Analysis of Idiom Variations in English for the Enhanced Automatic Look-up of Idiom Entries in Dictionaries

Kyo Kageura
Library and Information Science Course,
Graduate School of Education, University of Tokyo.
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan.
kyo@p.u-tokyo.ac.jp

Miwa Toyoshima
Graduate School of Arts and Sciences, University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902, Japan.
miwatysm@tj8.so-net.ne.jp

Abstract
In this paper, we report the results of our analysis and classification of variation patterns of English idioms. The results of the classification are to be used as a basis for developing enhanced automatic look-up functions for idiom entries in dictionaries. Although many high-quality dictionaries contain a sufficient number of idioms for their intended users, the methods available for looking up entries in both paper and electronic dictionaries and machine translation systems are not satisfactory. Providing adequate automatic look-up functions is complicated by the existence of idiom variants, which can often be very creative. In order to develop a high-quality look-up mechanism for idiom entries, therefore, it is necessary to clarify the variation patterns of idioms so that a computationally tractable mechanism for locating idioms and their variations can be developed. For our analysis, we used informants to obtain a range of possible idiom variations. Variations were first classified into paradigmatic and syntagmatic variations, then detailed patterns were defined from the syntactic and semantic point of view.

1 Introduction

Although many high-quality dictionaries contain a sufficient number of idioms for their intended users, the methods available for looking up idiom entries are far from satisfactory, not only in paper dictionaries but also in electronic dictionaries. In the case of paper dictionaries, the user has to guess the core constituent word of an idiom in order to locate the entry under which the idiom can be found. In the case of some electronic dictionaries, the situation is much better as the user can rely on "and" searches of the constituents of an idiom, but problems can still ocur when users are aware that there is an idiom within a certain syntagmatic chunk but cannot identify exactly which syntagmatic chunk constitutes the idiom.

Automatic look-up methods embodied in machine translation systems are not satisfactory, either. The automatic matching of occurrences of idioms in texts to idiom entries (which are given in standard forms) fails in many situations. Take, for instance, the following examples:

(1) He said it with his tongue in his cheek.

(2) He said it with his big fat tongue in his big fat cheek.

Many available machine translation systems successfully detect the idiomatic expression "with one's tongue in one's cheek" in (1), but none among those we checked were able to properly translate (2). Existing methods for looking up idioms cannot deal with the rich variations of idioms that are abundant in the ordinary texts we read.

In order to develop a high-quality look-up mechanism for idiom entries, therefore, it is necessary to clarify the variation patterns of idiomsso that a computationally tractable mechanism for locating idioms and their variations can be developed. This paper reports the results of our analysis of variation patterns of English idioms

2 The position of the present research

There are a few linguistic and lexicographical studies of the typology of idioms and their variations (e.g. Benson 1985; Moon 1998). In addition, many basic language references devote some space, though often scattered, to describing the nature of idioms and sometimes their variations (e.g. Quirk, et. al., 1985; Biber, et. al., 1999). The descriptions given in these studies are, however, for human use and not for machine use.

On the other hand, in the field of natural language processing (NLP), research in automatic extraction of collocations and idioms that is not based on the use of dictionaries exists (e.g. Smadja, 1993; Piao, et. al., 2003; Widdows & Dorow, 2005), but research in automatic matching of idiom entries to their occurrences in running texts does not.

The present work, which gives a basic description and classification of idioms and their variants, is similar to descriptive studies in linguistics and lexicography. The difference lies in the fact that our overall objective is to develop a computationally tractable mechanism for automatically matching idiom entries to occurrences in running texts. It is thus necessary in our case to restrict the levels of analysis and description to those which can be reliably dealt with by existing resources and computational methods. After reviewing the existing resources and their reliability, we decided on the following criteria:

1. To use POS-taggers and/or morphological analysers, but not parsers, as the performance of current parsers is not sufficient for robust and real-time use in dealing with actual texts. This means that we assume POS information and finite-state automata but no class of computation more powerful than that.

