Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary

Pornpimon Palingoon Pornchan Chantanapraiwan Supranee Theerawattanasuk
Thatsanee Charoenporn Virach Sornlertlumvanich

Information Research and Development Division
National Electronics and Computer Technology Center, NSTDA
112 Thailand Science Park, Paholyothin Rd.,
Klong 1, Klong Luang,
Pathumthani, Thailand 12120
Tel. +66-2-564-6900
Fax.: +66-2-564-6901-5
e-mail : pornpimon_palingoon@notes.nectec.or.th, pornchanc@notes.nectec.or.th,
supranee@notes.nectec.or.th, tcharoen@notes.nectec.or.th, virach@links.nectec.or.th

Abstract

In lexicography, it is currently evident that corpora data still provide lexical informaiton as objective criteria of language descriptions in dictionary-making, especially in assigning meanings to lexical items and describing actual use. At this point, it means that the quantitative approach can add up our understanding of linguistic behavior and give a basic representation of language, together with qualitative approach systemically. The aim of this paper is to analyze the principles of compiling bilingual English-Thai dictionary in two main issues: (1) defining lexical items in bilingual dictionary-making by employing translation process and corpus-based information, and also proposing bottom-up definition model for bilingual dictionary-making. (2) considering correlation between qualitative and quantitative approaches in assigning meanings to lexical items of bilingual corpus-based dictionary.

1 Introduction

Nowadays, it is widely acknowledged that Natural Language Processing by computer is an essential subject in building the idealistic Artificial Intelligence (AI), which is expected to analyze, understand, and generate all sorts of facts and complexities in any language use as well as native or non-native speakers who have manipulated them. Since the end of nineteenth century, corpus linguistics, the well-known term based on the "real life" examples of language use, has gradually been extended its scope and influence concerned with natural language processing. With respect to the efficacy of such corpora data in all sizes and representativeness, researchers and lexicographers can examine naturalistic languages by considerably employing both qualitative and quantitative approaches, which will be mentioned later in this paper.

This means that the human internal representation would be possibly uncovered to scholars, who work very hard in making natural language system come true. As a result, there are a lot of research projects in various fields such as machine translation and dictionary-making, exploiting corpora technologies such as statistics calculations, part-of-speech tagging, parsing, annotations, word sense disambiguation, and so on, in order to model the cognitive system of computer in the same way as humans do. With these modern acceptable techniques, for being time and in the future, we may concretely vastly describe some occurrences of unexplainable language phenomena such as semantic restrictions of lexical items so as to show human being's intuition systemically and to provide real language information as much and easy as

possible. However, there are still much more complexities to show a perfect representation of human being's competence and performance as described in many research papers as in works of Tony McEnery & Andrew Wilson, 1993 in Corpus Linguistics and Martin Chodorow, 1998 in Using Corpus Statistics and WordNet Relations for Sense Identification.

From the challenging aspect mentioned above, so what does this paper seek to achieve? This paper intends to propose the essential principles of compiling bilingual corpus-based dictionary in two main issues. **The first issue**, as a qualitative approach, concerns with defining lexical items in bilingual dictionary-making with support of translation process and corpus-based information. The bottom-up definition model for bilingual dictionary will be presented to serve the need of real language usages in NLP. Some practical examples in defining lexical items *noun, verb, and adj* of bilingual English-Thai dictionary will be raised up here as a group of representatives in both languages. **The second issue** shows whether we can measure the proficiency of corpora by considering the correlation between qualitative and quantitative approaches in assigning meanings to lexical items of bilingual corpus-based dictionary.

2 Qualitative approach: defining of lexical items in bilingual dictionary

In this section we will describe the common and efficient process of defining lexical items in bilingual dictionary-making, which requires much more attention to details, deep knowledge of languages involved, associated with difficult time consuming, and labour-intensive process. Two subsections are shown as follow.

2.1 General definition in lexicography

In general, a dictionary seems to be concerned with stating the meanings of words. Knowing the meaning of a word means that we can understand the concept reality underlying the pattern recognition of cognitive system in human brain, known as *top-down processing* or *top-down perception*, and *bottom-up processing* or *bottom-up perception* (Eskey, 1988). In both bilingual and multilingual dictionary, it is quite difficult to find a certain word in source language standing for a common or technical term in target language? Which English word will be worthy of บาดตา[bàattaa], บาดใจ[bàatjay], บาดหู[bàathŭu], บาดหมาง[bàatmăŋ] in Thai? And how do we describe an abstraction of semantic relatedness between languages? If these questions are clearly answered, then the problems of concept alignment by computer would be possibly solved.

