Strong Associations Can Be Weak: Some Thoughts on Cross-lingual Word Webs for Translation

Oi Yee Kwong
Department of Translation
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
oykwong@arts.cuhk.edu.hk

Abstract

This paper discusses the implications of human word association norms on the modelling of word associations from large corpora and the relevance of different types of associations in the process of translation, with a focus on adjectives. It is observed that the proportion of paradigmatic responses found in English norms tends to be higher, whereas a clear preference for syntagmatic associations is exhibited in Chinese norms. Further comparison with corpus-based extracted associations, using various functions in the Sketch Engine, shows that collocational associations might be more effectively extracted, but there is also considerable individual variation for different words. It is suggested that although free associations elicited in isolated context serve to reveal a wide range of potential lexical relations, their usefulness and relevance in real language applications should consider the actual task and its information demand. A purpose-based approach to construct cross-lingual word webs for computer-aided translation is thus proposed.

1 Introduction

Many online dictionaries, thesauri and other lexical resources are now capable of providing users with flexible modes of searching and displaying lexical information. In particular, access by meaning is recognised as even more important than access by form. As Zock et al. (2010) remarked, word access in a dictionary is a search problem. The storage of information does not guarantee successful access, and adequate navigational means have to be provided. In other words, while lexical databases tend to contain rich information about words, their usefulness (to humans or to computers) will actually depend on how readily the right information could be retrieved at the right time for the right purpose.

The onomasiological approach for organising and retrieving lexical items starts with concepts and leads to forms, which is typically what thesauri are designed for. Word finding in this way often assumes an extensive inter-connection of words, which is largely inspired by psychological models of the mental lexicon (e.g. Aitchison, 2003; De Deyne et al., 2016). Enhancement of word access in electronic dictionaries thus focuses on identifying, capturing and making available a wide range of word associations to enable words to be searched via multiple routes.

To this end, empirical evidence from psycholinguistic data, especially word association norms, offers valuable information about the variety of associative relations and their relative significance in the mental word web (e.g. Joyce and Srdanović, 2008; Kwong, 2013). At the same time, computational linguists and lexicographers have attempted to model such relations and even the corresponding associative strengths (e.g. Church and Hanks, 1990; Kilgarriff et al., 2004), not necessarily as ambitious as to reconstruct the

human mental lexicon, but often aiming to enhance lexical access with a mechanism taking advantage of the organisation of the mental word repository. For instance, even when a user fails to name the target word, as in the tip-of-the-tongue situation, he or she should be enabled to access the word by means of other closely associated words that can be thought of (e.g. Sinopaknikova and Smrž, 2006; Rapp and Zock, 2014; Zock et al., 2010).

A very wide range of associative relations have been revealed from word association norms, but as they are elicited in isolation, their readiness to be computationally modelled and their relevance in specific applications might vary. In this study, we further explore the implications from word association norms especially with respect to bilingual dictionary access. In Section 2, we first compare among several existing word association norms for the distribution of different associative types. In Section 3, we then investigate how thoroughly such associations could be modelled by various means and tools. In Section 4, we discuss the need and relevance of word associations in the context of a specific task, namely translation, and propose that word associations have to be flexibly utilised according to the nature of a task and thus its information demand. The study is concluded with future directions in Section 5.

The current investigation focuses on adjectives, which are relatively less addressed than nouns and verbs in related studies. In addition, the polysemy of adjectives bears significant implications on translation, and is worth studying for computer-aided translation.

2 Clues from Word Association Norms

The following word association norms were used: the Birkbeck Association Norms (Moss and Older, 1996) and the University of South Florida Association Norms (Nelson et al., 1998) for English, and the Hong Kong Chinese Association Norms (Kwong, 2013) for Chinese, labelled as BBK, USF, and HKC respectively.

Twenty adjectival stimuli found in both English datasets and with at least partial equivalents in the Chinese dataset were selected, as listed in the first column of Table 1 and Table 2 respectively.

2.1 Intra-lingual Comparison

In Table 1, the columns under BBK and USF show the number of responses appearing twice or more, referred to as non-single responses hereafter (F2), the number of responses appearing once only (F1), and the top response (Top 1) for individual stimuli in the two sets of norms.

