purpose.

Once a high-quality online dictionary has been produced and used for some time, such logfiles will provide reliable evidence of the items which dic-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

tionary user actually look up. Studies of logfiles show that there is not a complete correspondence between the most frequent words in a corpus and the words most frequently looked up in dictionaries. Bergenholtz and Norddahl (2012), for instance, have shown that some Danish words, which are very frequent in the corpus, are seldom or never looked up in a big online dictionary with more than hundred thousand lemmata whereas other words with a low corpus-occurrence are frequently consulted by the users after a total of more than 20 million lookups.

There is little doubt that logfiles will increasingly be used as an empirical basis for the selection of specific lexicographical data categories such as lemmata. In this respect, the frequency of the words appearing in the logfiles, or just the appearance itself, will become the basic criterion for lemma selection as it is currently the case in the Spanish–English–Spanish *Diccionarios Valladolid-Uva* (under production) which do not use corpora at all but only logfiles as the primary empirical basis (personal information). However, before logfiles can be used as a reliable empirical basis for future Chinese learner's dictionaries, a number of requirements have to be fulfilled. First of all, at least one high-quality learner's dictionary designed from scratch for the digital media should be produced and made available online. Then a statistically significant number of lookups should have been made, for instance 20 million. In addition, if the new dictionary is planned to serve foreigners learning Chinese in China, the logfiles used as empirical basis should make allowance for a distinction between users (learners) living inside and outside China. Finally, and in order to make an even better product, it should be possible to distinguish between lookups related to text production and text reception, respectively. In this respect, some lemmata could be given extra treatment with the inclusion of additional data categories in order to assist text production whereas others could focus on explanations with a view to supporting text reception. This would save time for lexicographers and result in a more focused lexicographical product.

Until this nearby future becomes reality, the combination of the criteria of frequency and relevance discussed above can be recommended when lemmata are selected to compile new digital learner's dictionaries for foreigners learning Chinese in China. But even so, the existing technology already allows for a gradual transition to new selection criteria as well as new publication methods; cf. Bergenholtz and Johnsen (2005), De Schryver (2013), Trap-Jensen et al. (2014), among others. The possibility of constant updating in the online media allows for a flexible publication process where the first version (or "edition") of a web-based dictionary can be made available to its users when a certain percentage of articles covering the most frequent and relevant lemmata have been finished, for instance, 20–30 percent. This could for example be the 4,000 words which Zhang (2015) recommends for a Chinese dictionary for foreign learners at the beginner's level. Once this number of articles has been completed, the lexicographers can continue working in two directions: (1) follow the established work plan and compile articles based on the selected lemma stock, and

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

(2) simultaneously study the logfiles (on a daily or weekly basis), detect words looked up by the users but still not treated in the dictionary and straightaway compile the corresponding dictionary articles, whether or not the words in question are included in the originally selected lemma stock. Such a methodological procedure will undoubtedly put real user needs at the centre of the lexicographical compilation process.

Finally, it can be said that the new disruptive computer and information technologies open new horizons to lexicography as a millennial cultural practice. Modern lexicographers — and publishers — should take full advantage of these technologies and adapt their methods accordingly. Lemma selection, from being a once-and-for-all decision in printed dictionaries, has been transformed into a dynamic endeavour which, in principle, can continue for years even after the first version of an online dictionary has been published. Continuous refinement and adaptation to the users' real needs should be the guiding principle also for online Chinese learner's dictionaries aimed at assisting non-native speakers living and learning Chinese in China.

Acknowledgements

Thanks are due to the China Scholarship Council for funding the project (Grant No. 201606435015) as well as to the Spanish Ministry of Economy and Competitiveness for funding the project "La Teoría Funcional de la Lexicografía: Diseño y Construcción de Diccionarios de Internet" (Ref. FFI2014-52462-P) in which this article is theoretically embedded.

References

Bergenholtz, H. and B. Norddahl. 2012. Ordbogsartikler som ingen læser. *LexicoNordica* 19: 207-223.

Bergenholtz, H. and M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes, Journal of Linguistics* 34: 117-141.