According to the traditional rules of lexical definition, lexicographers can identify word meanings by classifying a lexical item into a group of senses with support of lexicographers' deep knowledge and vast experiences, called as top-down definition model in this paper. So, the number of word senses compiled by different contributors are certainly differrent depending upon the goal and background knowledge of definers. Take the word *deep* for instance. By definition of Encarta Online Dictionary 2001, this word is provided fifteen senses for three categories *adjective*, *adverb*, and *noun* as well as seventeen senses of Online Oxford Dictionary and fifteen senses of Online WordNet 1.7 vocabulary helper. Differently, Random House Dictionary (Unabridge Version, 1993) provides a group of thirty-eight senses for *deep*. For general defining of lexical items in monolingual dictionary, the problems of definition can be solved by 1) paraphrase 2) amplification and 3) substitutity (see more details in Pornpilas, 1990, Landau, 1993)

In case of bilingual dictionary, translation process is an important part of compiling, and concept equivalence of senses between source language and target language is what lexicographers strive to arrive at. Methodologically, Nida (1976) has suggested that a suitable way to find a lexical meaning is to translate from source languages to target languages. The definer is to work backwards from the surface of the original text to its deep structure of the new language, and then generate a surface structure in the second language. His methodology focuses on decoding and recoding process as other practicing translators. Nida's model of translation process can be addressed as a diagram below.

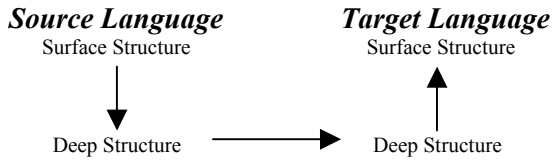

Figure 1 Traditional translation model

In Figure1, it is very hard to achieve the accurate interpretation of word or word combinations, especially in the translation of idioms, since they are heavily influenced by linguistic and cultural diversity between two languages. Roda P. Roberts (2000) has suggested that it is now generally believed that translation in bilingual dictionary is not only the translation of individual words of each language, but also the translation of a message outside a text. For example, *khazi* is one of British slangs equivalent to a general term *lavatory* or *toilet*, and *Bobby* is a slang for *policeman*. (These examples are picked up from www.Londonslang.com). The core notion of *khazi* can be attached to the only one concept of ห้องน้ำ [hɔ̂ɔŋnáam] or ส้วม [sûam] in Thai, and *Bobby* can be attached to the only one concept of ตำรวจ [tamrùat]. These words cannot be equally matched without considering all the knowledge and values shared by a society between two languages.

2.2 Bottom-up definition model for bilingual dictionary

The major problems of bilingual defining mostly results from unequal lexical concepts between languages. Besides some solutions for monolingual definition mentioned in 2.1, the bottom-up definition model is specifically proposed as a kind of corpus-based technique for bilingual dictionary, especially for LEXiTRON electronic dictionary (English<->Thai Version 2.0), to establish translated equivalences and to choose an appropriate lexical form in target language. Based on the naturalness of language, this model provides "real life" sample sentences for determining lexical meanings. The model operation is to consider the syntactic relationship of cooccurred words in a same sentence. This appears to be the most important strategy to predict the amount and accuracy of lexical meanings. That is, as seen in an equation below, if *x* occurs to *y* and *z* in a same sentence, then *y* and *z* must be semantically related to *x*.

$$s_n = y_1 + y_2 + \ldots y_n + x + z_1 + z_2 + \ldots z_n$$

where, x is the studied lexical form.

y_n *is the front lexical form of x.*

z_n *is the back lexical form of x.*

s_n *is the string of occurring x with y and z.*

Consequently, the key concept of model looks like a case frame or case relation— it is a relationship between verb and other lexical items in a same sentence— in the Case Theory. Two word senses, for instance, can be extracted from *besides* by looking at a group of sentences (a), (b), and (c) in BNC corpus (100 million words). Looking at "semantic coincidence" between their surrounding word, *besides* in (a) and (c) function as a preposition and its meaning equals to "in addition to", "apart from" in English, and "นอกจาก"[nɔ̂ɔkjàak] in Thai. The other (b) is an adverb of which meaning is the same as "moreover" or "ยิ่งกว่านั้น"[yîŋkwaanán], "นอกจากนั้น" [nɔ̂ɔkjàaknán].