Among the 20 stimuli, only 8 have the same top response in the two datasets (easy – hard, empty – full, good – bad, happy – sad, innocent – guilty, narrow – wide, obvious – clear, and strong – weak), all except one are antonym pairs. For the remaining cases, the top responses are more often syntagmatic in BBK, mostly the nouns that are typically modified by the corresponding adjectives (e.g. brittle – bone, precious – stone). In contrast, more paradigmatic top responses are found in USF, with many synonym pairs (e.g. broad – wide, calm – quiet, precious – valuable).

Among the non-single responses, overlapping items range from 2 to 5, with the percentage of overlap (with respect to BBK) reaching as much as 100% (Obvious) to 28.6% (Broad), averaging at 51.3%. There are also some unexpected observations. First, despite the vast difference in the number of participants, it is nevertheless natural to expect the bigger set of norms should more or less cover the smaller set, especially for the frequent responses. However, in 5 out of the 20 cases, the top response in BBK is not even found among the non-single responses in USF. Second, the distributions of the association types are also not uniform. As seen in Table 3, the proportions of adjectival and nominal responses in BBK are comparable, at 48.75% and 47.41% on average respectively. But in USF, adjectival responses almost double nominal ones, amounting to 61.30% and 32.65% on average respectively. This point will be further discussed in Section 2.3.

2.2 Cross-lingual Comparison

As mentioned, the corresponding stimuli selected from HKC are partial equivalents of the English stimuli. Hence the responses may only be associated with particular word senses possessed by the English words. As seen in Table 2, F2 and F1 for HKC stimuli are closer to BBK than USF, given the similar number of participants for the norming of individual stimuli in HKC and BBK.

English	BBK			USF			Overlapping Responses	
	F2	F1	Top 1	F2	F1	Top 1	N	Items
Active	5	22	Passive	21	40	Sports	2	Fit, Passive
Brittle	8	20	Bone	15	29	Peanut	4	Bone, Break, Fragile, Peanut
Broad	7	22	Bean	16	36	Wide	2	Shoulders, Wide
Calm	8	26	Water	14	38	Quiet	3	Peaceful, Quiet, Sea
Common	9	29	Land	23	40	Uncommon	3	Law, Place, Usual
Correct	4	21	Right	7	11	Wrong	2	Right, Wrong
Easy	5	18	Hard	8	23	Hard	4	Difficult, Hard, Rider, Simple
Empty	6	17	Full	11	22	Full	2	Box, Full
Good	3	24	Bad	8	17	Bad	2	Bad, Evil
Great	7	25	Weak	18	32	Big	2	Big, Good
Happy	6	22	Sad	8	19	Sad	2	Sad, Smile
Innocent	8	18	Guilty	16	9	Guilty	4	Bystander, Guilty, Man, Shy
Narrow	5	24	Wide	12	16	Wide	3	Mind, Thin, Wide
Obvious	5	24	Clear	19	45	Clear	5	Clear, Easy, Evident, Open, Obscure
Plain	7	53	Jane	20	45	Simple	3	Boring, Jane, Ordinary
Precious	6	19	Stone	23	32	Valuable	4	Gem, Jewel, Metal, Stone
Rare	7	33	Bird	27	37	Common	3	Extinct, Steak, Uncommon
Sharp	5	21	Knife	18	19	Point	4	Blunt, Edge, Knife, Point
Strong	5	25	Weak	11	20	Weak	3	Man, Muscle, Weak
Wise	7	16	Old	10	14	Smart	3	Knowledge, Old, Owl

Table 1 English Stimuli and Top Responses

Chinese	HKC		
	F2	F1	Top 1
積極 *ji1ji2 'active'*	8	29	進取 *jin4qu3 'aggressive'*
脆弱 *cui4ruo4 'brittle'*	5	25	心靈 *xin1ling2 'heart'*
廣泛 *guang3fan4 'broad'*	11	33	興趣 *xing4qu4 'interest'*
平靜 *ping2jing4 'calm'*	11	35	海 *hai3 'sea'*
普通 *pu3tong1 'common'*	9	28	平凡 *ping2fan2 'plain'*
正確 *zheng4que4 'correct'*	5	23	答案 *da2an4 'answer'*
容易 *rong2yi4 'easy'*	12	24	困難 *kun4nan2 'hard'*
空虛 *kong1xu1 'empty'*	5	23	寂寞 *ji4mo4 'lonely'*
良好 *liang2hao3 'good'*	12	26	表現 *biao3xian4 'performance'*
偉大 *wei3da4 'great'*	10	28	母親 *mu3qin1 'mother'*
快樂 *kuai4le4 'happy'*	8	35	開心 *kai1xin1 'joyful'*
單純 *dan1cun2 'innocent'*	13	29	天真 *tian1zhen1 'childlike'*
狹窄 *xia2zai2 'narrow'*	12	37	小巷 *xiao3xiang4 'alley'*
明顯 *ming2xian3 'obvious'*	5	45	突出 *tu1chu1 'outstanding'*
平凡 *ping2fan2 'plain'*	12	32	人 *ren2 'person'*
寶貴 *bao3gui4 'precious'*	4	21	時間 *shi2jian1 'time'*
罕見 *han4jian4 'rare'*	9	42	疾病 *zhi2bing4 'disease'*
尖銳 *jian1rui4 'sharp'*	7	28	問題 *wen4ti2 'question'*
強烈 *qiang2lie4 'strong'*	7	23	感覺 *gan3jue2 'feeling'*
明智 *ming2zhi4 'wise'*	5	21	選擇 *xuan3ze2 'choice'*