2. To assume that high-quality thesauri are available (e.g. Fallbaum, 1998), but not detailed conceptual classifications or resources with discourse information.

Note also that the described range of variation patterns, when incorporated into automatic matching algorithms, can be overgenerative though not excessively, because the computational problem is defined here as the problem of matching when both ends are given, rather than the problem of generating acceptable variations.

3 Preparation of the data

Given the lack of effective methods for collecting idiom variations under relevant idiom entries – the very mechanism that we intend to develop – the basic idiom variation data had to be collected manually. Because any data, whether collected from informants or from corpora, is limited, and we were concerned with the range of patterns of variations rather than the issue of which variations are used how many times, we decided to collect the idiom variation data using informants. To do so, we randomly selected idiom entries from a widely used English-Japanese idiom dictionary (McCaleb & Iwasaki, 2003), and asked three native English speakers (a Canadian, an American and an Australian, the latter two of whom are professional editors) to generate possible variations of these idioms. Table 1 gives the basic quantities of idioms and their variations. It should be noted that the data only provides basic variations; it gives neither the maximum types of possible variations nor negative examples. This limitation in the nature of the data can be compensated for to a large degree by the generalised description of variation patterns, but what is lacking from the analysis should be remedied at a later stage, after the basic look-up mechanism is developed.

Informants	# idioms	# variants	# variants/idioms	maximum	minimum
h	475	469	0.99	2	0
j	661	890	1.35	5	0
s	777	822	1.06	8	0
Total	1913	2181	1.14	8	0

Table 1. Basic quantities of idiom variation data

4 The result of the analyses

4.1 Patterns of idioms

Type	Example	# in the data	Tag
NP (pre modification)	poker face	215	a1
NP (post modification)	babe in arms	79	a2
NP (mixed)	a rotten apple in the barrel	55	a3
VP (general verbs)	take the plunge	905	b1
VP (“be” verbs)	be all balled up	64	b2
Adjectival phrases	straight as an arrow	112	c
Adverbial phrases	all along	26	d
Prepositional phrases	with open arms	201	e
Independent clause	it's your baby	195	f
Dependent clause	if you prefer	40	g
Other	popsicle	21	h

Table 2. Basic types of idioms.

The idioms we dealt with are classified into 11 basic types, as given in Table 2. Although fully understanding the fact that the essential nature of idioms cannot be properly represented by simple syntactic patterns of idiom forms, these basic types serve for our purpose of describing idiom variations as a basis for establishing a look-up mechanism, because the implementation of this mechanism, as explained, can rely only on syntactic and lexical information.

4.2 Patterns of idiom variations

Idiom variations can be classified into (1) the paradigmatic replacement of one or more constituents, (2) syntagmatic augmentation (i.e. modification and/or insertion of new elements) and (3) deletion. In our data, the first two patterns were the most common. In addition, we recognised (4) cases in which these variations co-occur with dependencies, and (5) a small number of complex variations which we simply put into a "miscellaneous" class. Table 3 summarises these patterns with examples. In some cases, paradigmatic replacement and syntagmatic augmentation cannot be so clearly distinguished. For instance, "cash and carry one's point" as a variation of "carry one's point" can be interpreted either as "carry" being replaced by "cash and carry" or as "cash and" being inserted before "carry". Our criteria in these cases are practical. In this case, for instance, it is unlikely that a thesaurus that shows the semantic relation between "carry" and "cash and carry" could exist as the latter unit is not completely lexicalised. For automatic processing, therefore, it is more convenient to treat this as a case of syntagmatic augmentation.

