Multiword Expressions

Thesis submitted to the Faculty of Engineering & Technology, Jadavpur University
In partial fulfillment of the requirements for the Degree Of

Master of Computer Science & Engineering

In the Department of Computer Science & Engineering

By

Tanmoy Chakraborty

Exam Roll No. – M4CSE11-23

Class Roll No. – 000910502033

Registration No. – 108432 of 2009-2010

Under the esteemed Guidance of

Dr. Sivaji Bandyopadhyay

Professor, Department of Computer Science and Engineering
Jadavpur University, Kolkata – 700032

Department of Computer Science & Engineering

Jadavpur University

Kolkata-700032

May, 2011

Department of Computer Science & Engineering
Faculty of Engineering & Technology
Jadavpur University

TO WHOM IT MAY CONCERN

I hereby recommend that the thesis entitled "**Multiword Expressions**" has been carried out by **Tanmoy Chakraborty** (Reg. No. 108432 of 2009-2010, Class Roll No. 000910502033 and Exam Roll No. M4CSE11-23), under my guidance and supervision may be accepted in partial fulfillment for the degree of **Master of Computer Science and Engineering** in the Faculty of Engineering and Technology, Jadavpur University.

(Prof. Sivaji Bandyopadhyay)
Thesis Supervisor
Department of Computer Science and Engineering,
Jadavpur University, Kolkata- 700032

Countersigned:

(Prof. Chandan Mazumdar)
Head of the Department,
Department of Computer Science and Engineering,
Jadavpur University, Kolkata- 700032

(Prof. Niladri Chakraborty)
Dean,
Faculty of Engineering and Technology,
Jadavpur University, Kolkata- 700032

FACULTY OF ENGINEERING & TECHNOLOGY
JADAVPUR UNIVERSITY
KOLKATA-700032

CERTIFICATE OF APPROVAL

The foregoing thesis is hereby accepted as a credible study of an engineering subject carried out and presented in the manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it is submitted.

FINAL EXAMINATION
FOR EVALUATION
OF THESIS

1. ____________________________

2. ____________________________

(Signature of Examiners)

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of his "**Multiword Expressions**" studies.

All information in this document have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name: Tanmoy Chakraborty

Examination Roll Number: M4CSE11-23

Thesis Title: **Multiword Expressions**

Signature with Date:

Dedicated to

My lovely

Parents

Who show me this world.

Acknowledgement

I express my sincere gratitude to **Prof. Sivaji Bandyopadhyay**, my guide for his affectionate and valuable guidance without whose help the present work could not have been a successful one. I am also indebted to him as a Professor who introduced me to the world of Natural Language Processing.

Also I thank **Prof. Chandan Mazumdar**, Head of the Department of Computer Science and Engineering, for his assistance in allowing me to work in the departmental laboratory without which my work would have been incomplete.

I also thank **Prof. Rana Duttagupta**, **Prof. Bijan Bihari Bhaumik** of Jadavpur University and **Mr. Kousik Dasgupta** of Kalyani Government Engineering College, Kalyani, Nadia, West Bengal for their assistance and constant support.

I would also like to convey my sincere gratitude to all my respected teachers and faculty members in this department for their invaluable suggestions and kind cooperation.

A great thank to my seniors **Dr. Sudip Kumar Naskar, Dr. Asif Eikbal , Partha Pakray, Anup Kumar Kolya, Sibabrata Paladhi, Dipankar Das, Pinaki Bhaskar, Amitava Das, Aniruddha Ghosh, Nongmeikapam Kishorjit Singh, Santanu Pal, Somnath Banerjee , Tanik Saikh** and **Tapabrata Mandol**, my friend **Bidhan Chandra Pal** and my junior **Soujanya Poria** in NLP laboratory in JU for their valuable suggestions and technical assistance. We are like a family in the technical world of NLP. Also I express my thanks to my class mate **Preetam Mukherjee** and all others friends in the class.

It would be unfair if I do not give thanks to our NLP Laboratory attendants **Rased Ali** and **Gopal Fakir** for their assistance at the time when we feel exhausted for a break and provide us with a friendly environment.

Last but not the least, the encouragement given by my mother **Mrs. Anuradha Chakraborty**, my father **Mr. Ranjit Chakraborty**, my uncle **Mr. Asim Bagish**, my aunt **Mrs. Kuheli Bagish** and my sweet little brother **Abhra Bagish** who have always been a constant source of inspiration and whose encouragement is beyond linguistic expression for me.

Jadavpur University, Kolkata

Tanmoy Chakraborty

Contents

Thesis advisor
Prof. Sivaji Bandyopadhyay

Author
Tanmoy Chakraborty

Multiword Expressions

Abstract

In natural languages, words can occur in single units called simplex words or in a group of simplex words that function as a single unit, called Multiword Expressions (MWEs). Although MWEs are similar to simplex words in their syntax and semantics, they pose their own sets of challenges. MWEs are arguably one of the biggest roadblocks in computational linguistics due to their high productivity and due to the bewildering range of syntactic, semantic, pragmatic and statistical idiomaticity they are associated with. In addition, the large numbers in which they occur in a text demand specialized handling. Moreover, dealing with MWEs has a broad range of applications, from syntactic disambiguation to semantic analysis in Natural Language Processing (NLP).

In this research, the main goal is to use computational techniques to shed light on the underlying linguistic processes that generate MWEs across constructions and languages; to generalize existing techniques by abstracting away from individual MWE types; and finally to exemplify the utility of MWE interpretation within general NLP tasks such as Machine Translation, Authorship identification and Stylometry Analysis.

In this thesis, we mainly target on Bengali MWEs and additionally take into account English MWEs as a continuation of parallel process. In particular, we focus on reduplicated phrases, noun compounds (NCs), verbal phrases (VPs) (complex predicates: compound verbs and conjunct verbs) in Bengali along with some other classes (Adjective-noun, verb-subject and verb-object combinations) of English MWEs due to their high productivity and frequency.

Besides resource-constrained and unsophisticated handling of Bengali language, the challenges dealing with the above mentioned phrases are manifold. For NCs, the challenges are: (1) identifying them from the corpus; (2) interpreting the semantic relation (SR) that represents the underlying connection between the head noun and modifier(s); (3) resolving syntactic ambiguity in NCs comprising three or more terms; and (4) analyzing the impact of word sense on noun compound interpretation. Our basic approach is to identify Bengali noun-noun (N-N) bigram MWEs from the Bengali using simple statistical approaches (Chapter 7). We also deal with the reduplicated phrases and try to explore the semantics using some traditional resources like English WordNet and Bengali monolingual and bilingual dictionaries (Chapter 6). Finally, identification task has been modified beyond the conventional treatment of MWEs due to the insufficiency of resources (i.e. corpus, WordNet etc.) in Bengali. This concept is highly motivated by the traditional definition of MWE in that the semantics of the composite phrase are likely to be unpredictable by the meaning of its parts. We call this approach as '*Semantic Clustering Approach*' (Chapter 7). Meanwhile, we have also proposed different taxonomies both for NCs and Reduplications in Bengali based on their linguistic evidences (Chapter 7 and Chapter 6).

We have two experiences dealing with the bulk amount of data within a short-fixed time boundary. Prior challenge was raised by the shared task named '*Semantic Evaluation (SemEval 2010): Task 5 – Automatic Keyphrase Extraction from Scientific Articles*' organized as part of the 48th Annual Meeting of the Association for Computational Linguistic (ACL 2010). We have been given the training and testing data which were related to computer Science domain and they were unformatted and noisy. For every article, we had to identify first fifteen relevant keyphrases with their stemmed forms. We have tackled two major issues in this regard: *Candidate selection* and *feature engineering*. To develop an efficient candidate selection method, first we take a supervised approach and analyze various properties of keyphrases which can be selected as the features of CRF based system like term frequency, Inverse term frequency, collective frequency, length, position, Part of speech, chunk and dependency and collect the outputs of the CRF as candidates. Secondly, we re-examine the existing features broadly and collect the nature and variation of keyphrases using regular expressions (Chapter 5).

The second challenge has been proposed by the share task named '*Distributional Semantics and Compositionality (DISCo)*' in conjunction with ACL 2011 (Chapter 8). This task focuses on the identification of compositionality of three types of English phrases (i.e. adjective-noun, verb-object and verb-subject combinations) from a large corpus. We have given statistical evidences against each phrase using traditional statistical methodologies of MWEs like frequency, Point-wise Mutual Information (PMI), T-score, Chi-square coefficient and perplexity (root and surface level). Each feature is cross-validated to check their individual impact and finally they are aggregated using average and weighted combination methods.

This thesis draws its conclusions showing some of the impacts of MWEs in Natural Language Processing applications. We have experimented with MWEs in two major application domains: (i) Stylometry and Authorship Identification (Chapter 9.1) and (ii) Textual alignment and Machine Translation (Chapter 9.2). The experiment concerning Stylometry was first for any Indian languages as far our knowledge is concerned. This domain is challenging because we want to analyze the impact of MWEs in the writer's style of writing and identifying the other influencing factors in Bengali writings by which the system may be able to identify the prospective author. We have experimented in two ways: (i) rule-based and (ii) machine learning approaches. Secondly, we have shown the impact of MWEs in textual alignment. Source and target side treatments of MEWs and considering them as single token in either or both side have led to the increase of the BLEU evaluation score of a Phrase based Statistical Machine Translation (SMT) system for English-Bengali machine translation system.

Finally, we conclude the thesis with a chapter-by-chapter summary and outline of the findings of our work, suggestions for potential NLP applications and a presentation of further research directions (Chapter 10).

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Motivation

Lexemes are the basic unit of natural (i.e., human) language. In a sentence, they combine together and interact to form structures and meaning. Lexemes can occur in single units called *simplex words*, which is the smallest lexical unit that contains meaning, or as multiple simplex words that function as a single lexical unit, called ***Multiword Expressions*** (MWEs). (1.1) – (1.4) show a number of MWEs (in bold) in context.

(1.1) The ***marketing manager*** can learn how to *take advantage of* the growing database...

(1.2) Most of the time it failed to ***make it out*** of the pit lane...

(1.3) They were ***by and large*** of the type postulate...

(1.4) You should also ***make a note*** of the ***serial number*** of your ***television video***...

Both simplex words and MWEs function as structural and conceptual units of language. However, MWEs often require deeper syntactic and semantic reasoning due to subtle interactions with the syntax and semantics of their component simplex words, or alternatively behavior which is completely at odds with their parts. In the following examples, the relationship between the MWEs and their component simplex words is relatively transparent.

(1.5) He immediately ***got on*** the bus.

(1.6) Everyone ***makes mistakes***.

(1.7) The ***bus driver*** accidentally hit the ***garbage bin***.

In (1.5)–(1.7), the MWEs are relatively easy to detect as their components occur continuously. The semantics of the MWEs in these examples is also predictable. The meaning of ***bus driver*** as "one who drives a bus" is easily accessible despite ***bus*** having meanings including "an electrical conductor that makes a common connection between several circuits" and "a car that is old and unreliable", and ***driver*** having meanings including "a golfer who hits the golf ball with a driver" and "a program that determines how a computer will communicate with a peripheral device"[1]. The process for disambiguating the semantics in context here is identical to that of determining the word sense, e.g., from among its many senses, based on analysis of the combinatory interaction between possible word senses of the lexemes in the sentence. However, while at this simplistic level MWEs are similar to simplex words in terms of their function within a sentence, they pose bigger challenges due to their syntactically and semantically unexpected behavior (Sag *et al.* 2002). (1.8) – (1.10) show more complicated MWEs where knowledge of the components alone is insufficient to predict the observed linguistic behavior.

(1.8) Kim ***took*** her pen ***out***.

(1.9) She likes to ***take a*** long ***bath*** for relaxation after exams.

(1.10) He will inherit when his grandfather ***kicks the bucket***.

(1.8) – (1.10) are MWE examples which are hard to recognize as a single unit due to their length or the fact that they are discontinuous. For example, although ***take out*** is an MWE, it is not immediately apparent that (1.8) includes a token instance of it since ***out*** is separated from the verb ***take***. Also, due to the internal modification by *long*, ***take a bath*** is not easily recognizable as a unit, or analogously, it is not immediately apparent that ***long*** is not a component of the MWE. In addition, MWEs are often confused with non-MWEs, e.g. the MWE vs. non-MWE usages of ***put on*** in "*put the coat on"* vs. "*put the coat on the table"*, respectively. As a result of such variations in the context of usage of MWEs, it is sometimes difficult to distinguish MWEs from compositional usages of the individual simplex words. Though often understated, understanding and processing language are overwhelmingly difficult without the means to syntactically recognize MWEs. MWEs are problematic semantically as well. The meaning(s) of an MWE cannot always be directly predicted from its component words. The contribution of the MWE components to its semantics can vary widely from no contribution from any single word

[1] Glosses taken from WordNet 2.1.

(e.g., *kick the bucket*), to a single component making the most significant contribution (e.g., *finish up*), to all words in an MWE contributing equally (e.g., *bus driver*). The meaning of *kick the bucket* as an MWE is “pass from physical life and lose all bodily attributes and functions necessary to sustain life”. However, unfortunately, neither *kick* nor *bucket* contains this meaning. Hence, estimating the exact meaning of *kick the bucket* from its parts is futile. It is also impossible to estimate the meaning of *by and large* as ‘mostly’ from the components *by* and *large*. Hence, semantically, some MWEs need a different treatment compared to simplex words.

The number of MWEs is estimated to be of the same order of magnitude as the number of simplex words in a speaker’s lexicon (Wood 1964; Gates 1988; Jackendoff 1973). To add to this, new types of MWE are continuously created as languages evolve (e.g. *shock and awe*, *cell phone*, *ring tone*) (Dias *et al.*1999). Regionally, MWEs vary considerably. For example, *take away* and *take out* have an identical meaning in the context of fast food outlets, but the former is the preferred expression in Australian English, while the latter is the preferred expression in American English. Another example is *mail box* and *post box* in the context of postal service, where the former is the preferred form in American English and the latter is the preferred form in Australian English. MWEs can also be used to represent information concisely (Levi 1978). For example, *winter school* is a compact way of expressing “a school which is held in the winter”. MWEs can also lend emphasis to language (Brinton 1985; Side 1990). For example, *up* in *finish up the food* adds the meaning of “completion”. That is, *finish up* has the meaning of *finish*, but it also contains the entailment that the *food* is completely consumed and emphasizes the completeness of the eating action.

There is a modest body of research on modeling MWEs which has been integrated into NLP applications, e.g., for the purpose of fluency, robustness or better understanding of natural language. Understanding MWEs has broad utility in tasks ranging from syntactic disambiguation to conceptual (semantic) comprehension. Explicit lexicalized MWE data helps to simplify the syntactic structure of sentences that include MWEs, and conversely, a lack of MWE lexical items in a precision grammar is a significant source of parse errors (Baldwin *et al.* 2004). Additionally, it has been shown that accurate recognition of MWEs influences the accuracy of semantic tagging (Piao *et al.* 2003), and word alignment in machine translation (MT) can be improved through a specific handling of the syntax and semantics of MWEs (Venkatapathy and Joshi 2006).

Syntactically, one of the major issues with MWEs is recognition, due to idiomatic and syntactically-flexible expressions. MWEs are often found in the form of semi- or non-fixed expressions. The components often inflect for number or tense (e.g., *family cars*, The plane has *taken off*). The occurrence of the components also varies with context. For example, modifiers can internally modify the components of MWEs (e.g. *make a **big** mistake*).

Semantically also, MWEs can cause difficulties for comprehension. MWEs can be semantically idiomatic, i.e., the meaning can be explicitly or implicitly derived from the components of MWEs or be completely unrelated to the semantics of the parts. It is also relatively common for the components of MWEs to combine compositionally to form competing analyses. For example, *a piece of cake* can be an MWE with meaning “any undertaking that is easy to do”, or alternatively it can be a simple compositional expression referring to a portion of cake. Moreover, MWEs are highly productive and their components are often used to generate novel MWEs. The verb *take*, for example, combines with a number of prepositions to form verb particle constructions including *take away*, *take off* and *take up*, each of which has distinctive semantics.

To add to these difficulties, MWEs occur in a bewildering array of syntactic and semantic types which are interrelated to varying degrees, such that neither is it possible to come up with a genuinely general-purpose analysis of all MWEs, nor is it adequate to try to document each individual MWE type independently. For example, while syntactically identifying instances of noun compounds such as *paper submission* and *chocolate bar* is relatively easy, it is much harder with other types of MWEs such as *in one’s shoes* and *break the ice*. Semantically, predicting the meaning of MWEs is relatively easy with some types of MWEs such as *take a walk* and *make a note (of)*, whereas with other MWEs such as *make out* and *kick the bucket* it is considerably more difficult.

MWEs pose significant challenges for NLP, and developing a framework for modeling MWEs both syntactically and semantically is vital to the furtherance of NLP.

1.2 Research Issues, Related Work and Focus

The major NLP tasks relating to MWEs are: (1) identifying and extracting MWEs from corpus data and disambiguating their internal syntax and (2) interpreting MWEs. Increasingly, these tasks are being pipelined with parsers and applications such as machine translation

(Venkatapathy and Joshi 2006). Depending on the type of MWE, the relative importance of these syntactic and semantic tasks varies. For example, with noun compounds, the identification and extraction tasks are relatively trivial, whereas interpretation is considerably more difficult.

Prior to detailing the computational tasks relating to MWEs, let us briefly define a number of MWE types which will recur in later discussions. Full details of the MWE types are described in Section 2.1.3. A *noun compound* (NC, e.g., *golf club* or *paper submission*) is a compound noun made up of two or more nouns. A *verb-particle construction* (VPC, e.g., *hand over* or *battle on*) is a verbal MWE made up of a verb and obligatory particle(s). A *light-verb construction* (LVC, e.g. *take a walk* or *make a mistake*) is a verbal MWE made up of a verb and (usually indefinite singular) object NP, where the verb has bleached semantics and the noun complement determines the semantics of the MWE to a large degree. A *determinerless prepositional phrase* (D-PP, e.g. *at school* or *on air*) is an adverbial MWE made up of a preposition and a singular noun without a determiner. Finally, an *idiom* (e.g., *kick the bucket* or *take a turn for the worse*) is an amalgam of words in a construction other than those explicitly identified above, which has different semantics to that of the combination of the individual components.

In the following sections, we discuss the primary research issues relating to MWEs, and prior work done in each area. In doing so, we offer our perspective on why these issues continue to pose a challenge for NLP.

1.2.1 Identification

Identification is the task of determining individual occurrences of MWEs in running text. The task is at the token (instance) level, such that we may identify 50 distinct occurrences of *pick up* in a given corpus. To give an example of an identification task, given the corpus fragment in (2) (taken from “The Frog Prince”, a children’s story), we might identify the MWEs in (2):

> (2) One fine evening a young princess **put on** her bonnet and clogs, and **went out** to **take a walk** by herself in a wood; ... she ran to **pick** it **up**; ...

In MWE identification, a key challenge is in differentiating between MWEs and literal usages for word combinations such as *make a face* which can occur in both usages (*Kim* ***made a face*** *at the policeman* [MWE] vs. *Kim* ***made a face*** *in pottery class* [non-MWE]). Syntactic ambiguity is also a major factor, e.g. in identifying VPCs in context. For example, in the sentence ‘Kim **signed in** the room’, there is ambiguity between a VPC interpretation (sign in = “check in/announce arrival”) and an intransitive verb + PP interpretation (“Kim performed the act of **signing in** the room”).

MWE identification has tended to take the form of customized methods for particular MWE construction types and languages (e.g. English VPCs, LVCs), but attempts have been made to develop generalized techniques, as outlined below. Perhaps the most obvious method of identifying MWEs is via a part-of-speech (POS) tagger, chunker or parser, in the case that lexical information required to identify MWEs is contained within the parser output. For example, in the case of VPCs, there is a dedicated tag for (prepositional) particles in the Penn POS tagset, such that VPC identification can be performed simply by POS tagging a text, identifying all particle tags, and further identifying the head verb associated with each particle (e.g. by looking left for the first main verb, within a word window of fixed size) (Baldwin and Villavicencio 2002; Baldwin 2005a). Similarly, a chunker or phrase structure parser can be used to identify constructions such as noun compounds or VPCs (McCarthy, Keller, and Carroll 2003; Lapata and Lascarides 2003). This style of approach is generally not able to distinguish MWE and literal usages of a given word combination, however, as they are not differentiated in their surface syntax. Deep parsers which have lexical entries for MWEs and disambiguate to the level of lexical items are able to make this distinction, however, via supertagging or full parsing (Baldwin et al. 2004).

Another general approach to MWE identification is to treat literal and MWE usages as different senses of a given word combination. This then allows for the application of word sense disambiguation (WSD) techniques to the identification problem. As with WSD research, both supervised (Patrick and Fletcher 2005) and unsupervised (Birke and Sarkar 2006; Katz and Giesbrecht 2006; Sporleder and Li 2009) approaches have been applied to the identification task. The key assumption in unsupervised approaches has been that literal usages will be contextually similar to simplex usages of the component words (e.g. *kick* and *bucket* in the case of *kick the bucket*). Mirroring the findings from WSD research, supervised methods tend to be more accurate, but have the obvious drawback that they require large numbers of annotated literal and idiomatic instances of a given MWE to work. Unsupervised techniques are therefore more generally applicable.

A third approach, targeted particularly at semantically idiomatic MWEs, is to assume that MWEs occur: (a) in canonical forms, or (b) only in particular syntactic configurations, and do not undergo the same level of syntactic variation as literal usages. This relates to the prediction of the non-decomposable VNICs, where the prediction is that VNICs such as *kick the bucket* will not be passivised or be internally modifiable. If we have a method of identifying the limits of syntactic variability of a given MWE, therefore, we can assume that any usage which falls outside these (e.g. *kicked a bucket*) must be literal. The problem, then, is identifying the degree of syntactic variability of a given MWE. This can be performed manually, in flagging individual MWE lexical items with predictions of what variations a given MWE can undergo (Li, Zhang, Niu, Jiang, and Srihari 2003; Hashimoto, Sato, and Utsuro 2006). An alternative which alleviates the manual overhead associated with hand annotation is to use unsupervised learning to predict the "canonical" configurations for a given MWE, which can optionally be complemented with a supervised model to identify literal usages which are used in one of the canonical MWE configurations (e.g. *Kim* ***kicked the bucket*** *in frustration, and stormed out of the room*) (Fazly, Cook, and Stevenson 2009).

In research to date, good results have been achieved for particular MWEs, especially English VPCs. However, proposed methods have tended to rely heavily on existing resources such as parsers and hand-crafted lexical resources, and need to be tuned to particular MWE types.

1.2.2 Extraction

MWE extraction is a type-level task, wherein the MWE lexical items attested in a predetermined corpus are extracted out into a lexicon or other lexical listing. For example, with a given verb *take* and preposition *off*, we wish to know whether the two words combine together to form a VPC (i.e. *take off)* in a given corpus. This contrasts with MWE identification, where the focus is on individual token instances of MWEs, although obviously extraction can be seen to be a natural consequence of identification (in compiling out the list of those attested MWEs). The underlying assumption in MWE extraction is that there is evidence in the given corpus for each extracted MWE to form an MWE in some context, without making any claims about whether there also exist simple compositional combinations of those same words. The motivation for MWE extraction is generally lexicon development or expansion, e.g., in recognizing newly-formed MWEs (e.g. *ring tone* or *shock and awe*) or domain-specific MWEs (e.g. *bus speed* or *boot up* in an IT domain). In general, MWE extraction pulls MWEs out of context as standalone lexical items, although this generally involves analysis of the context of a given combination of words. However, as stated above, extraction often takes advantage of the results of MWE identification. For example, Baldwin (2005a) extracted English VPCs based on identifying VPC candidates using resources including a parser and chunker. Extracting MWEs is relevant to any lexically-driven application, such as grammar development or information extraction. In addition, it is particularly important for productive MWEs or domains that have distinctive MWE content. MWE extraction is as difficult as MWE identification in terms of syntactic flexibility and ambiguity. The bulk of research on MWE extraction has focused on extracting English verb-particle constructions, light-verb constructions and idioms (Baldwin and Villavicencio 2002). Despite a healthy body of research on MWE extraction, however, the results have not been as compelling as for MWE identification. Baldwin (2005a) achieved high accuracy on an English VPC extraction task, whereas others such as verb-noun pair extraction (Venkatapathy and Joshi 2005; Fazly and Stevenson 2007) still have considerable room for improvement. Part of the complexity here is that the target lexical resource for the MWE extraction often introduces its own constraints or requirements for extra lexical properties.

The motivation for MWE extraction is generally lexicon development and expansion, e.g., recognizing newly-formed MWEs (e.g., *ring tone* or *shock and awe*) or domain-specific MWEs. Extracting MWEs is relevant to any lexically-driven application, such as grammar engineering or information extraction. Depending on the particular application, it may be necessary to additionally predict lexical properties of a given MWE, e.g., its syntactic or semantic class. In addition, it is particularly important for productive MWEs or domains which are rich in technical terms (e.g., *bus speed* or *boot up* in the IT domain). There has been a strong focus on the development of general-purpose techniques for MWE extraction, particularly in the guise of collocation extraction. The dominating view here is that extraction can be carried out via association measures such as Point-wise Mutual Information (PMI) or the T-test, based on analysis of the frequency of occurrence of a given word combination, often in comparison with the frequency of occurrence of the component words (Church and Hanks 1989; Smadja 1993; Frantzi, Ananiadou, and Mima 2000). Association measures provide a score for each word combination, which forms the basis of a ranking of MWE candidates. Final extraction, therefore, consists of determining an appropriate cut-off in the ranking, although evaluation is often carried out over the full ranking. Collocation extraction techniques have been applied to a wide range of extraction tasks over a number of languages, with the general finding that it is often unpredictable which association measure will work best for a given task. As a result, recent research has focused on building supervised classifiers to combine the predictions of a number of association measures, and has shown that this leads to consistently superior results than any one association measure (Pecina 2008). It has also been shown that this style of approach works most effectively when combined with POS tagging or parsing, and strict filters on the type of MWE that is being extracted (e.g., adjective–noun or verb–noun: Justeson and Katz (1995)). It is worth noting that association measures have generally been applied to continuous word n-grams, or less frequently, pre-determined dependency types in the output of a parser. Additionally, collocation extraction techniques tend to require a reasonable number of token occurrences of a given word combination to operate reliably, which we cannot always assume (Fazly 2007).

A second approach to MWE extraction, targeted specifically at semantically and statistically idiomatic MWEs, is to extend the general association measure approach to include substitution (Lin 1999; Schone and Jurafsky 2001; Pearce 2001). For example, in assessing the idiomaticity of *red tape*, explicit comparison is made with lexically-related candidates generated by component word substitution, such as *yellow tape* or *red strip*. Common approaches to determining substitution candidates for a given component word are (near-) synonymy—e.g. based on resources such as WordNet—and distributional similarity. Substitution can also be used to generate MWE candidates, and then check for their occurrence in corpus data. For example, if *clear up* is a known (compositional) VPC, it is reasonable to expect that VPCs such as clean/tidy/unclutter/... up are also VPCs (Villavicencio 2005). That is not to say that all of these occur as MWEs, so an additional check for corpus attestation is usually used in this style of approach.

A third approach, also targeted at semantically idiomatic MWEs, is to analyze the relative similarity between the context of use of a given word combination and its component words (Schone and Jurafsky 2001; Stevenson, Fazly, and North 2004; Widdows and Dorow 2005). Similar to the unsupervised WSD-style approach to MWE identification, the underlying hypothesis is that semantically idiomatic MWEs will occur in markedly different lexical contexts to their component words. A bag of words representation is commonly used to model the combined lexical context of all usages of a given word or word combination. By interpreting this context model as a vector, it is possible to compare lexical contexts, e.g., via simple cosine-similarity (Widdows 2005). In order to reduce the effects of data sparseness, dimensionality reduction is often carried out over the word space prior to comparison. The same approach has also been applied to extract LVCs, based on the assumption that the noun complements in LVCs are often deverbal (e.g. *bath, proposal, walk*), and that the distribution of nouns in PPs post-modifying noun complements in genuine LVCs (e.g. (make a) proposal of marriage) will be similar to that of the object of the underlying verb (e.g. *propose marriage*) (Grefenstette and Teufel 1995). Here, therefore, the assumption is that LVCs will be distributionally similar to the base verb form of the noun complement, whereas with the original extraction method, the assumption was that semantically idiomatic MWEs are dissimilar to their component words.

A fourth approach is to perform extraction on the basis of implicit identification. That is, (possibly noisy) token-level statistics can be fed into a type-level classifier to predict whether there have been genuine instances of a given MWE in the corpus. An example of this style of approach is to use POS taggers, chunkers and parsers to identify English VPCs in different syntactic configurations, and feed the predictions of the various preprocessors into the final extraction classifier. Alternatively, a parser can be used to identify PPs with singular nouns, and semantically idiomatic D-PPs can be extracted from them based on distributional (dis)similarity of occurrences with and without determiners across a range of prepositions (van der Beek 2005).

A fifth approach is to use syntactic fixedness as a means of extracting MWEs, based on the assumption that semantically idiomatic MWEs undergo syntactic variation (e.g. passivization or internal modification) less readily than simple verb–noun combinations (Bannard 2007; Fazly, Cook, and Stevenson 2009).

In addition to general-purpose extraction techniques, linguistic properties of particular MWE construction types have been used in extraction. For example, the fact that a given verb–preposition combination occurs as a verb (e.g. *take off, clip-on*) is a strong predictor of the fact that the combination is occurring as a VPC. One bottleneck in MWE extraction is the token frequency of the MWE candidate. With a few notable exceptions (e.g. (Baldwin 2005a; Fazly, Cook, and Stevenson 2009)), MWE research has tended to ignore low-frequency MWEs, e.g., by applying a method only to word combinations which occur at least N times in a corpus.

1.2.3 Measuring Compositionality

In the context of MWEs, compositionality denotes the degree to which the properties of the MWE are inherited directly from those of the components. While there are various definitions of *compositionality*, for the purposes of this thesis, we focus specifically on semantic compositionally and consider compositional MWEs to be those where the meaning of the MWE is fully or largely derived from the semantics of its components. Conversely, with non-compositional MWEs there is a marked difference in the semantics of the MWE vs. the semantics of the components. For example, with the two VPCs *spill down* and *conk out*, we would claim that *spill down* is compositional whereas *conk out* (means 'fail' where '*conk*' means 'informal term for the nose') is non-compositional. That is, *spill* and *down* determine the full semantics of *spill down*, while *conk* has little or no bearing on the semantics of *conk out*. While we will tend to refer to compositionality as a binary distinction, in practice there is a continuum of compositionality from complete compositionality to complete non-compositionality. Modeling/measuring the compositionality of MWEs is the task of predicting the semantic association between an MWE and its components under the assumption that, to a certain degree, the meanings of MWEs and the components can be semantically resolved using WordNet or other semantic classes. We consider compositionality modeling to be a type-level task and to be invariant across individual senses (i.e., meaning) of the MWE. This is clearly an oversimplification and there are certainly cases of different senses having different degrees of compositionality. Hence, the task aim is to find whether the components of a given MWE semantically contribute to the semantics of the MWE, and if so, how much. The task of modeling compositionality, i.e., whether the components contribute to the meaning of the MWE, is a binary decision. Measuring compositionality, on the other hand, is a more semantically intensive task, where we not only predict whether a MWE is compositional, but we also estimate the degree of compositionality.

Studying compositionality has its own benefits. It provides information for improving output quality in NLP applications such as machine translation and text generation (Nunberg *et al.* 1994; Sag *et al.* 2002; Venkatapathy and Joshi 2006). It is also a prerequisite task for semantic interpretation over compositional MWEs. Previous research on modeling/measuring the compositionality of MWEs has primarily focused on English noun compounds and verb-particle constructions (Venkatapathy and Joshi 2005; Piao *et al.* 2006; Kim and Baldwin 2007a). Recently, MWE compositionality has been studied not only to detect or measure the degree of the compositionality, but also to utilize this in NLP applications. Venkatapathy and Joshi (2006) successfully showed the utility of MWE compositionality in a word alignment task between English and Hindi. However, since the task setup was supervised, large amounts of training data were necessary. There is a gap in the research literature on measuring the degree of MWE compositionality and also on the utility of compositionality in NLP applications.

1.2.4 Semantic Classification

Semantic classification is the task of specifying the semantics of MWEs based on a generalized semantic inventory (compatible with both simplex words and MWEs). It tends to presuppose the ability to classify the degree of compositionality of MWEs and apply only to compositional MWEs. That is, the task focus of MWE semantic classification is to specify the meanings of MWEs according to predefined semantic categories such as WordNet. Figure 1.1 illustrates an example task in the context of VPCs.

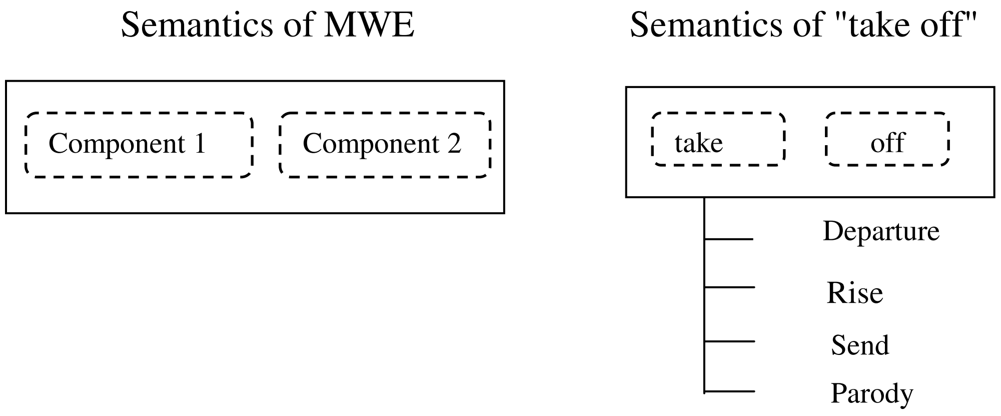


Figure 1.1: The semantic classification task

In Figure 1.1, the target is to determine the semantics of a given MWE. Often the meaning of the components is employed to specify the semantics of the whole. Hence, compositionality is a very useful clue in estimating the meaning of compositional MWEs. In our example, the target is to determine the different senses of *take off* (i.e., “departure, rise, send up, parody”). This can be performed based on individual analysis of *take* and *off* to some degree. WordNet is commonly used as a sense inventory for semantic classification tasks, although there are instances of user-defined sense inventories (e.g., *particle semantics* in Bannard (2003) and Cook and Stevenson (2006)).

Semantic classification in the context of MWEs is non-trivial due to the varying degrees of opacity in MWEs. The contribution of the individual components can vary (e.g., *eat up* and *start over*, where the verb is the primary determinant of the semantics). Sometimes none of the parts contribute to the semantics of the MWE (i.e., in fully non-compositional VPCs such as *make out*). Prior work related to the semantic classification of MWEs has been undertaken from both the linguistic and computational perspectives (Fraser 1976; Bame 1999; Gries 1999; Bannard 2003; ÒHara and Wiebe 2003; Patrick and Fletcher 2004; Cook and Stevenson 2006). Most of the research on the semantic classification of MWEs has focused on English VPCs. The relatedness between semantic classification and measuring the compositionality of MWEs is not well understood, warranting further study.

