CLASSIFICATION OF MWES IN HINDI USING ONTOLOGY

RAKHI JOON AND ARCHANA SINGHAL
Department of Computer Science, University of Delhi, New Delhi, India
Email: [1]rjoon30@gmail.com, [2]singhal_archana@yahoo.com

ABSTRACT: Multi Word Expressions (MWEs) are very important aspect of Natural Language Processing tasks. In Hindi there are some variations in MWE types as compared to English and other languages. Proposed work provides a brief overview of various MWE in Hindi, their usage and significance. Some of these types are proposed by many authors and some have not received proper attention of researchers. In proposed work some of new categories of multi words are included in the previous classification of Hindi MWE to reveal new concepts related to MWEs. An Ontology is designed based on the proposed classification. Since Ontology itself is a classifier so it becomes easy to understand and use this new classification and their applications in NLP tasks like Question Answering, Information Retrieval, Machine translation, and so on.

KEYWORDS: MWE, Ontology, Idioms, Adverbs, Named Entities.

INTRODUCTION

Various authors described the Multi Word Expression (MWE) concept as "Idiosyncratic concepts that cross word boundaries" (Sag et al., 2002). As per the linguistic properties, "*MWE are lexical items that can be decomposed into multiple lexemes and display lexical, syntactic, semantic, pragmatic and statistical idiomaticity"* (Baldwin and Kim, 2010). In the above definition the first part describes that MWEs are made up of two or more words delimited by white space, the second part covers all the necessary properties for a word to become a MWE. All these are based on one property i.e. idiomaticity in context of MWE. Idiomaticity refers to deviation from the basic properties of the words and applies at lexical, syntactic, semantic, pragmatic and statistical level. Basically, MWEs is any phrase that cannot be *entirely* predicted on the basis of standard grammar rules and lexical entries (http://mwe.stanford.edu/reading-group.html). The meaning of MWEs cannot be directly given by its components (Venkatapathy and Agrawal, 2007). Some examples of MWE in Hindi are रेल गाडी (*rail gaadi,* Train), जल प्रपात (*jal prapaat,* waterfall), and so on. Various researchers have worked on MWE in English but in Hindi MWEs have not gained proper attention of researchers. Proposed work is a modest attempt to explore various types of MWEs present in Hindi and classify them on the basis of their properties and usage using Ontology (Martin, 2011, Knublauch et al., 2004). The paper is organized in the following manner: The next section gives a brief review of related work. In the coming section, a brief description of Existing types of MWEs in Hindi is given, which is followed by the proposed work section, which mainly covers a new classification of MWEs in Hindi, the ontology of Hindi MWEs and discussion then finally the paper is concluded in the conclusion and future work section.

RELATED WORK

Present section discusses about work done in the area of MWE by eminent researchers in past few years. Basic terms of MWEs are described in (Sag et al., 2002, Baldwin and Kim, 2010). A three stage statistical approach for identifying multiword terms based on the co-related text-segments