Table 2 Chinese Stimuli and Top Responses

As reported in Kwong (2013), collocational responses are abundant in the Hong Kong Chinese Association Norms, especially for abstract nouns and verbs. Also, there are quite a constant proportion of non-linguistic associations. It was thus suggested that the top responses for individual stimulus words may serve to inform the design of semantic lexicons, but the majority and infrequent responses may not even be properly qualified as weak associations. Nominal responses also make up the majority of responses in general, even for adjectival stimuli, although they also elicited relatively more adjectival, and paradigmatic, responses than stimuli of other parts of speech.

With respect to the selected stimuli in this study, Table 2 shows that the top responses are adjectives in only 7 out of the 20 cases (積極 active – 進取 aggressive, 普通 common – 平凡 plain, 容易 easy – 困難 hard, 空虛 empty – 寂寞 lonely, 快樂 happy – 開心 joyful, 單純 innocent – 天真 childlike, 明顯 obvious – 突出 outstanding). All other top responses are nouns (e.g. 偉大 great – 母親 mother, 狹窄 narrow – 小巷 alley). This is an interesting distribution especially when compared with the English norms.

Table 3 shows the proportions of non-single responses in the various association norms by part of speech (POS), with N for noun, A for adjective, and V for verb. As reported in Section 2.1, USF has many more adjectival responses than nominal responses compared to BBK, although both English norms show the dominance of adjectival or paradigmatic responses. For HKC, however, nominal responses dominate, followed by adjectives and verbs, with average proportion at 59.04%, 23.66% and 13.79% respectively.

		N (%)	A (%)	V (%)
BBK	**Avg**	**47.41**	**48.75**	**1.25**
	Max	88.89	100.00	25.00
	Min	0.00	11.11	0.00
USF	**Avg**	**32.65**	**61.30**	**4.07**
	Max	63.64	100.00	13.33
	Min	0.00	33.33	0.00
HKC	**Avg**	**59.04**	**23.66**	**13.79**
	Max	100.00	55.56	50.00
	Min	25.00	0.00	0.00

Table 3 POS Distribution of Responses

2.3 Word Associations across Languages

It can be readily observed from the above comparisons that English and Chinese speakers exhibit different patterns in what one might consider "strong" associations in their mental lexicons. Based on the selected adjectival samples, apparently English speakers tend to come up with more paradigmatic responses as the most strongly associated words, while syntagmatic responses (mostly nouns which are typically modified by the adjectives) are dominant among Chinese speakers. Considering all non-single responses, still more adjectival responses were elicited from English speakers than Chinese speakers. The adjectival responses, corresponding to paradigmatic relations, could be the relatively narrow synonymy or antonymy relations (e.g. good – bad), or broader conceptual semantic relations and contextual collocations (e.g. wise – old, innocent – shy). The differences and characteristics revealed from the association norms can be attributed to polysemy to a certain extent. It happens that the English stimuli are relatively more polysemous while the Chinese stimuli are often their partial equivalents only. For instance, "innocent" may mean "not guilty" or "simple-minded", while 單純 only covers the latter sense. The morphological properties of the two languages may also make a difference. The disyllabic Chinese words are often formed with two individual morphemes. When they are combined to form a word, very often the resulting word will have more specific meanings. With such additional constraints on the word sense, it may somehow limit the paradigmatic relations, making them less readily available than their syntagmatic or collocational counterparts. The grammatical system is apparently better defined in English where word classes or POS categories are relatively more clearly distinguished. Given the lack of morphology and various specific word formation mechanisms, categorial fluidity is more common in Chinese, and POS groups are less homogenous. For instance, Chinese adjectives may often function like adverbs to modify verbs (e.g. 廣泛 broad – 傳播 *chuan2bo1* 'communicate', which actually means "widely spread" when used together). This probably explains for the much higher proportion of verbal responses for the adjectival stimuli in the Chinese norms than the English norms. The above comparison thus

suggests that modelling of word associations has to consider language difference and weigh various associative types accordingly.