1.2.5 Semantic Interpretation

As MWEs are made up of two or more simplex words, syntactic and semantic associations arise between the components. The semantic interpretation or semantic role labeling of MWEs is the task to determine the semantic relation between the components, in the form of a relation set which is specific to an MWE construction type. Note that the semantic interpretation, once again, relates closely to compositionality, in that compositionality is a claim on whether the semantic association between the components is transparent or not, whereas semantic interpretation seeks to unearth a precise description of the semantic relation between those components. For example, the knowledge that *bus driver* is fully compositional provides us the means to infer the semantics of the components, but semantic interpretation seeks to specify exactly how *bus* and *driver* relate to each other, e.g. in predicting that the *driver* is the agent of control of the *bus*. If we knew that *bus driver* were non-compositional, however, we would know not to attempt to semantically interpret it based on the components. In this sense, modeling/measuring compositionality is a prerequisite for semantic interpretation. Figure 1.2 depicts the task of semantic interpretation with an example.

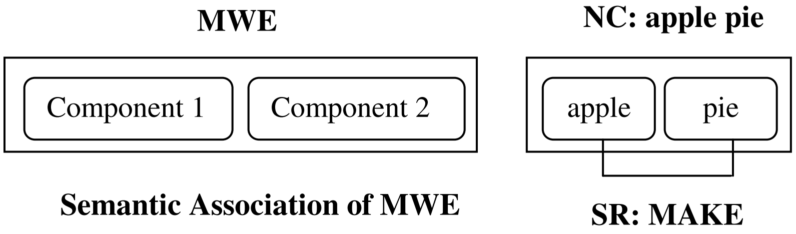


Figure 1.2: The semantic interpretation task

In Figure 1.2, the target is to interpret the semantic relation between the components. For example, *apple pie* can be interpreted as "pie made from apple". The semantic relation between *apple* and *pie* is specified as **MAKE**, where the head noun is made from the modifier.

Semantic relations (or associations) are most commonly used to interpret noun compounds and determiner-less prepositional phrases. The semantic relation used to interpret a given MWE varies with the components. For example, the semantic relation in *morning juice* is 'TIME' ("juice in the morning") whereas that in *orange juice* is 'MAKE' ("juice made from orange(s)"). Another example with D-PPs is *by car/bus/plane..*, where a mode of transportation combined with the method/manner preposition *by* leads to the semantic relation manner, whereas other nouns such as *day* lead to specific temporal interpretations. The majority of past research on semantic interpretation has focused on interpreting noun compounds (Vanderwende 1994; Copestake and Lascarides 1997; Lapata 2002; Moldovan *et al.* 2004; Kim and Baldwin 2005; Nastase *et al.* 2006) and D-PPs (Van Der Beek 2005; Baldwin *et al.* 2006). This research, particularly that on NC interpretation, has been suggested to be relevant for the NLP applications for QA and IR (Moldovan *et al.* 2004), although there is no definitive empirical evidence to support this claim.

In all prior work, however, a major difficulty in semantic interpretation has been the design of a standard set of semantic relations with which to perform the interpretation. For interpreting noun compounds, the scalability and portability to novel domains/NC types is questionable, as methods make specific assumptions about the domain or range of NC interpretation. The current level of accuracy of NC interpretation over open domain data is not high enough to utilize the acquired data for NLP applications. Also, lack of agreement on the semantic relations used for MWE interpretation makes it hard to incorporate NC interpretation into applications.

Another point is that much of the work on semantic interpretation is based on supervised methods, which raises questions on the amount of training data and effective learning algorithms for a particular method or set of semantic relations.

1.2.6 Internal Syntactic Disambiguation

Duringthe process of MWE identification and extraction, for some MWE types it is necessary to disambiguate the internal syntax of individual MWEs. A prominent case of this in English is noun compounds with 3 or more terms. For example, *glass window cleaner* has two possible interpretations,[2] corresponding to the two possible bracketing of the compound: (1) “a cleaner of glass windows” (= [[glass window] cleaner]), and (2) “a cleaner of windows, made of glass” (= [glass [window cleaner]]. In this case, the first case (of left bracketing) is the correct analysis, but *movie car chase*, e.g., is right bracketing (= (movie (car chase))). The process of disambiguating the syntax of an NC is called **bracketing**. The most common approach to bracketing is based on statistical analysis of the components of competing analyses. In the adjacency model, for a ternary NC N1 N2 N3, a comparison is made of the frequencies of the two modifier–head pairings extracted from the two analyses, namely N1 N2 and N1 N3 in the left bracketing case, and N2 N3 and N1 N3 in the right bracketing case; as N1 N3 is common to both, in practice, N1 N2 is compared directly with N2 N3. A left bracketing analysis is selected in the case when N1 N2 is judged to be more likely, otherwise a right bracketing analysis is selected (Marcus 1980). In the dependency model, the NC is instead decomposed into the dependency tuples of N1 N2 and N2 N3 in the case of left bracketing, and N2 N3 and N1 N3 in the case of right bracketing; once again, the dependency N2 N3 is common to both, and can be ignored. In the instance that N1 N2 is more likely than N1 N3, the model prefers a left bracketing analysis, otherwise a right bracketing analysis is selected (Lauer 1995). While the dependency model tends to outperform the adjacency model, the best-performing models take features derived from both along with various syntactic and semantic features (Nakov and Hearst 2005; Vadas and Curran 2008).

[2] More generally, for an n item noun compound, the number of possible interpretations is defined by the Catalan number : $C_n = \left(\frac{1}{n+1}\right)({}^{2n}C_n)$

1.2.7 MWEs and Machine Translation

The identification of multiword expressions (MWE) and their appropriate handling is necessary in constructing professional tools for language manipulation. MWEs are a problem in the word alignment of parallel corpora, and various strategies for improving the result have been suggested (Ney and Popovic 2004). Machine translation (MT) and automatic dictionary compilation (ADC) are examples of such applications, where MWEs play a major role. The identification of MWEs in running text is a complex problem that requires more than one solution (Mendes et al. 2006). Although some MWEs can be isolated in the tokenizer, and then analysed as a single cluster, most of them cannot.

Phrase-based machine translation (PBMT) model has proved itself as a great improvement over the initial word based approaches (Brown et al., 1993). Recent syntax-based models perform even better than phrase-based models. However, when syntax-based models are applied to new domain with few syntax-annotated corpora, the translation performance would decrease. A *multiword expression* can be considered as a word sequence with relatively fixed structure representing special meanings. For *bilingual multiword expression (BiMWE)*, Zhixiang et al. (Zhixiang et al., 2009) defined a bilingual phrase as a bilingual MWE if (1) the source phrase is a MWE in source language; (2) the source phrase and the target phrase must be translated to each other exactly, i.e. there is no additional (boundary) word in target phrase which cannot find the corresponding word in source phrase, and vice versa. Since MWE usually constrains possible senses of a polysemous word in context, they can be used in many NLP applications such as information retrieval, question answering, word sense disambiguation and so on.

For machine translation, Piao et al. (2005) have noted that the issue of MWE identification and accurate interpretation from source to target language remained an unsolved problem for existing MT systems. This problem is more severe when MT systems are used to translate domain-specific texts, since they may include technical terminology as well as more general fixed expressions and idioms. Although some MT systems may employ a machine-readable bilingual dictionary of MWE, it is time-consuming and inefficient to obtain this resource manually. Therefore, some researchers have tried to use automatically extracted bilingual MWEs in SMT. Tanaka and Baldwin (2003) described an approach of noun-noun compound machine translation, but no significant comparison was presented. Lambert and Banchs (2005) presented a method in which bilingual MWEs were used to modify the word alignment so as to improve the SMT quality. In their work, a bilingual MWE in training corpus was grouped as one unique token before training alignment models. They reported that both alignment quality and translation accuracy were improved on a small corpus. However, in their further study, they reported even lower BLEU scores after grouping MWEs based on part-of-speech on a large corpus (Lambert and Banchs, 2006). Nonetheless, since MWE represents linguistic knowledge, the role and usefulness of MWE in full-scale SMT is intuitively positive. The difficulty lies in how to integrate bilingual MWEs into existing SMT system to improve SMT performance, especially when translating domain texts.

1.2.8 Authorship Identification and Stylometry Analysis

Contemporary stylistic and stylometric studies usually focus on an author with a distinctive style and often characterize that style by comparing the author’s texts to those of other authors. When an author’s works display diverse styles, however, the style of one text rather than the style of the author becomes the appropriate focus. Because authorship attribution techniques are founded upon the premise that some elements of authorial style are so routinized and habitual as to be outside the author’s control, extreme style variation within the works of a single author seems to threaten the validity of the entire enterprise. This apparent contradiction is only apparent, however, for the tasks are quite different. Successful attribution of a diverse group of texts to their authors requires only that each author’s texts be more similar to each other than they are to texts by other authors, or, perhaps more accurately, that they be less different from each other than from the other texts. The successful separation of texts or sections of texts with distinctive styles from the rest of the works of an author takes for granted a pool of authorial similarities and isolates whatever differences remain.

Stylometry, which may be considered as an investigation of “Who was behind the keyboard when the document was produced?” or “Did Mr. X wrote the document or not?” is a long term study mainly in forensic investigation department that started from late Nineties. In the past, where Stylometry emphasized the rarest or most striking elements of a text, contemporary techniques can isolate identifying patterns even in common parts of speech. The pioneering study on authorship attributes identification using word-length histograms appeared at the very end of nineteen century (Malyutov 2006). After that, a number of studies based on content analysis

(Krippendorf 2003), computational stylistic approach (Stamatatos et al. 1999), exponential gradient learn algorithm (Argamon et al. 2003), Winnow regularized algorithm (Zhang 2002), Support Vector Machine based approach (Pavelec et al. 2007) etc have been proposed for various languages like English and Portuguese. Recently, research has started to focus on authorship attribution on larger sets of authors: 8 (Halteren 2005), 20 (Argamon et al. 2003), 114 (Madigan et al. 2005), or up to thousands of authors (Koppel et al. 2007). The use of computers regarding the extraction of Stylometrics has been limited to auxiliary tools (i.e., simple program for counting user-defined features fast and reliably). Hence, authorship attribution studies so far may be looked like *computer-assisted*, not *compute-based*.

Recent work has shown that the same techniques that are able to attribute texts correctly to their authors even when some of the authors' styles are quite diverse do a good job of distinguishing an unusual passage within a novel from the rest of the text (Hoover 2003). Other quite subtle questions have also been approached using authorship attribution techniques. More than 20 years ago, Burrows showed that Jane Austen's characters can be distinguished by the frequencies of very frequent words in their dialogue (1987). More recent studies have used authorship techniques to investigate the sub-genres and varied narrative styles within Joyce's *Ulysses* (McKenna and Antonia 2001), the styles of Charles Brockden Brown's narrators (Stewart 2003), a parody of Richardson's *Pamela* (Burrows 2005), and two translations of a Polish trilogy made a hundred years apart (Rybicki 2005). Hugh Craig has investigated chronological changes in Ben Jonson's style (1999a, 1999b), and Burrows has discussed chronological changes in the novel genre (1992a).

1.3 Focus of the Thesis

1.3.1 Our Aims and Approaches

The goals of this MWE study are to shed light on underlying linguistic processes giving rise to MWEs across constructions and languages, to generalize techniques for analyzing MWEs, MWE classifications, and finally to exemplify the utility of MWE within general NLP tasks. To develop a framework for modeling MWEs in this thesis, our principal approach is to employ statistical approaches, and furthermore to integrate symbolic approaches and some supervised applications wherever possible to build from richer syntactic and semantic representations.

1.3.2 Scope of Research

In this thesis, we exclusively deal with Bengali MWEs and in some cases, handling English MWEs. Our motivation in targeting Bengali MWEs relates largely to the challenges behind the resource constraint and the unavailability of past experiments. As per as the application of MWEs is concerned, our experiments focus the pioneering approach in Bengali language. English MWEs relate largely to resource availability. There is currently a large number of lexical resources (e.g.WordNet and CoreLex) and tools/software (e.g. RASP and WordNet::Similarity) available for English. Resources such as WordNet and RASP have been widely used as a means of syntactic and semantic analysis for various NLP tasks in English. But in Bengali, only lexical resource which is publicly available is the Shallow Parser[3] developed by Indian Institute of Information technology, Hydrabad. In some cases, we have utilized English resources for evaluating our experimental results. We focus our attention primarily on noun compounds (NCs) in Bengali and Adjective-noun combination, Verb-object and Verb-subject combinations in English. The selection of these classes is due to their high frequency and productivity. We will focus predominantly on binary NCs, i.e., NCs made up of two nouns (e.g., *computer science* and *golf club* in English; ***taser ghar*** (*house of cards, fragile*) in Bengali).

[3] http://ltrc.iiit.ac.in/analyzer/bengali

Chapter 2

The Linguistics of Multiword Expressions

Multiword expressions can be syntactically and semantically categorized into various types, including noun compounds and idioms. Each type of MWE has distinctive linguistic features which we will describe in Section 2.1.1. Due to these differences, for distinct MWEs, we have specific objectives for knowledge acquisition and different obstacles to overcome. For example, interpreting semantic relations in noun compounds is a hard task while extracting or identifying them is relatively trivial. On the other hand, extracting or identifying verb-particle constructions is challenging since there is often ambiguity with a verb-PP analysis. Also, measuring compositionality is an important task for VPCs as there is a more uniform distribution of VPCs across the spectrum of compositionality, whereas it is less of an issue for noun compounds as they are mostly compositional[1]. In this chapter, we will survey the linguistics of the major types of English MWE.

2.1 Overview of Multiword Expressions

Multiword expressions are lexical items that can be decomposed into multiple simplex words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag *et al.* 2002).

[1] That is noun compound *types* are mostly compositional; noun compound *tokens* are arguably not.

2.1.1 Linguistic Properties of Multiword Expressions

In languages such as English, the conventional interpretation of the requirement of decomposability into lexemes is that MWEs must in themselves be made up of multiple whitespace-delimited words. For example, *marketing manager* is potentially a MWE as it is made up of two lexemes (*marketing* and *manager*), while fused words such as *lighthouse* are conventionally not classified as MWEs (in practice, a significant subset of research on English noun compounds has considered both fused and whitespace-separated expressions). In languages such as German, the high productivity of compound nouns such as *Kontaktlinse* “contact lens” (the concatenation of *Kontakt* “contact” and *Linse* “lens”), without whitespace delimitation, means that we tend to relax this restriction and allow for single-word MWEs. In non-segmenting languages such as Japanese and Chinese (Baldwin and Bond 2002; Xu, Lu, and Li 2006), we are spared this artificial consideration. The ability to decompose an expression into multiple lexemes is, however, still applicable, and leads to the conclusion, e.g., that *fukugô-hyôgen* “multiword expression” is a MWE (both *fukugô* “compound” and *hyôgen* “expression” are standalone lexemes), but *buchô* “department head” is not (*bu* “department” is a standalone lexeme, but *chô* “head” is not).

The second requirement on a MWE is for it to be idiomatic. Baldwin and Kim (Baldwin and Kim 2010) provided a detailed account of idiomaticity in its various manifestations. They are described in the following section.

- **Idiomaticity**

Idiomaticity is defined as lexico-syntactic, semantic, pragmatic and statistical markedness (Katz and Postal 2004; Chafe 1968; Cruse 1986; Jackendoff 1997). Lexico-syntactic idiomaticity means that the MWE has surprising syntax given the syntax of the individual simplex words. For example, *apple pie* is an entirely unsurprising combination of the nouns *apple* and *pie*, whereas *by and large* is a coordination of a preposition and an adjective to form an adverbial phrase, an effect which is not predicted by Standard English grammar rules. As such, *apple pie* is not lexico-syntactically idiomatic while *by and large* is. Semantic irregularity commonly happens in idioms such as *in one’s shoes*, where the semantics is not immediately predictable from the simplex semantics of *shoes*. Pragmatic idiomaticity occurs in situated expressions such as *good morning* and *all board*. That is, these MWEs are associated with very particular situations and are anomalous in other contexts (e.g., *good morning* when finishing a meal, or *all aboard* when watching a soccer match). Statistical idiomaticity occurs with MWEs such as *black and white* where they occur with uncommonly high frequency in contrast to alternative forms of the same expression. It is perfectly acceptable to say *white and black*, but the skew towards the first form is sufficiently great that *white and black photograph*, e.g., is marked in English.

- **Lexical Idiomaticity**

Lexical idiomaticity occurs when one or more components of an MWE are not part of the conventional English lexicon. For example, *ad hoc* is lexically marked in that neither of its components (*ad* and *hoc*) are standalone English words. Lexical idiomaticity inevitably results in syntactic and semantic idiomaticity because there is no lexical knowledge associated directly with the parts from which to predict the behavior of the MWE. As such, it is one of the most clear-cut and predictive properties of MWEhood.

- **Syntactic Idiomaticity**

Syntactic idiomaticity occurs when the syntax of the MWE is not derived directly from that of its components (Katz and Postal 2004; Chafe 1968). For example, *by and large*, is syntactically idiomatic in that it is adverbial in nature, but made up of the anomalous coordination of a preposition (by) and an adjective (large). On the other hand, *take a walk* is not syntactically marked as it is a simple verb–object combination which is derived transparently from a transitive verb (take) and a countable noun (walk). Syntactic idiomaticity can also occur at the constructional level, in classes of MWEs having syntactic properties which are differentiated from their component words, e.g., verb-particle constructions and determinerless prepositional phrases described in later section.

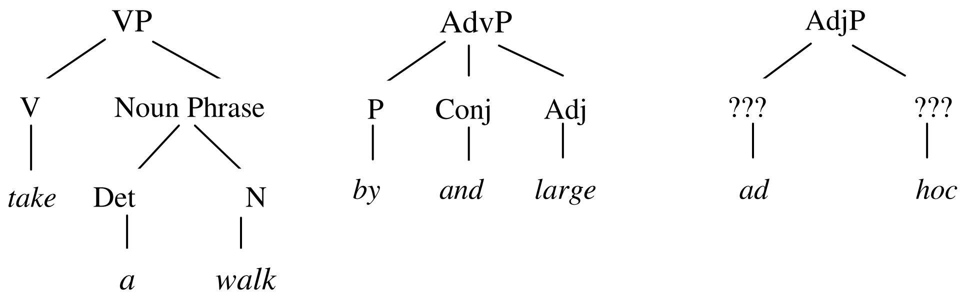


Figure 2.1: Examples of syntactic non-markedness vs. markedness

- **Semantic Idiomaticity**

Semantic idiomaticity is a reflection of the meaning of a MWE not being explicitly or implicitly derivable from its parts (Katz and Postal 2004; Chafe 1968; Bauer 1983). For example, *birds of a feather* usually indicates "people with similar interests", which we can not predict from either *birds* or *feather*. On the other hand, *all aboard* is not semantically marked as its semantics is fully predictable from its parts. Many cases are not as clear cut as these, however. The semantics of *blow hot and cold* ("constantly change opinion"), for example, is partially predictable from *blow* ("move" and hence "change"), but not as immediately from *hot and cold*. There are also cases where the meanings of the parts are transparently inherited in the MWE but there is additional semantic content which has no overt realization. One such example is *bus driver* where *bus* and *driver* both have their expected meanings, but there is additionally the default expectation that a *bus driver* is "one who drives a bus" and not "one who drives *like* a bus". *Semantic idiomaticity* is directly related to compositionality, as the degree of semantic contribution of the components indicates the *semantic idiomaticity* as well as the compositionality.

Related to the issue of *semantic idiomaticity*, there has been discussion of the notions of *non-identifiability* and *figuration* (Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994). We roughly classify these properties under our definition of *semantic idiomaticity* for the purposes of this thesis.

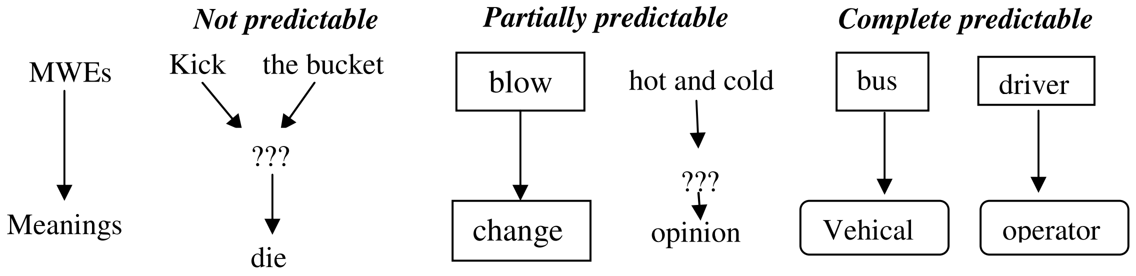


Figure 2.2: Examples of semantic idiomaticity

Non-identifiability (Nunberg *et al.* 1994) is the notion of the meaning of an MWE not being easily predictable from the surface form (components), much like our definition of *semantic idiomaticity*. For example, the meaning of *kick the bucket* ("die") cannot be derived from either *kick* or *bucket*. Another example is *make out*, where the parts (i.e., *make* and *out*) do not semantically contribute to the meaning of the whole. This property relates closely to compositionality. That is, when MWEs are compositional, the meaning of MWEs can be predicted from the parts. Hence, non-identifiability coincides with non-compositionality (other examples of non-identifiable and non-compositional MWEs are *on ice*, *cock up*, *chicken out* and *by and large*).

Figuration (Fillmore *et al.* 1988; Nunberg *et al.* 1994) is an attribute of encoded expressions such as metaphors (e.g., *take the bull by the horns*), metonymies (e.g., *lend a hand*) and hyperboles (e.g., *not worth the paper it's printed on*). It is defined as the property of the components of an MWE having some metaphoric or hyperbolic meaning in addition to their literal meaning. That is, the semantics of the MWE is derived from the components through a process of metaphor, hyperbole or metonymy, although the precise nature of the figuration may be more or less obvious. Hence, *figuration* involves subtle interactions between idiomatic and literal meaning. We return to touch on the relationship between *figuration* and *semantic idiomaticity* below.

- **Pragmatic idiomaticity**

Pragmatic idiomaticity is the condition of a MWE being associated with a fixed set of situations or a particular context (Kastovsky 1982; Jackendoff 1997; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). *Good morning* and *all aboard* are examples of pragmatic MWEs: the first is a greeting associated specifically with mornings[2] and the second is a command associated with the specific situation of a train station or dock, and the imminent departure of a train or ship. Pragmatically idiomatic MWEs are often ambiguous with (non-situated) literal translations; e.g., *good morning* can mean "pleasant morning" (c.f. Kim had a good morning).

- ***Statistical Idiomaticity***

Statistical idiomaticity occurs when a combination of words occurs with surprising frequency, relative to the component words or alternative phrasings of the same expression (Pawley and Syder 1983; Cruse 1986; Sag *et al.* 2002). Cruse (1986:p281) provides some nice examples of *statistical idiomaticity* in the matrix of adjectives and nouns presented in Table 2.1.

[2] Which is not to say that it can't be used ironically at other times of the day!

	unblemished	***spotless***	***flawless***	***immaculate***	***Impeccable***
performance	-	-	+	+	+
argument	-	-	+	-	?
complexion	?	?	+	-	-
behavior	-	-	-	-	+
kitchen	-	+	-	+	-
record	+	+	-	?	+
reputation	?	+	-	?	?
taste	-	-	?	?	+
order	-	-	?	+	+
credentials	-	_	_	_	+

Table 2.1: Examples of statistical idiomaticity ("+" = strong lexical affinity, "?" = marginal Lexical affinity, "–" = negative lexical affinity) (Cruse 1986)

The adjectives are largely synonymous, and yet different nouns have particular preferences for certain subsets of the adjectives as modifiers, as indicated by the cells in the matrix ("+" indicates a strong lexical affinity, "?" indicates a marginal lexical affinity, and "–" indicates a negative lexical affinity). Note that the statistical idiomaticity (i.e., the alternative phrasing) can be in terms of alternative orderings of the components. For example, *black and white* is much more common in English than *white and black*, while the reverse holds in the case of other languages such as Japanese and Spanish (see Table 2.1). For the purposes of this thesis, we will follow Sag *et al.* (2002) in referring to MWEs which are only statistically idiomatic (i.e., not also lexico-statistically, semantically or pragmatically idiomatic) as *collocations*.

Statistical idiomaticity relates to the notion of *institutionalization/conventionalization*, i.e. a particular word combination coming to be used to refer a given object (Fernando and Flavell 1981; Bauer 1983; Nunberg *et al.* 1994; Sag *et al.* 2002). For example, *traffic light* is the conventionalized descriptor for "a visual signal to control the flow of traffic at intersections". There is no reason why it shouldn't instead be called a *traffic director* or *intersection regulator*, but the simple matter of the fact is that it is not referred to using either of those expressions; instead, *traffic light* was settled on as the canonical term for referring to the object. Similarly, it is an arbitrary fact of the English language that we say *many thanks* and not **several thanks*, and *salt and pepper* in preference to *pepper and salt*.[3]

Nunberg *et al.* (1994) consider *collocation* (*conventionality* in their terms) to be a mandatory property of MWEs. We consider conventionality to relate to semantic, pragmatic and statistical idiomaticity, but consider that MWEs do not have to have any one of these three forms of markedness (e.g., MWEs which are strictly lexico-syntactically idiomatic are classified as MWEs in this research). Collocations are most apparent when observed in contrast with *anti-collocations*. *Anti-collocations* are lexico-syntactic variants of collocations which have unexpectedly low frequency (Pearce 2001). For example, *pepper and salt* is an anti-collocation for *salt and pepper*, and *traffic director* is an anti-collocation for *traffic light*.

It is important to acknowledge that our use of the term *collocation* differs from the mainstream usage in computational linguistics, where a collocation is often defined as an arbitrary and recurrent word combination that co-occurs more often than would be expected by chance (Choueka 1988; Lin 1998b; Evert 2004).

Above, we described four different forms of *idiomaticity*. We bring these together in categorizing a selection of MWEs in Table 2.2. In Table 2.2, some examples such as *kick the*

	Lexico-syntactic	***Semantic***	***Pragmatic***	***Statistical***
all aboard	-	-	+	+
black and white	–	?	_	+
by and large	+	+	-	-
kick the bucket	-	+	-	-
Social butterfly	-	+	-	+
make out	-	+	-	-
shock and awe	-	-	+	+
to and fro	+	-	-	+
bus driver	-	+	-	+
traffic light	-	-	-	+

Table 2.2: Classification of MWEs in terms of different forms of idiomaticity

[3] Which is not to say there wasn't grounds for the selection of the canonical form at its genesis, e.g., for historical, crosslingual or phonological reasons.

bucket, *make out* and *traffic light* are marked with only one form of idiomaticity, which is sufficient for them to be classified as MWE. On the other hand, other MWEs such as *shock and awe* and *to and fro* are idiosyncratic in more ways than one. We analyze *shock and awe* as being pragmatically idiomatic because of its particular association with the bombardment of Baghdad at the commencement of the Iraq War, and *to and fro* as being lexicosyntactically idiomatic because of the relative syntactic opacity of the antiquated *fro*.

Other Properties

Other common properties of MWE are: single-word paraphrasability, proverbiality and prosody. Unlike idiomaticity, where some form of idiomaticity is a necessary feature of MWEs, these other properties are neither necessary nor sufficient. Prosody relates to semantic idiomaticity, while the other properties are independent of idiomaticity as described above.

• Crosslingual Variation

There is remarkable variation in MWEs across languages (Villavicencio, Baldwin, and Waldron 2004). In some cases, there is direct lexicosyntactic correspondence for a cross-lingual MWE pair with similar semantics. For example, *in the red* has a direct lexico-syntactic correlate in Portuguese with the same semantics: *no vermelho*, where *no* is the contraction of *in* and *the*, *vermelho* means red, and both idioms are prepositional phrases (PPs). Others have identical syntax but differ lexically. For example, *in the black* corresponds to *no azul* ("in the blue") in Portuguese, with a different choice of colour term (*blue* instead of *black*). More obtusely, *Bring the curtain down* corresponds to the Portuguese *botar um ponto final em* (lit. "put the final dot in"), with similar syntactic make-up but radically different lexical composition. Other MWEs again are lexically similar but syntactically differentiated. For example, *in a corner* (e.g., The media has him in a corner) and *encurralado* ("cornered") are semantically equivalent but realized by different constructions – a PP in English and an adjective in Portuguese.

There are of course many MWEs which have no direct translation equivalent in a second language. For example, the Japanese MWE *zoku-giiN*, meaning "legistors championing the causes of selected industries" has no direct translation in English (Tanaka and Baldwin 2003). Equally, there are terms which are realised as MWEs in one language but single-word lexemes in another, such as *interest rate* and its Japanese equivalent *riritsu*.

- **Single-word paraphrasability**

Single-word paraphrasability is the observation that significant numbers of MWEs can be paraphrased with a single word (Chafe 1968; Gibbs 1980; Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994). While some MWEs are single-word paraphrasable (e.g., *leave out* = "omit"), some are not (e.g., *look up* = ?). Also, MWEs with arguments can sometimes be paraphrasable (e.g., *take off clothes* = "undress"), just as multi-word non-MWEs can be single-word paraphrasable (e.g., *not sufficient* = "insufficient").

- **Proverbiality**

Proverbiality is the ability of an MWE to "describe and implicitly to explain a recurrent situation of particular social interest in the virtue of its resemblance or relation to a scenario involving homely, concrete things and relations" (Nunberg *et al.* 1994). For example, VPCs and idioms are often indicators of more informal situations (e.g., *piss off* is an informal form of *annoy*, and *kick the bucket* is an informal form of *die, demise*). Nunberg *et al.* (1994) treat *informality* as a separate category, where we combine it with *proverbiality*.

- **Prosody**

MWEs can have distinct *prosody*, i.e., stress patterns, from compositional language (Fillmore *et al.* 1988; Liberman and Sproat 1992; Nunberg *et al.* 1994). For example, when the components do not make an equal contribution to the semantics of the whole, MWEs can be prosodically marked, e.g., *soft spot* is prosodically marked (due to the stress on *soft* rather than *spot*), although *first aid* and *red herring* are not. Note that *prosodic* marking can equally occur with non-MWEs, such as *dental operation*.

2.1.2 Collocations and MWEs

A common term in NLP which relates closely to our discussion of MWEs is collocation. A widely-used definition for collocation is "an arbitrary and recurrent word combination" (Benson 1990), or in our terms, a statistically idiomatic MWE (esp. of high frequency). While there is considerable variation between individual researchers, collocations are often distinguished from "idioms" or "non-compositional phrases" on the grounds that they are not syntactically idiomatic, and if they are semantically idiomatic, it is through a relatively transparent process of figuration or metaphor (Choueka 1988; Lin 1998; McKeown and Radev 2000; Evert 2004). Additionally, much work on collocations focuses exclusively on predetermined constructional templates (e.g. adjective-noun or verb-noun collocations). In Table 2.2, e.g., *social butterfly* is an uncontroversial instance of a collocation, but *to and fro* would tend not to be classified as collocations. As such, collocations form a proper subset of MWEs.

2.1.3 Types of Multiword Expression

English MWEs can be syntactically and semantically categorized in various ways. In this thesis, we adopt the classification and terminology of Bauer (1983) and Sag *et al.* (2002), as outlined in Figure 2.3. The classification of MWEs into *lexicalized phrases* and *institutionalized phrases* hinges on whether the MWE is lexicalized (i.e., explicitly encoded in the lexicon) on the grounds of lexico-syntactic or semantic idiomaticity, or a simple collocation (i.e., only statistically idiosyncratic). Note that we will largely ignore pragmatic idiomaticity for the remainder of this thesis. *Lexicalized phrases* are MWEs in which the components have idiosyncratic syntax or semantics in part or in combination. Lexicalized phrases can be further split into: *fixed expressions* (e.g., *by train*, *at first*), *semi-fixed expressions* (e.g., *spill the beans*, *car dealer*, *Chicago White Socks*) and *syntactically-flexible expressions* (e.g., *add up*, *give a demo*).

- *Fixed expressions* are fixed strings that undergo neither morphosyntactic variation nor internal modification. For example, *by and large* is not morpho-syntactically modifiable (e.g., **by and larger*) or internally modifiable (e.g., **by and very large*). Non-modifiable determinerless prepositional phrases such as *on air* and *by car* are also fixed expressions.
- Semi-fixed expressions are lexically-variable MWEs that have hard restrictions on word order and composition, but undergo some degree of lexical variation such as inflection (e.g., kick/kicks/kicked/kicking the bucket vs. *the bucket was kicked), variation in reflexive pronouns (e.g., in her/his/their shoes) and determiner selection (e.g., *The Beatles* vs. a *Beatles album*). Non-decomposable VNICs (e.g., *kick the bucket, shoot the breeze*) and nominal MWEs (e.g., *attorney general, part of speech*) are also classified as semi-fixed expressions.
- *Syntactically flexible expressions* are MWEs which undergo syntactic variation, such as verb-particle constructions, light-verb constructions and decomposable idioms. The nature of the flexibility varies significantly across construction types. Verb-particle constructions, for example, are syntactically flexible with respect to the word order of the particle and NP in transitive usages: *hand in the paper* vs. *hand the paper in*. They are also usually compatible

with internal modification, even for intransitive VPCs: *the plane took right off*. Light-verb constructions (e.g., *give a demo)* undergo full syntactic variation, including passivization (e.g., *a demo was given*), extraction (e.g., *how many demos did he give?*) and internal modification (e.g., *give a clear demo*). Decomposable idioms are also syntactically flexible to some degree, although the exact form of syntactic variation is hard to predict (Riehemann 2001).

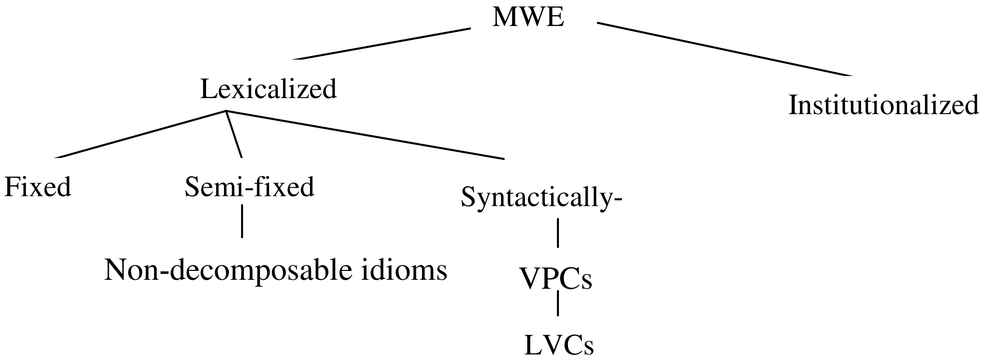


Figure 2.3: MWE types (Sag et al. 2002)

As described in Section 2.1.1, *collocations* (or *institutionalized phrases*) are MWEs that occur with surprising frequency, relative to the component words or alternative phrasings of the same expression (i.e., they are strictly statistically idiosyncratic), but which are otherwise unmarked. Examples include *peanut butter and jam*, *salt and pepper*, *telephone booth*, *many thanks* and *traffic light*.