existing in a group of documents is presented in (Chen et al., 2006), and it is shown that information retrieval (Gune, 2010) plays a big role in extracting MWE from corpus. A good review of almost all types of MWEs covered by the various researchers is given in (Baldwin and Kim, 2010). Many types of MWEs are discussed along with their linguistic properties and research issues. Basic classification of MWEs is also included by the authors. In (Sinha, 2011), the author examined various types of MWE encountered in Hindi. Many of these types were not given proper attention by other researchers for e.g. 'vaalaa' construct, doublets, replication etc. These types are examined from machine translation viewpoint. Many of these are frequently used in day to day activities but are not given place in formal textual corpus. The author presented a stepwise mining of various MWEs in Hindi and their machine translation perspectives. To represent MWEs in Static Machine Translation (SMT) pipeline, a phrase based SMT framework is adopted in (Ghoneim and Diab, 2013). Various types of MWEs are considered for the proposed approach: Verb-based MWEs (VNC, VPC, and LVC), Noun-based MWEs (NNC, and NE), Adjective (AJ) and Adverb (AV) based MWE. A list of MWE extracted from English WordNet database 3.0 is also used and named entities are also considered as a type of MWE e.g. Abraham Lincoln. In (Mukerjee et al., 2006, Chakrabarti, 2008, Venkatapathy, 2005, Venkatapathy and Agrawal, 2007, Kunchukuttan and Damani, 2008, Gune, 2010) various complex predicates are discussed. Five types of V+V sequences are observed in (Chakrabarti, 2008) and in (Venkatapathy, 2005), the authors integrated all the existing features and investigated a range of classifiers for their suitability for recognizing the non compositional V-N collocations. V-N type of MWEs covers a very large percentage of all MWEs and are important for NLP applications like machine translation, etc. The N+V expressions in Hindi are analyzed in (Venkatapathy and Agrawal, 2007) and an approach to measure their relative compositionality using Maximum Entropy Model (MaXEnt) is proposed. Some of the features are computed by mapping the N+V expressions in Hindi to V-N expressions in English. A system for extracting compound noun MWE for Hindi from the given corpus is presented in (Kunchukuttan and Damani, 2008). On the basis of linguistic and psycholinguistic principles the authors has also mentioned the major categories of compound noun MWEs. Statistical methods are used for ranking of collocations and thus the results are presented in form of precision, recall and other measures. Various other classifications of MWEs include Hindi morphemes (Sinha, 2009), replicating words (Sinha and Thakur, 2005), Idioms (Priyanka and Sinha, 2014), and named entities (Saha et al., 2008, Srivastava et al., 2011). The concepts of ontology are discussed in (Martin, 2011, Knublauch et al., 2004).

BACKGROUND

Since Hindi grammar contains a rich set of various kinds of words and method of their usage, the major properties of MWE, which have received particular position in Hindi MWE literature, are discussed here. A Hindi sentence itself exhibit various properties and predicates like noun, pronoun, verb, adjective, adverb and so on. In formation of a multi word or extraction of MWEs from a corpus, first it is detected whether the combination of words which is to be categorized as MWE exhibits the necessary and sufficient conditions of MWE which mainly include that words have to be separated by space or delimiter, non-compositionality of meaning and idiomaticity. for e.g. अस्त्र शस्त्र (*aastra shastra*, Weapons). Some of the MWEs types got attention of researchers and some remain hidden, so in our classification some new types are also included along with existing classification of MWEs for Hindi (Sinha, 2011). Various existing types of Hindi MWEs are as follows.

Acronyms and Abbreviations

In Hindi the acronyms and abbreviations are different from that in English and other languages. For e.g. the name of 11th president of India is 'अवुल पकिर जैनुलाअबदीन अब्दुल कलाम '(*Abul Pakir JanulAabdin Abdul Kalam*) and is abbreviated as 'ए पी जे अब्दुल कलाम '(*A. P. J. Abdul Kalam*). In the same way Hindi acronym for 'इंडियन नेशनल लोकदल '(*Indian National Lokdal*) is 'इनेलो '(*INeLo*). Since acronyms are single words but they represent MWEs (Sinha, 2011).

Complex Predicates

Various types of complex predicates are there which can be included in the list of MWEs for Hindi. The Adjective-Verb combination (Mukerjee et al., 2006) for e.g. उपलब्ध है (*uplabdh hai*, available), Adverb+Verb combination like वापस लेना (*vaapas lena*, return back), compound nouns (Kunchukuttan and Damani, 2008) like रेल गाडी (*rail gaadi*, train), compound verbs (Chakrabarti, 2008) like मार डालना (*maar daalna*, to kill) and Noun+Verb (Venkatapathy, 2005, Venkatapathy and Agrawal, 2007) like जीत दिलाना (*jeet dilaana*, to draw victory) , अवसर देना (give an opportunity), and so on.

Foreign Words and Terms

Now-a-days it is very common to mix foreign words in Hindi conversation (Sinha, 2011) and many of these foreign words may appear as MWEs with absolute combination. These may include institutionalized phrases (Ghoneim and Diab, 2013) like '*traffic light*'. The foreign root word may undergo morphological variations, phonetic variations according to Hindi grammar rules. Some of foreign MWEs are: सीनियर डॉक्टर (*Senior Doctor*), पेइंग गेस्ट (Paying Guest), and so on.