3 Modelling of Word Associations

It is generally realised that while word association norms are important resources not only for understanding the mental lexicon but also as linguistic resources for a variety of applications, they are expensive to obtain, especially in large scale with reliable sampling. Computational modelling with large corpora is a natural way out. Most typically, Church and Hanks (1990) measured associative strength with mutual information. Wettler and Rapp (1993) relied on co-occurrence frequencies to model word associations, which tend to be biased toward syntagmatic associations. Lin (1998) extracted paradigmatically related words based on contextual similarity.

While human word association norms exhibit a wide range of associative relations, some of which are even non-linguistic and personal, we try to investigate the extent to which the linguistic ones can be effectively modelled. In this study, we make use of the Word Sketch function and the Thesaurus function in the Sketch Engine for the comparison. The Word Sketch function shows a one-page summary of the grammatical and collocational behaviour of words (Kilgarriff et al., 2004). The Thesaurus function produces a list of words occurring in similar contexts as the query word (Rychlý and Kilgarriff, 2007).

3.1 Corpus-based vs Human Associations

Collocations fall between free combinations and idioms (McKeown and Radev, 2000). Typically they refer to grammatically bound co-occurring words (e.g. modifier-head constructions). A more inclusive view will also consider broader semantic relations or topical associations. Thus collocations might involve words of the same or different word classes. On the contrary, paradigmatic associations always involve words of the same POS. Human responses in free word association norms, as seen above, encompass a wide range of both linguistic and non-linguistic relations.

In this comparison, we used the Word Sketch function and the Thesaurus function in the Sketch Engine (SkE), and compared the collocations and similar words extracted with the non-single responses for the selected stimuli in the various association norms. For English, we tested with the British National Corpus (BNC) and the ukWaC corpus. For Chinese, we used the ChineseTaiwanWaC (twWaC) corpus for the current purpose. The top 50 similar words returned by the Thesaurus function were considered, and all default *gramrel* relations in the corresponding Sketch Grammars were included for the Word Sketch function. Other parameters were kept at the default settings.

Table 4 shows the results for comparing the SkE extractions with the association norms. The first figure in each cell is the number of overlapping words, and the figure in brackets is the percentage of non-single responses in the association norms found in the SkE extraction results.

In general, the Thesaurus function tends to generate fewer words matching the association norms. For instance, with ukWaC, the Thesaurus function produces 3.55 words on average which are found among the association responses for a particular stimulus in USF, whereas the Word Sketch function produces 5.25 matching words on average. It should be noted that the number and the percentage presented in Table 4 are not necessarily linked to the same stimulus. Since the number of non-single responses is different across the stimuli, the one with most matching words are not always the one with the highest percentage of overlap. The two figures are presented to give a different reference point only.

Modelling with different corpora may make a difference, but with respect to the results in this study, the difference does not seem to be drastic. For instance, despite the considerable size difference between the two corpora, using BNC or ukWaC leads to similar overlapping with USF associations, although with slight variations.

One interesting observation from Table 4 is that while the Word Sketch function is in general more effective than the Thesaurus function in extracting word matching the association norms, the difference is more pronounced in the Chinese data. As discussed earlier, syntagmatic associations are more abundant for the adjectival stimuli in the Chinese norms, whereas the English norms exhibit a relatively higher proportion of paradigmatic responses, which probably accounts for the better modelling results by Word Sketch for Chinese.