(a) It was already late afternoon, but it was the only German greeting I knew ***besides*** `Heil Hitler'.
(b) ***Besides***, we know where they're going.
(c) The only person there ***besides*** Olivia who did not was Dr. Saunders.

This bottom-up model consists of three components: 1) lexical form (surface structure) 2) lexical concept (deep structure) and 3) instruments.

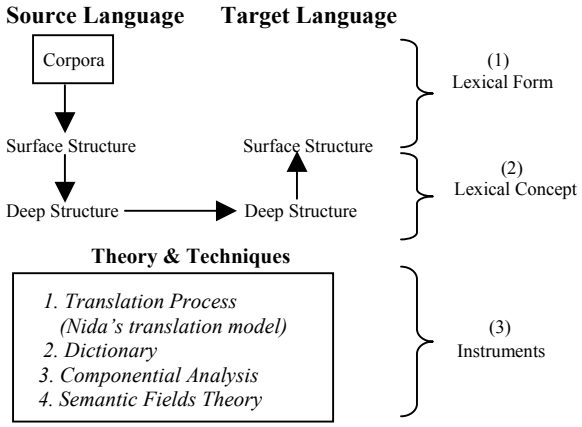

Figure 2 Bottom-up definition model for LEXiTRON

As shown in Figure 2, the means of connection between source language and target language can be acheived by the last component, (3) instruments, especially by the translation process. To represent the efficiency of the model, several examples are extracted from LEXiTRON, which is compiled on the basis of the sentence samples provided by English/Thai ORCHID Corpus (400K words). In this corpus, for example, the word *champion* belongs to two categories: *noun* and *verb*. They are related to a sense domain of "fighting". There are three source-lexical concepts for a category of noun—*winner, defender, warrior* and a category of verb—*to defend*. According to componential analysis, the definer would analyze fundamental semantic distinctions of each source-lexical concept (e.g., ±human, ±alive, ±win, ±war). In order to choose appropriate lexical forms, each source-lexical concept must be matched with target-lexical concepts which covers the same semantic distinctions as of source concepts. By the semantic fields theory, at last, *champion* agrees with three Thai lexical concepts and nine lexical forms as shown below.

Source	**Target**	
champion	ผู้ชนะเลิศ[phûuchanálêət]	**lexical forms**
	ผู้ชนะ[phûuchaná]	
	คนชนะ[khonchaná]	
	คนชนะเลิศ [khonchanálêət]	
	ผู้ยอดเยี่ยม[phûuyôɔtyîəm]	
	นักรณรงค์[nákronnaroŋ]	
	นักต่อสู้[náktɔ̀ɔsûu]	
	นักรบ[nákróp]	
	อัศวิน[?àtsawin]	
winner	ผู้ชนะเลิศ[phûuchanálêət]	**lexical concepts**
defender	นักต่อสู้[náktɔ̀ɔsûu]	
warrior	นักรบ[nákróp]	

However, although the bottom-up model is dominated by actual language use from large corpora data, the top-down model is needed to complete the bottom-up activity, especially to judge semantic distinctions and lexical forms in target language. The instruments 1, 2, 3, and 4 need to be carried out by definers' individual knowledge in the top-down model. This means most part of this definition model still depends on a major role of human decisions as generally seen in general-purpose bilingual dictionaries in Thailand. The next section may provide an appropriate response for the real bottom-up model.

3. The correlation between qualitative and quantitative approaches

The qualitative approach focuses on the introspective judgement of human beings which mentioned as top-down model in section2, while the quantitative approach is on the text judgement. So, if we can measure the proficiency of these approaches, it is closer to the state-of-the-art automatic lexicography. This section is an attempt to consider the correlation between qualitative and quantitative approaches only of bilingual English-Thai dictionary. Particularly, the positive and negative relationship between these approaches are described as a guideline to improve the corpus methodology such as representation, manipulation, and retrieval of corpora data. Several samples from small experiments may be helpful in evaluating the use of bilingual dictionary as a tool for computational language studies in the future.