Idioms

As described in (Priyanka and Sinha, 2014), the meaning of Idioms like other cannot be derived from the meanings of individual elements and usually refer to some event, story or action behind it. Idioms are not only language specific but also specific to ethnicity and cultural relations. Not all idioms can be included in the list of MWEs, it depends on the necessary and sufficient conditions. Some example idioms in Hindi which can be included in MWEs list are: अंगूठा दिखाना: इंकार करना (*to Show thumb*, to refuse), ऊँगली उठाना: आलोचना करना (*to lift finger*, to blame), कान काटना: पराजित करना (*to cut ear*, to defeat), and so on.

Morphemes

Morphemes are Hindi word-elements used to form full words. These are mainly of two types: a) Prefix (Affixes attached to the beginning of Hindi words), b) Suffix (Affixes attached to the end of Hindi words). A detailed use of one of Hindi morpheme 'वाला (*vaalaa*)' is described in (Sinha, 2009), for e.g. सितारों वाली चुनरी (stole with shinning work), पीने वाला पानी (*pine vaalaa pani*, drinking water), दूध वाला (*dudh vaalaa*, milkman), and so on. Various other morphemes can be included like the 'वाला 'morpheme.

Replicating words

Many languages contain replicating words for general usage and part of speech. Generally, these types of words retain the non-compositionality property of MWEs. Some of replicating words from Hindi are: जगह जगह (*place place*, everywhere), अभी अभी (*now now*, recently), etc. sometimes a particle can come in between the replicating words and thus change their meaning (Sinha, 2011) e.g. पैसा ही पैसा (*paisa hi paisa*, a lot of money), and so on. It should be remembered that not all replicating words form MWE.

PROPOSED WORK

Proposed Classification of Hindi MWEs

The proposed work identifies various new types of Hindi MWEs, mainly the proverbs, named entities and adverbs types, since already existing classification do not cover these types of Hindi MWEs. The motivation to identify these types comes from the grammar rules of Hindi and literature survey. Ontological classification of MWEs types in Hindi is also a new concept which helps in finding the relationship between the two or more MWE types. Ontological representation also helps to add new categories in the classification with ease if some more additions are to be done in future. The new classification and ontology creation is a step towards the direction of giving a standard classification for Hindi MWEs types like English (Ghoneim and Diab, 2013). The new types which are added in the proposed research work are listed below:

Proverbs

A Proverb is a short popular saying which represents a general truth. In Hindi it is called 'कहावत ' (*Kahavat*). Some of popular examples of proverbs which can be included in MWEs list are: जिस की लाठी उस की भैंस (*Whoever owns the lathi (a huge cane/stick) owns the buffalo*, Might is right), दूर के ढोल सुहावने लगते हैं (*The drums sound better at a distance*, The grass is always greener on the other side), and so on.

Named Entities

Named Entities (NE) can be in form of person, designation, organization, abbreviations, brand, location names, etc. Not every named entity form a MWE. Some of the examples of Hindi are taken from the NEs which form MWE e.g. मनमोहन सिंह (Manmohan Singh), नई दिल्ली (New Delhi), उच्च श्रेणी लिपिक (*Uacch shreni lipik*, Upper Division Clerk), and so on.

Adverbs

As per grammar definition, an adverb is a word that simplifies the meaning of a grammar constructs like verb, adjective, other adverb, clause, or sentence and express in form of manner, place, time, or degree. Since from the generic types of adverbs three types are found to contain some MWEs and these are described as:

*Adverb of Place(*स्थानवाचक*)*

The words in which verb appears as attribute of place, are called adverbs of place. E.g. बाहर भीतर (*bahar bhitar,* in and out), आगे पीछे (*aage piche,* front and back), ऊपर नीचे (*upper niche,* up and down), and so on

*Adverb of Quantity (*परिमाणवाचक*)*

The words in which verb appears as attribute of quantity, are called adverbs of quantity. E.g. थोडा थोडा (*thoda thoda,* very less), बारी बारी (*baari baari*, one by one), and so on.

*Adverb of Time (*कालवाचक*)*

The words in which verb appears as attribute of time, are called adverbs of time. E.g. अभी अभी (*abhi abhi,* recently), पहली बार (*pehli bar,* first time), आज कल (*aaj kal,* now-a-days), and so on.

Adverb of Manner (रीतिवाचक)

The words in which verb appears as the manner of actions being done, are called adverbs of manner. E.g. धीरे धीरे (*dhire dhire,* very slowly), भली भांति (*bhali bhanti,* clearly), and so on.