2.2 Noun Compound

Nominal MWEs are one of the most common MWE types, in terms of token frequency, type frequency, and their occurrence in the world's languages (Tanaka and Baldwin 2003; Lieber and Ṡtekauer 2009). In English, the primary type of nominal MWE is the noun compound (NC), where two or more nouns combine to form a nominal compounds, such as *golf club* or *computer science department* (Lauer 1995; Sag, Baldwin, Bond, Copestake, and Flickinger 2002; Huddleston and Pullum 2002); the rightmost noun in the NC is termed the head noun (i.e., *club* and *department*, respectively) and the remainder of the component(s) modifier(s) (i.e., *golf* and *computer science*, respectively). Within NCs, there is the subset of compound nominalizations, where the head is deverbal (e.g., *investor hesitation* or *stress avoidance*). There is also the broader class of nominal MWEs where the modifiers are not restricted to be nominal, but can also be verbs (usually present or past participles, such as *connecting flight* or *hired help*) or adjectives (e.g., *open secret*). To avoid confusion, we will term this broader set of nominal MWEs nominal compounds. In Romance languages such as Italian, there is the additional class of complex nominals which include a preposition or other marker between the nouns, such as *succo di limone* "lemon juice" and *porta a vetri* "glass door". One property of noun compounds which has put them in the spotlight of NLP research is their underspecified semantics. For example, while sharing the same head, there is little semantic commonality between *nut tree*, *clothes tree* and *family tree*: a nut tree is a tree which bears edible nuts; a clothes tree is a piece of furniture shaped somewhat like a tree, for hanging clothes on; and a family tree is a graphical depiction of the genealogical history of a family (which can be shaped like a tree). In each case, the meaning of the compound relates (if at times obtusely!) to a sense of both the head and the modifier, but the precise relationship is highly varied and not represented explicitly in any way. Furthermore, while it may be possible to argue that these are all lexicalised noun compounds with explicit semantic representations in the mental lexicon, native speakers generally have reasonably sharp intuitions about the semantics of novel compounds. For example, a *bed tree* is most plausibly a tree that beds are made from or perhaps for sleeping in, and a *reflection tree* could be a tree for reflecting in/near or perhaps the reflected image of a tree. Similarly, context can evoke irregular interpretations of high-frequency compounds (Downing 1977; Spārck Jones 1983; Copestake and Lascarides 1997; Gagn´e, Spalding, and Gorrie 2005). This suggests that there is a dynamic interpretation process that takes place, which complements encyclopedic information about lexicalised compounds.

One popular approach to capturing the semantics of compound nouns is via a finite set of relations. For example, *orange juice*, *steel bridge* and *paper hat* could all be analysed as belonging to the *make* relation, where head is made from modifier. This observation has led to the development of a bewildering range of semantic relation sets of varying sizes, based on abstract relations (Vanderwende 1994; Barker and Szpakowicz 1998; Rosario and Hearst 2001; Moldovan), direct paraphrases, e.g. using prepositions or verbs (Lauer 1995; Lapata 2002; Grover, Lapata, and Lascarides 2004; Nakov 2008), or various hybrids of the two (Levi 1978; Vanderwende 1994). This style of approach has been hampered by issues including low inter-annotator agreement (especially for larger semantic relation sets), coverage over data from different domains, the impact of context on interpretation, how to deal with "fringe" instances which do not quite fit any of the relations, and how to deal with interpretational ambiguity (Downing 1977; Spärck Jones 1983). An additional area of interest with nominal MWEs (especially noun compounds) is the syntactic disambiguation of MWEs with 3 or more terms. For example, glass window cleaner can be syntactically analyzed as either (glass (window cleaner)) (i.e., "a window cleaner made of glass", or similar) or ((glass window) cleaner) (i.e., "a cleaner of glass windows"). Syntactic ambiguity impacts on both the semantic interpretation and prosody of the MWE. The task of disambiguating syntactic ambiguity in nominal MWEs is called ***bracketing***.

2.3 Verb-particle constructions

Verb-particle constructions (i.e. *VPCs*) are made up of a verb and obligatory particle (s) such as *hand in* and *take off* (Bolinger 1976b; Jackendoff 1997; Huddleston and Pullum 2002; Sag *et al.* 2002). The obligatory particles are usually intransitive prepositions, adjectives or verbs, as shown in (2.1)–(2.3).

(2.1) verb + intransitive prepositions: *battle on*, *take off*

(2.2) verb + adjectives: *cut short*, *band together*

(2.3) verb + verbs: *let go*, *let fly*

Generally, VPCs are both idiosyncratic and semi-idiosyncratic combinations although some are adverbial and/or non-lexical particle cases (Dehe *et al.* 2001). VPCs often involve subtle interactions between the verb and particle (Bolinger 1976b; Jackendoff 1973; Fraser 1976; Lidner 1983; Kayne 1985; Svenonius 1994; Dehe *et al.* 2001; Dehe 2002). For example, the particle can impact on various properties of the verb, including: aspect (e.g. *eat* vs. *eat up*), reciprocity (e.g. *ring* vs. *ring back*) and repetition (e.g. *start* vs. *start over*).

Note that VPCs are termed *phrasal verbs* by some researchers (Bolinger 1976b; Side 1990; Dirven 2001; McCarthy *et al.* 2003) and verb-particle constructions by others (Dehe *et al.* 2001; Bannard *et al.* 2003; Bannard 2003; Baldwin *et al.* 2003a; Cook and Stevenson 2006; Kim and Baldwin 2007a). In this thesis, we will refer to them exclusively as VPCs. One MWE type which relates closely to VPCs is *prepositional verbs* (Jackendoff 1973; O'Dowd 1998; Huddleston and Pullum 2002; Baldwin 2005b), which are similarly made up of a verb and preposition, but the preposition is transitive and selected by the verb (e.g., *refer to, look for*). It is possible to differentiate transitive VPCs[4] from prepositional verbs via their respective linguistic properties (Bolinger 1976b; Jackendoff 1973; Fraser 1976; Lidner 1983; Ö̈Dowd 1998; Dehe *et al.* 2001; Jackendoff 2002; Huddleston and Pullum 2002; Baldwin 2005b):

- in the case that the object NP is not pronominal, transitive VPCs can occur in either the joined or split word order (c.f. (2.4)), while prepositional verbs must always occur in the joined form (c.f. (2.7));
- in the case that the object NP is pronominal, transitive VPCs must occur in the split word order (c.f. (2.5)), while prepositional verbs must occur in the joined form (c.f. (2.8));
- manner adverbs cannot occur between the verb and particle in VPCs (c.f. (2.6)), while they can occur with prepositional verbs (c.f. (2.9)). In this thesis, we will focus exclusively on VPCs where the particle is prepositional.

Verb-particle constructions

(2.4) Non-pronominal object: optional joined/split word order

- *Put on the sweater.*
- *Put the sweater on.*

(2.5) Pronominal object: obligatory split word order

- *Finish it up.*
- **Finish up it.*

(2.6) With manner adverb

- *Quickly eat up the food.*
- **Eat quickly up the food.*

Prepositional verbs

(2.7) Non-pronominal object

- *Look for a word.*
- **Look a word for.*

(2.8) Pronominal object

- *Look for it.*

[4] Prepositional verbs are obligatorily transitive, so there is no ambiguity with intransitive VPCs.

- **Look it for.*

(2.9) With manner adverb

- *Come with me quickly.*
- *Come quickly with me.*

VPCs undergo morphological, syntactic and semantic variation. Morphologically, VPCs inflect for tense and number (e.g., *take/takes/took/have taken/is taken/... off*). Syntactically, VPCs undergo word order variation, and are internally modifiable by a small set of adverbs (e.g., *right*, *back*, *way* and *all the way*). Semantically, VPCs populate the spectrum of compositionality relative to their components (Lidner 1983; Brinton 1985; Ishikawa 1999; Olsen 2000; Jackendoff 2002; Bannard *e1t al.* 2003; Cook and Stevenson 2006). According to the view of Bannard *et al.* (2003), VPCs can be sub-classified into four compositionality classes based on the independent semantic contribution of the verb and particle: (1) both the verb and particle contribute semantically, (2) only the verb contributes semantically, (3) only the particle contributes semantically, and (4) neither the verb nor the particle contributes semantically. Other researchers such as McCarthy *et al.* (2003) employ a one-dimensional classification of VPC compositionality (over a cline or a number of discrete sub-classes): compositional vs. non-compositional. Table 2.3 details the two classification systems, with examples.

Semantic Compositional	Contribution	Examples
verb & particle	Yes	*get down, take off*
verb	Yes	*lie down, eat up*
particle	Yes	*close off, be away*
none	No	*chicken out, make out*

Table 2.3: Classification of the compositionality of VPCs (Bannard et al. 2003 vs. McCarthy et al. 2003)

2.4 Light-Verb Constructions

Light-verb constructions (i.e., LVCs) are made up of a verb and a noun complement, often in the indefinite singular form (Jespersen 1965; Abeillė 1988; Miyagawa 1989; Grefenstette and Teufel 1994; Hoshi 1994; Sag *et al.* 2002; Huddleston and Pullum 2002; Butt 2003; Stevenson *et al.* 2004). The name of the construction comes from the verb being semantically bleached or "light", in the sense that their contribution to the meaning of the LVC is relatively small in comparison with that of the noun complement. LVCs are also sometimes termed *verb-complement pairs* (Kan and Cui 2006) or support verb constructions (Calzolari *et al.* 2002). Our definition of *light-verb constructions* is in line with that of Huddleston and Pullum (2002). The principal light verbs are *do, give, have, make, put* and *take*, for each of which we provide a selection of LVCs in (2.10)–(2.15). English LVCs generally take the form verb+*a/an*+object, although there is some variation here.

(2.10) do: *do a demo, do a drawing, do a report*

(2.11) give: *give a wave, give a sigh, give a kiss*

(2.12) have: *have a rest, have a drink, have (a) pity (on)*

(2.13) make: *make an offer, make an attempt, make a call*

(2.14) put: *put the blame (on), put an end (to), put stop (to)*

(2.15) take: *take a walk, take a bath, take a photograph (of)*

There is some disagreement in the scope of the term LVC, most notably in the membership of verbs which can be considered "light". Calzolari *et al.* (2002), e.g., argued that the definition of LVCs (or support verb constructions in their terms) should be extended as follows: (1) when the verbs combine with an event noun (deverbal or otherwise) and the subject is a participant in the event most closely identified with the noun (e.g., *take an exam, ask a question, make a promise*); and (2) when the subject of these verbs belongs to some scenario associated with the full understanding of the event type designated by the object noun (e.g., *pass an exam, survive an operation, answer a question, keep a promise*).[5]

Morphologically, LVCs inflect but the noun complement tends to have fixed number and a preference for determiner type (Wierzbicka 1982; Alba-Salas 2002; Kearns 2002; Butt 2003; Folli *et al.* 2003; Stevenson *et al.* 2004). For example, *put an end (to)* undergoes full verbal inflection (*put/puts/putting an end (to)*), but the noun complement cannot be pluralized or modified derivationally (e.g. **put an ending (to)*, **put ends to*).[6]

As described above, there is little constraint on the syntax of LVCs. Semantically, although the meaning of the verb in LVCs is bleached, a given noun will usually have strong constraints on which light verb(s) it combines with to form an LVC (e.g., *put blame (on)* vs.

[5] All examples are taken from Calzolari *et al.* (2002).

[6] But also note other examples where the noun complement can be pluralized, e.g. *take a bath* vs. *take baths*.

**do/give/have/make blame*), and different light verbs can lead to VPCs with different semantics (Butt 2003). For example, *put blame (on)* and *take blame* are both LVCs but having very different semantics: the subject of *put blame (on)* is the Agent of the blaming and the object of the PP headed by *on* is the Patient, while the subject of *take blame* is the Theme. Also, what light verb a given noun will combine with to form an LVC is often consistent across semantically-related noun clusters (e.g., *give a cry/moan/howl* vs. **take a cry/moan/howl*[7]).

2.5 Idioms

An *idiom* is an MWE whose meaning is fully or partially unpredictable from the meanings of its components (e.g., *kick the bucket*, *blow hot and cold*) (Nunberg *et al.* 1994; Potter *et al.* 2000; Sag *et al.* 2002; Huddleston and Pullum 2002). Huddleston and Pullum (2002) identified subtypes of idioms such as *verbal idioms* (e.g., *jump off, get out, run ahead*) and *prepositional idioms* (e.g., *for example, in person, under the weather*) which we classify as VPCs/prepositional verbs and determinerless PPs, respectively. In our terms, therefore, idioms are those non-compositional MWEs not included in the named construction types of VPCs, prepositional verbs, noun compounds and determinerless PPs. While all idioms are non-compositional (to varying degrees), we further categorize them into two groups: decomposable and non-decomposable (Nunberg *et al.* 1994). With *decomposable idioms*, given the interpretation of the idiom, it is possible to associate components of the idiom with distinct elements of the idiom interpretation based on semantics not immediately accessible from the components in isolation. Assuming an interpretation of *spill the beans* such as reveal'(X, secret'), e.g., we could analyze *spill* as having the semantics of reveal' and *beans* having the semantics of secret', and hence arrive at a post hoc explanation for the interpretation of the idiom via the reverse-engineered semantics of the components (through figuration of some description). Note that the interpretations of the components (*spill* as reveal' and *beans* as secret') are removed from those for the simplex words, and it is on this basis that we consider the idiom non-compositional. Other examples of decomposable idioms are *pull one's leg* and *pull strings*. Examples of non-decomposable idioms where a post hoc semantic decomposition is not accessible are *break a leg* and *kick the bucket*. Decomposable idioms tend to be syntactically flexible, as defined by the nature of the semantic decomposition, whereas non-decomposable idioms tend not to be syntactically flexible (Katz and

[7] Examples are from Stevenson *et al.* (2004).

Postal 2004; Wood 1964; Chafe 1968; Kastovsky 1982; Pawley and Syder 1983; Cruse 1986; Jackendoff 1997; Sag *et al.* 2002). For example, *spill the beans* can be passivized (*It's a shame the beans were spilled*) and internally modified (*AT&T spilled the Starbucks beans*).

2.6 Determinerless-Prepositional Phrases

Determinerless prepositional phrases (i.e., D-PPs) are MWEs that are made up of a preposition and a singular noun without a determiner (Quirk *et al.* 1985; Huddleston and Pullum 2002; Sag *et al.* 2002; Baldwin *et al.* 2006). Syntactically, D-PPs are highly diverse, and display differing levels of syntactic markedness, productivity and modifiability (Chander 1998; Ross 1995). That is, some D-PPs are non-productive (e.g., *on top* vs. **on edge*) and non-modifiable (e.g., *on top* vs. **on table top*), whereas others are fully-productive (e.g., *by car/foot/bus/...*) and highly modifiable (e.g., *at high expense*, *on summer vacation*). In fact, while some D-PPs are optionally modifiable (e.g., *on vacation* vs. *on summer vacation*), others require modification (e.g., **at level* vs. *at eye level*, and *at expense* vs. *at company expense*) (Baldwin *et al.* 2006).

Syntactically-marked D-PPs can be highly productive (Ross 1995; Grishman *et al.* 1998). For example, the preposition *by* combines with a virtually unrestricted array of countable nouns (e.g., *by bus/car/taxi/...*), but does not combine with uncountable nouns (e.g., **by information/ linguistics/...*).

Semantically, D-PPs have a certain degree of semantic markedness on the noun (Haspelmath 1997; Mimmelmann 1998; Stvan 1998; Bond 2001; Borthen 2003). For example, *in* combines with uncountable nouns which refer to a social institution (e.g., *school, church, prison* but not *information*) to form syntactically-*un*marked D-PPs with marked semantics, in the sense that only the social institution sense of the noun is evoked (e.g., *in school/church/prison/...* vs. **in information*) (Baldwin *et al.* 2006).

Class	Examples
institutional	*at school, in church, on campus, in gaol*
media	*on TV, on record, off screen, in radio*
metaphor	*on ice, at large, at hand, at liberty*
temporal	*at breakfast, on holiday, on break, by day*
means/manner	*by car, by hammer, by computer, via radio*

Table 2.4: A semantic classification of D-PPs

Note that some D-PPs with *in* combine with countable nouns such as *pub* and *hospital* but they do not refer to social institution. In general, D-PPs have been categorized into five semantic groups by Stvan (1998). These classes often correlate with a particular compositionality, e.g., metaphorical D-PPs are non-compositional while the other classes are compositional.

2.7 Chapter Summary

In this chapter, we have described the linguistic properties of MWEs, provided a classification of English MWEs, and provided details of a number of key English MWE types.

First, we defined the following linguistic properties in the context of MWEs: *idiomaticity*, *non-identifiability*, *situatedness*, *figuration*, *single-word paraphrasability*, *proverbiality* and *prosody*. We have identified *idiomaticity* as a primary defining property of MWEs, and described the relevance of the various properties to it. In particular, we subclassified idiomaticity according to the five areas of: lexical, syntactic, semantic, pragmatic and statistical idiomaticity. *Lexico-syntactic idiomaticity* was defined to be a mismatch in the syntax of the MWE relative to the properties of the simplex components. *Semantic idiomaticity* was defined to be (semantic) non-compositionality, i.e., a mismatch in the semantics of the MWE and that of its components. *Pragmatic idiomaticity* was defined to be the situatedness of an MWE, or association of the MWE with a particular situation. Finally, *statistical idiomaticity* was defined to occur when the frequency of the MWE is unusually high compared to that of its components or alternative phrasings of the same expression. From this, we then defined a *collocation* to be an MWE which was strictly *statistically idiomatic*. Other properties we identified were *single-word paraphrasability*, *proverbiality* and *prosody*. *Single-word paraphrasability* is the ability of an MWE to be paraphrased with a single simplex word; *proverbiality* is the property of MWEs to represent a recurrent situation of particular social interest; and *prosody* relates to the observation that certain MWEs occur with abnormal stress patterns.

We also provided a detailed description of the syntax and semantics of the following MWE types: *noun compounds*, *verb-particle constructions*, *light-verb constructions*, *idioms* and *determinerless prepositional phrases*. *Noun compounds* are comprised solely of nouns. *Verb-particle constructions* are combinations of a verb and one or more particle. *Light-verb constructions* are made up of one of a small subset of verbs with bleached semantics, and a noun complement. *Idioms* are non-compositional MWEs which do not fall into any of the identified

MWE types. We have further sub-classified idioms into *decomposable* and *non-decomposable* idioms. Finally, *determinerless PPs* are made up of a preposition and a singular noun without a determiner.

Chapter 3

Statistical Frameworks for MWE Extraction and Related Work

3.1 Introduction

In this chapter, we will look at the underlying methods commonly used in statistical approaches to MWE extraction: co-occurrence properties, substitutability, distributional similarity, semantic similarity and linguistic properties. We will take a look at how these methods are used for computational tasks relating to MWE extraction, and weigh up the advantages and disadvantages of each approach. We will also look at prior approaches, and provide an overview and comparison of the methods used in this thesis.

3.2 Co-occurrence Properties

3.2.1 Overview of Co-occurrence Properties

The use of *co-occurrence properties* in modeling MWE involves analyzing the co-occurrence of the components of an MWE under the assumption that two or more words occur together with markedly high frequency iff they form an MWE. This basic approach forms the basis of a plethora of association measures and has been used extensively for collocation extraction (in the standard use of the term) (Choueka *et al.* 1983; Smadja 1993; Lin 1998b; Pearce 2001; Evert 2004; Pecina 2005). This property has been found to be highly effective for extracting statistically-marked MWEs such as *shock and awe* as their co-occurrence tends to have abnormally high frequency relative to the alternative ordering. Example (3.1) is a sample of such high frequents such binomials (relative to their alternative ordering) while example (3.2) is a sample of such high-frequent binomials where both orderings have approximately the same frequency. This method can also be paired with analysis of alternative wordings for a given phrase in the form of substitutability (see Section 3.3). Note that when we say co-occurrence we refer to the co-occurrence of the parts rather than co-occurrence with any specific context, which is the basis of distributional similarity in Section 3.4.

(3.1) MWEs: *black and white, by and large, salt and pepper, shock and awe*

(3.2) Non-MWEs: *blue and red, small and large, salt and sugar*

Note that the underlying mechanism driving co-occurrence is statistical idiomaticity, as most MWEs are statistically idiomatic to some degree. In (3.1), for example, the method can be seen to have extracted statistically-marked MWEs (*by and large*) as well as semantically- (*black and white*) and pragmatically-marked MWEs (*shock and awe*).

Co-occurrence properties are often measured by association measures such as pointwise/ specific mutual information (Church and Hanks 1989), the Dice coefficient (Church and Hanks 1989), the student's T-test, Pearson's chi-square (Dunning 1993) and log likelihood (Dunning 1993). The measurement of co-occurrence properties is useful when the components combine together with markedly high frequency relative to the components, or alternatively relative to an alternative form of the same MWE. However, quantitatively measuring co-occurrence properties via a given association measure has its limitations. As most of the measures rely on lexicalized corpus frequencies, they are vulnerable to the effects of data sparseness.

Furthermore, it is often difficult to predict which association method will perform best over a given MWE type and corpus (Pecina 2005). Co-occurrence properties have been used widely in tasks such as extracting collocation and MWEs (Smadja 1993; Grefenstette and Teufel 1995; Villavicencio *et al.* 2004; Baldwin 2005b; Fazly *et al.* 2005; Villada Moiron 2005; Pecina 2005; Widdows and Dorow 2005; Kan and Cui 2006), modeling the compositionality of MWEs (Bannard 2003; McCarthy *et al.* 2003; Venkatapathy and Joshi 2005; Fazly and Stevenson 2007; Kim and Baldwin 2007a), and classifying MWE semantics (Fraser 1976; Lapata and Keller 2004). Below, we outline a representative selection of papers on the co-occurrence properties of

MWEs, in the context of extraction, compositionality modeling and semantic classification tasks, respectively. Note that in some instances, the original research uses the term *collocation* in the broader sense of the term to mean MWE. In our description of the research, we will use the terms MWE and collocation as outlined in Chapter 2.

3.2.2 Co-occurrence Properties for Extraction

Smadja (1993) proposed the XTRACT system for extracting MWEs from raw text, building on a number of ideas from previous work (Choueka *et al.* 1983; Church and Hanks 1989). The basis for XTRACT is that the components of MWEs co-occur with unexpectedly high frequency, and also that they tend to occur in fixed word positions relative to each other (an assumption which clearly falls down with VPCs in the split configuration). The method is made up of three steps. The first (similar to Church and Hanks (1989)) is to extract binary MWE candidates within a 5-word window based on strength (frequency of collocation) (identification of a particular word order/positioning which is notably more frequent than others). For example, for a given target word *takeover*, occurring with words *pill*$_{+2}$, *make*$_{+2}$ and *attempt*$_{+2}$, *attempt*$_{-1}$ (where *word*$_{N}$ is an occurrence of *word N* words to the left of the target word, considering each combination of word and position as a distinct data point), the method may filter out all but *attempt*$_{-1}$, corresponding to *takeover attempt* (i.e. *attempt* -1 words to the left = one word to the right of the target word).

The second step (similar to Choueka *et al.* (1983)) is to combine binary MWE candidates into multiple-word combinations and complex expressions. That is, from binary collocations extracted in the first stage, the method generates *n*-gram collocations from individual occurrences of the two words and analyzes the distribution of words and POS in surrounding context, and identifies any extra components which commonly co-occur with the elements of the bigram. For example, *chip stocks* may be expanded into *blue chip stocks*, and *price index* may be expanded *into the consumer price index.*

The third step involves syntactic analysis of the binary or larger MWEs from the second step to ensure they follow constituent boundaries and correspond to common syntactic configurations, e.g. modifier-modifiee, subject–verb or verb–object. In more recent work, Pecina (2005) tested a large number of co-occurrence-based extraction methods proposed in previous work in an MWE extraction task. The aim of the work was to empirically evaluate a comprehensive list of automatic MWE extraction methods using precision–recall curves, and to propose a new approach for combining individual extraction methods using supervised learning methods. Pecina used a total of 84 association measures based on occurrence frequencies (i.e., co-occurrence properties) over binary MWEs. As association measures, he used simple probabilities, mutual information and derived measures, statistical tests of independence, likelihood measures, and various heuristic association measures and coefficients. He also used context association measures based on syntactic and semantic units, with a more sound linguistic foundation. The final conclusion of this work was that the combination of multiple independent measures is superior to any one individual extraction method at MWE extraction.

Grefenstette and Teufel (1995) developed a method for extracting light verbs and their complements (i.e. *LVCs*) using co-occurrence properties. The basic idea behind this work is that the noun complements in LVCs are often deverbal (e.g., *proposal*), and that the distribution of nouns in PPs post-modifying noun complements in genuine LVCs (e.g., *(make a) proposal of marriage*) will be similar to that of the object of the underlying verb (e.g., *propose marriage*). Grefenstette and Teufel collected verbs and their nominalized forms, along with verb–object relations for the verbs and verb–noun–PP relations for the nouns, based on a low-level parser and heuristics.[1] From this, they selected the most common verb supporting the structure NP PP where the given nominalization heads the NP and the prepositional head of the PP is most similar to that of the underlying verb of the nominalization. In this case, therefore, multiple co-occurrences are considered (verb–noun and noun–preposition) to predict the light verb associated with a given nominalization. Baldwin (2005b) employed several statistical tests to extract prepositional verbs (see Section 2.3). The main idea in this work is that the verb and preposition in prepositional verbs co-occur more frequently than for simple verb–preposition combinations. Baldwin proposed a number of unsupervised methods to extract prepositional verbs based on statistical tests such as chi-square and Dice’s coefficient, as well as substitutability with highly frequent verbs and transitive prepositions (see Section 3.3). The method also adopted linguistic features of prepositional verbs, and demonstrated that co-occurrence properties were effective in

[1] Note that a parser has been employed in several MWE extraction methods, including Baldwin (2005a) in the context of English VPC extraction. However, in Baldwin (2005a), the parser(s) are used extensively not only to extract VPC candidates but also to analyze the argument structure of the VPC.

the extraction task, but that the combination of all extraction method strategies was superior overall.

3.2.3 Co-occurrence Properties for Compositionality

McCarthy *et al.* (2003) proposed a method to measure the compositionality of English VPCs based on the intuition that compositional MWEs are more likely to occur in similar contexts to their component words, than is the case for non-compositional MWEs. In detail, McCarthy *et al.* used distributional similarity (see Section 3.4) and statistical tests to model the compositionality of English VPCs. First, the authors identified VPC, verb and preposition instances from the output of the RASP parser, and from these calculated context vectors. They then calculated the distributional similarity between different combinations of VPCs and verbs, and used six different methods to estimate VPC compositionality. One such method was *overlap*, which is overlap in the top *X* neighbors of the VPC (not including the simple verb itself) and the same number of neighbors of the simplex verb. Another is *same particle* − *simplex* which is the number of neighbors in the top *X* which share the same particle as the VPC, minus the number of neighbors in top *X* for the simplex verb which share the same particle as the VPC. In addition to distributional similarity, the authors employed several statistical tests to measure compositionality, both based on corpus statistics and dictionary occurrence. In evaluation, they found high correlation between the best of the distributional methods (same particle − simplex) and the human-annotated compositionality values, and that simple co-occurrence in the form of statistical tests performed very badly over the target task.

Venkatapathy and Joshi (2005) proposed a method for measuring the relative compositionality of a verb–noun pair such as *take place* or *feel safe*. Verb–noun pairs often occur with high frequency, making them suited to co-occurrence-based analysis. The proposed methods are based on various types of collocation and context. The authors used five different co-occurrence tests, namely frequency, point-wise mutual information, least mutual information difference with similar collocations, distributed frequency of object and distributed frequency of object using the verb information. They also used distributional similarity based on the approach of Baldwin *et al.* (2003a) to model the compositionality of English MWEs. They evaluated the proposed methods using correlation, following the methodology of McCarthy *et al.* (2003). The authors concluded that collocation features are better for measuring the relative compositionality of verb–noun pairs than distributional similarity, and that the correlation between the combined features and the human ranking was much better than that using individual features.

3.2.4 Co-occurrence Properties for Semantics

Lapata and Keller (2004) used co-occurrence properties in a variety of NLP tasks, including bracketing of compound nouns and interpreting compound nouns. The main motivation for this research was to evidence the usefulness of web data by employing it for probabilistic modeling. The probabilistic models they used for NC bracketing and interpretations were very simplistic, and based on simple co-occurrence of parts of the NC (in the first instance) and parts with different prepositions (in the second). That is, for the bracketing task, they tested 10 different probabilistic models integrating the frequencies of bracketed candidates (e.g. *((back up compiler) disk)* vs. *(backup (compiler disk))*). For NC interpretation, they tested the method proposed in Lauer (1995) based on the co-occurrence of nouns with different prepositions (e.g. *night flight* paraphrased as *flight at night*). Their research demonstrated that simple web frequencies were highly successful when applied to these two (and other) tasks.

3.3 Substitutability

3.3.1 Overview of Substitutability

Substitutability is the ability to replace parts of MWEs with alternative lexical items, and involves comparison of the target MWE with anti-collocations. Also, this method is directly related to *single-word paraphrasability* described in Section 2.1.1. This approach is effective when parts of an MWE occur with unusually high frequency relative to lexical alternatives, i.e. their collocational association is high. In this thesis, we consider substitutability to be a subset of cooccurrence properties.

Substitutability can be applied to either compositional or non-compositional MWEs. Substitutability is closely related to anti-collocation, as when parts of the MWE are replaced, the new lexical items are generally no longer MWEs. Note that in substitutability, we always consider the whole MWE (in the form of the original or the anti-collocation), while in co-occurrence properties, we sometimes compare the whole to a variant word order, and sometimes compare the whole to its parts. Analysis of substitutability tends to be based on the same inventory of statistical tests as for co-occurrence, as outlined in Section 3.2.1.

In generating substitution candidates, we often replace components of the original MWE with synonyms, sister words or antonyms, depending on the task and approach. This is based on the assumption of institutionalization, i.e. that a particular word combination has been established as an MWE to the exclusion of other plausible possibilities based on related words. Table 3.1 gives details examples where substitution leads to syntactically and/or semantically anomalous word combinations.

MWE		***Non-MWE***
frying pan	→	*frying pot*
salt and pepper	→	*salt and sugar*
many thanks	→	*several thanks*
red tape	→	*yellow tape*

Table 3.1: MWEs and Non-MWEs based on substitution

In Table 3.1, when parts such as *pan* and *many* are replaced with related words, the newly-formed word combinations (i.e. *frying pot* and *several thanks*, respectively) are no longer MWEs. Similarly, *yellow tape*, formed by substituting *red* with *yellow* in *red tape*, does not preserve the original meaning of “bureaucracy” (non-elective government officials). Substitutability can also be used to investigate the limits of productivity of MWEs such as VPCs and NCs. Despite various semantic restrictions, certain MWEs are highly productive. Hence, substitutability can be employed in order to construct new MWEs while maintaining the original “semantic collocation” (e.g. the same verb synset combined with the same particle).

(3.3) *call up* -> *phone/ring up* vs. * telephone *up*

(3.4) *lemon juice* -> *orange/fruit/lime juice*

In (3.3), *call up* is the basis for generating the VPCs *phone up* and *ring up*, but anomalously not *telephone up*, despite *telephone* being a lexical variant of *phone*. Starting with *lemon juice* in (3.4), we form the three NCs *orange juice*, *lime juice* and *fruit juice*, based on substituting *lemon* with a synonym, hypernym and sister word, respectively.

In a computational context, substitutability is broadly used to classify word combinations as MWEs or non-MWEs (Lin 1998b; Lin 1998d; Lin 1999; Pearce 2001). Substitutability is also applicable to the modeling of MWE compositionality (Bannard *et al.* 2003; Bannard 2003; McCarthy *et al.* 2003; Kim and Baldwin 2007a), the generation of MWEs with related semantics or compositionality (Stevenson *et al.* 2004; Baldwin 2005b; Turney 2005; Kim and Baldwin 2007b; Kim and Baldwin 2007c), and semantic classification (Villavicencio *et al.* 2004; Villavicencio 2005; Uchiyama *et al.* 2005).

3.3.2 Substitutability for Extraction

Lin (1999) proposed a method for classifying word combinations as non-compositional and compositional using the substitution method. The idea behind this work is that when a phrase is non-compositional, substitution candidates will tend to have markedly different frequencies of occurrence. For example, *red tape* occurs with much higher frequency than *yellow tape* or *orange tape*, indicating that it is non-compositional. On the other hand, *economic impact* has similar frequency to alternative wordings such as *financial impact* and *economic effect* and is hence predicted to be compositional.

As the source of the substitution candidates, Lin used a distributional thesaurus (Lin 1998a), which was pre-computed from the output of the Minipar dependency parser (Lin 1993). He also used the output of Minipar over a large-scale corpus to compute frequencies of different word combinations in particular syntactic configurations, from which he calculated the degree of association via a variant of point-wise mutual information. He compared the degree of association of the target word combination with substitution candidates via a *Z*-score, which provides an indication of the relative differential in the association values. Only if the differential is high over all substitution candidates is the target word combination considered to be an MWE. The study found that substitutability was a successful means of predicting the non-compositionality of word combinations.

Pearce (2001) proposed a method for extracting MWEs using substitution over WordNet. The motivation for the substitution method is that parts of compositional MWEs can be substituted with related words such as synonyms and hypernyms while maintaining the same basic semantics. Similar to Lin (1999), if the substitution candidates occur markedly less frequently than the original, it is an interpreted to be an indication that the original was an MWE. Table 3.2 illustrates examples of MWEs and corresponding *anti-collocations* generated by this method.

MWE	Anti-collocations
emotional ***baggage***	*emotional* ***luggage***
many *thanks*	***several*** *thanks*
strong *coffee*	***powerful*** *coffee*

Table 3.2: Examples of MWEs and anti-collocations (Pearce 2001)

In Table 3.2, *emotional baggage* is an MWE whereas *emotional luggage* is not, despite *baggage* and *luggage* being synonyms. That is, in terms of the MWE properties described in Section 2.1.1, the MWEs in Table 3.2 are institutionalized, as indicated by their unusually high frequency relative to their anti-collocations. In evaluation, Pearce classified the test instances into three classes: MWE, potential and unknown. The experimental results were promising, and demonstrated the power of the rich hierarchical structure of WordNet.

3.3.3 Substitutability for Semantic Classification

Turney (2005) proposed a method for measuring the relational similarity between a pair of nominal phrases, for use in analogical reasoning. For example, the noun pair *cat:meow* is analogous to the pair *dog:bark*, because both represent an animal and its sound. Likewise, *milk:drink* and *pie:eat* form a relational pair in which the relation would be of the type food and how to consume it. The particular task Turney is interested in is the SAT test, where given a target noun pair such as *quart:volume* and a set of 5 candidate noun pairs, such as in (3.5), the task is to select the candidate noun pair that is most relationally similar to the target pair.