	Thesaurus N(%)				Word Sketch N(%)	
	BNC vs USF	*ukWaC vs USF*	*ukWaC vs BBK*	*twWaC vs HKC*	*ukWaC vs USF*	*twWaC vs HKC*
Avg	3.40 (23.50)	3.55 (24.53)	1.15 (20.05)	1.65 (20.97)	5.25 (33.67)	3.35 (42.49)
Max	8.00 (50.00)	9.00 (50.00)	3.00 (60.00)	4.00 (50.00)	13.00 (60.00)	8.00 (85.71)
Min	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	2.00 (11.11)	1.00 (8.33)

Table 4 Comparing Corpus-based Associations and Human Responses

3.2 Extracting Associations from Corpora

One of the most closely related studies worked on Japanese. Joyce and Srdanović (2008) compared the lexical relationships observed in word association norms and those in the collocational and thesaural data extracted with the Sketch Engine. Six Japanese lexical items were selected, including two verbs, one adjective and three nouns. As expected, a rich variety of associative relations have been observed from the word association norms. While there was considerable overlap between the two resources, attention was drawn to the relations which were only found in the association norms but absent from the associations extracted by the Sketch Engine. More fine-grained sub-categories for typical associations, including even encyclopedic and cultural specific ones, were distinguished. The value of word association norms as linguistic resources was highlighted and it was suggested that they be incorporated in electronic dictionaries for a more comprehensive coverage to enhance association-based lexical access as has often been aspired.

In this study, we have focused on adjectives as the stimulus words. Grammatically they are supposed to form a homogenous group sharing most distributional features. However, when it comes to associations, individual variations are more than obvious. On the one hand, the computational extraction of associations is not equally or comparably effective for all stimuli. For instance, with Word Sketch on ukWaC and compared with USF, the number of matching words vary from 2 to 13. In the best case, 60% of overlap was found (e.g. for "brittle", matched associations include "peanut", "hard", "fragile", "dry", "crack", "hair", "stiff", "weak" and "bone"), whereas in the worst case, only 11.11% overlap could be achieved (e.g. for "great", only "big" and "little" could be matched). Similarly for the Chinese data, the overlapping ranges from 1 to 8, and in terms of percentage, it could be as poor as 8.33% (e.g. quite unexpectedly, for 容易 'easy', only 簡單 'simple' could be matched) to as good as 85.71% (e.g. the matched words for 強烈 'strong' include 反對 *fan3dui4* 'oppose', 慾望 *yu4wang4* 'desire', 要求 *yao1qiu2* 'request', 氣味 *qi4wei4* 'smell', 建議 *jian4yi4* 'suggest' and 感受 *gan3shou4* 'feeling').

Meanwhile, the relative association strengths found in the association norms and the extracted words are seldom in concord. Sometimes the results could be quite counter-intuitive as the following example.

If we try the Thesaurus function in the Sketch Engine, with "strong" as the query word, it turns out that "weak" is not a strongly associated item. With BNC, "weak" appears at the 34th position in the ordered list of similar words. With ukWaC, "weak" even comes at the 67th position. Hence larger corpora may not always produce the expected and desired results. Nevertheless, with both corpora, generating a thesaurus with "weak" as the query word unanimously gives "strong" as the foremost associated word.

4 Purpose-based Word Webs

Human word association norms contain many possible kinds of lexical relations. Some can be conveniently defined by linguistic means, such as paradigmatic relations and some syntagmatic relations. Broad conceptual relations need to be topically situated. In addition, there is always a considerable amount of personal associations. These are often single responses, and although they cannot be analysed linguistically, they are still cognitively salient at least to some individuals.

The last type of associations aside, the others can potentially be modelled from large corpora by various means. However, as seen from the above discussion, the effectiveness of such modelling varies. On the one hand, humans do not generate

similar types of responses even for similar types of stimuli, or the strength of a particular type of response could be different across stimuli. On the other hand, associated words extracted from corpora do not always substantially match human responses, and even when there is overlap, the relative association strengths could find little correlation, if any at all.

What does such discrepancy imply on the modelling of word associations? Even norms with large samples and participants could only show the tip of an iceberg within the whole lexical repository. The issue is therefore not whether one could model the responses found in association norms. The more important question is what purpose the modelling is supposed to serve, and whether the results really serve the purpose. Free word association norms are elicited in isolation, but in real language applications one often works in a context. Hence amidst a sea of free associations, according to the task purpose and information demand, some associations must be more relevant and useful than others, and it is this subset of associations that the modelling should settle on. In other words, we need effective means to filter enormous word webs to allow flexible utilization of the word associations.

4.1 Enhancement of Dictionary Access

Studies in dictionary access have drawn on association norms, which inspire many attempts to provide adequate navigational means for dictionary users to access what they want, especially when they could only start with some fuzzy query. One such scenario is the tip-of-the-tongue problem, as Zock et al. (2010) suggested, in which case an extensively linked lexicon and making these links available is particularly essential.