The above examples of proverbs, named entities and the types of adverbs can be considered as MWEs, and this is a new classification added in the list of Hindi MWEs. Next, the designed ontology based on proposed classification of MWEs in Hindi is described.

Ontology of MWEs in Hindi

The existing classification of Hindi MWEs given by various researchers is done by considering various grammar constructs. A specific tool was not used for classifying MWEs and thus ontology has been designed based on the existing and proposed classification of MWE in Hindi to give a new direction for classification. Ontology consists of concepts, attributes and properties representing relationships between concepts (Martin, 2011). Ontology mainly exhibits following properties:

a) Names for important concepts in the domain,
b) Background knowledge of the domain.

The main elements of ontology are the Classes (which represent the concepts) along with their hierarchy, concept properties also called as slots or attributes, restrictions on these properties (like type, cardinality, domain, etc.), relations between concepts (disjoint, equality, etc.) and Instances. To build ontology for specific application the domain and scope related to that application should be known. Then the important terms are find out and the next step is to define classes and their hierarchies then slots are defined and finally restrictions on the slots are defined (like cardinality, value-type, and so on).

Protégé is used for ontology development in our work. Various other tools are also available like NeOn Toolkit, SWOOP, Vitro, Ontofly, etc. but due to its popularity and pluggable feature it is preferred by most of researchers. Protégé architecture is divided into a “model” part and a “view” part. Protégé’s *model* is the internal representation mechanism for ontology and knowledge bases. Protégé s *view* components provide a user interface to display and manipulate the underlying

model (Knublauch et al., 2004). The above described classification is represented here in form of ontology as shown in Fig 1.

In the proposed ontology, some of the types are taken from previous classification like acronyms and abbreviations (Sinha, 2011), complex predicates (Mukerjee et al., 2006, Chakrabarti, 2008, Venkatapathy, 2005, Venkatapathy and Agrawal, 2007, Kunchukuttan and Damani, 2008, Gune, 2010), foreign words and terms (Sinha, 2011), morphemes ('vaalaa') (Sinha, 2011) and replicating words (Sinha and Thakur, 2005). Some more types which are described by various authors but are not considered as MWEs in Hindi are idioms (Priyanka and Sinha, 2014) and named entities (Saha et al., 2008, Srivastava et al., 2011). Hindi grammar also contain some more constructs which can be considered as MWEs like Proverbs, morphemes types (Postfix and Prefix), Adverbs types and so on. These are the new categories which are considered along with existing types of Hindi MWEs.