Cai, Y.Q. 2011. The User-friendly Principle in the Compilation of Dictionary for Chinese Language Learning. *Lexicographical Studies* 2: 67-77.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Available at: www.coe.int/lang. Accessed 9 June 2017.

Cowie, A.P. 1999. Learners' Dictionaries in a Historical and a Theoretical Perspective. Herbst, T. and K. Popp. (Eds.). 1999. *The Perfect Learners' Dictionary (?):* 3-13. Tübingen: Max Niemeyer.

De Schryver, G.-M. 2013. The Concept of Simultaneous Feedback. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 548-556. Berlin: Walter de Gruyter.

Gellerstam, M. 1999. LEXIN — lexikon för invandrare. *LexicoNordica* 6: 3-18.

Guo, X.L., X.S. Ma and K.T. Li. 2014. Comparative Study of Chinese Characters and Words Based on Language Situation in China. *Journal of Beihua University (Social Sciences)* 15(3): 10-13.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.

Hao, Y.X. and Z.J. Wang. 2013. A Study on the Requirements of Learners in L1 Environment for Chinese Language Learners' Dictionary. *TCSOL Studies* 3: 50-57.

Hernández, H.H. 1989. *Los diccionarios de orientación escolar. Contribución al estudio de la lexicografía monolingüe española.* Tübingen: Max Niemeyer.

Hult, A-K. 2016. Ordboksanvändning på nätet. En undersökning av användningen av Lexins svenska lexicon. Gothenburg: Institutionen för svenska språket.

Jiang, L.Sh. 2006. Preface. Lu, J.J. and W.H. Lü (Eds.). 2006. *The Commercial Press Learner's Dictionary of Contemporary Chinese*. Beijing: The Commercial Press.

Jin, P.P. 2015. Thirty-year Researches on Chinese Learner's Dictionaries. *The Journal of Yunnan Normal University: Foreign Language Teaching and Research* 13(3): 27-37.

Lew, R. and A. Adamska-Sałaciak. 2015. A Case for Bilingual Learners' Dictionaries. *ELT Journal* 69(1): 47-57.

Li, Q.H. 1999. The Issue of Quantity of Vocabulary on the Outline of Chinese. *Language Teaching and Research* 1: 50-59.

Li, Y. 2013. On the Compilation of General-purpose Chinese Dictionaries for Foreign Learners of Chinese. *Lexicographical Studies* 5: 34-39.

Liu, L.L. 2000. *A Dictionary of Chinese Usage: 8000 Words Chinese Proficiency Test Vocabulary Guideline*. Beijing: Beijing Language and Culture University Press.

Liu, S.T. 2014. Systematic Research on the Structural Features of Chinese Learner's Dictionaries for Foreigners. *Social Sciences in China,* 17 February: A8.

Liu, Y. 2016. It is Not Easy to Find Jobs in China. *People's Daily Overseas Edition*, October 17, 2016. Retrieved from http://paper.people.com.cn/rmrbhwb/html/2016-10/17/content_1719150.htm.

Lu, J.J. and W.H. Lü. 2006a. *The Commercial Press Learner's Dictionary of Contemporary Chinese.* Beijing: The Commercial Press.

Lu, J.J. and W.H. Lü. 2006b. The Compilation of a Monolingual Learner's Dictionary of Chinese as a Foreign Language: A Venture and Some Considerations. *Chinese Teaching in the World* 1: 59-69.

Lü, S.H. 1980. *800 Words of Modern Chinese*. Beijing: The Commercial Press.

Martín, I.L. (Ed.). 2011. *Bienvenidos. El primer diccionario de español.* Madrid: Octaedro.

National Office for Teaching Chinese as a Foreign Language. 2001. *The Outline of Chinese Vocabulary and Chinese Character Level*. Beijing: Jingji Kexue Press.

Pálfi, L.-L. and S. Tarp. 2009. Lernerlexikographie in Skandinavien — Entwicklung, Kritik und Vorschläge. *Lexicographica* 25: 135-154.