Positively, the higher cooperation between qualitative and quantitative approaches help supplement the idea of corpus-based NLP. Palmer (1981) proposed that linguistics is the "scientific" study of language. Mostly, a scientific study is empirical; it may be possible to test and verify the statements made within it. Being a quantitative science for language studies, corpus linguistics is an empirical study based on language-statistical calculations which provides the object criteria for assigning meanings to lexical items and also indicates the semantic distinction of lexical items.

One of the most traditional quantitative approach appears to be statistically the observation of context environments from a large corpus such as lexical collocations and grammatical structures. For example, the word *participate* is regularly used with two prepositions *in* and *with* and the expression "*It's no use*" is always followed by *gerund* or *v-ing*. The frequency of their coocurrences in corpora can reliably determine the naturalness of language uses.

There are several examples, which show the real condition of useful lexical information in corpora. The first example illustrates the corpus potential that one can easily define the word

“ปอด”[pɔ̀ɔt] as “lung” in English, but if it is “ปอดแหก”[pɔ̀ɔthɤ̀k], the context from Thai ORCHID corpus can separate its meaning as “coward” or “fearful” from a gerneral term “ปอด”[pɔ̀ɔt]. Moreover, there are five times of word combinations โคตรเหง้า[khôotŋâw] occurred in corpus data, but it does not appear in general-purpose monolingual dictionary such as Thai Royal Institute Dictionary (2525), พจนานุกรมไทยฉบับทันสมัย (2543), พจนานุกรมนอกราชบัณฑิตยสถาน (2544), and also in Thai-English dictionary (Wit Thiengburanathum, 1998; Domnern Garden and Sathienpong Wannapok; 1999).

It can assume that quantitative data in concordance program help confirm the truth of language use via statistic values such as frequency of lexical co-occurrences, Z-score, the chi-squared test, correlation analysis and so forth. The idea of statistic information of lexical co-occurrences is important to determine which pairs of words have a statistically significant relation between them.

Looking at 50 times of *salmonella* occurrences in BNC, we can gain the coocurrence between *salmonella*, a medical technical term, and four groups of lexical forms— 28 times of pathological environments (food, place, etc), 2 times of salmonella properties (shape, colour, etc), 13 times of cause and effect terms (infection, contamination, etc), and 13 times of treatments (test, clinical techniques, etc)— in a same sentence. It is not an accident, but the *salmonella* sentences give an obvious evidence of concept linking between words qualitatively and quantitatively.

The simple experiment of co-occurrence statistics between medical disease (*i.e gonorrhea, leukaemia, and bronchitis*) and some negative verbs in English (*i.e eradicate, infect, die, and suffer*) are chosen to show the probability of lexical occurrences by employing the correlation analysis. The experimental samples were chosen from two large corpora— COBUILD and ORCHID. The experimental result of correlation (r=1) indicates that the more strongly connected two lexical items are, the more positively completely qualitative approach correlates with quantitative approach as shown in graph below.

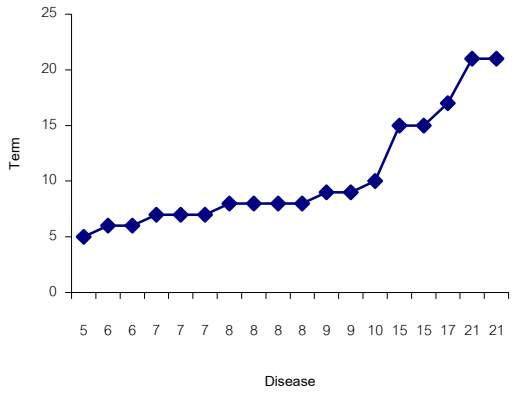

Figure 3 Graph represents correlation betweeen concept linking of words qualitatively and quantitatively

In addition, the result of alternative hypothesis in the experiment gave evidence of correlation between qualitative and quantitative approach in which these two variables have significantly statistical correlation at α=0.05 as shown in Figure 4.

