Extracting and Annotating Extended Lexical Units of Culinary Terms for Korean Culinary Manuscripts of *Joseon* Period[1]

Kil-Im Nam, Hyeon-Ju Song, Jun Choi & Young-Hee Hyun

Keywords: *ELUs (Extended Lexical Units), culinary terms, terminological lexicography, semantic annotation.*

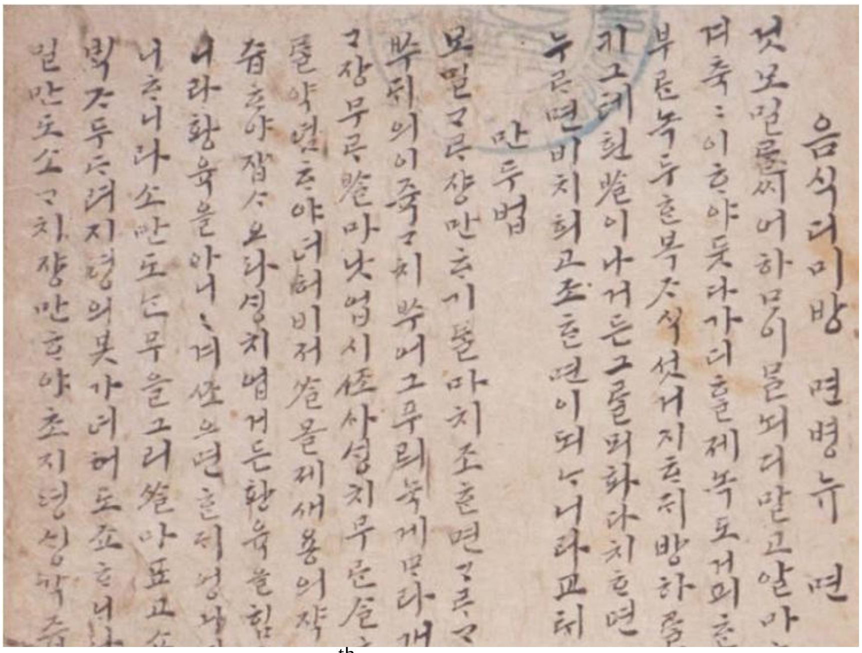
Abstract

This is a follow up study of the previously reported project conducted from 2007 to 2009. In the previous project, a corpus of culinary manuscripts was constructed with rich morphological and semantic annotations. However, the morpheme based annotation was not sufficient for extracting traditional culinary terms since many terms are in the form of so-called 'extended lexical units (ELUs).' To tackle the limitations of the original annotations, This research attempted to apply phrase level semantic annotation. By extracting ELUs of culinary terms, firstly, richer information of the expressions could be obtained. Secondly, more accurate annotation has been achieved in the current research. Lastly, the products attained from this study can be applied to compile domain-specific dictionaries (in this case, culinary domain) and contribute to extend lemma status to multi-word items.

1. Introduction

This is a follow up study of the previously reported project *Compiling and Developing a Corpus of 17-19th Century Korean Culinary Manuscripts and a Customized Corpus Browser* conducted from 2007 to 2009. In the previous project, we constructed a corpus of culinary manuscripts with rich morphological and semantic annotations. A customized corpus browser was also implemented to maximize the usability of the corpus. The project provided valuable and specialized source of information on Korean traditional dishes and recipes for general public as well as linguists and culinary experts. It also effectively demonstrated a novel method of terminological lexicography utilizing historical material.

However, the morpheme based annotation was not sufficient for extracting traditional culinary terms since many terms are in the form of so-called ‘extended lexical units (ELUs). To tackle the limitations of the original annotations, we attempted to apply phrase level semantic annotation in current project.



음식디미방 면병뉴 면
것모밀ᄅᆞᆯ ᄡᅥ허 ᄆᆞ이 ᄇᆞᆯ되디 말고 알마
겨 축〻이 ᄒᆞ야 둣다가 ᄃᆡ흘제 녹도거피ᄒᆞᆫ
부ᄅᆞᆫ녹두 ᄒᆞᆫ복ᄌᆞ식 섯거 지ᄒᆞ되 방하 즁
키 그ᄅᆡ 힌ᄡᆞᆯ이 나거든 그ᄅᆞᆯ 펴 화다치ᄒᆞᆫ 면
누ᄅᆞᆫ 면비치 희고 조ᄒᆞᆫ 면이 되ᄂᆞ니라 교퇴
만두법
모밀ᄀᆞᄅᆞ 장만ᄒᆞ기ᄂᆞᆫ 마치 조ᄒᆞᆫ 면 ᄀᆞᄅᆞ ᄆᆞ
ᄡᆞᆨ 뒤의 이 죽ᄀᆞ치 ᄡᆞᆨ어 그무리 ᄂᆞᆨ게 ᄆᆞ다 개
ᄀᆞ장 무ᄅᆞᆫ ᄡᆞᆯ 마낫 업시 ᄉᆞ셔 사섬치 무ᄅᆞᆫ 슴
폴 약혀 ᄒᆞ야 너허 비저 ᄆᆞᆯ올제 새옹의 작
쳡 ᄒᆞ야 잡ᄉᆞ오 다셩치 엄거든 환육ᄋᆞᆯ 힘
나라 황육ᄋᆞᆯ 아나 〻 겨샤ᄋᆞ면 ᄒᆞᆫ디 셩나
너ᄎᆞ니라 소만도ᄂᆞᆫ 무ᄋᆞᆯ 그리 ᄡᆞᆯ아 묘고 소
ᄇᆞᆨ 깃두ᄃᆞ려 지령의 봇가 너허도 죠ᄒᆞ니
길만도소ᄀᆞ치 쟝만ᄒᆞ야 초지령 싱박 즙