Rundell, M. 2010. What Future for the Learner's Dictionary? Kernerman, I.J. and P. Bogaards. (Eds.). 2010. *English Learners' Dictionaries at the DSNA 2009*: 169-175. Jerusalem: Kdictionaries.

Rundell, M. 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.

Rundell, M. and A. Kilgarriff. 2011. Automating the Creation of Dictionaries: Where Will It All End? Meunier, F., S. de Cock, G. Gilquin and M. Paquot. (Eds.). 2011. *A Taste for Corpora. In Honour of Sylviane Granger*: 257-281. Amsterdam/Philadelphia: John Benjamins.

Shi, G.H. and S.X. Wang. 2011. *A Chinese Dictionary for Learners and Teachers*. Beijing: The Commercial Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

Song, Q.C. 2013. A Demographic Sociological Analysis of Foreigners and Hong Kong, Macao, and Taiwan Residents in Mainland China. *The Journal of Shandong University: Philosophy and Social Sciences* 2: 89-99.

State Language Commission of China. 1988. *List of Frequently Used Characters in Modern Chinese*. Beijing: Language & Culture Press.

Tarp, S. 1999. Lørnerordbøger for indvandrere og andet godtfolk. *LexicoNordica* 6: 107-132.

Tarp, S. 2004. Basic Problems of Learner's Lexicography. *Lexikos* 14: 222-252.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.

Tarp, S. 2011. Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. Fuertes-Olivera, P.A. and H Bergenholtz (Eds.). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 54-70. London/New York: Continuum.

Tarp, S. 2018. The Concept of Dictionary. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 237-249. London: Routledge.

Tarp, S., K. Fisker and P. Sepstrup. 2017. L2 Writing Assistants and Context-Aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.

Trap-Jensen, L., H. Lorentzen and N.H. Sørensen. 2014. An Odd Couple — Corpus Frequency and Look-up Frequency: What Relationship? *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 2(2): 94-113. https://doi.org/10.4312/slo2.0.2014.2.94-113. Accessed 4 July 2018.

Tseng, W.H. 2014. Classification on Chinese 8,000 Vocabulary. *Teaching Chinese as a Second Language* 16: 23-35.

Wang, H.Y. 2009. What Chinese Dictionaries are Expected by Foreign Learners. *Chinese Teaching in the World* 4: 567-575.

Wang, X. and S.D. Liu. 2014. The Study on Lemma Selection for Export-oriented Learner's Dictionaries. *Ludong University Journal: Philosophy and Social Sciences* 31(3): 69-74.

Wei, J.X. and H.L. An. 2014. The Review of Compilation and Research on Export-oriented Chinese Learning Dictionary. *Journal of Guangdong Ocean University* 34(5): 70-75.

Wei, X.Q., Y.D. Geng and D.B. Wang. 2014. *Lexicography in China (1978–2008)*. Beijing: The Commercial Press.

Welker, H.A. 2008. *Panorama geral da lexicografia pedagógica*. Brasilia: Thesaurus Editora.

Wiegand, H.E. (Ed.). 1998. *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von* Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Tübingen: Niemeyer.

Wu, Y.M. (Ed.). 2008. *My Chinese Picture Dictionary*. Beijing: The Commercial Press.

Xie, H.J. and L. Li. 2012. An Investigation into the Publication and Using Condition of CFL Chinese Learner's Dictionaries. *Ludong University Journal: Philosophy and Social Sciences* 29(1): 62-68.

Yang, H. 2015. *The Investigation of Use of the Chinese Dictionary Status Quo of International Students in China*. Unpublished Master's thesis, Chongqing Normal University, Chongqing, China.

Yang, J.H. 2016. On the Four Principles of Compiling Chinese Dictionaries for Foreign Learners. *Lexicographical Studies* 1: 45-51.

Yang, N. and X.D. Zhang. 2017. The Mania of Learning Chinese: A Bonus to Overseas Chinese. *People's Daily Overseas Edition*, April 17, 2017. Retrieved from http://world.people.com.cn/n1/2017/0417/c1002-29214843.html.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

Zhang, B.L. 2015. Compiling Design of Chinese as A Second Language Learning Dictionary Series. *Bilingual Education Studies* 2(1): 37-44.