Type	Example	#	Tag
Paradigmatic replacement	one's head screwed on right -> screwed on wrong	759	x
Syntagmatic augmentation	make allowance for -> make unholy allowance for	1203	y
Deletion	not get to first base -> got to first base	3	s
Dependent replacements	more dead than alive -> more alive than dead	39	xx
Dependent augmentations	There's no accounting for preferences -> One can't account for preferences	101	yy
Replace & augment	weak as a baby -> strong as a baby ox	91	xy
Deletion & replacement	go back to the basics -> plunge into the basics	20	sx
Deletion & augmentation	people will talk -> people happily talk	8	sy
Others	take it from me -> rely on me	95	z

Table 3. Broad classification of variations of idioms

Paradigmatic replacement (x) was further classified by the part-of-speech (i.e. n – noun, v – verb, ad – adjective, av – adverb, p – preposition, det – articles, cj – conjunction, aux – auxiliary verb, dg – change between determiner and genitive case nouns) of the elements that are replaced as well as by the type of relations between the replacing and replaced elements. Table 4(a) shows the numbers of paradigmatic replacements in the data by subtypes. Note that a single variation in our data can be complex and can contain more than one variation, so the total number of variations by individual variation patterns is larger than the number of variants given in Table 1. In Table 4(a), "indirect" means that the replacing and replaced elements are connected through the third element in the idiom, "context sliding" refers to non-semantic replacement by sound similarity etc., "phrasal" is the replacement of the same phrasal unit as a whole, and "singular/plural" is, of course, relevant only for the replacement of noun elements.

Replacement by antonymous, synonymous, equal-status or generic/specific elements constitutes about 75% (245 out of 327) of noun replacements, about 69% (133 out of 192) of verb replacements, about 80% (113 out of 141) of adjective replacements and 65% (26 out of 40) of adverb replacements. Although the quantities here do not directly reflect the actual occurrence of variation patterns, they indicate that we can deal with the major types of idiom variations generated by paradigmatic replacements by using standard thesauri, which contain this information. Among "others," we observed creative replacement taking advantage of sound similarities or other features, such as "back and forth" → "buck and forth". Though it is easy to see the relationship between the original idiom and these kind of variations for human readers, it would be rather complex for machines to deal with such replacements systematically.

		x-n	x-v	x-ad	x-av	x-p	x-det	x-cj	x-aux	x-dg
antonymous	20	16	22	12	8	-	-	-	-	
synonymous		133	79	61	12	6	-	-	-	-
equal-status		88	38	30	2	-	-	-	-	-
replacement with addition	4	0	13	0	-	-	-	-	-	
indirect		2	3	3	0	-	-	-	-	-
context sliding	3	2	1	1	-	-	-	-	-	
phrasal		23	5	1	3	-	-	-	-	-
generic/specific	4	-	-	-	-	-	-	-	-	
singular/plural		3	-	-	-	-	-	-	-	-
others		47	49	10	10	21	8	5	4	7
total	327	192	141	40	35	8	5	4	7	

Table 4(a). Number of paradigmatic replacements by subtypes

Syntagmatic augmentation (y) was classified into subcategories by the part-of-speech and the formal/semantic role of the augmenting elements, as well as by the positions they augment. Table 4(b) shows the number of syntagmatic augmentations by subtypes. Note that pp stands for prepositional phrase, cl stands for clause, p stands for preposition and other notations are the same as in Table 4(a). The positional information is not given for the sake of succinctness.

		y-n	y-v	y-ad	y-av	y-pp	y-cl	y-det	y-p	y-aux
negation	0	0	6	13	-	-	-	-	-	
magnification		2	0	163	265	-	-	-	-	-
marked coordination	9	3	7	6	-	-	-	-	-	
unmarked coordination	0	1	15	0	-	-	-	-	-	
general modification		44	-	402	167	-	-	-	-	-
genitive	8	-	-	-	-	-	-	-	-	
others	20	2	16	21	25	2	2	0	4	
total	83	6	609	472	25	2	2	0	4	