Figure 1. An example of a 17th century hand-written culinary manuscript.

2. Implication of extracting extended lexical units

Sinclair (1991:109-110) claims that a considerable part of human language use is based on the 'idiom principle' instead of the 'open-choice' principle. The idiom principle, according to Sinclair's definition, is the tendency that native language speakers use 'semi-preconstructed phrases'. The major factors constructing the idiom principle are (a) typical register in which language is used, (b) language-external situation in which language is recurrently used reflecting human affairs, and (c) economy of effort.

However, it has to be mentioned that each component of the high frequency phrases does not presuppose semantic non-compositionality. In other words, these components frequency occur only because they are conventionally used together frequently. In addition, as Hanks (2010) points out, even though technical terms are used infrequently, they tend to include interesting clusters of, if not phraseological, domain-specific expressions.

The main purpose of this study is to prove that the semantic unit of specific terms should be treated on the phrase level. To fulfill this purpose, we have analyzed ELUs from a particular set of texts, *the Korean culinary manuscripts of Joseon period*.

ELU normally function as a specific semantic unit. There are some ELUs that are not semantically idiomatic but occur in clusters very frequently in some particular registers. Some examples of these ELUs are *cwumek-makom* 'as much as a fist' QUANTITY, *payksyeycakmalhaye* 'to wash clearly' COOKING PROCEDURE, *panmannikkyesikiko* 'halfway to let cool off' COOKING STATE, *senulha-n tey* 'in a cool place' STORING PLACE. This kind of semantic units are meaningful and useful for culinary experts who searching for practical and technical information from recipe books. We also examined the quantitative and qualitative properties of ELUs to detect unit in the particular text.

According to Sinclair et al. (2004:24), the ultimate dictionary should describe a comprehensive list of semantic units in a given language. This has an important implication on compiling dictionaries of technical terms or expressions used in recipes, for instance. A recipe is a kind of *appeal text* (Brinker, 1992), which delivers accurate information for the recipients to be able to perform certain actions. Thus information users would browse texts for semantic information, and they will search for phrase level semantic units in most cases. Our attempt to extract and annotate ELUs in culinary texts is based on this practical need.

Table 1. Enhanced semantic classes in ELU level annotation.

Lexical item	Previous annotation		Present annotation	
	Morphemes	Semantic Classes	ELUs	Semantic Classes
(a) *senulha-n tey* Cool-RL place 'in a cool place'	*senulh*	-	*senulha-n tey*	STORING PLACE
	n	-		
	tey	-		
(b) *khongcwukkachi* bean porridge like 'like bean porridge'	*khongcwuk*	PORRIDGE	*khongcwukkachi*	COOKING STATE
	kachi	-		
(c) *himcwulepsansal* Stringy there not be-RL meat 'soft meat'	*himcwul*	-	*himcwulepsansal*	INGREDIENTS
	epsan	-		
	sal	INGREDIENTS		

In Table 1, the expression (a) *senulha-n tey* was not semantically annotated in the previous study. However, the whole cluster of *senulha-n tey* can be annotated as an independent unit in a semantic class STORING PLACE. Similarly, the semantic class for expression (b) *khongcwukkachi* was changed from PORRIDGE to COOKING STATE. In case of (c) *himcwulepsansal*, more accurate information has been added even though the semantic class remains unchanged. The whole cluster *himcwulepsansal* annotated with a semantic class INGREDIENTS in ELU level yields far more concrete and useful information than *sal* which was previously annotated in morpheme level.

3. ELU extraction strategy

To overcome the limitation of morpheme based analysis described above, we have annotated expressions in ELU level using the classifying system of the culinary terms (Paek et al., 2008). Our annotation and extraction scheme is composed of two procedural steps.