(3.5) *day:night*, *mile:distance*, *decade:century*, *friction:heat*, *part:whole*

In this case, the answer would be *mile:distance* on the basis that the first noun is a specific measurement of the second noun.

While the noun pairs are not in fact MWEs, they are closely related to NCs, and the methodology proposed by the author is closely related to methods used for interpreting NCs.

To measure the similarity between a giving combination of two noun pairs, the author employs substitution relative to the target noun pair A: B, replacing a word at a time based on the top 10 related words using synonymy, hypernymy and sister words. He then filters generated word pairs based on frequency, and measures the similarity of phrases based on clustering to confirm that they preserve the same relational semantics.

Two notable aspects of this research are that: (1) it is based on substitutability; and (2) it makes use of clustering and not classification, and as such does not attempt to resolve the exact relation between the nouns in a given pair.

Uchiyama *et al.* (2005) used the co-occurrence properties of Japanese compound verbs to predict their semantics. Japanese compound verbs are made up of a verb in the continuative form (*V*1) and an auxiliary verb (*V*2), as in *tabe-sugiru/eat too much*. Japanese compound verbs are highly productive and semantically ambiguous, and are subject to semantic constraints between the first verb and the second verb. (3.6)– (3.8) show examples of Japanese compound verbs and a classification according to the semantics of the *V*2 (i.e. spatial, aspectual and adverbial), which also correspond to distinct translation strategies into English (as indicated). Note that the translation between Japanese and English has been carried out base on the fact that they have a semi-similarity due to their loose connection.

(3.6) Spatial compound verbs: *V* 2 is translated as a verb in English.

nage-ageru ´ throw (a ball) up
keri-ageru ´ kick (a ball) up

(3.7) Aspectual compound verbs: *V* 2 is translated as a particle in English.

yude-ageru ´ finish boiling (vegetables)
musi-ageru ´ finish steaming (vegetables)

(3.8) Adverbial compound verbs: *V* 2 is translated as a adverb in English.

donari-ageru ´ shout
odosi-ageru ´ threaten

Uchiyama *et al.* (2005) proposed a novel machine learning method to disambiguate the semantics of *V*2, based on the co-occurrence of *V*1 and *V*2. The method is based on a matrix analysis of *V*1–*V*2 combinatory. That is, the features used to classify a given combination of *V*1 and *V*2 are based on the semantic classes of each $V2'$ which co-occurs with *V*1, and each $V1'$ which co-occurs with *V*2, based on the row containing *V*1 and column containing *V*2.

3.4 Distributional Similarity

3.4.1 Overview of Distributional Similarity

Distributional similarity is a method for estimating semantic similarity based on the analysis of the contexts in which two lexical items are used. The basic idea behind this method was popularized by Firth (1957), and states that when two words are similar, they will occur in similar contexts (i.e. their neighboring words within a word window will be similar). In the context of MWEs, distributional similarity is frequently used to compare the token occurrences of an MWE with the token occurrences of its components outside of the MWE. For example, when *kick the bucket* is used as an idiom, it may occur commonly with words such as *mourn*, *sad* and *bury*, while *kick* and *bucket* may occur commonly with very different words such as *water*, *accident* and *container*. This suggests that the semantics of *kick the bucket* differs from that of its parts, and that it is therefore a non-compositional MWE. A common window size used to model contextual similarity is 25 words to either side of a given lexical item token. The similarity between the context vectors associated with two lexical items is commonly measured with *cosine similarity* (Salton *et al.* 1975). (3.9) is an example of the idiom *kick the bucket*, where a context window of 5 words has been indicated via underlining; (3.10) is a literal usage of *kick the bucket* with its corresponding 5-word context window.

(3.9) The old man requested, “When I ***kick the bucket***, bury me on top of that mountain.”

(3.10) When we were about to enter the room, Kim accidentally ***kicked the bucket*** next to the door.

Comparing distributional similarity with the previous two methods, it is similar to co-occurrence properties in that it compares word combinations, with the big distinction that distributional similarity analyses the context of token occurrences of a given lexical item, whereas co-occurrence properties analyses the frequencies of components.

Distributional similarity is a more powerful method in that there is greater scope for parameterization/reformulating in terms of: how the context window is defined, how token counts are translated into feature vectors, and how context vectors are compared. In the context of translating token counts into feature vectors, e.g. a considerable amount of work has been done on dimensionality reduction, such as with *latent semantic analysis* (LSA) (Landauer *et al.* 1998) to overcome data sparseness.

Co-occurrence properties, on the other hand, are based fundamentally on token counts of components/re-orderings of the original lexical item, with the only place for innovation in the numeric interpretation of those numbers. One way in which researchers have extended the basic distributional similarity method is by redefining the context window to look at the second-order co-occurrence of words. Here, rather than using the neighboring words of the target lexical item's neighboring words across multiple contexts as a direct representation of the target expression, the neighboring words of a specific token occurrence of the target expression are in turn modeled via their neighboring words. For example, let's assume that the target word *bank* has neighboring words *money*, *stock* and *savings* in a given context window. Rather than represent these directly as a 3-term (sparse) vector, we look to see what words each of them co-occurs with across the sum total of their usages. For example, *money* might co-occur with terms such as *banking* and *market* across all of its token occurrences, giving us a rich vector with which to present that one context term. We similarly generate individual vectors for the other two context terms and use the combination of the three to represent the original context. If we were then to compare the original token instance of *bank* with a single token instance of *financial institution*, say, although the immediate context words may not overlap, there is a good chance hat the context vectors for each of the context words will. Second-order co-occurrence therefore provides a powerful mechanism for performing token-level analysis of context, e.g. in disambiguating individual occurrences of word sequences (such as *kick the bucket*) as either MWEs or simple compositional combinations.

The main weakness of distributional similarity is that it relies on large amounts of corpus data to operate effectively. Distributional similarity has been employed to model the compositionality of MWEs (Schone and Jurafsky 2001; Bannard 2003; Baldwin *et al.* 2003a; Venkatapathy and Joshi 2005; McCarthy *et al.* 2007), to identify MWEs (Katz and Giesbrecht 2006), and to classify the semantics of MWEs (Stevenson *et al.* 2004).

3.4.2 Distributional Similarity for Compositionality

Bannard *et al.* (2003) used the distributional semantics of English VPCs to measure their compositionality and to model the contribution of the verb and particle in the overall semantics of the VPC. The basic idea behind this work is that if an MWE is compositional, then it will occur in the same lexical context as its components. The authors assumed that VPCs populate a continuum between fully compositional and fully non-compositional structures. Bannard *et al.* used four different classification methods: the method of Lin (1999), the context space model of Schutze (1998), a substitution method, and distributional similarity between each of the components and the overall VPC. The authors found that the mixed methods performed best, and the third and fourth methods outperformed the first and second methods. Significantly, this paper showed that distributional semantics can be applied to the analysis of particles and MWEs, where previous work had tended to focus exclusively on simplex content words.

Baldwin *et al.* (2003a) used distributional similarity to compare MWEs with their components, focusing on NCs and VPCs. The proposed method was based on the context space model of Schutze (1998), which incorporates LSA[2]. (3.11) illustrates the outputs of the method for the VPCs *cut out* and *cut off* with the component verb *cut*. Based on the similarity values, the model is predicting that *cut out* is more compositional than *cut off*.

(3.11) similarity (*cut*, *cut out*) = .433 vs. similarity (*cut*, *cut off*) = .183

To evaluate their method, the authors compared the predicted similarity between VPCs and their component verbs, and NCs and their component nouns, with similarities generated from WordNet. They found a weak correlation between the two, and once again demonstrated the potential for distributional semantics to model the compositionality of MWEs.

3.4.3 Distributional Similarity for Identification

Katz and Giesbrecht (2006) used second-order distributional similarity to identify non-compositional MWEs (i.e. idioms) in German. As outlined above, the intuition behind the method is that non-compositional MWEs will co-occur with significantly different words to their components, as can be captured in their second-order cooccurrence. For example, when *kick the bucket* is used as an idiom (meaning “die”), then the context words around it will be very different to those for both *kick* and *bucket* in isolation, whereas when it is used compositionally, it will be more similar in usage to the component words. To measure the similarity between

[2] http://infomap.stanford.edu/

German MWEs and their components, they once again employed the context space model of Schutze (1998).

Figure 3.2 shows the context vector associated with an idiomatic usage of *den loffel Abgeben* (corresponding to *kick the bucket* in German, and literally meaning "to eat the spoon"), compared to each of its component words vs. a paraphrase for the MWE (*sterben*, meaning *die*). Here, therefore, the prediction would be that the usage is idiomatic rather than literal. The authors concluded that it is possible to identify MWEs in context using distributional similarity.

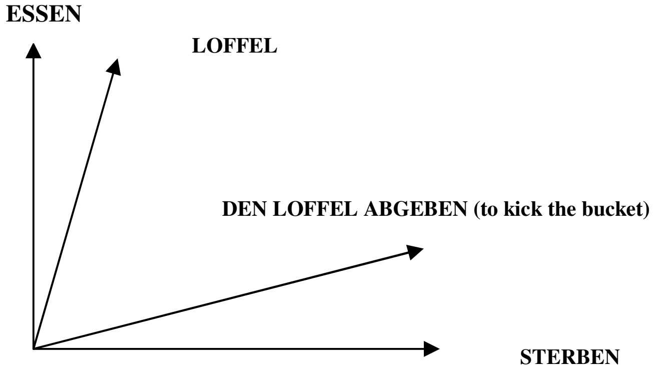


Figure 3.1: Distributional semantics of the German idiom *den loffel Abgeben* (Katz and Giesbrecht 2006)

3.5 Semantic Similarity

3.5.1 Overview of Semantic Similarity

Semantic similarity uses a direct model of the semantics of the parts (and possibly the whole) of an MWE to measure compositionality. The underlying assumption is that with compositional MWEs, the semantics of the whole MWE can be decomposed into the semantics of the parts. For example, we would expect the semantics of *add up* to be closely related to that of *add*, and to a lesser degree *up*. Similarly, we would expect *sum up* to have similar properties to *add up* based on them both incorporating the same particle and *sum* and *add* being similar (Villavicencio 2005; Kim and Baldwin 2007a). Compared with *distributional similarity*, the main difference is that *semantic similarity* employs the semantics from the MWE parts whereas *distributional similarity* uses the information from the target word's neighboring words.

One application of semantic similarity is in the interpretation of MWEs. That is, when the corresponding components of a pair of MWEs are similar (such as with *sum up* vs. *add up* above), it is generally the case that they have a similar interpretation, e.g. via a semantic relation. This gives rise to a method for interpreting MWE semantics (Rosario and Marti 2001; Moldovan *et al.* 2004; Kim and Baldwin 2005; Nastase *et al.* 2006; Girju 2007; Kim and Baldwin 2007c). (3.12) and (3.13) show how to interpret the semantic relations in NCs using semantic similarity.

(3.12) modifier = fruit, head noun = liquid -> SR = make

e.g. *apple juice, orange juice, grapes nectar, chocolate milk*

(3.13) modifier = location, head noun = liquid -> SR = location

e.g. *Fuji apple, California orange, Bordeaux wine*

In (3.12) and (3.13), despite different combinations of lexical items, NCs such as *apple juice* and *chocolate milk* are predicted to have the same SR of make, as the modifier and head noun, respectively, have similar semantics.

The advantage of this method comes from the ability to use existing similarity measures for simplex words (e.g. based on lexical resources such as WordNet or CoreLex) to accurately interpret MWEs, although such methods are limited by the coverage of the underlying similarity measures (and hence the coverage of any base lexical resources).

This method is employed in computational tasks such as interpreting NCs (Rosario and Marti 2001; Moldovan *et al.* 2004; Kim and Baldwin 2005; Girju 2007), and modeling the compositionality of MWEs (Piao *et al.* 2006; Kim and Baldwin 2007a).

3.5.2 Semantic Similarity for Compositionality

Piao *et al.* (2006) proposed the use of semantic similarity to test the compositionality of MWEs. The basic idea was that there is a correlation between compositionality and the relative similarity between the semantics of an MWE and its parts. To model semantics, they used a field taxonomy based on the Lancaster English Semantic Lexicon[3], which is derived from the McArthur (1981) Longman Lexicon of Contemporary English. The lexicon has 21 major semantic fields, further divided into 232 subcategories, and contains nearly 55,000 single-word entries and over 18,800 MWEs entries. (3.14) shows the semantic hierarchy for *food & farming*, e.g.

(3.14) F: FOOD & FARMING

Food ⊃ Drinks ⊃ Cigarettes & Drugs ⊃ Farming & Horticulture

[3] www.comp.lancs.ac.uk/ucrel/usas/

The paper proposes a novel method for measuring the semantic distance between an MWE and its component words based on hand-tagged hierarchical semantic information. Piao *et al.* evaluated the proposed method over 89 MWEs, scoring each from 0 (least compositional) to 10 (completely compositional). They used Spearman's correlation coefficient to measure the correlation between the automatic and manual rankings, and claimed results comparable to human performance.

3.5.3 Semantic Similarity for Semantic Relations

Rosario and Marti (2001) used semantic similarity to interpret NCs in the medical domain based on the CUI and MeSH medical ontologies. CUI is part of UMLS (Humphreys *et al.* 1998), and is comprised of three resources: a meta-thesaurus, semantic network and specialist lexicon. MeSH is one of the source vocabularies of UMLS, where concepts are identified by unique concept identifiers in hierarchical structures. (3.15) shows the hierarchical classes for the modifier and head noun in *flu vaccination*.

(3.15) *flu vaccination* ⟶ SR = purpose

- CUIs: C0016366 I C0042196
- MeSH: D4.808.54.79.429.154.349 I G3.770.670.310.890

To classify NCs which are manually tagged with medical classes, Rosario and Marti used a neural network. They found that the domain-specific lexical hierarchy successfully captured the semantic similarity of NCs to interpret SRs, but also that coverage is a significant bottleneck for the medical domain. Moldovan *et al.* (2004) used word sense collocations in NCs to interpret SRs.

The basic idea is that when NCs have the same sense collocation, i.e. corresponding components are semantically similar, then they most likely have the same SR. For example, *car factory* and *automobile factory* have identical sense collocation, and as such, have the same SR make. Moldovan *et al.* proposed a probabilistic model called semantic scattering to implement their sense collocation-based interpretation method. Semantic scattering is based on Equation 3.16 and Equation 3.17.

$$P\left({}^{r}/_{f_{ij}}\right) = \frac{n(r,f_{ij})}{n(f_{ij})} \tag{3.16}$$

where f_{ij} is a simplified feature pair $f_i\,f_j$ (i.e. the word senses of the modifier and head noun in an NC) and r is the semantic relation. The preferred SR $r*$ for the given word sense combination is that which maximizes the probability:

$$\begin{aligned} r^* &= argmax_{r€R} P\left(r/f_{ij}\right) \\ &= argmax_{r€R} P\left(f_{ij}/r\right) P(r) \qquad (3.17) \end{aligned}$$

In evaluation, the authors found that their method performed at about 43% accuracy over open domain NCs.

3.6 Linguistic Properties

3.6.1 Overview of Linguistic Properties

A final method is the use of *linguistic properties* to analyze MWEs. The assumption is that certain linguistic properties correlate with MWE compositionality, as well as particular syntactic and semantic types. Linguistic properties are generally considered in combination with other computational methods, rather than forming a standalone computational method, as they tend to suffer from data sparseness, i.e. have high precision but low recall over a given set of MWE types. (3.22) – (3.24) show an example of using linguistic properties to extract VPCs from corpus data.

(3.22) Particle position

- *lead (the donkey/them) on*
- **draw (inner strength/it) on*

(3.23) Particle modifiability

- *pick (back/right back) up the pencil*
- **draw (back/right back) on (inner strength/it)*

(3.24) Nominalization

- *feedback, backup*
- **drawon*

In (3.22), we are able to rule out *draw on* as a VPC based on the fact that the preposition is not compatible with the split word order. In (3.23), we are once again able to rule out *draw on* as a VPC on the grounds that it is not possible to modify the preposition *on* with *back* or *right back.* Finally, in (3.24), the fact that *draw on* does not nominalize to **drawon* is suggestive of the fact that it is not a VPC (although in this last case, it is a sufficient but not necessary condition on VPCs).

Linguistic properties often take the form of highly-specific syntactic features of MWEs, either in context or at the type-level. While linguistic properties can rely on context, they differ from distributional similarity in that they are very selective and fine-tuned to particular construction types. As shown in the examples above, linguistic properties can provide very reliable features for identifying or otherwise classifying MWEs. Their main drawbacks are that they do not easily generalize, and rely on the occurrence of very particular usages/contexts. Example tasks where linguistic properties have been employed are MWE extraction (Baldwin and Villavicencio 2002; Baldwin 2005a; Nakov and Hearst 2005), identification (Patrick and Fletcher 2004; Van Der Beek 2005; Kim and Baldwin 2006a) and semantic classification (O'Hara and Wiebe 2003; Stevenson *et al.* 2004; Cook and Stevenson 2006).

3.6.2 Linguistic Properties for Extraction

Baldwin (2005a) used linguistic properties (among many other features) to extract fully-specified English VPCs from raw text corpora. The basic approach in this work was to boost the precision of more general-purpose features with linguistic properties, based on the output of various preprocessors (e.g. parsers and chunkers). Specific linguistic properties used by the author were analysis of the word order of the object NP and preposition in transitive VPC candidates, particularly when the object NP is pronominalized. That is, transitive English VPCs undergo the particle alternation, producing the joined and split word orders. Also, pronominal objects must be expressed in the split configuration, and manner adverbs cannot occur between the verb and particle in either transitive or intransitive VPCs. These properties are illustrated in (3.25)–(3.27).

(3.25) Particle alternation

- joined: *Kim handed in the paper.*
- split: *Kim handed the paper in.*

(3.26) Pronominalized object word order

- *hand it in.*
- **hand in it.*

(3.27) Manner adverb word order

- *Hand it in promptly.*
- ?**Hand it promptly in.*

The task in Baldwin (2005a) was undertaken with no assumptions about corpus annotation, using only information from pre-processors such as a part-of-speech tagger, chunker and RASP. It also evaluated VPC extraction as both shallow and deep lexical acquisition tasks, that is either as the simple task of determining what combinations of verb and preposition can form a VPC, or as the harder task of determining what combinations of verb and preposition can form an intransitive and transitive VPC (e.g. for the purposes of a deep grammar lexicon).

The proposed method was tested over three corpora (Brown Corpus, Wall Street Journal and British National Corpus), and linguistic properties were shown to provide valuable evidence in the extraction task, especially over low-frequency VPCs. Nakov and Hearst (2005) employed linguistic properties in a probabilistic model for bracketing NCs with 3 or more terms in the medical domain, building on the work of Marcus (1980) and Lauer (1995). Nakov and Hearst extended this earlier work by integrating linguistic features into their model, based on analysis of surface features in web data as illustrated in (3.28)–(3.30).

(3.28) Dash or hyphen

- *left bracketing* : *cell-cycle analysis* ⟶ *((cell-cycle) analysis)*
- *right bracketing* : *donor T-cell* ⟶ *(donor (T-cell))*

(3.29) Genitive ending or possessive marker

- *left bracketing* : *brain stem's cells* ⟶ *((brain stem) cells)*
- *right bracketing* : *brain's stem cells* ⟶ *(brain (stem cells))*

(3.30) Capitalization

- *left bracketing* : *Plasmodium vivax Malaria* ⟶ *((Plasmodium vivax) Malaria)*
- *right bracketing* : *brain Stem cells* ⟶ *(brain (Stem cells))*

Based on these and other features, Nakov and Hearst (2005) developed an unsupervised method for NC bracketing using chi-square, and achieved 89.34% bracketing accuracy.

3.6.3 Linguistic Properties for Semantics

Cook and Stevenson (2006) classified particle semantics in English VPCs using linguistic properties of VPCs. The authors observed the following facts: (1) semantically similar verbs combine with a similar range of target particles; and (2) what verbs can combine with a given particle is an indicator of the semantics of the target particle. Based on these observations, the authors used slot features to encode the relative frequencies of the syntactic slots (i.e. subject, direct and indirect object, and object of a preposition), and particle features to encode the relative frequency of the verb co-occurring with high frequency particles. Cook and Stevenson classified the particle *up* into four different semantic classes, as illustrated in (3.31)–(3.34).

(3.31) Vertical: *The price of gas jumped up.*

(3.32) Goal-oriented: *The deadline is coming up quickly.*

(3.33) Completive: *Finish up your dinner quickly.*

(3.34) Reflexive: *Roll up the curtain.*

The paper also used co-occurrence features to classify particle semantics. In their experiments, the authors found that the method based on linguistic properties outperformed that using word co-occurrence features in the task of classifying particle semantics.

3.7 Collocations

3.7.1 Overview of Collocation

A Collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. Or in the words of Firth (1957: 181): “Collocations of a given word are statements of the habitual or customary places of that word.” Collocations include noun phrases like *strong tea* and *weapons of mass destruction*, phrasal verbs like *to make up*, and other stock phrases like *the rich and powerful*. Particularly interesting are the subtle and not-easily-explainable patterns of word usage that native speakers all know: why we say *a stiff breeze* but not **a stiff wind* (while either *a strong breeze* or *a strong wind* is okay), or why we speak of *broad daylight* (but not **bright daylight* or **narrow darkness*).

Collocations are characterized by limited *compositionality*. We call a natural language expression compositional if the meaning of the expression can be predicted from the meaning of the parts. Collocations are not fully compositional in that there is usually an element of meaning added to the combination. In the case of *strong tea*, *strong* has acquired the meaning *rich in some active agent* which is closely related, but slightly different from the basic sense *having great* *physical strength*. Idioms are the most extreme examples of non-compositionality. Idioms like *to kick the bucket* or *to hear it through the grapevine* only have an indirect historical relationship to the meanings of the parts of the expression. We are not talking about buckets or grapevines literally when we use these idioms. Most collocations exhibit milder forms of non-compositionality, like the expression *international best practice* that we used as an example earlier in this book. It is very nearly a systematic composition of its parts, but still has an element of added meaning. It usually refers to administrative efficiency and would, for example, not be used to describe a cooking technique although that meaning would be compatible with its literal meaning.

Collocations are important for a number of applications: natural language generation (to make sure that the output sounds natural and mistakes like *powerful tea* or *to take a decision* are avoided), computational lexicography (to automatically identify the important collocations to be listed in a dictionary entry), parsing (so that preference can be given to parses with natural collocations), and corpus linguistic research (for instance, the study of social phenomena like the reinforcement of cultural stereotypes through language (Stubbs 1996)).

There is much interest in collocations partly because this is an area that has been neglected in structural linguistic traditions that follow Saussure and Chomsky. There is, however, a tradition in British linguistics, associated with the names of Firth, Halliday, and Sinclair, which pays close attention to phenomena like collocations. Structural linguistics concentrates on general abstractions about the properties of phrases and sentences. In contrast, Firth's *Contextual Theory of Meaning* emphasizes the importance of context: the context of the social setting (as opposed to the idealized speaker), the context of spoken and textual discourse (as opposed to the isolated sentence), and, important for collocations, the context of surrounding words (hence Firth's famous dictum that a word is characterized by the company it keeps). These contextual features easily get lost in the abstract treatment that is typical of structural linguistics. A good example of the type of problem that is seen as important in this contextual view of language is Halliday's example of strong vs. powerful tea (Halliday1966: 150). It is a convention in English to talk about *strong tea*, not *powerful tea*, although any speaker of English would also understand the latter unconventional expression. Arguably, there are no interesting structural properties of English that can be gleaned from this contrast. However, the contrast may tell us something interesting about attitudes towards different types of substances in our culture (why do we use *powerful* for drugs like heroin, but not for cigarettes, tea and coffee) and it is obviously important to teach this contrast to students who want to learn idiomatically correct English. Social implications of language use and language teaching are just the type of problem that British linguists following a Firthian approach are interested in.

In this chapter, we will describe the principal approaches to finding collocations: selection of collocations by frequency, selection based on mean and variance of the distance between focal word and collocating word, hypothesis testing, and mutual information.

3.7.2 Different Methods for Collocation Extraction

- **Frequency**

Surely the simplest method for finding collocations in a text corpus is counting. If two words occur together a lot, then that is evidence that they have a special function that is not simply explained as the function that results from their combination. Since MWEs generally get institutionalized, the frequency of the collocation is a good first indicator of MWEness, given a large enough corpus. Hence candidate collocations are ranked by the frequency of occurrence in the corpus. The drawback of measuring simply the frequency of a phrase is that it needs a large corpus because fewer occurrence of a phrase in a corpus does not imply any measurable conclusion of the behavior of the phrase as MWE.

- **Mean and Variance**

Frequency-based search works well for fixed phrases. But many collocations consist of two words that stand in a more flexible relationship to one another. Consider the verb *knock* and one of its most frequent arguments, *door*. Here are some examples of knocking on or at a door:

(3.35) she knocked on his door.

(3.36) they knocked at the door.

(3.37) 100 women knocked on Donaldson's door.

(3.38) a man knocked on the metal front door.

The words that appear between *knocked* and *door* vary and the distance between the two words is not constant so a fixed phrase approach would not work here. But there is enough regularity in the patterns to allow us to determine that *knock* is the right verb to use in English for this situation, not *hit*, *beat* or *rap*.

A short note is in order here on collocations that occur as a fixed phrase versus those that are more variable. To simplify matters we only look at fixed phrase collocations in most of the cases, and usually at just bigrams. But it is easy to see how to extend techniques applicable to bigrams to bigrams at a distance. We define a collocational window (usually a window of 3 to 4 words on each side of a word), and we enter *every* word pair in there as a collocational bigram. We then proceed to do the calculations as usual on this larger pool of bigrams. However, the mean and variance based methods described in this section by definition look at the pattern of varying distance between two words. If that pattern of distances is relatively predictable, then we have evidence for a collocation like *knock . . . door* that is not necessarily a fixed phrase.

The mean is simply the average offset. The variance measures how much the individual offsets deviate from the mean. We estimate it as follows.

$$\sigma^2 = \frac{\sum_{1=1}^{n}(d_i - \mu)^2}{n-1} \tag{3.18}$$

where n is the number of times the two words co-occur, d_i is the offset for co-occurrence *i*, and μ is the mean. If the offset is the same in all cases, then the variance is zero. If the offsets are randomly distributed (which will be the case for two words which occur together by chance, but not in a particular relationship), then the variance will be high. As is customary, we use the *standard deviation* $\sigma=\sqrt{\sigma^2}$, the square root of the variance, to assess how variable the offset between two words is.

• Hypothesis Testing

One difficulty that we have glossed over so far is that high frequency and low variance can be accidental. If the two constituent words of a frequent bigram are frequently occurring words, then we expect the two words to co-occur a lot just by chance, even if they do not form a collocation. What we really want to know is whether two words occur together more often than chance. Assessing whether or not something is a chance event is one of the classical problems of statistics. It is usually couched in terms of hypothesis testing. We formulate a *null hypothesis* H_0 that there is no association between the words beyond chance occurrences, compute the probability p that the event would occur if H_0 were true, and then reject H_0 if p is too low (typically if beneath a *significance level* of p < 0:05, 0:01, 0:005, or 0:001) and retain H_0 as possible otherwise (Significance at a level of 0:05 is the weakest evidence that is normally accepted in the experimental sciences. The large amount of data commonly available for statistical NLP tasks means that we can often expect to achieve greater levels of significance).

- **T-test**

We need a statistical test that tells us how probable or improbable it is that a certain constellation will occur. A test that has been widely used for collocation discovery is the t test. The t test looks at the mean and variance of a sample of measurements, where the null hypothesis is that the sample is drawn from a distribution with mean μ. The test looks at the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely one is to get a sample of that mean and variance (or a more extreme mean and variance) assuming that the sample is drawn from a normal distribution with mean μ. To determine the probability of getting our sample (or a more extreme sample), we compute the t statistic:

$$t(x,y) = \frac{mean\big(P(X,Y)\big) - mean\ (P(X))mean(P(Y))}{\sqrt{\big(\sigma^2 P(X,Y)\big) + \sigma^2\big(P(X)\big)\sigma^2\ (P(Y)}\,)}$$

$$\approx \frac{C(X,Y) - \frac{C(X)C(Y)}{N}}{\sqrt{C(X,Y)}} \tag{3.19}$$

Here *C(x)* and *C(y)* are respectively the frequencies of word *X* and word *Y* in the corpus, *C(X,Y)* is the frequency of bigram <X Y> and N is the total number of tokens in the corpus. The bigram count can be extended to the frequency of word X when it is followed or preceded by Y in the window of K words (here K=1). If the t statistic is large enough we can reject the null hypothesis.

- **Pearson's chi-square test**

Use of the t-test has been criticized because it assumes that probabilities are approximately normally distributed, which is not true in general. An alternative test for dependence which does not assume normally distributed probabilities is the χ^2-test (pronounced "chi-square test"). In the simplest case, this 2 test is applied to a 2-by-2 table as shown in Table 3.3. The essence of the test is to compare the observed frequencies in the table with the frequencies expected for independence. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

	X = *new*	X ≠ *new*
Y= *companies*	n_{11} (*new companies*)	n_{12} (e.g., *old companies*)
Y ≠ *companies*	n_{21} (e.g., *new machines*)	n_{22} (e.g., *old machines*)

Table 3.3: A 2-by-2 table showing the dependence of occurrences of *new* and *companies*

Each variable in the Table 3.3 depicts its individual frequency e.g. n_{11} denotes the frequency of the phrase "new companies". The χ^2 statistic sums the differences between observed and expected values in all squares of the table, scaled by the magnitude of the expected values, as shown in equation (3.20)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \tag{3.20}$$

where $O_{ij} = \frac{\sum_k n_{ik}}{N} \times \frac{\sum_k n_{kj}}{N} \times N \; and \; N$ is the number of tokens in the corpus.

This result is the same as we got with the t statistic. In general, for the problem of finding collocations, the differences between the t statistic and the chi square statistic do not seem to be large. However, this test is also appropriate for large probabilities, for which the normality assumption of the T-test fails. This is perhaps the reason that the χ^2 test has been applied to a wider range of problems in collocation discovery.

- ## Point-wise Mutual Information

An information-theoretically motivated measure for discovering interesting collocations is point *wise mutual information* (Church et al. 1991; Church and Hanks 1989; Hindle 1990). Fano (1961: 27–28) originally defined mutual information between particular events x and y, in our case the occurrence of particular words, as follows:

$$PMI(x\ y) = \log_2 \frac{P(x,y)}{P(x).P(y)} \approx \log_2 \frac{NC(x,y)}{C(x).(y)} \tag{3.21}$$

The explanation of the variables of the above equation is described later. PMI represents the amount of information provided by the occurrence of the event represented by X about the occurrence of the event represented by Y.

3.8 Chapter Summary

In this chapter, we have provided an overview of the statistical approaches most commonly used in modeling MWEs. In detail, we presented five different statistical approaches: co-occurrence properties, substitutability, distributional similarity, semantic similarity and linguistic properties. For each approach, we described the basic ideas and reviewed a small sample of related work in the context of MWE tasks such as *MWE identification/extraction*, *semantic classification* and *interpreting semantic relations* (as defined in Section 1.2).

Co-occurrence properties are the use of the frequencies of the components or an alternative word ordering of a given MWE to analyze whether the MWE is statistically marked. These frequencies are then plugged into a variety of statistical tests to measure the cohesion among the components. In addition to being effective at detecting statistical idiomaticity, it has been employed in modeling compositionality.

Substitutability is analysis of the effect on the MWE of substituting components with related terms. The frequency or other features of the target MWE are then compared to the *anti-collocations* generated through substitution. Substitutability is considered a special case of co-occurrence properties and is particularly suited to the modeling of compositionality, and allows for more fine-grained analysis than co-occurrence properties. This method is often employed to extract MWEs and to classify the semantics of MWEs.

Distributional similarity involves analysis of the context of use of different lexical items. Based on the assumption that similar words will occur in similar contexts, the more similar the context vectors of a given pairing of lexical items, the more similar they are predicted to be. Distributional similarity can be calculated based on the contexts of use of a lexical item across multiple usages, or alternatively based on the context vectors of each of the context words surrounding a lexical item in a given usage (second-order co-occurrence). Unlike co-occurrence properties, distributional similarity is based on co-occurring words rather than component words of the MWE. It has been used to model compositionality, classify semantics and to identify MWEs.

Semantic similarity is the process of modeling the semantics of the whole via the semantics of the parts, notably in comparing corresponding components of a pairing of MWEs and inferring that the MWEs are similar in the instance that the components are similar. This approach is effective for interpreting the semantics of MWEs, and has been applied to the tasks of semantic interpretation and compositionality modeling.

Linguistic properties of MWEs can be used to model MWEs, e.g. based on the output of a parser. They tend to be highly construction-specific, and are high-precision but low-recall. As a result, they tend to be combined with other approaches rather than form standalone methodologies. This approach has been applied to MWE extraction and semantic classification.

Finally, we discuss collocation as a subset MWE and analyze different statistical methodologies used to identify co-occurrence property of certain phrase in a corpus based on the frequency of occurrence of individual as well as entire phrase.

Chapter 4

Resources and Tools used

This chapter describes the resources used in our research, including corpora, lexical resources, dictionaries and software. The resources vary in coverage, usage and language, and not only provide fundamental knowledge to understand context, but are also used in some cases to evaluate the proposed models.

4.1 Corpus

4.1.1 Bengali Corpus

We use the Bengali Corpus[1] developed under the joint collaborration of Bengal Engineering –Science University (BESU)[2], Indian institute of Technelogy-Kharagpur (IIT-KGP)[3] and www.rabindrasangit.org. This site containing Unicode characters of the documents was the first attempt by the Ministry of Information Technology, Government of West Bengal with the

[1] http://www.rabindra-rachanabali.nltr.org

[2] http://www.becs.ac.in

[3] http://www.iitkgp.ac.in

motivation to spread the invention of Bengali writings of the great Indian Noble laureate Rabindranath Tagore.

This site contains a huge number of poems, short and long stories, novels, dramas which are the part of Rabindrarachanaboli, the collected works of Rabindranath Tagore. This site contains hundreds of lyrics of Bengali songs wrriten by Rabindranath Tagore which are well-known as Rabindra Sangit in Bengali. The developers also attempt to include the unpublished articles and letters of Rabindranath Tagore. The corpus can be potentially used to identify the writing style of Rabindranath Tagore using NLP techniques.

This site does not contain any direct link to download the articles of Rabindranath Tagore. Even the site is dymanic in nature. The articles are not possible to be crawled by the crawler. We extracted the articles using copy-paste approach and created separate files for each article. When pasting the documents, the words and letters were scattered and sometime few letters were not supported in the Microsoft word document file. So, using them properly in our exprements was itself a challenging task.

4.1.2 English Corpus

We use English Wacky Corpora[4] which was developed by a community of linguists and information technology specialists. The resources contain four very large corpora, comparable in terms of size, sampling strategy and format (Baroni, Bernardini, Ferraresi and Zanchetta, 2009).