Table 4(b). Number of syntagmatic augmentations by subtypes

Syntagmatic augmentation basically follows ordinary syntactic rules. Upon closer inspection, it was observed that the modifications and augmentations can be divided into three major types, i.e. (a) omnipresent modifications by a few notable lexical items such as "very," "fucking," "reall," (b) ordinary modifications that follow standard semantic relationships between the modifier and the modified, and (c) creative modifications through sound similarity and other mechanisms, such as "blind as a bat" → "blind as a baseball bat" (note that this was also observed in paradigmatic replacement). The fact that the majority of syntagmatic augmentations follow standard syntactic rules is a promising sign for automatic matching of idiom entries with variations, but the extent to which we can narrow down the search space by incorporating systematic restrictions on augmentation patterns is a point to be examined in future research.

s-n	s-av	s-p	s-det	s-gen	xx	xy	yy	sx	sy	z
0	2	0	1	0	39	91	101	20	8	95

Table 4(c). Number of deletions and other variations.

Deletion (s) was classified by means of the part-of-speech of the deleted elements. Table 4(c) shows the number of deletions as well as the number of dependent combinations of different variation generating operations. The number of variations made by combinations of different variation generating mechanisms is rather high. Many of these can simply be dealt with by ignoring the dependencies and breaking down the variations into basic variation types. How to deal with the others requires further analysis.

4.3 Summary

All in all, we observed that the majority of variations fall into the category of paradigmatic replacement or syntagmatic augmentation. What is positive from the point of view of automatic processing is that in both cases, it seems possible to properly deal with the major types of variations by means of standard thesauri and the definition of basic syntactic patterns. Although further analysis and actual experiments to test the performance of an actual automatic look-up mechanism based on the framework given here are needed, our analysis shows that there is a realistic possibility that such a mechanism could be a significant improvement on the mechanisms currently available.

5 Conclusions

We have shown the result of the analyses of variation patterns of English idioms, which are described at the levels of (1) syntactic patterns when syntagmatic variations are concerned and (2) thesaural relations when paradigmatic variations are concerned. As idioms are thought to be quintessentially a unit consolidated within a given context and within discourse, the descriptive level we adopted in the present study is not necessarily sufficient for qualitatively characterising the essential range of variations of idioms. The level of analysis we adopted, however, is important and useful for realising a computationally tractable method of matching idiom entries to idiom occurrences in texts automatically. We are currently developing a look-up system, starting from an excessively over-generative variant matching method and limiting the variant generations gradually by incorporating the results of the analyses reported here.

Acknowledgement

This research was partly supported by the National Institute of Information and Communication Technology (NiCT) and by grant-in-aid (A) 17200018 "Construction of online multilingual reference tools for aiding translators" from the Japan Society for the Promotion of Sciences (JSPS).

References

A. Dictionaries

McCaleb, J. G., Iwasaki, M. (2003), *All-Purpose Dictionary of English Idioms and Expressions*, Tokyo, Aasahi Publishing.

B. Other Literature

Benson, M. (1985), 'Collocations and idioms', in Ilson, R. (ed.) *Dictionaries, Lexicography and Language Learning*. Oxford, Pergamon Press, pp. 61-68.

Biber, D. et al. (1999), *Longman Grammar of Spoken and Written English*, London, Longman.

Fallbaum, C. (ed.) (1998), *WordNet*, Cambridge, Mass, MIT Press.

Moon, R. (1998), *Fixed Expressions and Idioms in English*, Oxford, Clarendon Press.

Piao, S. S. L. et al. (2003), 'Extracting multiword expressions with a semantic tagger' in *Proceedings of the ACL2003 Workshop on Multiword Expressions*, pp. 48-53.

Quirk, R. et al. (1985), *A Comprehensive Grammar of the English Language*, London, Longman.

Smadja, F. (1993), 'Retrieving collocations from text: Xtract', *Computational Linguistics*, 19 (1), pp. 143-177.

Widdows, D., Dorow, B. (2005), 'Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns', in *Proceedings of the ACL2005 SIGLEX Workshop on Deep Lexical Acquisition*, pp. 48-56.