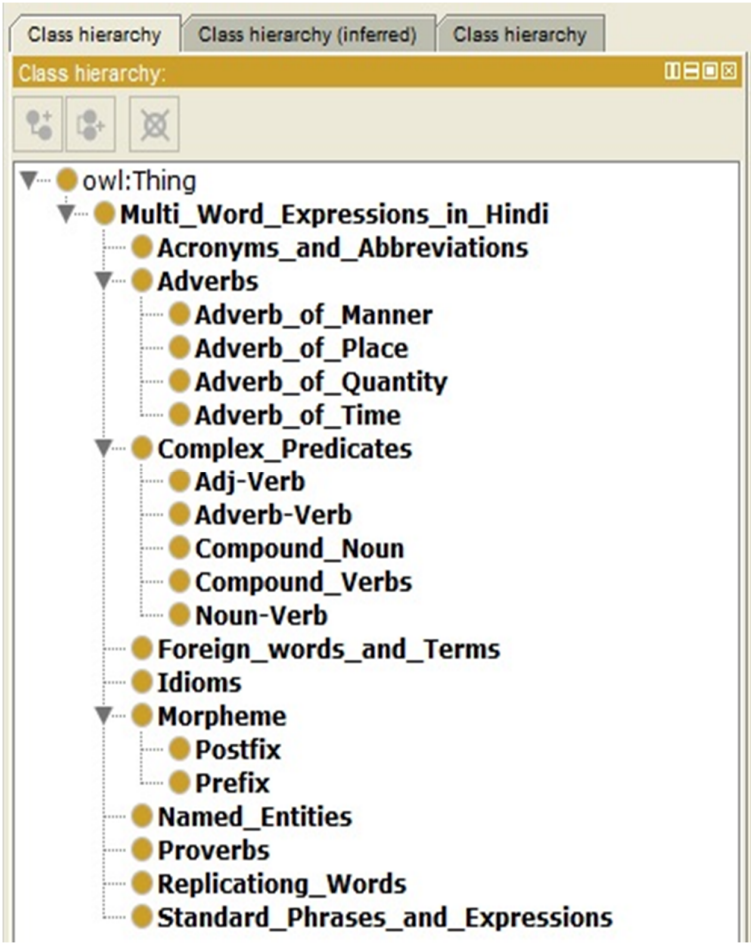

Figure 1. Ontology class hierarchy of Hindi MWEs

Testing and Discussion

A sample paragraph of Hindi is taken from Wikipedia for testing purpose as shown in Fig. 2. It can be extended for large corpus and further performance evaluation can be done. A comparison between the existing and proposed types of MWEs in Hindi is done.

Using the existing categories of MWEs in Hindi as explained in section III, following types are found in the above sample: 'ए पी जे अब्दुल कलाम' (Abbreviation), 'भारतीय राजनीती', 'भारतीय गणतंत्र' and तीव्र इच्छा (Complex predicates), 'गाइडेड मिसाइल्स' (Foreign word), 'धाक जमाई' (Idiom), 'करने वाले' (Morpheme) and 'साथ साथ' (Replicating word).

डॉक्टर ए पी जे अब्दुल कलाम (अवुल पकिर जैनुलाअबदीन अब्दुल कलाम) भारतीय गणतंत्र के ग्यारहवें निर्वाचित राष्ट्रपति हैं। वे भारत के पूर्व राष्ट्रपति, जानेमाने वैज्ञानिक और अभियंता के रूप में विख्यात हैं। उनके शिक्षक इयादुराई सोलोमन ने उनसे कहा था कि 'जीवन मे सफलता तथा अनुकुल परिणाम प्राप्त करने के लिए तीव्र इच्छा, आस्था, आपेक्षा इन तीन शक्तिओ को भली भांति समझ लेना और उन पर प्रभुत्व स्थापित करना चहिये। भारतीय राजनीती के साथ साथ उन्होंने शोध के शेत्र में भी अपनी धाक जमाई। डॉक्टर कलाम ने स्वदेशी लक्ष्य भेद करने वाले अस्त्र (गाइडेड मिसाइल्स) की खोज की। इनका कार्य काल 25 जुलाई 2007 को समाप्त हुआ। आज कल डॉक्टर कलाम भारतीय अंतरिक्ष विज्ञान एवम् प्रोधोगिकी संस्थान के कुलपति है। हम सब इनके लिए यही कामना करते है "तुम जियो हजारों साल, साल के दिन हो पचास हजार"।

Figure 2. Sample Hindi paragraph taken from Wikipedia

In the proposed work some more categories are added in the existing classification for accuracy as these new types were not covered in existing classification. Following are the types found using proposed classification in the above sample: '**तुम जियो हजारों साल, साल के दिन हो पचास हजार' (proverb),** 'अवुल पकिर जैनुलाअबदीन अब्दुल कलाम', 'पूर्व राष्ट्रपति', इयादुराई सोलोमन and '25 जुलाई 2007' (Named entities) and 'भली भांति', 'आज कल' (adverbs). The highligted words represents the existing types and the underlined words represent the proposed types.

It can be observed from above discussion that proposed classification is giving more accurate types of Hindi MWEs as compared to existing classification, and it will lead to good results in terms of high performance.

CONCLUSION AND FUTURE WORK

In this paper, a review of all types of Hindi MWEs is carried out and some additions to the existing types are done along with an ontological framework representation of Hindi MWEs types. As there is a lot of research done on English MWEs, but MWEs in Hindi has not gained proper attention of

researchers. Therefore in proposed work Hindi MWEs types are analyzed and discussed. Since the main aim of this study is to develop an accurate classification of MWEs in Hindi, some new types like adverbs (adverb of time, place, manner and quantity), proverbs and named entities, have been added to the existing classification for accuracy. Ontology is also developed for the proposed classification.

Since there may be other types which might be hidden in Hindi MWEs, the proposed ontology can be further refined for accuracy. Evaluation of performance of the proposed Hindi MWEs types in terms of Precision, Recall, F-1 measure is also suggested.