- **deWaC**: a 1.7 billion word corpus constructed from the Web limiting the crawl to the **.de** domain and using medium-frequency words from the SudDeutsche Zeitung corpus and basic German vocabulary lists as seeds. The corpus was POS-tagged and lemmatized with the TreeTagger.
- **frWaC**: a 1.6 billion word corpus constructed from the Web limiting the crawl to the **.fr** domain and using medium-frequency words from the Le Monde Diplomatique corpus and basic French vocabulary lists as seeds. The corpus was POS-tagged and lemmatized with the TreeTagger.

[4] http://wacky.sslmit.unibo.it/

- **itWaC**: a 2 billion word corpus constructed from the Web limiting the crawl to the **.it** domain and using medium-frequency words from the Repubblica corpus and basic Italian vocabulary lists as seeds. The corpus was POS-tagged with the TreeTagger, and lemmatized using the Morph-it! Lexicons.
- **ukWaC**: a 2 billion word corpus constructed from the Web limiting the crawl to the **.uk** domain and using medium-frequency words from the BNC as seeds. The corpus was POS-tagged and lemmatized with the TreeTagger.

4.2 Lexical Resources

4.2.1 WordNet

WordNet (Fellbaum 1998)[5] is a large-scale lexical database of English developed at Princeton University under the direction of George A. Miller. It groups English words (nouns, verbs, adjectives and adverbs) into sets of synonyms called synsets. WordNet provides short, general definitions for each synsets and records various conceptual-semantic and lexical relations between pairings of synsets. Initially, it was developed to produce a combination of dictionary and thesaurus to support automatic text analysis and NLP applications. It contains both simplex words and multiword expressions. As described in Table 4.1, the total of all unique noun, verb, adjective, and adverb lexical items is 155,327, contained in 117,597 unique synsets (based on version 2.1). Many lexical items have a unique synset classification within a given syntactic category, but are described under more than one syntactic category. WordNet has been used in various natural language processing tasks such as lexical semantics (McCarthy *et al.* 2004; Moldovan *et al.* 2004; Nastase *et al.* 2006), PP-attachment (Kim and Baldwin 2006a) and question answering (Prager and Chu-Carroll 2001; Hermjakob *et al.* 2002), and has become a mainstream language resource in NLP. The current version of WordNet is 3.0, although most of our experiments were carried out using WordNet 2.1 as it was the current version at the time. Table 4.1 summarizes the total number of words and multiword expressions contained in WordNet 2.1.

[5] http://wordnet.princeton.edu/

POS	# of lexical entries	# of MWEs
noun	117,097	59,876
verb	11,488	2,777
adjective	22,141	571
adverb	4,601	117

Table 4.1: Composition of WordNet 2.1

4.3 Tools

4.3.1 Bengali Shallow Parser

Bengali Shallow pareser[6] is the first shallow parser of Bengali language developed by Language Technologies Research Center, Indian Institute of Information Technology (IIIT), Hydrabad. It gives the analysis of a Bengali sentence at various levels. The analysis begins at the morphological level and accumulates the results of POS tagger and chunker. The final ouput combines the results of all these levels and shows them in a single representation (called Shakti Standard Format). The details of the tool are given to the documentation part of the downloaded software.

4.3.2 Conditional Random Field (CRF)

Relational data has two characteristics: first, statistical dependencies exist between the entities we wish to model, and second, each entity often has a rich set of features that can aid classification. For example, when classifying Web documents, the page’s text provides much information about the class label, but hyperlinks define a relationship between pages that can improve classification (Taskar et al. 2002). Graphical models are a natural formalism for exploiting the dependence structure among entities. Traditionally, graphical models have been used to represent the joint probability distribution p(y, x), where the variables y represent the attributes of the entities that we wish to predict, and the input variables x represent our observed

[6] http://ltrc.iiit.ac.in/analyzer/bengali

knowledge about the entities. But modeling the joint distribution can lead to difficulties when using the rich local features that can occur in relational data, because it requires modeling the distribution p(x), which can include complex dependencies. Modeling these dependencies among inputs can lead to intractable models, but ignoring them can lead to reduced performance.

A solution to this problem is to directly model the conditional distribution p(y|x), which is sufficient for classification. This is the approach taken by conditional random fields (Lafferty et al. 2001). A conditional random field is simply a conditional distribution p(y|x) with an associated graphical structure. Because the model is conditional, dependencies among the input variables x do not need to be explicitly represented, affording the use of rich, global features of the input. For example, in natural language tasks, useful features include neighboring words and word bigrams, prefixes and suffixes, capitalization, membership in domain-specific lexicons, and semantic information from sources such as WordNet. Recently there has been an explosion of interest in CRFs, with successful applications including text processing (Taskar et al. 2002), bioinformatics (Sato and Sakakibara 2005), and computer vision (Kumar and Hebert 2003).

In this section, we define CRFs with general graphical structure, as they were introduced originally (Lafferty et al. 2001). Although initial applications of CRFs used linear chains, there have been many later applications of CRFs with more general graphical structures. Also, although CRFs have typically been used for across-network classification, in which the training and testing data are assumed to be independent, we will see that CRFs can be used for within-network classification as well, in which we model probabilistic dependencies between the training and testing data. The generalization from linear-chain CRFs to general CRFs is fairly straightforward. We simply move from using a linear-chain factor graph to a more general factor graph, and from forward-backward to more general (perhaps approximate) inference algorithms.

First we present the general definition of a conditional random field. Let G be a factor graph over Y. Then p(y|x) is a conditional random field if for any fixed x, the distribution p(y|x) factorizes according to G. Thus, every conditional distribution p(y|x) is a CRF for some, perhaps trivial, factor graph. If F = { ψ_A} is the set of factors in G, and each factor takes the exponential family form, then the conditional distribution can be written as

$$p(y|x) = \frac{1}{Z(x)} \prod_{\psi_A € G} exp \left\{ \sum_{k=1}^{K(A)} \lambda_{Ak} f_{Ak}(y_A, x_A) \right\} \tag{4.1}$$

In addition, practical models rely extensively on parameter tying. For example, in the linear-chain case, often the same weights are used for the factors $\psi_t(y_t, y_{t-1}, x_t)$ at each time step. To denote this, we partition the factors of G into C = {C_1,C_2, . . .C_P }, where each Cp is a clique template whose parameters are tied. This notion of clique template generalizes that in Taskar et al. (2002). Each clique template Cp is a set of factors which has a corresponding set of sufficient statistics {$f_{pk}(x_p, y_p)$}. Then the CRF can be written as

$$p(y|x) = \frac{1}{Z(x)} \prod_{C_p € C} \prod_{\psi_C € C_p} \psi_C(x_c , y_c; \theta_p) \tag{4.2}$$

where each factor is parameterized as

$$\psi_C(x_c , y_c; \theta_p) = \exp \left\{ \sum_{k=1}^{K(p)} \lambda_{pk} f_{pk} (x_c , y_c) \right\} \tag{4.3}$$

and the normalization function is

$$Z(x) = \sum_{y} \prod_{C_p € C} \prod_{\psi_C € C_p} \psi_C(x_c , y_c; \theta_p) \tag{4.4}$$

For example, in a linear-chain conditional random field, typically one clique template C = {$\psi_t(y_t,y_{t-1}, x_t)$} is used for the entire network. Several special cases of conditional random fields are of particular interest. First, dynamic conditional random fields (Sutton et al. 2004) are sequence models which allow multiple labels at each time step, rather than single labels as in linear-chain CRFs. Second, relational Markov networks (Taskar et al. 2002) are a type of general CRF in which the graphical structure and parameter tying are determined by an SQL-like syntax. Finally, Markov logic networks (Richardson and Domingos 2005; Singla and Domingos 2005) are a type of probabilistic logic in which there are parameters for each first-order rule in a knowledge base.

4.3.3 WordNet::Similarity

We used the open-source WordNet::Similarity package (Patwardhan *et al.* 2003)[7] to compute word similarities. WordNet::Similarity is developed at the University of Minnesota, and provides various methods to measure the similarity or relatedness between a pair of concepts or word senses. It contains implementations of a variety of comparison methods, of three basic types: similarity, relatedness and random.

The similarity methods are categorized into two groups: path-based (LCH (Leacock and Chodorow 1998) and WUP (Wu and Palmer 1994)) and information-content based (RES (Resnik 1995), JCN (Jiang and Conrath 1997), and LIN (Lin 1998c)), as summarized in Figure 4.1. Path-based methods compute lexical similarity based on the shortest path between two target synsets based on the WordNet *is-a* hierarchy. The difference between LCH and WUP is in the calculation of path length. LCH calculates the path length between two target concepts (*c*1 and *c*2) based on Equation 4.5:

$$Similarity_{lch}(C_1, C_2) = -\log\left(\frac{P}{2\times depth}\right) \tag{4.5}$$

where *p* is the number of nodes in the shortest path connecting *c*1 and *c*2, and *depth* is the maximum depth of WordNet hierarchy.

WUP, on the other hand, is based on the path length to the root node from the least common subsumer (LCS) of the two target concepts (*c*1 and *c*2). The LCS is defined as that concept at greatest depth in the WordNet hierarchy that subsumes both *c*1 and *c*2. The calculation of similarity is based on Equation 4.6.

$$similarity_{wup}(C_1, C_2) = \frac{2\times P3}{P1+P2+2\times P3} \tag{4.6}$$

where *p*1 and *p*2 are the number of nodes on the path from *c*1 to *c*2 respectively and *p*3 is the number of nodes on the path between LCS and root.

RES, JCN and LIN augment the calculation of path length with the information content (IC) of the LCS, calculated as follows:

[7] www.d.umn.edu/~tpederse/similarity.html

$$IC(c) = -log\frac{freq(c)}{freq(root)} \tag{4.7}$$

where *freq*(*c*) is the frequency of a given concept *c*, and *freq*(*root*) is the frequency of the root of the hierarchy.

RES calculates the similarity of two concepts by the information of their LCS:

$$Similarity_{res} = IC(lcs(c_1, c_2)) \tag{4.8}$$

JCN is an extension of RES, where the path length between the two concepts is included in the calculation, based on:

$$Similarity_{jcn} = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \tag{4.9}$$

LIN is a further variant of RES, based on the Dice coefficient:

$$similarity_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1,c_2))}{IC(c_1)+IC(c_1)} \tag{4.10}$$

The relatedness measures use additional relations such as *has-part, is-made-of* and *is-an-attribute-of* in addition to the *is-a* relation. There are three relatedness measures: HSO (Hirst and St-Onge 1998), LESK (Banerjee and Pedersen 2003) and VECTOR (Patwardhan 2003). HSO is based on path similarity, and takes into consideration sequences of lexical relations connecting synsets in the WordNet hierarchy that are likely to be indicative of word-level (rather than sense) relatedness. LESK is based on the weighted word overlap of different pairings of synset glosses, over a variety of relation types.

VECTOR is a corpus-based measure. Each word is represented as a multi-dimensional vector of co-occurring words. The similarity of a word pair is measured by the cosine similarity of the two vectors. In Equation 4.11, $\overrightarrow{V1}$ and $\overrightarrow{V2}$ are the vectors of the two target words:

$$Relation_{vector}(c_1, c_2) = \frac{\overrightarrow{v1} \times \overrightarrow{v2}}{\left\|\overrightarrow{v1}\right\| \times \left\|\overrightarrow{v2}\right\|} \tag{4.11}$$

Finally, RANDOM measures similarity by random assignment.

Chapter 5
Automatic Extraction of Keyphrases

This chapter contains the detailed approach of automatic extraction of Keyphrases, which is a word or a set of words that describe the close relationship of *contain* and *context* in the documents using Conditional Random Fields (CRF). Keypharase sometime shows their multi-word characteristics in that they act as special meaning-bearing unit. Named-entities also belong to the class of MWE and they definitely act as keywords in the document. The system is trained using 144 scientific articles and tested on 100 scientific articles. Different combinations of features have been used. With reader and author assigned keyphrases, the system shows a precision of 17.80%, recall of 18.21% and F-measure of 18.00% with top 15 candidates. Automatic keyphrase extraction from document can be used in summarization, where keywords or query words are not available. The extracted keyphrases can be used as keywords to generate the summary.

5.1 Introduction

Keyphrase is a word or set of words that describe the close relationship of *contain* and *context* in the document. Keyphrases are simplex nouns or noun phrases (NPs) that represent the key ideas of the document. Keyphrases can serve as a representative summary of the document and also serve as high quality index terms (Kim and Kan, 2009). Keyphrases can be used in various natural language processing (NLP) applications such as summarization, information retrieval (IR), question answering (QA) etc. Keyphrase extraction also plays an important role in

Search engines. With the advancement of research, many attempts of automatic keyphrase extraction have been made and this attempt is one among them. The need of CRF is discussed in Section 5.2 followed by the system design, experimental results and conclusion in Section 5.3, 5.4 and 5.5 respectively.

5.2 Conditional Random Field (CRF)

Current NLP techniques cannot fully understand general natural language articles. However, they can still be useful on restricted tasks. One example is Information Extraction. For example, one might want to extract the title, authors, year, and conference names from a researcher's Web page. Or one might want to identify person, location, organization names from news articles (NER, named entity recognition). These are useful to automatically turn free text on the Web into knowledge databases, and form the basis of many Web services. The basic Information Extraction technique is to treat the problem as a text sequence tagging problem. The tag sets can be {title, author, year, conference, other}, or {person, location, organization, other}, for instance. Therefore Hidden Markov Model (HMM) has been naturally and successfully applied to Information Extraction. However, HMMs have difficulty modeling overlapping, non-independent features of the output. For example, an HMM might specify which words are likely for a given state (tag) via $p(x|z)$. But often the part-of-speech of the word, as well as that of the surrounding words, character n-grams, capitalization patterns all carry important information. HMMs cannot easily model these, because the generative story limits what can be generated by a state variable.

CRF has been discussed in detail in the Chapter 4.3. Here the implementation details are discussed as much as possible. Conditional Random Field (CRF) can model these overlapping, non-independent features. A special case, linear chain CRF, can be thought of as the undirected graphical model version of HMM. It is as efficient as HMMs, where the sum-product algorithm and max-product algorithm still apply.

5.2.1 The CRF Model

Let $x_{1:N}$ be the observations (e.g., words in a document), and $z_{1:N}$ the hidden labels (e.g., tags). A linear chain Conditional Random Field defines a conditional probability (whereas HMM defines the joint probability)

$$p(z_1:N|x_1:N) = \frac{1}{Z}\exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i f_i\ (z_{n-1}, z_n, x_1: N, n)\right) \tag{5.1}$$

Let us walk through the model in detail. The scalar Z is the normalization factor, or partition function, to make it a valid probability. Z is defined as the sum of exponential number of sequences,

$$Z = \sum_{z_1:N} exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F}\lambda_i f_i\,(z_{n-1}, z_n, x_1: N, n)\right) \tag{5.2}$$

Therefore is difficult to compute in general. It may be noted that Z implicitly depends on $x_{1:N}$ and the parameters λ. The big exponential function is there for historical reasons, with connection to the exponential family distribution. For now, it is sufficient to note that λ and *f()* can take arbitrary real values, and the whole exp function will be non-negative. Within the exp() function, we sum over $n = 1, \ldots, N$ word positions in the sequence. For each position, we sum over $i = 1, \ldots, F$ weighted features. The scalar λ_i is the weight for feature $f_i()$. The λ_i's are the parameters of the CRF model, and must be learned, similar to $\theta = \{\pi, \phi, A\}$ in HMMs.

5.2.2 Feature Functions

The feature functions are the key components of CRF. In our special case of linear-chain CRF, the general form of a feature function is $f_i(z_{n-1}, z_n, x_{1:N}, n)$, which looks at a pair of adjacent states z_{n-1}, z_n, the whole input sequence $x_1{:}N$, and where we are in the sequence (n). These are arbitrary functions that produce a real value.

For example, we can define a simple feature function which produces binary values: it is 1 if the current word is *John*, and if the current state z_n is PERSON:

$$f_1(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1\ if\ z_n{=}PERSON\ and\ x_n{=}\ John \\ o\ \ otherwise \end{cases} \tag{5.3}$$

How is this feature used? It depends on its corresponding weight λ_1. If $\lambda_1 > 0$, whenever f_1 is active (i.e. we see the word John in the sentence and we assign it tag PERSON), it increases the probability of the tag sequence $z_1{:}N$. This is another way of saying the CRF model should prefer the tag PERSON for the word John. If on the other hand $\lambda_1 < 0$, the CRF model will try to avoid the tag PERSON for John. Which way is correct? One may set λ_1 by domain knowledge (we know it should probably be positive), or learn λ_1 from corpus (let the data tell us), or both (treating domain knowledge as prior on λ_1). It may be noted that λ_1, $f_1()$ together is equivalent to (the log of) HMM's ϕ parameter $p(x$ = John|z = PERSON).

As another example, consider

$$f_2(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1\ if\ z_n{=}PERSON\ and\ x_{n+1=}\ said \\ o\ \ otherwise \end{cases} \tag{5.4}$$

This feature is active if the current tag is PERSON and the next word is 'said'. One would therefore expect a positive λ_2 to go with the feature. Furthermore, functions f_1 and f_2 can be both active for a sentence like "John said so." where z_1 = PERSON. This is an example of overlapping features. It boosts up the belief of z_1 = PERSON to λ_1+ λ_2. This is something HMMs cannot do: HMMs cannot look at the next word, nor can they use overlapping features. The next feature example is rather like the transition matrix A in HMMs. We can define

$$f_3(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1\ if\ z_{n-1}{=}OTHER\ and\ z_{n=}\ said \\ o\ \ otherwise \end{cases} \tag{5.5}$$

This feature is active if we see the particular tag transition (OTHER, PERSON). Note it is the value of λ_3 that actually specifies the equivalent of (log) transition probability from OTHER to PERSON, or OTHER, PERSON in HMM notation. In a similar fashion, we can define all K_2 transition features, where K is the size of tag set. Of course the features are not limited to binary functions. Any real-valued function is allowed.

5.2.3 Undirected Graphical Models (Markov Random Fields)

CRF is a special case of undirected graphical models, also known as Markov Random Fields. A *clique* is a subset of nodes in the graph that are fully connected (having an edge between any two nodes). A maximum clique is a clique that is not a subset of any other clique. Let *Xc* be the set of nodes involved in a maximum clique c. Let $\psi(Xc)$ be an arbitrary non-negative real-valued function, called the *potential function*. In particular $\psi(Xc)$ does not need to be normalized. The Markov Random Field defines a probability distribution over the node states as the normalized product of potential functions of all maximum cliques in the graph:

$$p(X) = \frac{1}{Z}\prod_{c}\psi(X_c) \tag{5.6}$$

where Z is the normalization factor. In the special case of linear-chain CRFs, the cliques correspond to a pair of states z_{n-1}, z_n as well as the corresponding x nodes, with

$$\psi = exp\ (\lambda_f) \tag{5.7}$$

This is indeed the direct connection to factor graph representation as well. Each clique can be represented by a factor node with the factor $\psi(X_c)$, and the factor node connects to every node in X_c. There is one addition special factor node which represents Z. A welcome consequence is that the sum-product algorithm and max-sum algorithm immediately apply to Markov Random Fields (and CRFs in particular). The factor corresponding to Z can be ignored during message passing.

5.2.4 CRF Training

Training involves finding the λ parameters. For this we need fully labeled data sequences $\{(x^{(1),}z^{(1)}), \ldots, (x^{(m)}, z^{(m)})\}$, where the first observation sequence, and so on[1]. Since CRFs define the conditional probability p(z|x), the appropriate objective for parameter learning is to maximize the conditional likelihood of the training data

$$\sum_{j=1}^{m} \log p(z^{(j)}|x^{(j)}) \tag{5.8}$$

Often one can also put a Gaussian prior on the λ's to regularize the training (i.e., smoothing). If λ ~N (0, σ2), the objective becomes

$$\sum_{j=1}^{m} \log p\big(z^{(j)}\big|\, x^{(j)}\big) - \sum_{i}^{F} \frac{\lambda_i^2}{2\sigma^2} \tag{5.9}$$

The good news is that the objective is concave, so the λ's have a unique set of optimal values. The bad news is that there is no closed form solution[2].

The standard parameter learning approach is to compute the gradient of the objective function, and use the gradient in an optimization algorithm like L-BFGS. The gradient of the objective function is computed as follows:

[1] Unlike HMMs which can use the Baum-Welch (EM) algorithm to train on unlabeled data x only, CRFs training on unlabeled data is difficult

[2] If this reminds you of logistic regression, you are right: logistic regression is a special case of CRF where there are no edges among hidden states. In contrast, HMMs when trained on fully labeled data have simple and intuitive closed form solutions.

$$\frac{d}{d\lambda_k}\sum_{j=1}^{m}\log p\left(z^{(j)}\middle|x^{(j)}\right)$$

$$-\sum_{i}^{F}\frac{\lambda_i^2}{2\sigma^2}= \frac{d}{d\lambda_k}\sum_{j=1}^{m}\left(\sum_{n}\sum_{i}\lambda_i f_i\left(z_{n-1}^{(j)}, z_n^{(j)}, x^{(j)}, n\right)-\log Z^{(j)}\right)-\sum_{i}^{F}\frac{\lambda_i^2}{2\sigma^2} \tag{5.10}$$

$$= \sum_{j=1}^{m}\sum_{n} f_k\left(z_{n-1}^{(j)}, z_n^{(j)}, x^{(j)}, n\right) - \sum_{j=1}^{m}\sum_{n} E_{z'_{n-1},z'_n}\left[f_k\left(z'_{n-1}, z'_n, x^{(j)}, n\right)\right]-\frac{\lambda_i^2}{\sigma^2} \tag{5.11}$$

where we used the fact

$$\frac{d}{d\lambda_k}\log Z = E_{z'}\left[\sum_{n} f_k(z'_{n-1}, z'_n, x, n)\right]=\sum_{n} E_{z'_{n-1},z'_n}[f_k(z'_{n-1}, z'_n, x, n)] \tag{5.12}$$

$$=\sum_{n}\sum_{z'_{n-1},z'_n} p(z'_{n-1}, z'_n|x)\, f_k(z'_{n-1}, z'_n, x, n) \tag{5.13}$$

Note the edge marginal probability $p(z'_{n-1}, z'_n \mid x)$ is under the current parameters, and this is exactly what the sum-product algorithm can compute.

The partial derivative in (12) has an intuitive explanation. Let us ignore the term λ_k/σ^2 from the prior. The derivative has the form of (observed counts of feature f_k) minus (expected counts of feature f_k). When the two are the same, the derivative is zero, and there is no longer an incentive to change λ_k. Therefore we see that training can be thought of as finding λ's that match the two counts.

5.2.5 Feature Selection

A common practice in NLP is to define a very large number of candidate features, and let the data select a small subset to be used in the final CRF model in a process known as feature selection. Often the candidate features are proposed in two stages:

1. Atomic candidate features. These are usually a simple test on a specific combination of words and tags, e.g.(*x* =John, *z* =PERSON), (*x* =John, *z* =LOCATION), (*x* =John, *z* = ORGANIZATION), etc. There are V_K such "word identity" candidate features, which is obviously a large number. Although it is called the word identity test, it should be

understood as *in combination with each tag value*. Similarly one can test whether the word is capitalized, the identity of the neighboring words, the part of speech of the word, and so on. The state transition features are also atomic. From the large number of atomic candidate features, a small number of features are selected by how much they improve the CRF model (e.g., increase in the training set likelihood).

2. "Grow" candidate features. It is natural to combine features to form more complex features. For example, one can test for current word being capitalized, the next word being "Inc.", and both tags being ORGANIZATION. However, the number of complex features grows exponentially. A compromise is to only grow candidate features on selected features so far, by extending them with one atomic addition, or other simple Boolean operations.

Often any remaining atomic candidate features are added to the grown set. A small number of features are selected, and added to the existing feature set. This stage is repeated until enough features have been added.

5.3 Preparing the System

5.3.1 Features Identification for the System

Selection of features is important in CRF. Features used in the system are,

F = { ***Dependency, POS tag(s), Chunk , TF range, Title, Abstract, Body, Reference, Stem of word, Wi-m, …, Wi-1, Wi, Wi+1,… , Wi-n*** **}**

(5.14)

The features are detailed as follows:

i) **Dependency parsing:** Some of the keyphrases are multiword. So relationship of verb with subject or object is to be identified through dependency parsing and thus used as a feature.

ii) **POS feature:** The Part of Speech (POS) tags of the preceding word, the current word and the following word are used as a feature in order to know the POS combination of a keyphrase.

iii) **Chunking:** Chunking is done to mark the Noun phrases and the Verb phrases since much of the keyphrases are noun phrases.

iv)**Term frequency (TF) range:** The maximum value of the term frequency (max_TF) is divided into five equal sizes (*size_of_range)* and each of the term frequency values is mapped to the appropriate range (0 to 4). The term frequency range value is used as a feature. i.e.

$$size\_of\_range = \frac{\max\_TF}{5}$$

Thus Table 5.1 shows the range representation:

Class	Range
0 to size_of_range	0
*size_of_range + 1 to2*size_of_range*	1
*2*size_of_range + 1 to 3*size_of_range*	2
*3*size_of_range + 1 to 4*size_of_range*	3
*4*size_of_range + 1 to 5*size_of_range*	4

Table 5.1: Term frequency (TF) range

This is done to have uniform values for the term frequency feature instead of random and scattered values.

v) **Word in Title:** Every word is marked with T if found in the title else O to mark other. The title word feature is useful because the words in title have a high chance to be a keyphrase.

vi) **Word in Abstract:** Every word is marked with A (Abstract) if found in the abstracts else O to mark other. The abstract word feature is useful because the words in abstracts have a high chance to be a keyphrase.

vii) **Word in Body:** Every word is marked with B (*Body)* if found in the body of the text else O to mark other. It is a useful feature because words present in the body of the text are distinguished from other words in the document.

viii) **Word in Reference:** Every word is marked with R if found in the references else O to mark other. The reference word feature is useful because the words in references have a high chance to be a keyphrase.

ix) **Stemming:** The Porter Stemmer algorithm is used to stem every word and the output stem for each word is used as a feature. This is because words in keyphrases can appear in different inflected forms.

x) **Context word feature:** The preceding and the following word of the current word are considered as context feature since keyphrases can be a group of words.

5.3.2 Corpus Preparation

Automatic identification of keyphrases is our main task. In order to perform this task the data provided by the SEMEVAL-2[3] Task Id #5 is being used both for training and testing. In total 144 scientific articles or papers are provided for training and another 100 documents have been marked for testing. All the files are cleaned by placing spaces before and after every punctuation mark and removing the citations in the text. The author names appearing after the paper title was removed. In the reference section, only the paper or book title was kept and all other details were deleted.

5.4 CRF based Keyphrase Extraction System

5.4.1 Extraction of Positional Feature

One algorithm has been defined to extract the title from a document. Another algorithm has been defined to extract the positional feature of a word, i.e., whether the word is present in title, abstracts, body or in references.

- **Algorithm 1:** Algorithm to extract the title.

Step 1: Read the line one by one from the beginning of the article until a '.'(dot) or '@' found in the line. ('.'(dot) occurs in author's name and '@' occurs in author's mail id).

Step 2**:** If '.' found first in a line then each line before it is extracted as Title and returned.

Step 3: If '@' found first in a line then extract all the line before it.

Step 4: Check the extracted line one by one from beginning.

Step 5**:** Take a line, extract all the words of that line. Check whether all the words are not repeated in the article (excluding the references) or not. If not then stop and extract all the previous lines as Title and return.

- **Algorithm 2:** Algorithm to extract the Positional Features.

Step 1**:** Take each word from the article.

Step 2**:** Stem all the words.

[3] http:// semeval2.fbk.eu/ semeval2.php? loction=data

Step 3**:** Check the position of the occurrence of the words.

Step 4**:** If the word occurs in the extracted title (using algorithm 1) of the article then mark it as 'T' else ‘O’ in title feature column.

Step 5**:** If the word occurs in between the word ABSTRACT and INTRODUCTION then mark it as 'A' else ‘O’ in abstracts feature column.

Step 6**:** If the word occurs in between the word INTRODUCTION and REFERENCES then mark it as 'B' else ‘O’ in body feature column.

Step 7**:** If the word occurs after the word REFERENCES then mark it as 'R' else ‘O’ in references feature column.

5.4.2 Generating Feature File for CRF

The features used in the keyphrase extraction system are identified in the following ways.

Step 1: The dependency parsing is done by the Stanford Parser[4]. The output of the parser is modified by making the word and the associated tags for every word appearing in a line.

Step 2: The same output is used for chunking and for every word it identifies whether the word is a part of a noun phrase or a verb phrase.

Step 3: The Stanford POS Tagger[5] is used for POS tagging of the documents.

Step 4**:** The term frequency (*TF*) range is identified as defined before.

Step 5: Using the algorithms described in Section 5.3.1 every word is marked as *T* or *O* for the title word feature, marked as *A* or *O* for the abstract word feature, marked as *B* or *O* for the body word feature and marked as *R* or *O* for the reference word feature.

Step 6**:** The Porter Stemming Algorithm[6] is used to identify the stem of every word that is used as another feature.

Step 7: In the training data with the combined keyphrases, the words that begin a keyphrase are marked with *B-KP* and words that are present intermediate in a keyphrase are marked as *I-KP*. All other words are marked as *O*. But for test data only *O* is marked in this column.

[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] http://nlp.stanford.edu/software/tagger.shtml

[6] http://tartarus.org/~martin/PorterStemmer/

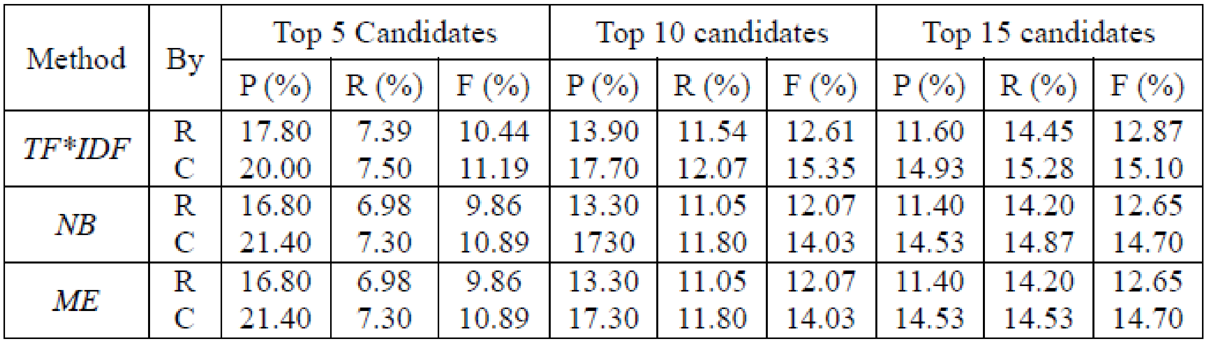
5.4.3 Training the CRF and Running Test Files

A template file is created in order to train the system using the feature file generated from the tanning set following the above procedure described in the previous section. After training the C++ based CRF++ 0.53 package[7] which is readily available as open source for segmenting or labeling sequential data, a model file is produced. The model file is required to run the test files.

The feature file is again created from the test set using the above steps as outlined in Section 5.3.2 except the step 7. For test set the last feature column i.e. Keyphrase column, is marked with ‘O’. This feature file is used with the C++ based CRF++ 0.53 package. After running the Test files into the system, the system produce the output file with the keyphrases marked with *B-KP* and *I-KP*. All the Keyphrases are extracted from the output file and stemmed using Porter Stemmer.

5.5 Evaluation and Error Analysis

The results of the baseline model provided by the task organizers are shown in Table 5.2.



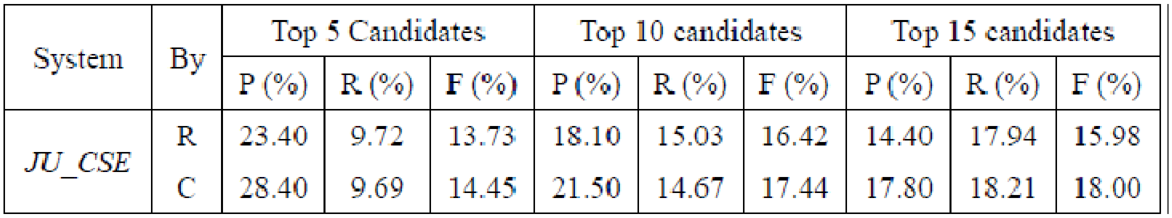
Method	By	Top 5 Candidates			Top 10 candidates			Top 15 candidates		
		P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
*TF*IDF*	R C	17.80 20.00	7.39 7.50	10.44 11.19	13.90 17.70	11.54 12.07	12.61 15.35	11.60 14.93	14.45 15.28	12.87 15.10
NB	R C	16.80 21.40	6.98 7.30	9.86 10.89	13.30 1730	11.05 11.80	12.07 14.03	11.40 14.53	14.20 14.87	12.65 14.70
ME	R C	16.80 21.40	6.98 7.30	9.86 10.89	13.30 17.30	11.05 11.80	12.07 14.03	11.40 14.53	14.20 14.53	12.65 14.70

Table 5.2: The Baselines provide by the organizer

In all tables, *P, R* and *F* mean micro-averaged precision, recall and F-scores. For baselines, they have used 1, 2 or 3 grams as candidates and TF·IDF as features. In Table 5.2, TF·IDF is an unsupervised method to rank the candidates based on TF·IDF scores. *NB* and *ME* are supervised methods using Naïve Bayes and maximum entropy in WEKA. In second column, *R* denotes the use of the reader-assigned keyword set as gold-standard data and *C* denotes the use of combined keywords i.e. both author-assigned and reader-assigned keyword sets as answers. There are three sets of score. First set of score i.e. Top 5 candidates, is obtained by evaluating only top 5

[7] http://crfpp.sourceforge.net/

keyphrases from evaluated data. Similarly Top 10 candidates set is obtained by evaluating top 10 keyphrases and Top 15 Candidates set result is obtained by evaluating all 15 keyphrases. The evaluation results of the CRF based keyphrase extraction system are shown in Table 5.3.



System	By	Top 5 Candidates			Top 10 candidates			Top 15 candidates		
		P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
JU_CSE	R	23.40	9.72	13.73	18.10	15.03	16.42	14.40	17.94	15.98
	C	28.40	9.69	14.45	21.50	14.67	17.44	17.80	18.21	18.00

Table 5.3: Result for JU_CSE system with CRF

The scores for the top 5 candidates and top 10 candidates of keyphrases extracted show a better precision score since the keyphrases are generally concentrated in the title and abstracts. The recall shows a contrast improvement from 9.69% to 18.21% as the number of candidate increases since the coverage of the text increases.

