Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 1317–1339

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing☆

Toshifumi Tanabe[a,*], Masahito Takahashi[b], Kosho Shudo[a]

[a] *Fukuoka University, 8-19-1, Nanakuma, Jonan-ku, Fukuoka 814-0180, Japan*
[b] *Kurume Institute of Technology, 2228-66, Kamitsu, Kurume 830-0052, Japan*

Received 11 October 2012; received in revised form 16 July 2013; accepted 6 September 2013
Available online 14 September 2013

Abstract

Since Sag et al. (2002) highlighted a key problem that had been underappreciated in the past in natural language processing (NLP), namely idiosyncratic multiword expressions (MWEs) such as idioms, quasi-idioms, clichés, quasi-clichés, institutionalized phrases, proverbs and old sayings, and how to deal with them, many attempts have been made to extract these expressions from corpora and construct a lexicon of them. However, no extensive, reliable solution has yet been realized. This paper presents an overview of a comprehensive lexicon of Japanese multiword expressions (Japanese MWE Lexicon: JMWEL), which has been compiled in order to realize linguistically precise and wide-coverage natural Japanese processing systems. The JMWEL is characterized by significant notational, syntactic, and semantic diversity as well as a detailed description of the syntactic functions, structures, and flexibilities of MWEs. The lexicon contains about 111,000 header entries written in kana (phonetic characters) and their almost 820,000 variants written in kana and kanji (ideographic characters). The paper demonstrates the JMWEL's validity, supported mainly by comparing the lexicon with a large-scale Japanese N-gram frequency dataset, namely the LDC2009T08 generated by Google Inc. (Kudo and Kazawa, 2009). The present work is an attempt to provide a tentative answer for Japanese, from outside statistical empiricism, to the question posed by Church (2011): "How many multiword expressions do people know?"

Keywords: Natural language processing; Multiword expression (MWE); Lexicon; Linguistic idiosyncrasy; Non-compositionality; Dependency structure; Internal modification

1. Introduction

The fact that linguistically idiosyncratic multiword expressions occur in authentic sentences with unexpectedly high frequency causes serious problems in linguistically comprehensive and precise natural language processing (NLP) based on compositionality. Since Sag et al. (2002), NLP researchers have become aware that proper treatment of such idiosyncratic multiword expressions (MWEs) is one of the most central and intriguing problems in the field.

In principle, the nature of the idiosyncrasy of MWEs is twofold: one is idiomaticity, i.e., non-compositionality of meaning; the other is the strong probabilistic affinity between component words. Many attempts have been made to extract these expressions from corpora capturing these properties, mainly using automatic methods that exploit

☆ This paper has been recommended for acceptance by 'Dr. E. Briscoe'.

* Corresponding author. Tel.: +81 92 871 6631.
E-mail address: tanabe@fukuoka-u.ac.jp (T. Tanabe).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.09.001

statistical means. However, to our knowledge, no reliable, extensive solution has yet been made available, presumably because of the data-sparseness problem in the corpus-based approach and the difficulty of extracting correctly without human insight. Recognizing the crucial importance of such expressions, one of the authors of the current paper began in the late 1960s to construct for general use a Japanese computational lexicon with comprehensive inclusion of idioms, idiom-like expressions, and probabilistically idiosyncratic expressions. In this paper, we begin with an overview of this lexicon named JMWEL (Japanese MWE Lexicon). The lexicon contains about 111,000 head entries, written in the form of hira-gana (phonetic character) strings, and their almost 820,000 notational variants, i.e., the mixed strings of kana (phonetic characters) and kanji (ideographic characters) which are common in Japanese written texts.

As with simplex words, MWEs are generally categorized into function MWEs and content MWEs. Japanese has various types of function MWEs that play important roles similar to the syntacto-semantic roles of case-marking particles or auxiliary verbs. However, most previous MWE studies in NLP have been limited to particular subcategories of content MWEs, such as phrasal verbs, phrasal adverbs, compound verbs, compound nouns, light verb constructions (LVCs) and verb-object idioms. The present paper is partly based on our earlier work (Shudo et al., 2011), focusing on content MWEs, but has been substantially expanded to include function MWEs. Thus, the JMWEL is the combination of two sublexicons: the JMWEL/F containing function MWEs and the JMWEL/C comprising content MWEs. The combined JMWEL is thus a remarkably comprehensive resource awaited for decades in the field of NLP.

The most important features of the JMWEL are:

1. A large notational, syntactic, and semantic diversity of contained expressions.
2. A detailed description of the syntactic function, and syntactic structure of each entry expression.
3. An indication of the syntactic flexibility of entry expressions (i.e., the possibility of additional, internal modification of constituent words).
4. An all-in-one architecture with uniform encoding schemas for each MWE.

In Section 2, we outline the main features of the present study, first presenting a brief summary of significant previous work on this topic. In Section 3, we propose and describe the criteria for selecting MWEs and introduce a number of classes of multiword expressions. In Section 4, we illustrate some important statistical properties of the JMWEL mainly by comparing the lexicon with a large-scale Japanese N-gram frequency dataset, the LDC2009T08, generated by Google Inc. (Kudo and Kazawa, 2009). In Section 5, we outline the format and content of the JMWEL/F, discussing the information on notational variants, syntactic functions, syntactic structures, and semantic labels. In Section 6, we outline the format and content of the JMWEL/C, discussing the information on notational variants, syntactic functions, morphosyntactic structures, contextual conditions, inflectional variants of state-describing expressions, and the syntactic flexibility of MWEs. The paper ends with concluding remarks in Section 7.

2. Related work

Gross (1986) analyzed French compound adverbs and compound verbs. According to his estimate, the lexical stock of such words in French would be respectively 3.3 and 1.7 times greater than that of single-word adverbs and single-word verbs. Jackendoff (1997) notes that an English speaker's lexicon would contain as many MWEs as single-words. Sag et al. (2002) pointed out that 41% of the entries of WordNet 1.7 (Fellbaum, 1999) are multiword. Uchiyama and Ishizaki (2003) reported that 44% of Japanese verbs are VV-type compounds. These and other similar observations underscore the great need for a well-designed, extensive MWE lexicon for practical natural language processing.

In the past, attempts have been made to produce an MWE lexicon. Examples include the following: Gross (1986) reported on a dictionary of French verbal MWEs with description of 22 syntactic structures; Kuiper et al. (2003) constructed a database of 13,000 English idioms tagged with syntactic structures; Baptista et al. (2004) reported on a dictionary of 3500 Portuguese verbal MWEs with ten syntactic structures; Fellbaum et al. (2006) discussed corpus-based studies used in the development of German verb phrase idiom resources; Laporte and Voyatzi (2008) reported on a dictionary of 6800 French adverbial MWEs annotated with 15 syntactic structures; and Wang and Yu (2010) reported on a Chinese idiom database, containing 38,000 idioms (derived from 11,000 basic idioms), and its applications. Our JMWEL approach differs from these studies in that it covers a comprehensive range of MWEs types, namely