The salience and interest in this area of research is also evident from the series of workshops on Cognitive Aspects of the Lexicon (CogALex). In the most recent CogALex workshop, there was a shared task addressing the lexical access problem with a bag of associated words (Rapp and Zock, 2014). Significant implications were drawn from word association norms, and systems were designed to model the intended word among an ordered list of candidates. Several applications were suggested, one of which is association-based machine translation, by translating meaning vectors into the target language and selecting the target language meaning vector and its corresponding linguistic phrase which is most similar to the source language meaning vector.

Nevertheless, while the idea of having more entry points for dictionary access is plausible, it is not always clear what precise associative relations are to be included and how it is to be implemented. After all, dictionary usage in practice often carries a purpose. We therefore propose a more user-oriented and purpose-based approach to the design of features to facilitate dictionary access and thus the modelling and inclusion of word associations in the process. We use translation as an example, and discuss how computer-aided translation may benefit from the comparison of word association data in this study.

4.2 A Scenario in Translation

In practical lexicography, the user profile is deemed particularly important in dictionary design (Atkins and Rundell, 2008). The content and presentation of a dictionary should be grounded on the purposes and proficiency, and thus the information demand, of the target users. During the 1980s, when computer-aided translation started to gain attention, the Translator's Workstation was proposed (Melby, 1982), where translators can work in an integrated environment with different resources at hand, including automatic dictionary lookup and the use of translation memory among others. By now most people will agree that word-for-word lookup is not all satisfactory and will not be sufficient in a real translation setting.

Let us consider a more realistic scenario, such as when a translation student needs to look up a bilingual dictionary to decide on how the phrase "strong allegation" should be rendered in Chinese. The adjective "strong" can be used in a wide range of context, and will be expressed differently in Chinese for "strong coffee", "strong man", "strong economy" and "strong emotion". Table 5 shows some more examples, which have not yet included cases where a disyllabic Chinese word encompassing the meaning of the adjective and the noun can be used, such as 濃茶 *nong2cha2* 'strong tea' and 強風 *qiang2feng1* 'strong wind'. It happens that "strong allegation" is not listed in the monolingual Macmillan English Dictionary or the bilingual dictionary available in Cambridge Dictionaries Online. The Word Sketch function

does not list “allegation” for “strong” either. So is this a weak association, weak enough for it to be excluded from major lexical resources? But if this collocation is repeatedly found in real text, then it must be relatively stronger in some context. How can we enable the student to access the relevant lexical information then?

Strong	**N**
洪亮 *hong2liang4*	聲線 *sheng1xian4 ‘voice’*
有力 *you3li4*	證據 *zheng4ju4 ‘evidence’*
強健 *qiang2jian4*	體魄 *ti3po4 ‘body’*
深刻 *shen1ke4*	印象 *yin4xiang4 ‘impression’*
巨大 *ju4da4*	壓力 *ya1li4 ‘pressure’*
濃烈 *nong2lie4*	氣味 *qi4wei4 ‘smell’*

Table 5 Some Contexts for “Strong”

Two types of information demand thus arise from this scenario. First, the student will need to find out that in a different collocation, the word “strong” or the phrase “strong+N” will have to be expressed differently in Chinese, and what may be similar collocations as “strong allegation”. Second, considering the register and context of the source text, the student will need to know what alternative expressions or (near-)synonyms might be available for his or her choice to render that group of collocations. The first question thus involves mainly decoding usage, requiring mostly collocational information, and the second question involves mainly encoding usage, concerning mostly with paradigmatic associations.

4.3 Bilingual vs Cross-lingual Associations

In fact the Sketch Engine has recently developed the Bilingual Word Sketch function (Baisa et al., 2014). The function allows lexicographers to compare collocations across translation equivalents, but as the developers pointed out, they are not the source and target languages as understood by translators. Moreover, as remarked by McKeown and Radev (2000), a concept expressed by way of a collocation in one language may not have a corresponding collocation in another language. Hence instead of bilingual associations, we propose cross-lingual word webs. Here we outline the steps needed for such word webs, illustrated with the “strong allegation” example, for which indirect means are needed to draw an association.