The F-score is 18.00% when top 15 candidates are considered which is 2.90% better from the best baseline model. Different features have been tried and the best feature we have used in the system is:

F = {*Dependency*, *POS*$_{i-1}$, *POS*$_{i}$, *POS*$_{i+1}$, *chunking, TF range, Title, Abstract, Body, Reference, Stem of word*, *W*$_{i-1}$, *W*$_{i}$, *W*$_{i+1}$} (5.15)

Here, ***POS*$_{i-1}$, *POS*$_{i}$** and ***POS*$_{i+1}$** are the POS tags of the previous word, the current word and the following word respectively. Similarly ***W*$_{i-1}$, *W*$_{i}$** and ***W*$_{i+1}$** denote the previous word, the current word and the following word respectively. This ***POS*$_{i}$** and ***W*$_{i}$** give a contrasting result when only the word and the POS of the word is considered.

A better result could have been obtained if the multiplication of Term Frequency and Inverse Document Frequency (***TF*IDF***) range is included [50, 51]. *TF*IDF* measures the document cohesion. The maximum value of the *TF*IDF* (max*_TF_IDF)* can be divided into five equal size (*size_of_range)* and each of the *TF*IDF* values is mapped to the appropriate range (0 to 4) i.e.

$$size\ of\ range\ = \frac{\mathrm{max\_TF\_IDF}}{5}$$

We have used the Unigram template in the template file CRF++ 0.53 package but the use of bigram could have improved the score.

Unigram template:

The number of feature functions generated by a template amounts to (L * N), where L is the number of output classes and N is the number of unique string expanded from the given template.

Bigram template:

This is a template to describe bigram features. With this template, a combination of the current output token and previous output token (bigram) is automatically generated. Note that this type of template generates a total of (L * L * N) distinct features, where L is the number of output classes and N is the number of unique features generated by the templates.

5.6 Chapter Summary and Future work

Conditional Random Field (CRF) model is a state-of-the-art sequence labeling method, which can use features of documents more sufficiently and effectively. At the same time, keywords extraction can be considered as the string labeling. In this chapter, we have proposed as implemented CRF-based keyphrase extraction technique. Experimental results show that the CRF model is a promising method in labeling the sequence, and it can take full advantages of all the features of the document. As future work, we plan to make further improvement on the precision and recall of CRF-based keyphrase extraction model. For example, we will use the semantic relations between the words. We also plan to apply the keyword extraction approach on the Web pages, E-mail and other non-academic documents. Meanwhile we will apply this method on some standard documents. It will be interesting to apply the CRF-based model to a large number of text mining applications such as text classification, clustering, summarization, filtering and so on.

Chapter 6

Identification of Reduplications in Bengali

6.1 Introduction

In linguistic studies, the term reduplication is generally used to mean repetition of any linguistic unit such as a phoneme, morpheme, word, phrase, clause or the utterance as a whole. The study of reduplication at all these levels is very significant both from the grammatical as well as the semantic point of view. The identification of reduplication is a part of general task of identification of multiword expressions (MWE). In the present work, reduplications have been identified from the Bengali corpus of the articles of Rabindranath Tagore. The rule-based approach is divided into two phases. In the first phase, identification of reduplications has been done mainly at general expression level and in second phase, their structural and semantics classifications are analyzed. This system has been evaluated with average Precision, Recall and F-Score values of 92.82%, 91.50% and 92.15% respectively.

6.2 Reduplications in Bengali

In all languages, the repetition of noun, pronoun, adjective, and verb are broadly classified under two coarse-grained categories: repetition at the (a) ***expression level***, and repetition at the (b) ***contents or semantic level.*** This paper deals with the identification of reduplications at both levels in Bengali. Reduplication is a common feature of Bengali. Bengali is the richest Indian language with 2400 words (Chaudhuri *et al* 2005) in the onomatopoeic and idiophonic category of reduplication. In Telugu and Marathi, this number is less than 500. In Hindi, the number is large, but they do not take as many suffix-like extensions as Bengali (Apte 1968; Bhaskara Rao 1977). The repetition at both the levels is mainly used for emphasis, generality, intensity, or to show continuation of an act. In certain cases, the repetition of a particular linguistic unit is obligatory. For example,

রাস্তায় **চলতে চলতে** লোকটি হঠাৎ থেমে গেল। (*Rastay Chalte Chalte Lokti Hatat Theme Gelo.*)

Walking on street, the man suddenly stopped.

This reduplication is used to express the infiniteness of the auxiliary verb and cannot bear any meaning in single use. Reduplication carries various semantic meanings and sometime helps to identify the mental state of the speaker as well. For example,

বেশী **তেল-টেল** ব্যবহার করছিস কেন ? (*beshi tel-Tel byabohAr korchhis keno?*)

Why are you using oil too much?

মোটা-সোটা লোকেদের উত্তেজিত কোরো না । (*Mota-Sota Lokeder Uttajita Koro Na.*)

Don't make fatty men excited.

In the first example, the reduplication is used in unpleasant or undesirable sense but the second example expresses the softness or gentleness of the speaker. In linguistic analysis, it is seen that sometime only vowel in the root word is changed to form the reduplication (e.g. **চুপ-চাপ,** *chup-chap, silence*), sometime only consonant in the first position is changed (e.g. **রকম-সকম,** *rakom-sakam, various*) or the consonant in the first position and matra (modified vowel) attached with that consonant are changed (e.g **কাপর-চোপর,** *kapar-chopar, clothes*) leaving other letters unchanged. Some correlative words are used in Bengali to express the possessiveness, relative or descriptiveness. They are called '***secondary descriptive compounds***'. For example,

ছেলেরা **মারামারি** করছে । (*Chelera Maramari Korche.*)

The boys are fighting.

This example shows that before reduplication, a matra **('-আ')** at the beginning and a matra (**'-ই'**) at the end are attached with the root verb (**'মার্'**) (*mar, to fight*) to make a correspondence with the main verb **'করা'** (*kara*, *to do*). This kind of partial reduplication forms a verb compound with the second light verb and is aligned with a single verb word in English.

The study of reduplication is a general subtask of multiword expression identification. Multiword Expressions (MWE) are those whose structure and meaning cannot be derived from their component words, as they occur independently. A typical natural language system assumes each word to be a lexical unit, but this assumption does not hold in case of MWEs (Fillmore 2003; Becker 1995). They have idiosyncratic interpretations that cross word boundaries and hence are a '*pain in the neck for NLP'* (Sag et. al 2002). Reduplicated words are usually collocated MWEs which are fixed expressions and cannot be written separately. Some of the proverbs and quotations can also be considered as fixed expressions.

6.3 Related Work on Reduplication Identification

The works on MWE identification and extraction have been continuing in English (Fillmore 2003; Sag et al. 2002) because after tokenization, multiword expressions are important in understanding the meaning in various application like machine translation, Information Retrieval system etc. Some of the MWE extraction tasks in English can be seen in (Diab and Bhutada 2009; Enivre and Nilson 2004; Koster 2004; Odijik 2004). Among Indian languages, Hindi compound noun MWE extraction has been studied in (Kunchukuttan and Damani 2008). Manipuri reduplicated MWE identification is discussed in (Nongmeikapam and Bandyopadhyay 2010). But there are no published works on identifications of reduplicated MWEs in Bengali.

6.4 Formation of Reduplicated Words

Reduplications of words in Bengali are formed in these three ways:

- **Repetition of same word:** Same word is repeated twice to express repetitiveness (**বাড়ী-বাড়ী** যাওয়া, *bari-bari jaoa, moving door to door*), incompleteness (**শীত-শীত** ভাব, *sheet-sheet bhab*, *feeling cold*), hesitation (**মানে-মানে** করা, *mane-mane kara*, *uttering for* *meaning*), softness (**হাসি-হাসি** মুখ, *hasi-hasi mukh*, *smiling face*), similarity (**লাল-লাল** ফুল, *lal-lal phul, deep red rose*) or onomatopoeic expression (**খট্ খট্** করছে, *khat-khat kora, knocking*).

- **Synonym or antonym word of first word***:* The second word is generally the synonym (**লোক-জন** নাই, *lok-jon nai*, *no sign of life*) or antonym (**পাপ-পুন্য** বিচার, *pap-punya bichar, vice and virtue*) of the first word to express completeness of the meaning or sometime used idiomatically.
- **Imitation or partial copying of first word:** A word followed by its partially changed form is used to express the similarity (**বই-টই,** *boI-toI, books*), unpleasant (**লুচি-মুচি,** *luchi-muchi*), roughness (**তাস-ফাস,** *tas-phas, cards*), softness (**নরম-সরম,** *naram-saram, soften*).

6.5 Reduplication Identification

Identification of MWEs is done during the tokenization phase and is absolutely necessary during POS tagging as is outlined in (Thoudam and Bandyopadhyay 2008) because repeated words sometimes do not have a dictionary entry. So POS tagger identifies it as unknown word at token level. Bengali Shallow Parser (described in Section 4.3.1) can only identify hyphened reduplication and gives them separate tags like RDP (reduplication) or ECH (echo).

Another objective for identifying reduplicated MWEs is to extract correct sense of reduplicated MWEs. For example, when first the consonant is changed to 'ট' (e.g. আর **লুচি-টুচি** লাগবে? *Ar Luchi-Tuchi Lagbe, Can you have more luchis?*), it expresses the softness of the speaker. But if the same word is changed into 'ম' (বেশী **লুচি-মুচি** খেও না ।, *Beshi Luchi-Tuchi kheo na, Do not eat many luchis.*), it expresses speaker's disregardness or hardness.

Sometime, reduplication is used for sentiment marking to identify whether the speaker uses it in positive or negative sense. For example,

(i) এত **বড় বড়** আশা কিসের ? *(Eto Bara Bara Asha Kisher? Why are you thinking so high?*)

(ii) কি **বড় বড়** বাড়ী এখানে *?(Ki Bara Bara Bari Ekhane? Here, the buildings are very large.)*

The first example expresses negative senses of the speaker, but second one shows positive sense for the same reduplication (**'বড় বড়',** *bara-bara, big-big*)**.**

6.6 General Classification of Reduplication

Four classes of reduplications commonly occur in the Indian language (Bengali, Hindi, Tamil, Manipuri etc.) (Keane 2001). In Bengali, another type called ***correlated word*** is also classified as reduplication.

- **Onomatopoeic expressions:** In certain cases the sound sequence of the word denotes the particular meaning of the form. Such forms of lexical items are known as onomatopoeic words. The onomatopoeic words represent an imitation of a particular sound or imitation of an action along with the sound, etc. For example, **খট্ খট্** *(khat khat, knock knock).*
- **Complete Reduplication:** The individual words carry certain meaning, and they are repeated. e.g. **বড়- বড়** *(bara-bara, big big)*, **ধীরে-ধীরে** *(dheere dheere, slowly).* In some cases, both the speaker and the listener repeat certain clauses or phrases in long utterances or narrations. The repetition of such utterances breaks the monotony of the narration, allows a pause for the listener to comprehend the situation, and also provides an opportunity to the speaker to change the style of narration. For example:

  তারপর! তারপর কি হল। (*Tarpar! Tarpar Ki Hala .)*

  What happened after that?
- **Partial Reduplication:** Only one of the words is meaningful, while the second word is constructed by partially reduplicating the first word. There are various ways of constructing such reduplications, but the most common type in Bengali is one where the first letter or the associated matra or both is changed, e.g. **ঠাকুর-ঠুকুর** *(thakur-thukur, God)*, **বোকা-সোকা** (boka-*soka, Foolish*), **সেজে-গুজে** (*seje-guje, dressed up*) etc.
- **Semantic Reduplication:** The paired members are semantically related. The most common forms of relation between the words are *synonymy* (**মাথা-মুন্ডু**, *matha-mundu, head*), *antonym* (**দিন-রাত**, *din-rat*, *day and night*), *class representative* (**চা-পানি**, *cha-paani, snacks*)).

- **Correlative Reduplication:** To express a sense of exchange or barter or interchange, the style of corresponding correlative words is used just preceding the main verb. Before reduplication, the formative affixes '-আ' is added with the root to form the first word and '-ই' is added for the second and then both are agglutinated to make a single word. For example, '*মারামারি*' (maramari, *fighting*). Here, the above specified affixes are added with the root '*মার*' (*mar, to fight*) to form a single token.

Sometime partial reduplication is termed as *echo-word*. At the semantic level, echo-words give an additional meaning indicating 'generally' or the meaning of similar indicating, action, manner and quality etc., which is indicated by the original word stem. Therefore, we may add 'and the like' in the gloss of the echo-words. For example: 'জল' *(jal, water*), '*জল-টল*' (*jal-tal, water and the like*).

6.6.1 Reduplication at the Expression Level

The various types of reduplications at the expression level are defined as given below:

1. Non-sound Symbolic Word

a. Nouns and pronouns: A number of nouns and pronouns are repeated in utterances very frequently. For example:

(i) রাম সারাদিন **বাড়ি বাড়ি** ঘুরছে। (*Ram Saradin Bari Bari Ghurchhe.)*

Ram is moving door to door whole day.

(ii) সবসমই **আমি আমি** কর কেন? *(Sabsamai Aami Aami Karo Keno?)*

Why do you utter about yourself every time?

b. Adjectives: Reduplication of adjectives is used for emphasis or multiple senses.

(i) বাগানে প্রচুর **লাল লাল** ফুল আছে। *(Bagane Prachur Lal Lal Phul Aache.)*

There are lots of red flowers in the garden.

c. Verbs: The repetition of verbs may be obligatory or optional.

(i) কথা **বলতে বলতে** সে চুপ করে গেল। (*Katha Bolte Bolte Se Chup Kore gelo.)*

Talking about something, suddenly he stopped.

(ii) **ভেবে-চেন্তে** সব কাজ কর। (*Bhebe-chinte Sab Kaj Karo.*)

Always think before doing something.

In the first example, the use of reduplication is obligatory and in the second example the synonymous reduplication is optional.

d. Adverb: The repetition of some adverbs is compulsory or optional depending upon the situation:

(i) রাম **ধীরে ধীরে** হাঁটছে। (Optional) (*Ram Dhire Dhire Hantchhe .)*

Ram is walking slowly.

(ii) সে **মাঝে মাঝে** বাড়ি আসে। (Obligatory) (*Se Majhe Majhe Bari Aase .)*

He comes to the house quite often.

2. Sound words

Mainly sound words are onomatopoeic expressions. The constituent words imitate a sound, and the unit as a whole refers to that sound. For example:

(i) **ছল্ ছল্** করে জল পরছে। (*Chal Chal Kore Jal Porche, sound of water falling on a surface*)

Sometime, besides sound expression, reduplication can also be used to express feelings. For example:

(ii) ব্যথায় **টন্ টন্** করছে। (*Byathay Tan Tan Korche, feeling very pain)*

6.6.2 Reduplication at the Sense Level

Reduplication at the sense level is an important feature of Bengali as well as some other Indo-Aryan languages like Tamil, Manipuri (Becker 1975), Hindi etc. Different types of reduplication at sense level and their corresponding expression level classifications are described as follows:

(i) **Sense of repetition:** It expresses a sense of repetitiveness. Complete reduplications are mainly grouped into this class. For example:

বছর বছর এক কাজ কর। (*Bachar Bachar Ek Kaj Kara .)*

Do the same job every year.

(ii) **Sense of plurality:** Completely reduplicated word sometime expresses the plurality of the noun associated with it. They are mainly adjective and used before noun as a modifier. For example:

কি **বড় বড়** বাড়ী এখানে! (*Ki Bara Bara Bari Ekhane.)*

Here, the houses are very large.

(iii) **Sense of Emphatic or Modifying Meaning:** The use of complete or mimic reduplicated MWEs indicates the ‘degree’ or ‘very’ that carries emphatic or modifying meanings. For example:

লাল-লাল ফুল (*Lala-lala phul*, *Deep red rose*)

The sentence without the reduplication means only ‘*red rose’*, but after reduplication, emphasise on the meaning ‘*red*’ becomes deeper.

(iv) **Sense of completion:** Mainly partial or semantic reduplication belongs to this class. For example:

খেয়ে দেয়ে আমি শুতে যাব । (*Kheye Deye Ami Shute Jaba .)*

After eating, I shall go to sleep.

(v) **Sense of hesitation, incompleteness or softness:** Mainly noun, adjective, adverbial complete reduplications are included in this class. For example:

এত **হাসি হাসি** মুখ কেন ? (*Eta Hasi Hasi Mukh Kena ?)*

Why does your face smiling?

Sometime, sense of interest or intension is expressed when this complete reduplication is used just before *verbal root ‘করা’ (kara, to do),’ ভাবা’ (bhaba, to think) or other words like ’মতো’ (mato, like), ‘লাগা’ (laga, feel)* etc. For example:

দাদা দাদা করে পাগল। (*Dada Dada Kore Pagal.)*

Crazy about his (her) brother.

(vi) **Sense of incompleteness of the verbs:** Completely reduplicated infinite verbs are placed in this class. Mainly *‘ইয়া’ (‘-এ’)* or *‘ইতে’ (‘-তে’)* inflection is added with the verbal-adjective word to make this duplication.

কথা **বলতে বলতে** হটাৎ সে চুপ করে গেল। (*Katha Bolte Bolte Hatat Se Chup Kore Gelo.)*

Talking about something, suddenly he stopped.

(vii) **Sense of corresponding correlative words:** To express a sense of exchange or barter or interchange, the style of corresponding correlative words is used just preceding the main verb.

নিজেরা **মারামারি** কর না । (*Nijera Maramari Kara Na.)*

Don't fight among yourselves.

(viii) **Sense of Onomatopoeia:** This class includes mainly onomatopoeic expressions.

শ্যামল দরজা **খট খট** করছে। (*Shyamal Darja Khata khata Karchhe .)*

Shyamal is knocking at the door.

6.7 System Design

The system is designed in two phases. The first phase identifies mainly five cases of reduplication discussed in Section 6.6.1 and the second phase attempts to extract the associated meaning or semantics discussed in Section 6.6.2. This system uses a large number of Bengali articles written by the noted Indian Nobel laureate Rabindranath Tagore (discussed in Section 4.1.1). Most of the reduplications in the corpus are separated by space or agglutinated together. Moreover, hyphen is used in other places other than reduplication ('এ-রকম', *e-rakam*).

6.7.1 Algorithms for Identifying Reduplication

The system considers the starting word position as position 1. For **complete reduplication**, identification is done only by comparing two consecutive words w1 and w2. Some time, inflection or matra is added to w2 (**রকম রকমের** জিনিস**,** *rakam-rakamer ginis, various types of goods*)**.** The matra is removed from w2 before comparison. The following algorithm is followed:

Algorithm 1: Complete Reduplication

```
If length (w1) == length (w2) then
  Compare all characters of both from position 1 to end;
      If all is equal then Complete Reduplication;
```

In **partial reduplication**, three cases are possible- (i) change of the first vowel or the matra attached with first consonant, (ii) change of consonant itself in first position or (iii) change of both matra and consonant. Exception is reported where vowel in first position is changed to consonant and its corresponding matra is added. For example, আবোল-তাবোল (*abal-tabal, **incoherent*** **or** ***irrelevant***)**.** Here, none of the individual words has any dictionary entry. This special case is handled by checking the change of vowel to its corrosponding matra attached with the new consonent (here, vowel **'আ'** is changed to its corrosponding matra attached (**'-আ'**) with consonant (**'ত'**).These are due to the orthographic rules applied in Bengali**.** Partial reduplication is handled using algorithm 2 discussed below. Here also, inflection has been removed before comparison. Linguistic study reveals that (Chattopadhyay 1992) when only consonant can be changed to any of the following four consonants: 'ট', 'ফ', 'ম', 'স'. But any consonant can be produced when both the consonant and matra are changed.

Algorithm 2: Partial Reduplication

```
If 1st char in both words are consonants and length are same
    If 1st char in both words are similar then
        If 2nd char in both words are matra and dissimilar then
            Check for all chars of both from position 3 to end;
            If all similar then reduplication; //matra  change
                Else if 1st char of w1 is স, ম, ফ, ট then
                Check for other chars of both from position 2 to end;
            If all similar then reduplication; // Consonant change
        If 1st char of both are dissimilar and 2nd char of both are matra and dissimilar then
            Check for other chars of both from position3 to end;
             If all are similar then Reduplication; //both change
            If 1st char of w1 is vowel and 1st char of w2 is consonant with corresponding
        matra and length are same then
                Check for other chars of both from position2 to end;
          If all similar then reduplication; //Special case
```

For **onomatopoeic expression,** mainly words are repeated twice and may be with some matra (mainly এ-matra is added with the first word to make second word). After reduplication, w1 and w2 are agglutinated to make a single word. In this case, after removing inflection, words are divided equally and then the comparison is done. Sometime onomatopoeic expression is used to express feelings (কনকনে **শীত**, *kankane shit*, *extremely cold*) or to emphasise the adjectives related to the noun (e.g. টকটকে **লাল**, *taktake lal*, *Deep red*). After collecting the words tagged as adjective and removing matra attached with the last letter, the algorithm 3 is applied.

Algorithm 3: Onomatopoeic expression

```
Word is divided into two parts and assigned to two strings S1 and S2 separately;
Check all the chars of both S1 and S2 sequentially;
        If all is equal then reduplication;
```

For **correlative reduplication**, approaches are more or less same with the previous algorithm 3. Here, naturally matra is not added with w2 and before reduplication, the formative affixes *‘–আ’* and *‘-ই’* are added with the root to form the first word and second words respectively and agglutinated together to make a single word. The algorithm is described below.

Algorithm 4: Correlative expression

```
If length of the word is even then
        Word is divided into two parts and assigned to two strings S1 and S2 respectively;
            If last char of S1 is আ-matra and last char of S2 is ই-matra then
                Check all the chars of both s1 and s2 from position 1 to position (end-1);
                If all is equal then reduplication;
```

For **semantic reduplication,** a dictionary based approach has been taken. List of inflections identified for the semantic reduplication is shown in Table 6.1.

Set of identified inflections and matra
0(শূন্য), এ(-য়ে, -য়), -তে(-এতে), -কে, -রে(-এরে), -র, -এর(য়ের), এরা, -দের, -টা, -টি, -গুলো, -ও, -ই,

Table 6.1: Inflections identified for semantic reduplication

This system has identified those consecutive words having same part-of-speech (mainly noun, adjective and adverb). Then, morphological analysis has been done to identify the roots of both w1 and w2. In synonymous reduplication, w2 is the synonym of w1. So at first in Bengali monolingual dictionary; the entry of w1 is searched to have any existence of w2. If matching is found, it is considered as reduplicated word. For antonym words, the opposite word of w1 is difficult to identify. These opposite words are mainly *gradable opposites* (পাপ-পুন্য, *pap-purna, Vice and Virtue*) where the word and its antonyms are entirely different wordforms. The *productive opposites* (গররাজি, *gargaji, disagree* is the opposite of রাজি, *raji*, *agree)* are easy to identify because the opposite word is generated by adding prefix or suffix with the original. In dictionary based approach, English meaning of both w1 and w2 are extracted and opposite of w1 is searched in English WordNet for any entry of w2.

The first model for identifying the five types of reduplications is shown in Figure 1. The functions performed by the different parts of the proposed architecture are:

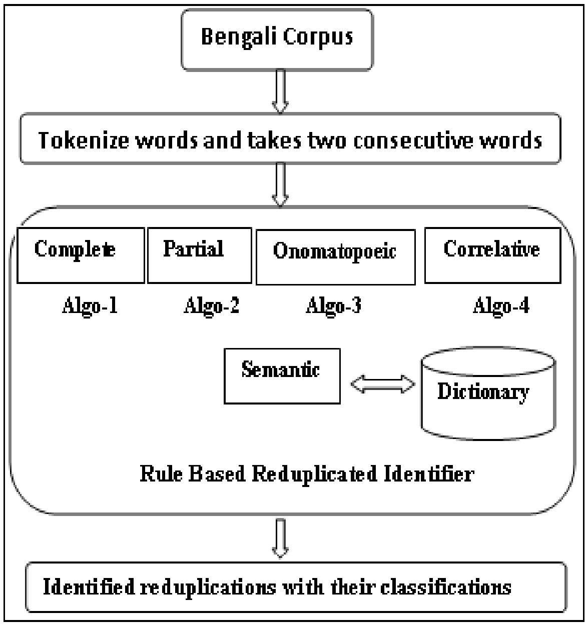

Bengali Corpus
Tokenize words and takes two consecutive words
Complete
Partial
Onomatopoeic
Correlative
Algo-1
Algo-2
Algo-3
Algo-4
Semantic
Dictionary
Rule Based Reduplicated Identifier
Identified reduplications with their classifications

Figure 6.1: System Architecture of first model

- **Tokenizer:** It separates the words based on blank space or special symbols (like hyphen, exclamation notation etc) to identify two consecutive words W_i and W_{i+1}.
- **Reduplication Identifier:** It is the main component of the first model. Consecutive tokens are passed to it to verify whether they are reduplicated words and to find the class they belong. Dictionary is sued for semantic reduplication.
- **Dictionary:** It includes the lexicon and the associated semantics. The system uses both Bengali-to-Bengali and Bengali-to-English dictionaries.

6.7.2 Semantics (sense) Analysis

Mainly eight types of semantic classifications are identified in Section 6.2.7 and their correspondence with expression level classifications has been mentioned. For example, if the reduplication is an onomatopoeic expression, its sense is easily identified as the sense of onomatopoeia. When infinite verb with complete reduplication is identified in a sentence, it obviously expresses the sense of incompleteness. The semantic or partial reduplicated words belong to the sense of completion. The correlative word is classified as the sense of corresponding correlative word because it is generally associated with the full verb in the sentence. The problem arises when grouping the complete reduplication. Sometime they are used as sense of repetition; plurality and sometime they express some kind of hesitation, incompleteness or softness. To disambiguate these senses, the system identifies some related words like ***‘করা’*** (*kara,* to do),*’ভাবা’* (*bhaba*, *to think*), *’মতো’* (*mato*, *like*), *‘লাগা’* (*laga, feel*) to classify them as incompleteness or softness or similarity. But these are not enough for disambiguating the sense of the phrase. For example,

ঘন ঘন মেঘ জমেছে। (*Ghana Ghana Megh Jamechhe.)*

Deep clouds gather.

Here, though it belongs to the sense of similarity, system cannot identify this using the above words. Sense disambiguation task has been identified as a future work.

6.8 Evaluation Metrics

The experiments are done on the corpus collected from some selected articles of Rabindranath Tagore (described in Section 4.1.1). The documents are cleaned automatically using rules and spelling mistakes and improper syntax are being checked manually. Dictionary is used for identifying semantic reduplications. As this is the first attempt in Bengali to identify reduplication, evaluation gold standard corpus is not available. Standard IR metrics like Precision, Recall and F-score are used to evaluate the system.

Total number of relevant reduplication is identified manually. For each type of structural classification, separate Precision, Recall and F-score are calculated. The overall system score is the average of these scores. Statistical co-occurrence measures are also calculated on each of the types. The following are the measures that have been used:

Frequency: Since MWEs generally get institutionalized, the frequency of the collection is a good indication of reduplication, given a large enough corpus.

Hyphen and closed form count: Orthographic representation of a collocation may provide clues about the collocation being reduplication. Words joined with hyphens (**হাত-টাত,** *hat-Tat, hands*) or occurring in closed form (**সেজেগুজে,** *sejeguje, dressed up*) are likely to denote a single concept or may be non-compositional.

6.8.1 Experimental Results

The collected corpus includes 14,810 tokens for 3675 distinct word forms at the root level. Precision, Recall, F-score are calculated for each class as well as for the reduplication identification system and are shown in Table 6.2 and Figure 6.2.

Reduplications	Precision	Recall	F-Score
Onomatopoeic	99.85	99.77	99.79
Complete	99.98	99.92	99.95
Partial	79.15	75.80	77.44
Semantic	85.20	82.26	83.71
Correlative	99.91	99.73	99.82
System	92.82	91.50	92.15

Table 6.2 Evaluation results for various reduplications

The scores of partial and semantic evaluation are not satisfactory because of some wrong tagging by the shallow parser (adjective, adverb and noun are mainly interchanged). Some synonymous reduplication (**ধীরে-সুস্থে,** *dhire-susthe, slowly and steadily, leisurely*) implies some sense of the previous word but not its exact synonym. These words are not identified properly.

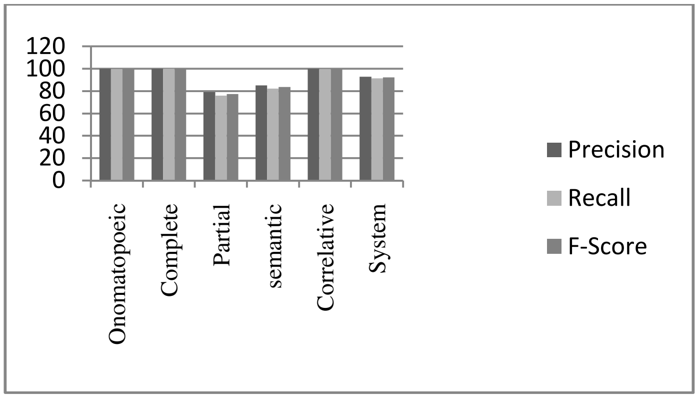


Figure 6.2 Bar graph of five types of reduplications and the system performances

Frequency is an important indication of whether a compound is a MWE. Figure 6.3 shows that in this corpus, the use of complete reduplication is more and hence a useful statistics has been developed for this corpus and the writing style of the author.

In this corpus maximum identified hyphened words are not reduplicated words and only 8.52% of reduplications are hyphened. This result shows that the trend of writing reduplications is the use of space as separator. Also the percentage of closed reduplications is 33.09% where maximum of them are onomatopoeic, correlative and semantic reduplications. 100% of correlative reduplications and maximum of onomatopoeic reduplications are closed.

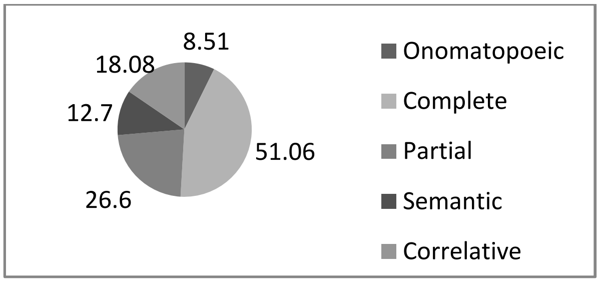


Figure 6.3 Frequency analysis of different reduplications

6.9 Conclusion and Future Work

As in other Indo-Aryan languages, reduplication is a very productive process at both the grammatical as well as semantic levels in Bengali. This paper has illustrated the phenomenon at the expression as well as at the semantic levels. As indicated above, the reduplication is mainly used for emphasis, generality, intensity, or to show continuation of an act. The semantics of the reduplicated words indicate some sort of sense disambiguation that basically cannot be bounded by only rule based approach. Sometimes it is observed in the present system that some word combinations are wrongly identified as reduplicated MWEs. This issues need to be studied further. Apart from this, more work needs to be done for identifying semantic reduplication using statistical and morphological approach. By gathering all these statistics, future research is planned on the field of Stylometry analysis or Plagiarism detection to identify the writing style of an author.

Chapter 7

Identification of MWEs in Bengali

7.1 Introduction

One of the key issues in both natural language understanding and generation is appropriate processing of *multiword expressions* (MWEs). Automatic extraction of MWEs from text corpora using linguistic and statistical tools is an alternative to manual creation of such databases. The existing techniques for automatic extraction of MWE rely upon accurate parts-of-speech (POS) taggers, shallow parsers and rich lexical resources such as WordNets. However, most of the Indian languages including Bengali cannot boast of even a good size representative corpus, let alone such sophisticated NLP tools. The chapter describes an approach for automatic extraction of certain categories of MWEs for Bengali in such a miserly resource scenario. The technique uses a morphological analyzer and a moderate size untagged text corpus. The results for Bengali are encouraging and the generic nature of the approach makes us believe that similar results can be expected for other languages as well.

7.2 Classification of Multiword Expressions in Bengali

Agarwal et al. (2004) proposed a different taxonomy for MWEs in Bengali based on syntactic flexibility and POS categories.

1. Words with Spaces

This class consists of MWEs which are syntactically rigid. No word can be inserted in between the expression; neither the words can be inflected except for the last one in some cases. Such expressions can be considered as a single lexical unit with spaces in between.

Agrawal et al. (2004) further classified them in terms of their morpho-syntactic category as follows.

(1.a) ***Cranberry*****:** No inflection allowed even to the last word and some individual words may not be a part of the standard Bengali vocabulary. E.g. যেন তেন প্রকারেন (*yena tena prakArena*, by any means), যার পর ন্যায় (*yAra para nAi,* ultimate), সোনায় সোহাগা(*sonAYe sohAgA,* an excellent combination) etc.

(1.b) ***Named Entities*****:** For examples names of people – কবিগুরু রবীন্দ্রনাথ ঠাকুর (*kaviguru rabindranAtha ThAkura,* Ravindranath Tagore), places – পশ্চিম মেদিনীপুর (*pashchima baNgal,* West Bengal) etc., where inflections can be added to the last word only.

(1.c) ***Idiomatic Compound Nouns*****:** These are noun-noun MWEs that are idiomatic or unproductive in nature and inflection can be added only to the last word. The formation of such compounds may be due to hidden conjunctions e.g. মা-বাবা (*mA bAbA, parents)* meaning *mA* (mother) and *bAbA* (father) or hidden inflections e.g. ললাটের লেখন (*lalATa lekhana)* meaning *lalATera* (of forehead) *lekhana* (writings) i.e. fate.

(1.d) ***Idiomatic Noun Groups with Inflections*****:** These are also noun-noun compounds with idiosyncratic meaning, but the first noun is in inflected (generally possessive) form. E.g. তাসের ঘর (*tAsera ghar,* house of cards - fragile), ঘোড়ার ডিম (*gho.DAra Dima,* horse's egg, absurd).

2. Semi-productive with Minor Syntactic Variations

This category includes MWEs that are either (semi-)productive in nature with its own grammar (like the numbers) or allow slight syntactic variations like inflections or a limited number of word insertion. They can be further classified as follows.

(2.a) ***Numbers*****:** Numerical expressions highly productive and can be expressed by a small grammar. However, no word can be inserted in between and inflections can be added only to the last word. E.g. এক হাজার চুরাশি (*eka hAjAra chaurAsI,* one thousand eighty four) etc.

(2.b) ***Kin Terms*****:** Bengali kin terms are normally two word MWEs such as মাসতুতো ভাই (*mAstuto bhAi,* maternal cousin).

(2.c) ***Productive Compound Nouns*****:** same as 1.c, except for the fact that the meaning is not idiomatic. These are also called institutionalized phrases. E.g. মাছ ভাজা (*mACha bhAjA,* fish fry), দুগ্গা পূজা (*durgA pujo,* Durga puja) etc.

(2.d) ***Noun – Noun collocations with inflections*****:** same as 1.d, except for the fact that these are semi-productive. E.g. মাটির মানুষ (*mATira mAnuSha,* man of earth, down-to-earth).

(2.e) ***Conjunct Verbs*****:** These are a pair of similar verbs used together to denote some other action. When used in inflected form, the same inflection (normally *e* or *te*) is added to both the verbs. E.g. *khAoYA dAoYA* (take food), *pa.DA shonA* ("read-hear" – study) etc.