Zhang, X.M. 2010. An Exploratory Discussion on Chinese Learners' Dictionaries in China. *Lexicographical Studies* 3: 27-37.

Zheng, D.O. 2004. On Chinese Learner's Dictionaries for Foreigners. *Chinese Teaching in the World* 4: 85-94.

Zheng, S.P. 2009. *A Learner's Chinese Dictionary*. Beijing: Foreign Language Teaching and Research Press.

Zöfgen, E. 1994. *Lernerwörterbücher in Theorie und Praxis. Ein Beitrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen.* Tübingen: Niemeyer.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

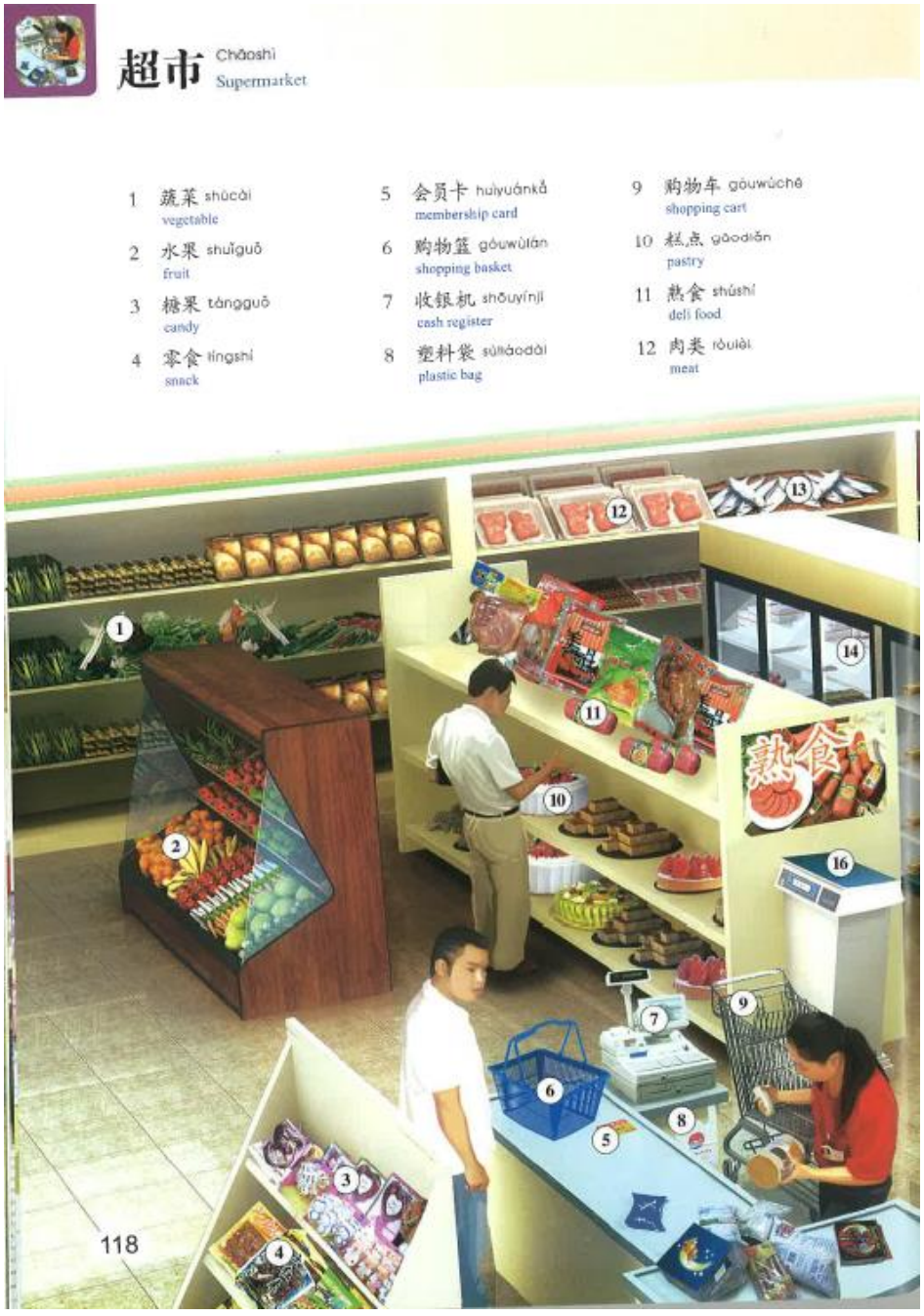

Figure 1: Illustration from *My Chinese Picture Dictionary*: Supermarket

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1470

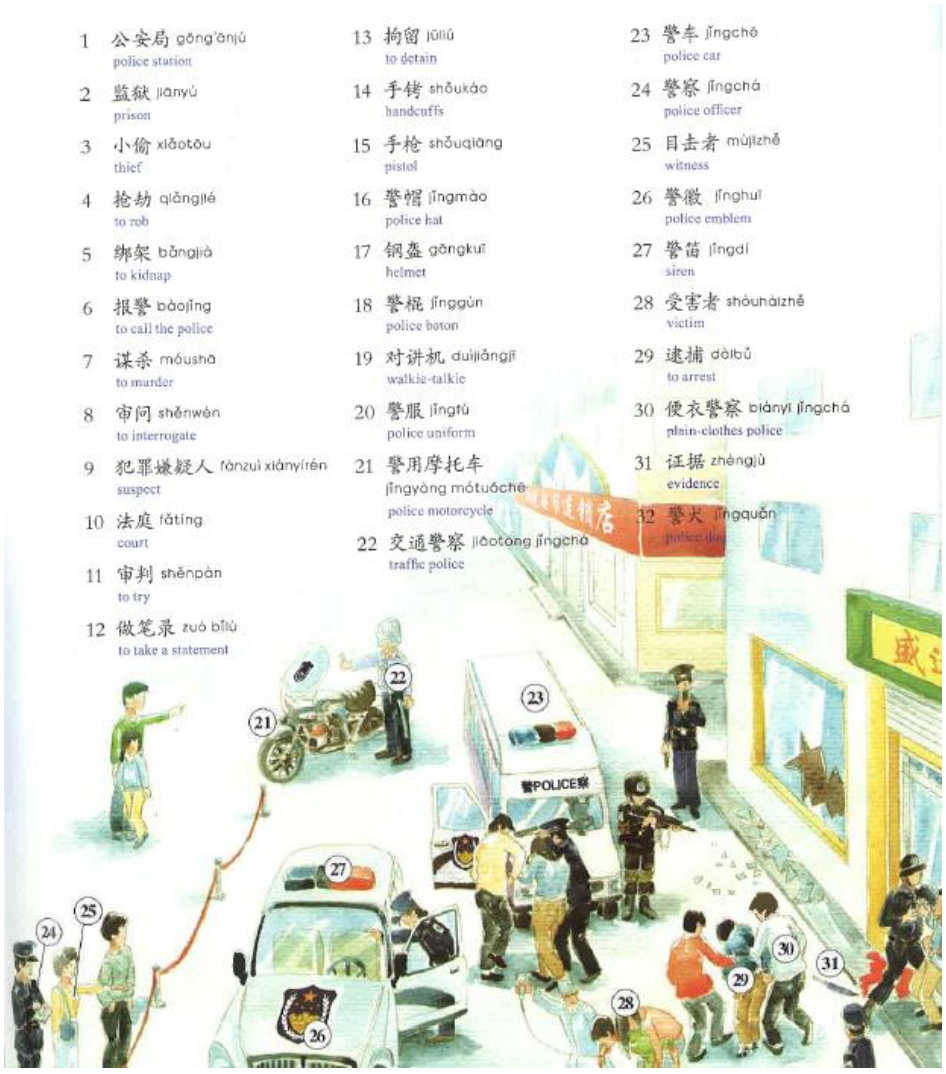

Figure 2: Illustration from *My Chinese Picture Dictionary*: Police Station