In the first step, we retrieved and annotated a selection of ELUs containing morphemes concerned with 'space', 'time', 'status', and 'unit'. As shown in Table 2, it is clear that more specific and richer information on COOKING and/or STORING PLACE(←SPACE), SEASON TO COOK(←TIME), COOKING STATE(←STATUS), etc. can be collected when the terms are examined in the group of ELUs. This high quality information is not accessible with lists of simple morphemes.

Table 2. Examples of ELUs including specific morphemes.

Morpheme Level		ELU Level		
morphemes	semantic classes	ELUs		semantic classes
tey place 'in a place'	PLACE	*senulha-n tey* cool-RL place 'in a cool place'	*tewu-n tey* warm-RL place 'in a warm place'	STORING PLACE

ttay, cey, nal time, time, day ‘at the time, on the day’	TIME	*paykossyenghiphi-l ttay* pear-flowers-well-bloom-RL time ‘when pear flowers are in bloom’	*tasaha-n cey* warm-RL when ‘when peach flowers are in bloom’	SEASON TO COOK
kachi, katko like, like ‘like, such as’	STATUS	*sanghwa cci-ki-kachi* sanghwa steam-NOM-like ‘like steaming sanghwa’	*pichitaynip-kasko* *color-NOM-bamboo leaf-like* ‘like the color of bamboo leaves’	COOKING STATE
kirey, makom as long as, as much as ‘unit of measure’	UNIT	*han chi kilay* *one-chi-unit* ‘3.03cm’	*yakkwa-makom* yakkwa-as ‘as much as yakkwa(Korea cookie)’	MEASURING UNIT

Secondly, we extracted lexical items manually whenever the items needed to be annotated as an ELU while reading the source material.

ELU level annotation has turned out to be beneficial both in qualitative and quantitative aspects. In quantitative aspect, 12% more items were detected in total. A handful of ELUs were extracted in the categories of COOKING STATE & TIME, INGREDIENTS, and STORAGE in particular. Table 3 shows the numbers of additionally discovered ELUs for each semantic class.

Table 3. Annotations by the morphemic units vs. the ELUs.

Semantic Classes	Number of morpheme units	Number of ELUs
NAME OF DISH	277	-
INGREDIENTS	911	320
SPICES AND GARNISH	270	-
KITCHENWARE	299	10
MEASURING UNIT	367	-
COOKING PROCEDURE	1561	30
COOKING STATE & TIME	375	125
STORAGE	6	7
TASTE	54	-
Total	4120	492

Analysis in ELU level has improved the quality of information and expanded its quantity by incorporating the context of the phrases. Identical results cannot be achieved through analysis in morpheme level. As a consequence of analyzing ELUs, we have obtained (a) additional information, which could not be detected by annotation in morpheme level, (b) more accurate semantic classes including the context, and (c) possibility of extracting more meaningful semantic units even when an expression is annotated in common categories both in morpheme and ELU levels.

From the comparison of these results, it seems obvious that the problems in our previous work were raised because the basic units of the semantic annotation were morphemes. Therefore, it is only natural that the semantic annotation in phrase level enabled us to harvest more accurate, specific and richer information for the same terms.

4. Summary

By extracting ELUs of culinary terms over the morpheme level terms, firstly, richer information of the expressions could be obtained. Secondly, more accurate annotation has been achieved in the current research since the context of each technical term was taken into consideration. Lastly, the products attained from this study can be applied to compile domain-specific dictionaries (in this case, culinary domain) and contribute to extend lemma status to multi-word items.

Note

[1] The Yale Romanization is used to transcribe the Korean data. The abbreviations are as follow: NOM- nominalizer, RL- relativizer.

References

Biber, D. et al. 1999. *The Longman grammar of spoken and written English*. London: Longman.

Biber, D., Y. Kim, and T. Nicole 2010. 'A Corpus-driven Approach to Comparative hraseology: Lexical Bundles in English, Spanish, and Korean.' *Japanese/Korean Linguistics* 17: 75–94.

Brinker, K. 1992. *Linguistic text analysis*. (Third Edition) Berlin: Erich Schmidt.

Hanks, P. 2010. 'Terminology, Phraseology, and Lexicography.' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*. Ljouwert: Fryske Akademy / Afuk.

Paek, D. H., K. I. Nam, M. H. Lee, E. J. Ahn and H. J. Song 2008. 'Compiling and Developing a Corpus of 17-19th Century Korean Culinary Manuscripts and a Customized Corpus Browser.' In E. Bernal and J. DeCesaris (eds.), *Proceedings of the XIII Euralex International Congress: Barcelona, 15-19 July 2008*. Barcelona: L'Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra, 741–746.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the text; Language, corpus and discourse*. London: Routledge.

Stubbs, M. 2009. 'The Search for Units of Meaning; Sinclair on Empirical Semantics.' *Applied Linguistics* 30.1: 115–137.