3. High Syntactic Variation but Fixed POS Categories

This class includes noun-verb, adverb-verb and adjective-verb collocations, where the syntactic structure is quite flexible. For example, the ordering and the inflections of the words can vary, and the words can be separated by arbitrarily large number of words. However, the POS category of the words involved in such collocations is restricted. This class has both unproductive and semi-productive sub-categories as described below.

(3.a) ***Do/Is Support Verbs*****:** This is a productive class where verbs are formed by addition of "do" (*karA*) or "is" (*haoYA*) to a noun. E.g. স্নান করা (*snAna karA,* take bath).

(3.b) ***Light Verb Constructions*****:** Some verbs like *deoYA* (to give) or *kATA* can have different senses in different context. They are often referred to as light verbs (Stevenson *et al*. 2004). E.g. *চুল কাটা (chula kATA,* to dress hair).

(3.c) ***Adjective-Verb and Adverb-Verb Collocations*****:** Might be idiomatic or compositional, but statistically marked. E.g. লজ্জায় লাল হওয়া (*lajjAYe lAla haoYA,* blush), মুষোল ধারায় বৃষ্টি (*musaladhAre bRRiShTi haoYA,* rain cats and dogs) etc.

4. Completely Flexible MWEs

This category includes idioms and proverbs for which neither the word ordering, nor the POS category of the expression is fixed. High degree of syntactic variation and even synonym substitution is allowed. E.g. উলু বোনে মুক্তো ছড়ানো (*ulu bane mukto Cha.DAno,* useless attempt) etc.

7.3 Related work

A number of research activities have been carried out regarding MWE in various languages like English, German and other European languages. Various statistical co-occurrence measurements like Mutual Information (MI) (Church and Hans 1989), Log-Likelihood (Dunning 1993), Salience (Kilgarriff and Rosenzweig 2001) have been suggested for identification of MWEs. In Indian languages like Hindi, a considerable approach in compound noun MWE extraction (Kunchukuttan and Damani 2008) and a classification based approach for N-V collocations (Venkatapathy and Joshi 2009) have been done. In Bengali, works on automated extraction of MWEs are limited in number. One method of automatic extraction of Multi-word expression in Bengali (Agarwal et al. 2004) focusing mainly on Noun-Verb MWE has been carried out using significance function. In this experiment, we have taken five association measures like Pointwise Mutual Information (PMI), Log-likelihood ratio (LLR), Co-occurrence measure, Phi-coefficient and Significance function for automatic extraction of N-N Multi-word expressions and a combined weighted measurement technique has been proposed for final evaluation. The association measures used can be computed using only bigram collocation statistics. The frequency of each nominal MWE is very small in a corpus. We have seen in a comparative study that the results, obtained using PMI or LLR, can not identify MWEs in top ranking. So, instead of emphasizing much on frequency and its related measurements like MI, PMI, closed count, effective frequency, our system has tried to focus on probability of co-occurrence of component words in terms of their lexical affinity to each other. We have used weighted combination of features instead of Machine Learning (ML) because ML approach is language dependent and fails for narrow domain (Dias 2003). Furthermore, we have also proposed a clustering technique to identify Bengali MWEs using semantic similarity measurement. It is worth noting that the conducted experiments are useful for identifying MWEs from the electronically resource constrained languages like Bengali which are unable to collect reasonable size of corpus for statistical observations.

7.4 Statistical Identification of Noun-Noun (N-N) Collocated MWEs

In the past few years noun compounds have received increasing attention as researchers work towards the goal of full text understanding. Compound nouns are nominal compound where two or more nouns are combined to form a single phrase such as *'golf club'* or '*computer science department'* (Baldwin and Kim 2010). There is also a broader class of nominal MWEs where the modifiers are not restricted to be nominal, but can also be verbs (*'hired help'*) or adjectives (*'open secret'*). To avoid confusion, we have termed this broader set as "nominal compounds". Compound noun MWEs can be defined as a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the others(s) in other contexts and which shows some phonological and/or grammatical isolation from normal syntactic usage. One propertyof compound noun MWEs is their underspecified semantics. For example, while sharing the same head, there is little semantic commonality between *'nut tree*', *'cloths tree'* and *'family tree'* (Baldwin and Kim 2010). In each case, the meaning of the compound relates to a sense of both the head and the modifier, but the precise relationship is highly varied and not represented explicitly in any way. In English, Noun-Noun (NN) compounds occur with high frequency and high lexical and semantic variability. A summary examination of the 90 million word written component of the British National Corpus unearthed over 400,000 NN compound types, with a combined token frequency of 1.3 million; that is, over 1% of words in the BNC are NN compounds (Tanaka and Baldwin 2003). Bengali is a language consisting of high morpho-syntactic variation at surface level. The use of compound noun multi-word expressions in Bengali is quite a common practice mainly in the literature. Examples are discussed in Section 7.2. They are very frequently used in Bengali literature. Another common term in NLP application, which relates closely to our discussion of MWEs is 'collocation'. A widely used definition for collocation is "an arbitrary and recurrent word combination" (Benson 1990), or in our terms, a statistically idiomatic MWE. Collocations are often distinguished from "idioms" or "non-compositional phrases" on the grounds that these are not syntactically idiomatic and if they are semantically idiomatic, it is through a relative transparent process of figuration[1] and metaphor.

In this work, we mainly investigate the noun-noun collocated compounds from Bengali corpus which are the subset of compound nouns and they are separated by space or hyphen. In Bengali, some compounds which are formed by two or more different words acting as a single entity are also the part of compound nouns, but morphological analysis is needed to separate their components (Dasgupta et al. 2005). So the compounds, n-grams (n>2) and named-entities are beyond the scope of our investigation. They require much larger corpus for accurate estimation of the association measures. Reduplication is another term very frequently used in Bengali and is sometime tagged by 'NN'. These are also not considered here as they are easy to identify because of their immediate co-occurrence and no (for complete and onomatopoeic reduplication) or minor syntactic variation in the components (for partial and correlative reduplication).

7.4.1 Classification of Bengali Compound Noun MWEs

As mentioned earlier, compound noun consists of more than one free morpheme and when acts as MWE, components lose their individual literal meaning and act as a single semantic unit. Compound noun MWEs can occur in open, closed or hyphenated forms and satisfy semantic non-compositionality, statistical co-occurrence or literal phenomena like reduplication etc (Kunchukuttan and Damani 2008). Agarwal et al. (2004) have classified Bengali MWEs in three main classes consisting of twelve different fine-grained subclasses which is discussed in Section 7.2. However, taking this classification as reference and focusing on compound noun, we have classified it in seven different subclasses:

(i) **Named-Entities (NE):** Name of the people (***Rabindranath Thakur***, *Rabindranath Tagore*), name of the location (***Bharat-barsa***, *India*), name of the organization (***Pashchim Banga Siksha Samsad***, *West Bengal Board of Education*) etc. where inflection can be added to the last word only.

(ii) **Idiomatic Compound Nouns:** These are unproductive and idiomatic in nature and inflection can be added only to the last word. The formation of this type is due to the hidden conjunction

[1] **Figuration** is the property of the components of a MWE having some metaphoric, hyperbolic or metonymic in addition to their literal meaning.

between the components or extinction of inflection from the first component (***maa-baba****, mother and father*).

(iii) **Idioms:** They are also compound nouns with idiosyncratic meaning, but first noun is generally in possessive form (***taser ghar***, *fragile*). Sometime, individual components may not carry any significant meaning and can not be a part of dictionary (***gadai laskari chal,*** *indolent habit*). For them, no inflection is allowed even to the last word.

(iv) **Numbers:** They are highly productive, impenetrable and allow slight syntactic variations like inflections. Inflections can be added only to the last component (***soya sat ghanta****, seven hours and fifteen minutes*).

(v) **Relational Noun Compounds**: They are mainly kin terms and bigram in nature. Inflection can be added with the last word (***pistuto bhai,*** *maternal cousin*).

(vi) **Conventionalized Phrases:** Sometime they are called as **'Institutionalized phrase'**. They are not idiomatic and a particular word combination coming to be used to refer to a given object. They are productive and have unexpectedly low frequency and in doing so, contrastively highlight the statistical idiomaticity of the target expression (***bibhha barshiki****, marriage anniversary*).

(vii) **Simile Terms:** They are analogy term in Bengali and sometime similar to the idioms except the fact that they are semi-productive (***hater panch****, remaining resource*).

(viii) **Reduplicated Terms:** Reduplications are non-productive and tagged as noun phrase. They are further classified as onomatopoeic expressions (***khat khat****, knock knock*), complete reduplication (***bara-bara****, big big*), partial reduplication *(**thakur-thukur**, God),* semantic reduplication (***matha-mundu****, head),* Correlative Reduplication (***maramari***, *fighting*).

A number of research activities in Bengali Named Entity detection have been carried out (Ekbal et al. 2008), but there is no such standard tool to detect this. Here we have manually identified NE. Though numbers and kin terms can be captured by some lexicons, the use of lexicons during development phase is not at all a very acceptable way. Our work mainly focuses on the extraction of productive and semi-productive bigram MWEs like idioms, idiomatic compound nouns, simile terms, numbers, relational terms, and conventionalized phrases.

7.4.2 Corpus Used

Resource acquisition is one of the most challenging obstacles to work with electronically resource constrained languages like Bengali. However, this system has used a large number of Bengali articles written by the noted Indian Nobel laureate Rabindranath Tagore (discussed in Section 4.1.1). While we are primarily interested in single document term affinity, document information need not be maintained and manipulated by the experiment and document length normalization need not be considered. The order of the documents within the sequence is not of major importance. After merging all the articles, a medium size raw corpus has been created. It consists of 393,985 tokens and 283,533 types. Actual motivation of choosing this domain is to develop a useful statistics and further work on the Stylometry analysis.

7.4.3 Experimental Details

Basic system architecture is shown in Figure 7.1. The complete extraction procedure has been divided mainly into three phases. In the first phase, after initial pre-processing, candidate selection has been done using some heuristics to feed them into the main extraction phase.

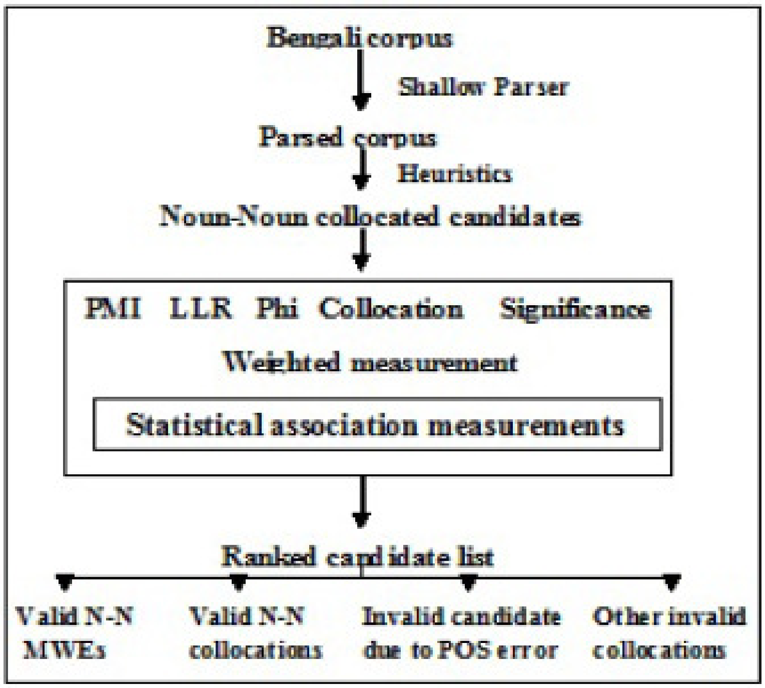

Bengali corpus
Shallow Parser
Parsed corpus
Heuristics
Noun-Noun collocated candidates
PMI LLR Phi Collocation Significance
Weighted measurement
Statistical association measurements
Ranked candidate list
Valid N-N MWEs
Valid N-N collocations
Invalid candidate due to POS error
Other invalid collocations

Figure 7.1 Basic system architecture

Mainly bigram collocations within same chunk have been extracted as candidates. In second phase, feature engineering consisting of various statistical co-occurrence parameters is applied on those candidates. Final decisions regarding a binary classification of MWE or non-MWE and Precision, Recall and F-score for each measurement are done in the final phase.

7.4.3.1 Initial Preprocessing

The crawled corpus is so scattered and unformatted that a basic semi-automatic pre-processing has been needed. Some of them are like sentence boundary detection and make the corpus suitable for parsing. Parsing using Bengali shallow parser has been done for identifying the POS, chunk, root and inflection of each token. Some of the tokens are misspelled due to typographic or phonetic error. For example, the token ***‘boi’*** *(book)* is written as ‘বই ’or sometime as ‘বৈ’. Shallow parser is not able to detect their actual root and inflection and the number of tokens is increased. Manual identification of these redundant synonymous phonetic words is done during this phase.

7.4.3.2 Candidate Selection

After pre-processing, bigram noun sequence within the same chunk is extracted from their POS and chunk categories. Shallow parser is confused with the two noun tags i.e. common noun (‘NN’) and proper noun (‘NNP’) because of the the continuous need for coinage of new terms for describing new concepts. For identifying all N-N MWEs, we have taken both of them and manual deletion of NEs has been done afterword. Although Chunking information helps to identify phrase boundary, some of the candidates belong to a chunk, which is formed by more than two nouns. Their frequency is also identified during evaluation phase. Bigram candidates can be thought of as <w1w2>. Total candidate selection phase is standing on the some heuristics described in Table 7.1. After first phase, a list of possible candidates is prepared. ‘NN’ and ‘NNP’ tags are mixed up and some of the consecutive nouns not belonging to a single chunk are also extracted by the parser. These parsing errors and NEs have been detected and filtered manually. A statistics of parsing error is calculated during evaluation phase.

Heuristics		
1.	POS	POS of each bigram must be either ‘NN’ or ‘NNP’
2.	Chunk	w1 and w2 must be in the same ‘NP’ chunk
3.	Inflection	Inflection[2] of w1 must be ‘-শুন্য’*(null)*, ‘-র’(-*r*), ‘-এর’(-*er*), ‘-এ’(-*e*), ‘-য়’(-*y*) or ‘-য়ের’(-*yr*) and for w2, any inflection is considered

Table 7.1 Heuristics applied in first phase.

7.4.3.3 Statistical Feature Engineering

We have said earlier that frequency information is not a reliable source of making any statistics especially for MWE because each MWE is too low in number in a medium size corpus. We have given a proof of this assumption taking directly frequency related measures like PMI and LLR. The following are the different association measures that we have taken for our analysis. Though these measures are discussed Section 7.3.2, they are briefly discussed hare:

- **Point-wise Mutual Information (PMI):** The PMI of a pair of outcomes *x* and *y* belonging to discrete random variables quantifies the discrepancy between the probability of their coincidence given their joint distribution versus the probability of their coincidence given only their individual distributions and assuming independence (Church et al.1990). Mathematically,

$$PMI(x, y) = \log \frac{P(xy)}{P(x)P(y)} \tag{7.1}$$

where, *P(xy)* = probability of the word *x* and *y* occurring together, *P(x)* = probability of *x* occurring in the corpus and *P(y)* = probability of *y* occurring in the corpus.

These probabilities can be assigned looking at the relative bigram and unigram frequency. This PMI is prone to highly overestimating the occurrence of rare events. This occurs since PMI does not incorporate the notion of support of the collocation (Kunchukuttan and Damani 2008).

[2] Linguistic study (Chattopadhyay, 1992) reveals that for compound noun MWE, considerable inflections of first noun are only those which are mentioned above.

- **Log-Likelihood Ratio (LLR):** The LLR is the ratio of the likelihood of the observations given the null-hypothesis to that of the alternate hypothesis (Dunning 1993). Generally, it is the ratio between the probability of observing one component of a collocation given the other is present and the probability of observing the same component of a collocation in the absence of other.

$$Log - Likelihood = -2\sum_{i,j} f(i,j)\log {}^{f(i,j)}\!\Big/\!_{f(i,j)} \tag{7.2}$$

Here the order of the words in the candidate collocation was irrelevant. We have adopted first probability using Baye's theorem by averaging the probability of w1 giving w2 and probability of w2 giving w1.

- **Phi-Coefficient:** In statistics, the Phi coefficient Φ is a measure of association for two binary variables. The Phi coefficient is also related to the chi-square statistic as:

$$\Phi = \sqrt{{}^{\chi^{2}}\!\Big/\!_{n}} \tag{7.3}$$

where n is the total number of observations and χ^2 is the chi-square distribution. Two binary variables are considered positively associated if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal. Here, the binary distinction denotes the positional information of the words. If we have a 2×2 table for two random variables x and y which denotes the presence of w1 and w2 respectively, we have following matrix:

	y=1	y=0	total
x=1	n_{11}	n_{10}	n_{x1}
x=0	n_{01}	n_{00}	n_{x0}
total	n_{y1}	n_{y0}	N

where, n_{11}=actual bigram <w1w2> count, n_{10}=frequency of bigram containing w1 but not w2, n_{01}=frequency of bigram containing w2 not w1, n_{00}=frequency of bigram not containing anyone of w1 and w2. n_{x1} and n_{x0} are the summation of their respective rows and n_{y1} and n_{y0} are the summation of their respective columns. Alternative words in place of absent w1 or w2 must be nouns. The phi coefficient that describes the association of x and y is

$$\varphi = \frac{n_{11}n_{00} - n_{01}n_{10}}{\sqrt{n_{x1}n_{x0}n_{y1}n_{y0}}} \tag{7.4}$$

- **Co-occurrence Measurement:** We have used co-occurrence measurement by using the formula adopted by Agarwal et al. (2004). It is defined as:

$$co(w1,w2) = \sum_{s \in S(w1,w2)} e^{-d(s,w1,w1)} \tag{7.5}$$

where, co(w1,w2)=co-occurrence frequency between the words (after stemming), S(w1,w2)=set of all sentences where both w1 and w2 occurs, d(s,w1,w2)=distance between w1 and w2 in a sentence s in terms of number of words. For every adjacent occurrence of w1 and w2, co(w1,w2) increases by 1, but if in a sentence they are largely separated, it increases only marginally. This measurement is used further in calculating significant function.

- **Significance Function:** Another effective co-occurrence measurement adopted by (Agarwal et al. 2004) is used in the present work. The definition of significance function for N-N collocation is as follows:

$$sig_{w1}(w2) = \sigma[\kappa_1.(1 - co(w1,w2).\frac{f_{w1}(w2)}{f(w1)})].\sigma[\kappa_2.\frac{f_{w1}(w2)}{\lambda} - 1] \tag{7.6}$$

$$sig(W1,W2) = sig_{W1}(W2).\exp[\frac{f_{W1}(W2)}{\max(f_{W1}(W2))} - 1] \tag{7.7}$$

where, $sig_{w1}(w2)$=significance of w2 with respect to w1. Here slightly modification has been done from the original by interchanging the roles of w1 and w2 in the first equation and averaging them. Same modification has been done for $f_{w1}(w2)$ which denotes number of w1 with which w2 has occurred. In the second equation, these modified values are used in their respective place. sig(w1,w2) denotes general significance of w1 and w2. σ(x) is the sigmoid function defined as [exp(-x)/(1+exp(-x))]. Two constants κ_1 and κ_2 define the stiffness of the sigmoid curve and for simplicity we have taken both of them as 5.0 (Agarwal et al. 2004). λ is defined as the average number of noun-noun co-occurrences. The value of significance function lies between 0 and 1.

- **Weighted Combination:** Final evaluation has been carried out by combining all the above-mentioned features. Experimental results show that Phi-coefficient, co-occurrence and significance functions actually based on the co-occurrence distribution has given more accurate results than the frequency-based measurement approaches like LLR, PMI in the higher ranks. So these three measures are considered and have been given certain weights after working with various weights. The final results are reported for the weighted triple <0.45, 0.35, 0.20> for co-occurrence, Phi and significance function respectively. The individual scores are normalized before assigning weights so that they are in the range 0 to 1.

7.4.4 System Evaluation

7.4.4.1 Evaluation Metrics

We have used standard IR metrics like Precision, Recall, F-score for evaluating our final weighted measurement as well as all the association measures. Manual identification of MWEs is done for evaluation purpose. Total candidates are divided into four classes: (1) valid N-N MWEs (M), (2) valid N-N semantic collocations but not MWEs (S), (3) invalid collocations due to considering bigram in a n-gram chunk where n>2 (B), (4) invalid candidates due to error in parsing like POS, chunk, inflection (E). For N number of candidates, three measuring approaches in percentage are calculated for each association measures.

Actual Precision (V) = M/N

Overall Precision (I) = (M+S)/N

Error rate due to B-type (O) =B/N

Precision for every measure is calculated as:

Precision (P) = (MWEs in top 1000 ranked candidates/ 1000)

Recall is defined as:

Recall (R) = (MWEs in top 1000 candidates/total N-N MWEs in the documents)

F-score (F) = (2*P*R) / (P+R)

Top 1000 ranked candidates are taken to evaluate each measure in the higher ranking.

7.4.4.2 Experimental Results

Four classes as discussed in Section 7.4.1 are identified manually and their frequencies are plotted in Figure 7.2.

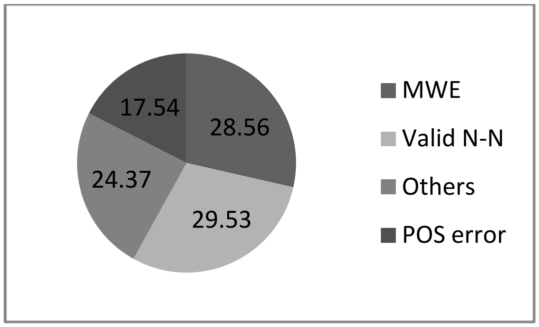


Figure 7.2 Frequencies of four types.

Maximum numbers of the candidates are erroneous due to parsing error. E-type candidates are filtered manually as it has produced erroneous statistics and the result might be biased. For each measurement, the scores have been sorted in descending order and the total range is divided into five ranks so that approximately equal scores fall within same rank. For every rank, three measures discussed in Section 5.1 are calculated and plotted in a graph. Table 7.2 depicts those results and Figure 7.3 gives a relative study of those measures.

The slope of each measure in Figure 7.3 is important in this purpose because the monotonously decreasing graph indicates the more number of MWEs in upper ranks rather than in lower ranks. PMI and LLR prove to be bad measures because graphs for LLR and PMI do not follow any significant alignment and slight upward slope have been noticed. This shows the presence of higher number of MWEs in the lower ranks.

Rank	**LLR**			**PMI**			**Co-occurrence**			**Phi Coefficient**			**Significance**		
	V	**I**	**O**	**V**	**I**	**O**	**V**	**I**	**O**	**V**	**I**	**O**	**V**	**I**	**O**
1	17.5	50.0	50.0	18.0	45.6	54.0	34.0	84.0	15.1	35.7	78.5	21.4	38.5	88.3	11.7
2	16.0	51.0	49.0	16.0	56.8	43.0	22.6	52.9	47.0	21.9	68.2	31.8	21.6	64.5	33.5
3	20.3	55.5	44.0	18.8	64.6	35.3	18.5	62.6	37.0	15.9	62.9	37.1	16.1	45.4	54.6
4	19.0	64.8	35.1	22.2	69.4	30.5	11.2	61.0	39.0	17.8	52.2	47.7	12.3	44.6	55.3
5	20.7	66.3	33.7	23.9	64.9	35.0	10.6	67.2	32.0	15.7	46.6	53.4	9.7	37.7	62.3

Table 7.2 Performance metrics of three different measures (in %) for each association approach.

Another important notification is that maximum of the lower ranked MWEs are reduplicated MWEs and they are filtered out when top 1000 ranked candidates are chosen. In weighted measured approach, maximum valid MWEs are listed in the top ranks. For this, V, I and O measures are shown in Table 7.3. It is clearly shown from the column (named as V) that the corresponding graph for valid N-N MWEs is decreasing in nature. The weighted combination approach improves upon each of the individual methods. If these association measures are combined using any ranking approach, it does not require any empirical settings of weights. But the problem is that there is no standard ranking methodology on these association measures.

Rank	**Weighted Measures**		
	V	**I**	**O**
1	46.54	89.23	10.77
2	30.27	72.28	27.72
3	13.43	59.98	40.02
4	7.66	36.62	63.38
5	5.09	23.39	76.61

Table 7.3 Results for weighted approach.

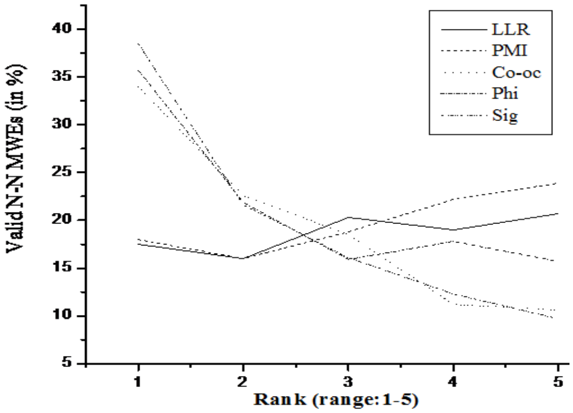


Figure 7.3 Valid N-N MWEs (in %) for each measure.

Top three candidates for each measure and its corresponding tags are shown in Table 7.4. Borda’s positional ranking that does an approximate aggregation of the ranked collocation list has been used as standard ranking function in previous studies (Kunchukuttan and Damani 2008).

Association Measures	**Top 3 Candidates**	**Tags (M/S/B)**
PMI	ghore dhuke	S
	payer dhula	M
	mukher pratibimba	S
LLR	ratdin moner	B
	barsar chata	S
	premer sagar	M
Co-occurrence	mathai hat	M
	maa-r mone	S
	mukher bhab	S
Phi-coefficient	khabarer kagaj	S
	haater panch	M
	dhaner haater	B
Significance	bayer din	S
	maran dasa	M
	haater panch	M
Weighted	galai dori	M
	garer maath	M
	nacher tale	S

Table 7.4 Top 3 candidates with their classifications.

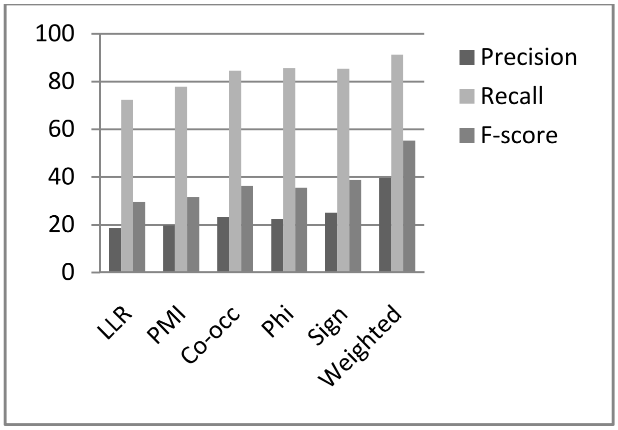


Figure 7.4 Precision, Recall and F-score (in %) for different measurements.

But the results were not satisfactory using this ranking and it did not serve as an effective MWE extraction technique.Precision, Recall and F-score are performed for all association measurements as well as for the weighted approach. These are measured among top 1000 candidates after manually deleting the parsing errors. The performance metrics for different measurements are shown in Figure 7.4.

Precision, Recall and F-score for weighted approach are 39.64%, 91.29% and 55.28% respectively, which are quite satisfactory in the first attempt. The present work does not focus on the increase of Precision. Our goal is to make a comparative study on the existing association measurements with our own weighted measurement and try to capture maximum number of the N-N collocated MWEs with in the top 1000 ranked candidates.

As an effort of developing a lexicon on N-N MWEs has been done simultaneously, we have observed the use of MWEs by the author in the documents. For this, we have chosen 10 novels of Rabindranath Tagore randomly and made a study using the following equation:

$$C_i=C_{i-1} \text{ U } (CL_i - C_{i-1}) \qquad (7.8)$$

where i=document id varied from 1 to 10, C_i=Combined list of N-N MWEs after ith document is processed, CL_i= Extracted list of N-N MWEs in ith document.

Value of (CL_i - C_{i-1}) is plotted against the document id in Figure 7.5, which indicates the use of new MWEs in the documents. The behavior of the graph is downward which indicates the saturation of use of MWEs with the increase of number of documents. Besides reduplications, some of the high frequent MWEs used through out the documents are shown in Table 5 according to their frequencies.

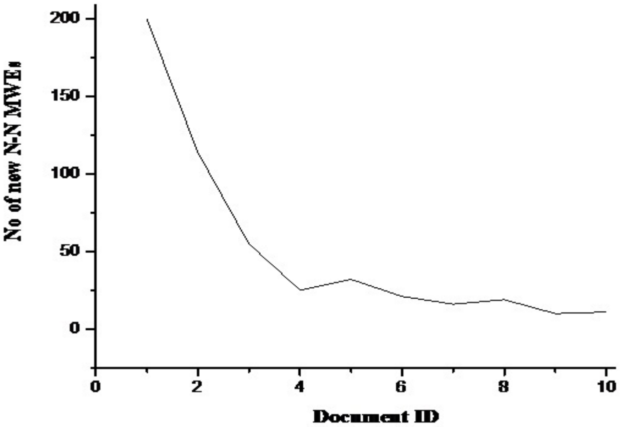


Figure 7.5 New N-N MWEs added for each document.

7.4.5 Observations and Discussions

Our approach for extraction of bigram noun-noun MWEs mainly focuses on the co-occurrence measurement of a bigram. From the experimental results it is quite evident that each idiomatic noun compound is not high in number and only frequency distribution measurement of these compounds is not an appropriate approach for any MWE measurement mainly for medium sized corpus.

Reduplications cause major variation of measurement in lower ranking. Though orthographic representation of collocation like hyphenation or closed form may provide clues about the collocation being a MWE (Kunchukuttan and Damani 2008), our experiment (Chakraborty and Bandyopadhyay 2010) has shown that in this corpus maximum identified hyphened words are not reduplicated words and only 8.52% of reduplications are hyphened. This result shows that the trend of writing reduplications is the use of space as separator. Also the percentage of closed reduplications is 33.09%.

High frequent N-N MWEs
abalar bol
galai dori
maran dasa
khamar chokhe
khider chote
charan darshan

Table 7.5 High frequent MWEs in 10 novels.

The presence of named-entities in the candidate list also affects the performance. While conceptually all named entities are MWEs, we do not include them in our research. We have manually filtered them at the beginning of the second phase.

Another important cause of taking the overall Precision (I) in consideration is that our basic goal is to build a statistics of different use of MWEs and compound in the articles by the writer and to identify the writing style or Plagiarism detection. Focusing on this, these semantically collocated compounds sometime express themselves as Institutionalized phrases in different text positions.

Significance function and co-occurrence measurement, which are used in this work, have been modified according to our need. Here, two binary variables used in Phi-coefficient are related to the positional information of constituent words w1 and w2. Weighted approach is basically a trial and error approach to find best triple.

Apart from being the first work of its kind for Bengali language, the contributions of this work are discussed as follows: (i) a morpho-syntactic classification of Bengali compound noun MWE classifications beyond the conventional classification of MWEs in English (Baldwin and Kim, 2010), (ii) new weighted approach for measuring MWEs which may be used for other types of collocation measurements, (iii) a list of Bengali N-N compound MWEs used as a lexical resource for developing synsets of MWEs, (iv) development of formatted corpus in Bengali for further study, (v) an initial study for the identification of Stylometry of Rabindranath Tagore.

7.4.6 Conclusion and Future Work

In the present work, we have identified nominal bigrams as MWEs using various statistical measurements. We have developed a list of bigram noun-noun candidates, annotated them and ranked them.

Complete identification of MWEs in Bengali is still far apart from the present work due to lack of lexical resources like WordNet. In English, two MWE types that are particularly well represented in WordNet are compound nouns (47,000 entities) and Multi-word verbs (2600 entries) (Baldwin et al. 2003). So verification using WordNet similarity is an easy approach in English. This is not possible for Bengali language. We are trying to develop such lexical resource for our purpose. Our weighted method has however given Precision of 39.64% and Recall of 91.29% at top 1000 candidates. Low Precision does not signify any bad conclusion because our main approach is to cover maximum number of MWEs in our list which has been satisfied by the high Recall. However, for our future study, we would like to apply this approach on the article of other writers and make a comparative study regarding the Stylometry of the writers. Further more, we will try to integrate Name-Entity Recognizer with the system to eliminate our manual filtering.

7.5 Identification of MWEs Using Semantic Clustering

Multi-word Expression is totally related to the semantics of a phrase, in that the meanings of its components do not contribute the actual semantics of the whole. This experiment deals with a reasonable approach in the task of identifying MWEs from a medium-size corpus by clustering semantically related nouns present in the corpus and use two similarity based measurements using vector space model. We also show our result using English WordNet::Similarity module. Here we have mainly focused on the bigram noun-noun compounds and developed a system that can make a binary distinction of the candidate phrase. The system also contributes to cluster the synonymous noun words present in a document using Bengali monolingual dictionary. Experimental results draw a satisfactory conclusion after showing precision, recall and F-score values in three cases.

7.5.1 Overview of Semantic Clustering

Semantic clustering aims to cluster semantically related words present in the document. Instead of word or sentence level, the context information of the surrounding tokens helps to manifest the underlying semantics. Identifying semantically related words for a particular token can be carried out by looking the surroundings token and finding the synonymous words of the surrounding words within a fixed context window. Higher number of occurrence of a particular expression is needed for statistical idiomaticity because one or few occurrences of a particular word cannot show all its meaning and in a medium-size corpus, it is hard to extract the synonymous word clustering. However, semantics of a word may be obtained by analyzing its similarity sets called *synset*. Semantic distance of two words can be measured by comparing their synsets. Higher value of the similarity comparison between two sets indicates more closeness of two words to each other.

Let W_1 and W_2 be the two component words of a bigram expression < W_1 W_2>. For individual components of the expression, semantically related words present in the documents are extracted by using the monolingual dictionariy:

$W_1= \{w^1_1, w^1_2, w^1_3,\ldots,w^1_m\}= \{w^1_i\}$

$W_2= \{w^2_1, w^2_2, w^2_3,\ldots,w^2_n \}=\{w^2_j\}$

Where, 1<=i<=m and 1<=j<=n.

Intersection of two synsets indicates the commonality of the two sets of the components. These common elements act as the dimensions of a vector space and the similarity based algorithm is applied to measure the semantic similarity between two components considering the component as vector in that semantic space.

7.5.2 Experimental Details

The complete experimental procedure moves through several phases. After acquisition and preprocessing of Bengali corpus, the system identifies noun-noun bigram candidates from the document based on some heuristics. Then the Bengali monolingual dictionary has been formatted which is used as the main clustering tool in this experiment. Based on this formatted dictionary, development of noun synsets and semantic clustering has been developed. The final decision has been made based on the proposed algorithm called as MWE-CHECKING. The main reason behind the usage of the monolingual dictionary is the lack of lexical resources like WordNet, large corpus, standard lemmatizer etc. in Bengali. This dictionary helped to make a closed set of synonyms with their lemmatized form which is used to compare two lexemes easily.