Ho: $\rho = 0$
H1: $\rho \neq 0$

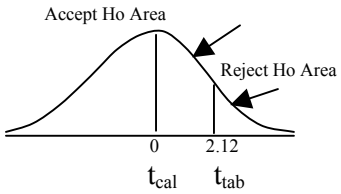

Figure 4 The alternative hypothesis of co-occurrence experiment

Another experiment, “การ”[kaan] and “ความ”[khwaam] function as a suffix *–al or -ness* in English which change from a lexical category to another one (e.g. arrive → arrival). This process is called the “nominalization” pattern in linguistics study. Despite their high frequency of “semantic and structural coincidences” with *noun, verb*, and *adjective*, it is hard to explain how they are used and learners of Thai often confuse them. When asked to choose between two sentences below, native speakers will only give some opaque reasons on their own intuitions that “ความเหมาะสม”[khwaammɔ̀sɔ̌m] in (a) is more suitable for making a natural sentence than in (b).

(a) **การเหมาะสม**น่าจะเป็นเดือนธันวาคมมากกว่า

[kanmɔ̀sǒm nâa jà pen dɨən thanwakhom mâkkhwà]

It should be in December.

(b) **ความเหมาะสม**น่าจะเป็นเดือนธันวาคมมากกว่า

[khammɔ̀sǒm nâa jà pen dɨən thanwakhom mâkkhwà]

It should be in December.

In fact, a linguist can describe that "การ" occurs frequently with several action verbs (e.g, กิน[kin], วิ่ง[wîŋ], เล่น[lên], ให้[hây], เที่ยว[thîew], โต้เถียง[tôthĭəŋ], ประท้วง[pràthúəŋ], ปรับ[pràp], บริการ[bɔɔ́rikan], and etc.) which never occurs to "ความ". On the other hand, "ความ" often occurs in conjunction with adjectives which rarely occur to "การ". It seems to be clearer when the empirical information of "การ" and "ความ" are shown in Table1.

Categories	N	V	Adj	Total
"การ"	6	93	1	100
"ความ"	4	46	50	100

Table1 Word classes occurring with "การ" and "ความ"

In this experiment, 100 "การ" and 100 "ความ" samples were gathered from Thai ORCHID corpus and compared their occurences with noun, verb, and adjective. The different amount of "การ" and "ความ" occurring with adjective forms significantly supports the semantic appropriateness of "ความเหมาะสม" used in sentence (b).

Negatively, consistencies of lexical representation in corpora data has become insufficient to confirm a real language use. For example, a native speaker can spontaneously tell that the polysemous word ซึ้ง[sɨ́ŋ] belongs to both *noun* as "*food container"* and *adjective* as *"profound or satisfied"*, but corpus data shows this word in only adjective form. Another quantitative findings of "การ" and "ความ" mentioned before are reliable to clarify the complexities of qualitative approach, but not at all practical for every event, especially in case of "การ", "ความ", and dynamic verbs. For example, both "ความเคลื่อนไหว"[khaamkhlɨ̂ənwăy] and "การเคลื่อนไหว"[kankhlɨ̂ənwăy] can be applied to the sentence (c) below.

(c) **การ/ความเคลื่อนไหว**ทางการเมืองของฝ่ายค้านจะยิ่งมีน้ำหนักยิ่งขึ้น

[kan/khamkhlɨ̂ənwăy thaŋ kanmɨəŋ khŏŋ fày khá̂an jà yîŋ mii námnàk yîŋ khɨ̂n]

The political movement of opposition will be more weighted.

The statistic descriptions, of course, can explain concretely, but all the frequencies cannot answer about semantic complexities between "การ" and "ความ" phenomenon in (c). However, this study may be clearly described at higher level such as discourse and pragmatic corpora. These samples prove that qualitative approaches are so much subtle that corpus technology cannot represent their richness.

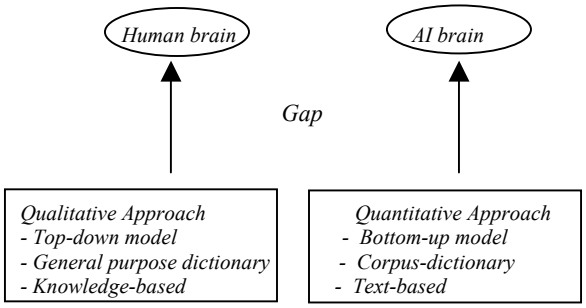

Figure 5 Gap exists between qualitative and quantitative approaches.