REFERENCES

I.A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, " Multiword expressions: A pain in the neck for NLP," *In: Proceedings of Third International Conference on Computational Linguistics and Intelligent Text Processing: CICLing-2002, Lecture Notes in Computer*, pp. 1-15, 2002.

J. Chen, C. Yeh, R. Chau, "Identifying Multi-Word Terms by Text-segments," *In: Proceedings of the Seventh International Conference on Web-Age Information Management Workshops (WAIMW'06)*, pp. 2–7, 2006.

T. Baldwin, and S.N. Kim, *Multiword Expressions Handbook of Natural Language Processing*, in Nitin Indurkhya and Fred J. Damerau (eds.) Second Edition, CRC Press, Boca Raton, USA, pp.267-292, 2010.

RMK Sinha, "Stepwise Mining of Multi-Word Expressions in Hindi," *In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA*, pp. 110–115, 23 June 2011.

M. Ghoneim, M. Diab, "Multiword Expressions in the Context of Statistical Machine Translation," *International Joint Conference on Natural Language Processing, Nagoya, Japan*, pp. 1181–1187, 14-18 October 2013.

A. Mukerjee, A. Soni, A.M. Raina, "Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora," *In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney*, pp. 28–35, July 2006

D. Chakrabarti, "Hindi Compound Verbs and their Automatic Extraction," *Coling 2008: Companion volume – Posters and Demonstrations, Manchester*; pp. 27–30, 2008.

S. Venkatapathy, "Relative compositionality of multi-word expressions : a study of verb-noun (V-N) collocations," *Lecture Notes in Computer Science. Natural Language Processing (IJCNLP 2005)*, vol. 3651, pp. 553–564, 2005.

S. Venkatapathy, P. Agrawal, "Relative Compositionality of Noun + Verb Multi-word Expressions in Hindi," *In: Proceedings of International Conference on Natural Language Processing (ICON), Kanpur*, November 2007.

A. Kunchukuttan, O.P. Damani, "A System for Compound Noun Multiword Expression Extraction for Hindi," *In: Proceedings of 6th International Conference on Natural Language Processing (ICON-2008*, 2008.

H. Gune, "Verbs are where all the action lies : Experiences of Shallow Parsing of a Morphologically Rich Language," *Coling 2010:Poster Volume, Beijing, August* 2010, pp. 347–355.

R.M.K. Sinha, “Learning Disambiguation of Hindi Morpheme "vaalaa’ with a Sparse Corpus,” *In: Proceedings of International Conference on Machine Learning and Applications,*pp. 653–657, 2009.

R.M.K. Sinha, A. Thakur, “Dealing with Replicative Words in Hindi for Machine Translation to English,” *In: proceedings of 10th Machine Translation summit (MT Summit X), Phuket, Thailand*; vol. 1992, pp. 157-164, 2005.

Priyanka, RMK Sinha, “A System for Identification of Idioms in Hindi,” *In: proceedings of Seventh International Conference on Contemporary Computing (IC3)*, pp. 467-472, August 2014.

S.K. Saha, S. Chatterji, S. Dandapat, S. Sarkar and P. Mitra, “A Hybrid Approach for Named Entity Recognition in Indian Languages,” *Proceedings of Workshop on NER for South and South East Asian Languages (IJCNLP-08), Hyderabad India*, pp. 17–24, January 2008.

S. Srivastava, M. Sanglikar and D.C. Kothari, “Named entity recognition system for Hindi language: a hybrid approach,” *International Journal of Computational Linguistics (IJCL)* vol. 2, no. 1, pp. 10-23, 2011.

A. Martin, “A framework for Business Intelligence application using Ontological classification,” *International Journal of Engineering Science and Technology (IJEST)*, vol. 3, no. 2, pp. 1213–1221, 2011.

H. Knublauch, R.W. Fergerson, N.F. Noy and M.A. Musen, “The Protégé OWL plugin: An open development environment for semantic web applications,” *In The Semantic Web–ISWC 2004* Springer Berlin Heidelberg, pp. 229-243, 2004.

T. Baldwin, “*Multiword Expressions*”, *Advanced course at the Australasian Language Technology Summer School (ALTSS 2004)*, Sydney, Australia, 2004.