The first question is which sense of “strong” is most relevant here. Suppose we start with the adjective “strong”, among the clusters of nouns which are typically modified by it, can we group “allegation” into one of these clusters? Using SkE to simulate the situation, we can get the nouns being modified with the Word Sketch function. At the same time, we use the Thesaurus function to find a list of similar words for “allegation”. Comparing the two sets of words, “evidence” and “argument” are found in common. The second question is how one should render the corresponding meaning of “strong” in the target language, in this case Chinese. Based on the set of similar words (allegation, evidence, argument, and possibly others), a corresponding Chinese word web can be built in the reverse direction. With the equivalents based on a bilingual dictionary, groups of similar words and collocated adjectives can then be extracted from a Chinese corpus (not necessarily parallel to the English one). The resulting associations, which may include 指控 ‘allegation’, 證據 ‘evidence’, 理據 ‘rationale’, 充分 ‘sufficient’, 有力 ‘strong’ and many others, could be of reference to the translator. The word web is expected to offer assistance by presenting possibilities like 有力的指控, 理據充足的指控 and 證據確鑿的指控 especially in the absence of “strong allegation” in dictionaries in the first place, and it is of course up to the translator to judge for their appropriateness in the specific context.

5 Conclusion and Future Work

Our comparison among word association norms and associations extracted from corpora has revealed discrepancy between (1) types of free association responses across languages, (2) words deemed closely related by humans and by statistics, and (3) relative association strengths in human responses and corpus-based associations. These observations bear important implications on modelling word associations and using them to enhance dictionary access. It is suggested that the usefulness and relevance of different associations depends on the actual task and its information demand, and purpose-based word webs are proposed. Future work includes more comparison of association norms, refinement of the modelling steps for cross-lingual word webs and their implementation for computer-aided translation.

Acknowledgments

The work described in this paper was supported by funding from the Faculty of Arts of the Chinese University of Hong Kong (Project No. 3132414).

References

Aitchison, J. (2003) *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell Publishers.

Atkins, B.T.S. and Rundell, M. (2008) *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Baisa, V., Jakubíček, M., Kilgarriff, A., Kovář, V. and Rychlý, P. (2014) Bilingual Word Sketches: the translate Button. In *Proceedings of the 16th EURALEX International Congress*, Bolzano, Italy, pp.505-513.

Church, K.W. and Hanks, P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics, 16(1)*:22-29.

De Deyne, S., Verheyen, S. and Storms, G. (2016) Structure and Organization of the Mental Lexicon: A Network Approach Derived from Syntactic Dependency Relations and Word Associations. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard and B. Job (Eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*. Springer Berlin Heidelberg.

Joyce, T. and Srdanović, I. (2008) Comparing Lexical Relationships Observed within Japanese Collocation Data and Japanese Word Association Norms. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, Manchester, U.K., pp.1-8.

Kilgarriff, A., Rychlý, P., Smrz, P. and Tugwell, D. (2004) The Sketch Engine. In *Proceedings of EURALEX 2004*, Lorient, France.

Kwong, O.Y. (2013) Exploring the Chinese Mental Lexicon with Word Association Norms. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27)*, Taipei.

Lin, D. (1998) Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, Canada, pp.768-774.

McKeown, K.R. and Radev, D.R. (2000) Collocations. In R. Dale, H. Moisl and H. Somers (Eds.), *A Handbook of Natural Language Processing*. Marcel Dekker.

Melby, A.K. (1982) Multi-Level Translation Aids in a Distributed System. In J. Horecký (Ed.), *COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics*, Academia, Prague and North-Holland, Amsterdam, pp.215-220.

Moss, H. and Older, L. (1996) *Birkbeck Word Association Norms*. Hove, U.K.: Psychology Press.

Nelson, D. L., McEvoy, C. L. and Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/.

Rapp, R. and Zock, M. (2014) The CogALex-IV Shared Task on the Lexical Access Problem. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon*, Dublin, Ireland, pp.1-14.

Rychlý, P. and Kilgarriff, A. (2007) An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Czech Republic, pp.41-44.

Sinopalnikova, A. and Smrz, P. (2006) Knowing a word vs. accessing a word: WordNet and word association norms as interfaces to electronic dictionaries. In *Proceedings of the Third International WordNet Conference*, Korea, pp.265-272.

Wettler, M. and Rapp, R. (1993) Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, Ohio, pp.84-93.

Zock, M., Ferret, O. and Schwab, D. (2010) Deliberate word access: an intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology, 13(4)*: 201-218.