Consequently, there is still a huge gap between qualitative and quantitative approaches as shown in Figure5. In practice, language descriptions of qualitative approach are recently concerned with the abstraction of semantic relatedness between word senses and also between languages. In contrast to qualitative analysis, quantitative approach must be based on statistical and methodological clarity as shown in several examples of this paper. In present language studies, a qualitative research is often a predecessor for quantitative analysis, provided that the gap between two approaches has been dimensionally fulfilled with both subjective and objective criterion. Up to the present, it means that the right side of diagram in Figure 5 will not reach a supreme goal of the left side, if

quantitative approach works against with qualitative one. The qualitative investigation in top-down model will never be sufficiently systematic and the quantitative one in bottom-up model never delicate enough to replace the other exactly.

In corpus-based technology, a very large and efficient corpora in Thai and English is to be extracted from many everyday language uses to improve the present bottom-up definition model. For corpus-based LEXiTRON, thus, the further research should emphasize on English<->Thai parallel corpus (see more details in Tony McEnery & Andrew Wilson, 1996) at all language levels and Thai sentence segmentation which is important to qualify the real bilingual corpus-based dictionary.

References

Boas, F. (1940). **Race, Language, and Culture.** New York: Macmillan.

Domnern, G. & Sathienpong, W. (1999). **พจนานุกรมไทย-อังกฤษ (Thai-English Dictionary).** (2nd ed.). Bangkok: Amarin Printing and Publishing.

Duangjai, V., Marasri, P., Suphap, D., & Sorachai, P. (1994). Correlation Analysis. In **สถิติธุรกิจ [Statistics for Business]**. (pp.175-186). Bangkok: Chulalongkorn University Press.

Eskey, D. E. (1988). Holding in the bottom: An interactive approach to the language problems of second language readers. In P. Block, (1986). The comprehension strategies of second language readers. **TESOL Quaterly, 20,** 463-494.

Isahara, Hitoshi. (1997). Cooperation for R&D of the natural language processing technology between Japan and Thailand. **Technical Report ORCHID Corpus, 1997-2001.** (pp. 1-3).

Louise, G. (1993). A Note on Lexical Disambiguation. In C. Souter & E. Atwell (Eds.), **The Coupus-based computational linguistics.** (pp. 225-237). Rodopi: Netherlands.

Nida, E. A. (1964). **Towards a science of translation.** Leiden: E.J. Brill.

Palmer, F. R. (1981). **Semantics.** Cambridge: Cambridge University Press.

Pornpilas, R. (1990). **การทำพจนานุกรม (Lexicography).** Master's Thesis. Chiengmai University, Thailand

Sydney, I. L. (1993). **Dictionary: the art and craft of lexicography.** Cambridge: Cambridge University.

Tony, M. E. & Andrew, W. (1996). **Coupus linguistics.** Edinburgh: Edinburgh University Press.

Wit, T. (1998). **Thai-English dictionary (Library edition).** Bangkok: Aksorn Pittaya Press.

The Royal Institute.**พจนานุกรมฉบับราชบัณฑิตยสถาน** [Dictionary]. (CD-ROM). (1982). The Royal Institute (Royal Institute & National Electronics and Computer Technology Center).

พจนานุกรมไทยฉบับทันสมัย [Up to dated edition Thai Dictionary]**.** (2000). Bangkok: Se-education Press.

พจนานุกรมนอกราชบัณฑิตยสถาน [None-Royal Institute Thai Dictionary]**.** (2001). Bangkok: Matichon Press.

British National Corpus(BNC). [On-line]: http://sara.natcorp.ox.ac.uk/lookup.html

Collins Cobuild. [On-line]: http://titania.cobuild.collins.co.uk/form.html

Encarta World English Dictionary. [On-line]. (2001). http://www.dictionary.msn.com/

London Slangs. [On-line]. http://www.londonslang.com/

WordNet 1.7 Vocabulary Helper. [On-line]. http://www.notredame.ac.jp/cgi-bin/wn

ORCHID Corpus. National Electronics and Computer Technology Center, Thailand.

.