7.5.2.1 Corpus acquisition and Candidate Selection

We have used the same Bengali corpus discussed in section 7.4.2 and the same heuristics are used to acquire the candidates from the corpus discussed in Section 7.4.3.2. After initial pre-processing and candidate extraction, we have a list of Noun-noun collocations which are to be distinguished as MWEs or Non-MWEs.

7.5.2.2 Dictionary Restructuring

This phase deals with the building of Bengali synsets that aim not only to identify the meaning of Multi-Word Expressions but also to step up towards the development of Bengali WordNet. The input monolingual dictionary (Samsada Bengali Abhidhana)[3] contains each word present with its parts-of-speech (*বি.* Noun, *বিণ.* Adjective, *সর্ব.* Pronoun, অব্য. Indeclinables, *ক্রি.* Verb), phonetics and synonymous sets. Synonymous sets are separated using distinguishable notations based on similar sense and differential sense. Then synonyms of different sense with

[3] http://dsal.uchicago.edu/dictionaries/biswas-bangala/

respect to a word entry is separated by a semicolon (;) and synonyms having same sense are separated by comma (,). An automatic technique is devised to identify the synsets for a particular word entry based on the clues (, and ;) of similar and differential senses. The symbol tilde (~) indicates that the suffix string followed by the tilde (~) notation makes another new word concatenating with the original entry word. A snapshot of the synsets for the Bengali word "অংশু" (Angshu) is shown in Figure 7.6. For each entry, identification number, synset entry numbers have been given to each entry. Though these identity numbers are not used directly in this experiment, they can help to separate the entries and track them for further operations.

Dictionary Entry:

অংশু [aṃśu] বি. 1 কিরণ, রশ্মি, প্রভা; 2 আঁশ, তন্তু, সুতোর সূক্ষ্ম অংশ ।[সং. অন্শ্+উ]। ~ ক বি. বস্ত্র , সূক্ষ্ম বস্ত্র ; রেশম পাট ইত্যাদিতে প্রস্তুত বস্ত্র (চীনাংশুক)। ~ জাল বি. কিরণরাশি, কিরণমালা। ~ ধর বি. অংশুর ধার; সূর্য ।~ মতী বিণ. (স্ত্রী) কিরণময়ী, জ্যোতির্ময়ী । ~ মান (-মত্) 1 সূর্য; 2 সূর্যবংশীয় সগর রাজার পৌত্র ।~ মালা বি. রশ্মিজাল, কিরণমালা । ~ মালী (-লিন্) বি. সূর্য । ~ ল বিণ. কিরণময় , কিরণবিশিষ্ট ।

Synsets:

অংশু কিরণ/রশ্মি/প্রভা_বি.#25_1_1 আঁশ/তন্তু/সুতোর_সূক্ষ্ম_অংশ_[_সং._অন্শ্+উ]_বি.#25_2_2
অংশুক বস্ত্র/সূক্ষ্ম_বস্ত্র_বি.#26_1_1 রেশম_পাট_ইত্যাদিতে_প্রস্তুত_বস্ত্র_(চীনাংশুক)_বি.#26_2_2
অংশুজাল কিরণরাশি/কিরণমালা_বি.#27_1_1
অংশুধর অংশুর_ধার_বি.#28_1_1 সূর্য_বি.#28_2_2
অংশুমতী (স্ত্রী)_কিরণময়ী/জ্যোতির্ময়ী_বিণ.#29_1_1
অংশুমান সূর্য_#30_1_1 __2_সূর্যবংশীয়_সগর_রাজার_পৌত্র#30_2_2
অংশুমালা রশ্মিজাল/কিরণমালা_বি.#31_1_1
অংশুমালী সূর্য_বি.#32_1_1
অংশুল কিরণময়/কিরণবিশিষ্ট_বিণ.#33_1_1

Figure 7.6 Monolingual Dictionary Entry and built synsets for the word "অংশু *(Angshu)*

The synonymous entries are separated by slash ("/") and spaces are replaces by underscore ("_") within same synonymous set to ignore the confusion of separation between the words in same sense and between two senses. In Table 7.6, the frequencies of different synsets according to their part-of-speech are shown.

Word Entries	33619
Synsets	63403
Noun	28485
Adjective	11023
Pronoun	235
Indeclinables	497
Verb	1709

Table 7.6 Frequency information of the synsets with their part-of-speech

7.5.2.3 Generation of Noun Synsets

This phase is the beginning of the main clustering approach. Here, we have tried to generate the synonymous sets for the nouns present in the corpus using the dictionary. As our main goal is to make an intersection of the synsets of two consecutive noun words, we have used the dictionary as our standard resource and the dictionary entries as the member of the synsets. Another reason behind the use of dictionary entries as the closer set of entries is that Bengali language is resource constraints. The lack of standard lemmatizer, stemmer in Bengali makes an obstacle of doing direct the comparison of the synsets of the two components of the candidate bigram present in the documents. We can imagine a dictionary as a close set of words W_1, W_2, W_3,……,W_n where

$$W^1=n_1^1, n_2^1, n_3^1, \ \ldots\ldots\ldots\ldots\ldots\ldots = \{n_i^1\}$$
$$W^2=n_1^2, n_2^2, n_3^2, \ \ldots\ldots\ldots\ldots\ldots\ldots = \{n_j^2\}$$
$$W^3=n_1^3, n_2^3, n_3^3, \ \ldots\ldots\ldots\ldots\ldots\ldots = \{n_k^3\}$$
….

…..

$$W^m=n_1^m, n_2^m, n_3^m, \ \ldots\ldots\ldots\ldots\ldots = \{n_p^m\}$$

Where, W^1, W^2, ….,W^m are the dictionary entries and i, j, k,…., p are the number which can vary from 1 to the number of synsets for the corresponding entries and make close sets of synonyms. Now, each noun entry identified by the shallow parser in the document is searched in the dictionary for its individual existence with or without inflection. Suppose N is a noun in the corpus and it is present in the synsets of the W^1, W^3 and W^5. Therefore, they become the entries of the synsets of N. Mathematically, the synset of the noun N can be defined as:

$$\text{SynSet (N)} = \{W^l\} \tag{7.8}$$

Where l € 1 to m, such that N € $\{n^l_a\}$. Here, a € {i, j, k, …, p} and m is the number of dictionary entries.

7.5.2.4 Semantic similarity between two nouns

After generating the synsets for all the noun words in the document, the main task is to identify the semantic similarity between two nouns. It has been done by the intersection of the synsets of the two words and a score has been given that shows the semantic affinity between each other. Suppose, N_i and N_j are the two noun words in the document and W_i and W_j are their corresponding synsets. Now, the semantic similarity of the two words can be defined as:

$$\text{Similarity}\ (N_i,N_j) = |W_i \cap W_j| \qquad (7.9)$$

From the above equation, it is clearly shown that this value is maximum when the similarity is measured with itself (i.e. Similarity (N_i, N_j) is maximum when i=j). This semantic similarity measurement approximately gives a similarity score according to the commonality of their synonymous sets. However, it is noted that this similarity measure is not the only procedure to measure the similarity between two words. The ultimate similarity measurement has been taken using two vector space model discussed in section 7.5.2.6.

7.5.2.5 Semantic clustering of Noun

Shallow Parser identifies all the nouns present in the document and tags them as 'NN' (common noun), 'NST' (noun denoting spatial and temporal expression) and 'NNP' (proper noun). For this experiment, we have used mainly the common noun. Now, according to the score obtained by the semantic similarity measurement, we have clustered all the nouns present in the document for a particular noun and give a score for each of the similarities. For example, suppose the nouns identified by the shallow parser in the document are W_1, W_2, …,W_i, W_j, W_k, W_l, W_m, W_n,….etc. Now, for each noun M belonging to that set, the semantic similarities of M with the nouns are shown in the figure 7.7. In this figure, the distances from the word M to other nouns are the weights associated with the edges (i.e. a, b, c, etc).

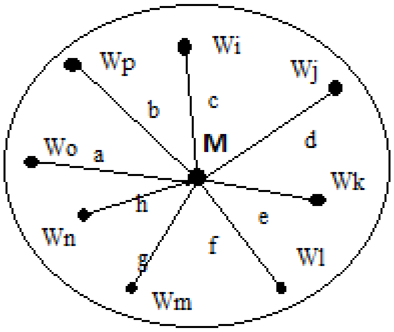


Figure 7.7 Semantic clustering of M with corresponding weights

7.5.2.6 Identification of Candidate Bigram as MWE

In this phase, the actual identification of candidates as MWE is done using the output obtained from the previous phase. If we consider noun-noun bigram as <M1 M2>, then the algorithm to identify the bigram as MWE is discussed below with proper example shown in Figure 7.8.

ALOGRITHM: MWE-CHECKING

INPUT: noun-noun bigram <M1 M2>

OUTPUT: True if MWE or false.

1. Extract semantic clusters of M1 and M2 using the procedure described in Section 7.5.2.3, 7.5.2.4 and 7.5.2.5.
2. Intersection of the synsets of both M1 and M2 (Figure 7.8 shows only the similar words common for both M1 and M2).
3. For measuring the similarity between M1 and M2:

    3.1. Identify the common elements of the similarity set of M1 and M2 (here common element is 8) with their scores (w_{ij}).

    3.2. In an n-dimensional vector space (here n=8), put M1 and M2 as two vectors and associated weights as their co-ordinates.

    3.3. calculate cosine-similarity measurement and Euclidean distance.

    3.4. Decision taken individually for two different measurements-

    3.4.1 If cosine-similarity > m, return false; Else return true; or

    3.4.2 If Euclidean distance > n, return false; Else return true;

(Where m and n are the pre-defined cut-off)

This algorithm looks little bit tricky especially in step 3. After identifying the common terms in both sets, a vector space model is used to identify the similarity between two components. In n-dimensional vector space, these common elements become the axes and each candidate acts as a vector in that space. The co-ordinate value of the vector in each direction is represented by the similarity measure between the candidate and the common term in that direction. The binary decision regarding the classification of a given candidate as MWE or not is a bit surprising (described in step 3.4). In the experiment, we have seen that the bigram MWEs mainly the idioms have shown low score of the similarity values between their constituents.

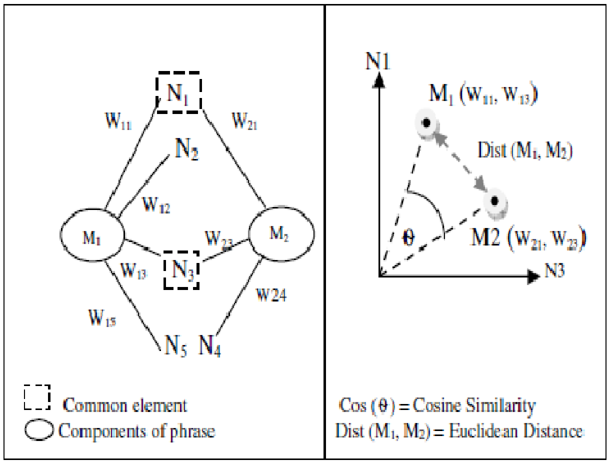


Figure 7.8.1 Intersection of the clusters of the constituents; Fig 7.8.2 Similarity between two constituents

If we take an example of a Bengali idiom ***hater panch*** *(remaining resource*), we have seen that WordNet defines two components of the idiom hat *(hand)* as a part of a limb that is farthest from the torso and ***panch*** *(five)* as a number which is one more than four. So from these two glosses it is quite clear that they are not at all semantically related in any sense. The synonymous sets for these two components extracted from the formatted dictionary are shown below –

Synset (হাত) = {হস্ত, কর, পাণি, বাহু, ভুজ, কৌশল, হস্তক্ষেপ, ধারণ, রেখা, লিখিত, হস্তাক্ষর, হস্তান্তর, হাজা}
Synset (পাঁচ) = {পঞ্চ, সংখ্যা, কর্ম, গঙ্গা, গব্য, কন্যা, গুণ, গৌড়, তন্ত্র, তীর্থ, পঞ্চত্ব, পনেরো, পূর্ণিমা, পঞ্চাশ}

So it is very clearly seen from the above synonymous sets that no one element of two sets is common and its similarity score is obviously zero. In this case, vector space model cannot be drawn using zero dimensions. For them, a marginal weight is assigned to show them as completely non-compositional phrase. To identify their non-compositionality, we have to show that their occurrence is not certain only in one case; rather they can occur side by side in several occasions. But this statistical proof can be determined better using a large corpus. Here, for those candidate phrases which show zero similarity, we have seen their existence more than one time in the corpus. Taking any decision using single occurrence may give incorrect result because they can be unconsciously used by the authors in their writings. That is why, the more the similarity between two components in a bigram, the less the probability to be a MWE.

7.5.2.7 WordNet::Similarity Measurement

As Aforementioned, there is no such lexical resource like WordNet in Bengali; we have tried to use English WordNet (discussed in Section 4.2.1) in this research to measure the semantic distance between two Bengali words after translating into English. WordNet::Similarity is an open-source package for calculating the lexical similarity between word (or sense) pairs based on variety of similarity measures. Basically, WordNet measures the relative distance between two nodes denoted by two words in the WordNet tree which can vary from -1 to 1 where -1 indicates total dissimilarity between two nodes.

In this experiment, we have first translated the root of two Bengali components in a phrase into their English forms using Bengali to English Bilingual Dictionary. Then these two words are passed into the WordNet based Similarity module for measuring the distance. A predefined cut-off value is determined to distinguish between MWE and simple compositional term. If the measured distance is less than that threshold, the similarity between them is less. But the candidate phrase consisting of these two words has a reasonable occurrence in the corpus. It concludes the phrase to be a MWE. Evaluation results are taken after varying the cut-off value.

7.5.3 Human Annotator’s Judgment

Three annotators identified as A1, A2 and A3 were engaged to carry out the annotation. The annotation agreement of 628 candidate phrases is measured using standard Cohen's *kappa* coefficient (κ) (Cohen 1960). It is a statistical measure of inter-rater agreement for qualitative (categorical) items. In addition to this, we also choose the measure of agreements on set-valued items (*MASI*) (Passonneau 2006) that was used for measuring agreement in the semantic and pragmatic annotation. Annotation results as shown in Table 7.7 are satisfactory.

MWEs	Agreement between Pair of annotators			
[# 628]	A1-A2	A2-A3	A1-A3	Average
KAPPA	87.23	86.14	88.78	87.38
MASI	87.17	87.02	89.02	87.73

Table 7.7 Inter-Annotator’s Agreement (in %)

The list of noun-noun collocations are extracted from the output of the parser for manual checking. It is observed that 39.39% error occurs due to wrong POS tagging or extracting invalid collocations due to considering bigram in a n-gram chunk where n > 2. We have separated these phrases from the final list.

7.5.4 Experimental Results

We have used standard IR matrices like Precision (P), Recall (R) and F-score (F) to evaluating the final results obtained from three modules. Human annotated list is used as the gold standard for the evaluation. The results are shown in Table 7.8. The predefined threshold has been varied to catch individual results in each case. Increasing recall in accordance with the increment of cut-off infers that maximum numbers of MWEs are identified in a wide range of threshold. But precision does not increase monotonously. It shows that higher cut-off degrades the performance. The reasonable results for precision and recall have been achieved in case of cosine-similarity at the cut-off value of 0.5 where Euclidean distance and WordNet Similarity give maximum precision at cut-off values of 0.4 and 0.5 respectively.

Baldwin et. al. (2003) suggested that WordNet::Similarity is effective to identify decomposability of Multiword Expression. We are surprisingly concluding the same for Bengali language. There are also candidates with very low value of similarity between their constituents (e.g. *ganer gajat*, 'earth of songs') and they are discarded from this experiment because of their low frequency of occurrence in the corpus.

Cut-off	**Cosine-Similarity**			**Euclidean Distance**			**WordNet Similarity**		
	P	R	F	P	R	F	P	R	F
0.6	70.75	64.87	67.68	70.57	62.23	66.14	74.60	61.78	67.58
0.5	78.56	59.45	67.74	72.97	58.79	65.12	80.90	58.75	68.06
0.4	73.23	56.97	64.08	79.78	53.03	63.71	75.09	52.27	61.63

Table 7.8 Precision, Recall and F-score for various measurements

7.5.5 Conclusion

We hypothesized that sense induction by analyzing synonymous set can assist the identification of Multiword Expression. We have introduced an unsupervised approach to explore the hypothesis and have shown that clustering technique along with similarity measures can be successfully employed to perform the task. This experiment additionally contributes to the followings- (i) Clustering of words having similar sense, (ii) Identification of MWEs for resource-constraint languages, (iii) Reconstruction of Bengali monolingual dictionary towards the development of Bengali WordNet. However complete identification of MWEs for Bengali is far apart from this study. Our intuition is that this algorithm is also applicable for other type of MEWEs like adjective-noun collocation, verbal MWEs. As our future work, we plan to apply it for other classes of MWEs as well as for other languages. Furthermore, we will try to integrate Name-Entity Recognizer with the system to extract all kinds of nominal MWEs from the corpus.

7.6 Identification of Complex Predicates in Bengali

This paper presents the automatic extraction of Complex Predicates (*CPs*) in Bengali with a special focus on compound verbs (*Verb* + *Verb*) and conjunct verbs (*Noun /Adjective* + *Verb*). The lexical patterns of compound and conjunct verbs are extracted based on the information of shallow morphology and available seed lists of verbs. Lexical scopes of compound and conjunct verbs in consecutive sequence of Complex Predicates (*CPs*) have been identified. The fine-grained error analysis through confusion matrix highlights some insufficiencies of lexical patterns and the impacts of different constraints that are used to identify the Complex Predicates (*CPs*). System achieves *F-Scores* of 75.73%, and 77.92% for compound verbs and 89.90% and 89.66% for conjunct verbs respectively on two types of Bengali corpus.

7.6.1 Introduction to Complex Predicate (CP)

Complex Predicates (*CPs*) contain [*verb*] + *verb* (*compound verbs*) or [*noun/ adjective/adverb*] +verb (*conjunct verbs*) combinations in *South Asian language*s (Hook 1974). To the best of our knowledge, Bengali is not only a language of South Asia but also the sixth popular language in the World [4] , second in India and the national language of Bangladesh. The identification of Complex Predicates (*CPs*) adds values for building lexical resources (e.g. WordNet (Miller *et al.* 1990; VerbNet (Kipper-Schuler 2005)), parsing strategies and machine translation systems.

Bengali is less computerized compared to English due to its morphological enrichment. As the identification of Complex Predicates (*CPs*) requires the knowledge of morphology, the task of automatically extracting the Complex Predicates (*CPs*) is a challenge. Complex Predicates

[4] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

(*CPs*) in Bengali consists of two types, compound verbs (*CompVs*) and conjunct verbs (*ConjVs*). The compound verbs (*CompVs*) (e.g. মেরে ফেলা *mere phela* 'kill', বলতে লাগল *bolte laglo,* 'started saying') consist of two verbs. The first verb is termed as *Full Verb* (*FV*) that is present at surface level either as conjunctive participial form -এ *(–e)* or the infinitive form -তে *(–te).* The second verb bears the inflection based on *Tense*, *Aspect* and *Person*. The second verbs that are termed as *Light Verbs* (*LV*) are polysemous, semantically bleached and confined into some definite candidate seeds (Paul, 2010).

On the other hand, each of the Bengali conjunct verbs (*ConjVs*) (e.g. ভরসা করা *bharsha kara* 'to depend', ঝক্ ঝক্ করা *jhakjhak kara* 'to glow') consists of noun or adjective followed by a *Light Verb* (*LV*). The *Light Verbs* (*LVs*) bear the appropriate inflections based on *Tense*, *Aspect* and *Person*. According to the definition of multi-word expressions (*MWEs*)(Baldwin and Kim 2010), the absence of conventional meaning of the *Light Verbs* in Complex Predicates (*CPs*) entails us to consider the Complex Predicates (*CPs*) as MWEs (Sinha 2009). But, there are some typical examples of Complex Predicates (*CPs*), e.g. দেখা করা *dekha kara* 'see-do' that bear the similar lexical pattern as *Full Verb* (*FV*)+ *Light Verb* (*LV*) but both of the *Full Verb* (*FV*) and *Light Verb* (*LV*) loose their conventional meanings and generate a completely different meaning ('to meet' in this case).

In addition to that, other types of predicates such as নিয়ে গেল *niye gelo* 'take-go' (took and went), দিয়ে গেল *diye gelo* 'give-go' (gave and went) follows the similar lexical patterns *FV+LV* as of Complex Predicates (*CPs*) but they are not mono-clausal. Both the *Full Verb* (*FV*) and *Light Verb* (*LV*) behave like independent syntactic entities and they belong to non-Complex Predicates (*non-CPs*). The verbs are also termed as *Serial Verb* (*SV*) (Mukherjee *et al.* 2006). Butt (1993) and Paul (2004) have also mentioned the following criteria that are used to check the validity of complex predicates (*CPs*) in Bengali. The following cases are the invalid criteria of complex predicates (*CPs*).

1. *Control Construction* (*CC*): লিখতে বলো *likhte bollo* 'asked to write', লিখতে বাধ্য করল likhte badhyo korlo 'forced to write'
2. *Modal Control Construction* (*MCC*): যেতে হবে *jete hobe* 'have to go' খেতে হবে *khete hobe* 'have to eat'

3. *Passives* (*Pass*) *:* ধরা পরল *dhora porlo* ‘was caught’, মারা হল *mara holo* ‘was beaten’

4. *Auxiliary Construction* (*AC*)*:* বসে আছে *bose ache* ‘is sitting’, নিয়ে চলো *niye chilo* ‘had taken’.

Sometimes, the successive sequence of the Complex Predicates (*CPs*) shows a problem of deciding the scopes of individual Complex Predicates (*CPs*) present in that sequence. For example the sequence, উঠে পরে দেখলাম *uthe pore dekhlam* ‘rise-wear-see’ (rose and saw) seems to contain two Complex Predicates (*CPs*) (উঠে পরে *uthe pore* ‘rose’ and পরে দেখলাম *pore dekhlam* ‘wore and see’). But there is actually one Complex Predicate (*CP*). The first one উঠে পরে *uthe pore* ‘rose’ is a compound verb (*CompV*) as well as a Complex Predicate (*CP*). Another one is দেখলাম *dekhlam* ‘saw’ that is a simple verb. As the sequence is not mono-clausal, the Complex Predicate (*CP*) উঠে পরে *uthe pore* ‘rose’ associated with দেখলাম *dekhlam* ‘saw’ is to be separated by a lexical boundary. Thus the determination of lexical scopes of Complex Predicates (*CPs*) from a long consecutive sequence is indeed a crucial task.

The present task therefore not only aims to extract the Complex Predicates (*CPs*) containing compound and conjunct verbs but also to resolve the problem of deciding the lexical scopes automatically. The compound verbs (*CompVs*) and conjunct verbs (*ConjVs*) are extracted from two separate Bengali corpora based on the morphological information (e.g. participle forms, infinitive forms and inflections) and list of *Light Verbs* (*LVs*). As the *Light Verbs* (*LVs*) in the compound verbs (*CompVs*) are limited in number, fifteen predefined verbs (Paul 2010) are chosen as *Light Verbs* (*LVs*) for framing the compound verbs (*CompVs*). A manually prepared seed list that is used to frame the lexical patterns for conjunct verbs (*ConjVs*) contains frequently used *Light Verbs* (*LVs*). An automatic method is designed to identify the lexical scopes of compound and conjunct verbs in the long sequences of Complex Predicates (*CPs*). The identification of lexical scope of the Complex Predicates (*CPs*) improves the performance of the system as the number of identified Complex Predicates (*CPs*) increases. Manual evaluation is carried out on two types of Bengali corpus. The experiments are carried out on 800 development sentences from two corpora but the final evaluation is carried out on 1000 sentences. Overall, the system achieves *F-Scores* of 75.73%, and 77.92% for compound verbs and 89.90% and 89.66% for conjunct verbs respectively. The error analysis shows that not only the lexical patterns but also the augmentation of argument structure agreement (Das 2009), the analysis of *Non-MonoClausal Verb* (*NMCV*) or *Serial Verb, Control Construction* (*CC*)*, Modal Control Construction* (*MCC*), *Passives* (*Pass*) and *Auxiliary Construction* (*AC*) (Butt, 1993; Paul, 2004) are also necessary to identify the Complex Predicates (*CPs*). The error analysis shows that the system suffers in distinguishing the Complex Predicates (*CPs*) from the above constraint constructions.

7.6.2 Related Work on Complex Predicates

The general theory of complex predicate is discussed in Alsina (1996). Several attempts have been organized to identify complex predicates in *South Asian languages* (Abbi 1991; Bashir 1993; Verma 1993) with a special focus to Hindi (Burton-Page 1957; Hook 1974), Urdu (Butt 1995), Bengali (Sarkar 1975; Paul 2004), Kashmiri (Kaul 1985) and Oriya (Mohanty 1992). But the automatic extraction of Complex Predicates (*CPs*) has been carried out for few languages, especially Hindi. The task described in (Mukherjee *et al.* 2006) highlights the development of a database based on the hypothesis that an English verb is projected onto a multi-word sequence in Hindi. The simple idea of projecting POS tags across an English-Hindi parallel corpus considers the Complex Predicate types, adjective-verb (AV), noun-verb (NV), adverb-verb (Adv-V), and verb-verb (VV) composites. A similar task (Sinha 2009) presents a simple method for detecting Complex Predicates of all kinds using a Hindi-English parallel corpus. His simple strategy exploits the fact that Complex Predicate is a multi-word expression with a meaning that is distinct from the meaning of the *Light Verb*. In contrast, the present task carries the identification of Complex Predicates (*CPs*) from monolingual Bengali corpus based on morphological information and lexical patterns. The analysis of V+V complex predicates termed as lexical compound verbs (*LCpdVs*) and the linguistic tests for their detection in Hindi are described in (Chakrabarti *et al.* 2008). In addition to compound verbs, the present system also identifies the conjunct verbs in Bengali. But, it was observed that the identification of Hindi conjunct verbs that contain noun in the first slot is puzzling and therefore a sophisticated solution was proposed in (Das 2009) based on the control agreement strategy with other overtly case marked noun phrases.

The present task also agrees with the above problem in identifying conjunct verbs in Bengali although the system satisfactorily identifies the conjunct verbs (*ConjVs*). Paul (2003) develops a constraint-based mechanism within HPSG framework for composing Indo-Aryan compound verb constructions with special focus on Bangla (Bengali) compound verb sequences. Postulating semantic relation of compound verbs, another work (Paul 2009) proposed a solution of providing lexical link between the *Full verb* and *Light Verb* to store the Compound Verbs in IndoWordNet without any loss of generalization. To the best of our knowledge, ours is the first attempt at automatic extraction of Complex Predicates (*CPs*) in Bengali.

7.6.3 Experimental Details

The compound verbs (*CompVs*) and conjunct verbs (*ConjVs*) are identified from the shallow parsed result using a lexical pattern matching technique.

7.6.3.1 Preparation of Corpora

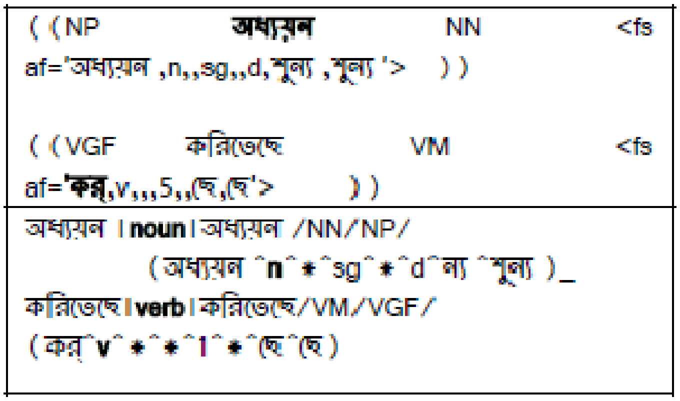
Two types of Bengali corpus have been considered to carry out the present task. One corpus is collected from a travel and tourism domain and another from an online web archive of Rabindranath Rachanabali (discussed in Section 4.1.1).The former EILMT travel and tourism corpus is obtained from the consortium mode project "Development of English to Indian Languages Machine Translation (EILMT)[5] System". The second type of corpus is retrieved from the web archive and pre-processed accordingly. Each of the Bengali corpora contains 400 and 500 development and test sentences respectively. The sentences are passed through an open source Bengali shallow parser. The shallow parser gives different morphological information (root, lexical category of the root, gender, number, person, case, vibhakti, tam, suffixes etc.) that help in identifying the lexical patterns of Complex Predicates (*CPs*).

7.6.3.2 Extracting Complex Predicates (*CPs*)

Manual observation shows that the Complex Predicates (*CPs*) contain the lexical pattern {[*XXX*] (**n/adj**) [*YYY*] (**v**)} in the shallow parsed sentences where XXX and YYY represent any word. But, the lexical category of the root word of XXX is either noun (**n**) or adjective (**adj**) and the lexical category of the root word of YYY is verb (**v**). The shallow parsed sentences are pre-

[5] The EILMT project is funded by the Department of Information Technology (DIT), Ministry of Communications and Information Technology (MCIT), Government of India.

processed to generate the simplified patterns. An example of similar lexical pattern of the shallow parsed result and its simplified output is shown in Figure 7.9.



```
( ( NP          অধ্যয়ন          NN          <fs
af='অধ্যয়ন ,n,,sg,,d,শূন্য ,শূন্য '>    ) )

( ( VGF       করিতেছে          VM          <fs
af='কর্,v,,,5,,ছে,ছে'>        ) )
```

```
অধ্যয়ন | noun | অধ্যয়ন /NN/NP/
          ( অধ্যয়ন ^n^*^sg^*^d^ন্য ^শূন্য )_
করিতেছে | verb | করিতেছে/VM/VGF/
( কর্^v^*^*^1^*^ছে^ছে )
```

Figure 7.9 Example of a pre-processed shallow parsed result.

The corresponding lexical categories of the root words অধ্যয়ন *adhyan* 'study' (e.g. *noun* for '**n**') and করা *kar*, 'do' (e.g. *verb* for '**v**') are shown in bold face in Figure 7.9. The following example is of conjunct verb (*ConjV*). The extraction of Bengali compound verbs (*CompVs*) is straightforward rather than conjunct verbs (*ConjVs*). The lexical pattern of compound verb is {[*XXX*](**v**) [*YYY*] (**v**)} where the lexical or basic POS categories of the root words of "*XXX*" and "*YYY*" are only verb. If the basic POS tags of the root forms of "*XXX*" and "*YYY"* are *verbs* (*v*) in shallow parsed sentences, then only the corresponding lexical patterns are considered as the probable candidates of compound verbs (*CompVs*).

Example 1:

শুইয়া |verb| শুইয়া/VM/VGNF/শো ^v^*^*^ইয়া^*^ইয়া^ইয়া)

#পড়িতাম |verb| পড়িতাম/VM/VGF/(পড়^v^*^*^1^*^ত^ত)

Example 1 is a compound verb (*CompV*) but Example 2 is not. In Example 2, the lexical category or the basic POS of the *Full Verb* (*FV*) is noun (*n*) and hence the pattern is discarded as non-compound verb (*non-CompV*).

Example 2:

লক্ষ্য |**noun**| লক্ষ্য /**NN**/NP/(লক্ষ্য ^**n**^*^*^*^*^*^poslcat="NM") #

করিয়া |**verb**| করিয়া/**VM**/VGNF/(কর^**v**^*^*^any^*^ইয়া^ইয়া)

aSa ‘come’	*dãRa* ‘stand’
rakha ‘keep’	*ana* ‘bring’
deoya ‘give’	*pOra* ‘fall’
paTha ‘send’	*bERano* ‘roam’
neoya ‘take’	*tola* ‘lift’
bOSa ‘sit’	*oTha* ‘rise’
jaoya ‘go’	*chaRa* ‘leave’

Table 7.9 List of *Light Verbs* for compound verbs.

Bengali, like any other Indian languages, is morphologically very rich. Different suffixes may be attached to a *Light Verb* (*LVs*) (in this case [*YYY*]) depending on the various features such as *Tense*, *Aspect*, and *Person*. In case of extracting compound verbs (*CompVs*), the *Light Verbs* are identified from a seed list (Paul, 2004). The list of *Light Verbs* is specified in Table 7.9. The dictionary forms of the *Light Verbs* are stored in this list. As the *Light Verbs* contain different suffixes, the primary task is to identify the root forms of the *Light Verbs* (*LVs*) from shallow parsed result. Another table that stores the root forms and the corresponding dictionary forms of the *Light Verbs* is used in the present task. The table contains a total number of 378 verb entries including *Full Verbs* (*FVs*) and *Light Verbs* (*LVs*). The dictionary forms of the *Light Verbs* (*LVs*) are retrieved from the Table.

On the other hand, the conjunctive participial form –এ/ইয়া *-e/iya* or the infinitive form –তে/-ইতে *–te/ite* are attached with the *Full Verbs* (*FVs*) (in this case [*XXX*]) in compound verbs (*CompVs*). ইয়া / *iya* and ইতে/ *ite* are also used for conjunctive participial form -এ *–e* or the infinitive form -তে *–te* respectively in literature. The participial and infinitive forms are checked based on the morphological information (e.g. suffixes of the verb) given in the shallow parsed results. In Example 1, the *Full Verb* (*FV*) contains -ইয়া *-iya* suffix. If the dictionary forms of the *Light Verbs* (*LVs*) are present in the list of *Light Verbs* and the *Full Verbs* (*FVs*) contain the suffixes of –এ/ইয়া - *e/iya* or –তে/-ইতে *–te/ite*, both verbs are combined to frame the patterns of compound verbs (*CompVs*).

deoya ‘give’	*kara* ‘do’
neoya ‘take’	*laga* ‘start’
paoya ‘pay’	*kata* ‘cut’

Table 7.10 List of *Light Verbs* for conjunct verbs.

The identification of conjunct verbs (*ConjVs*) requires the lexical pattern (*Noun / Adjective + Light Verb*) where a noun or an adjective is followed by a *Light Verb* (*LV*). The dictionary forms of the *Light Verbs* (*LVs*) that are frequently used as conjunct verbs (*ConjVs*) are prepared manually. The list of *Light Verbs* (*LVs*) is given in Table 7.10. The detection of *Light Verbs* (*LVs*) for conjunct verbs (*ConjVs*) is similar to the detection of the *Light Verbs* (*LVs*) for compound verbs (*CompVs*) as described earlier in this section. If the basic POS of the root of the first words ([*XXX*]) is either "*noun*" or "*adj*" (**n/adj**) and the basic POS of the following word ([*YYY*]) is "*verb*" (**v**), the patterns are considered as conjunct verbs (*ConjVs*). The Example 2 is an example of conjunct verb (*ConjV*).

For example, ঝক্ ঝক্ করা (*jhakjhak kara* ‘to glow’), তক্ তক্ করা (*taktak* ‘to glow’), চুপচাপ করা (*chupchap kara* ‘to silent’) etc are identified as conjunct verbs (*ConjVs*) where the basic POS of the former word is an adjective (**adj**) followed by করা *kara* ‘to do’, a common *Light Verb*.

Example 3:

ঝকঝক্ । adj । ঝকঝক্ /JJ/JJP/(ঝকঝক্ ^adj) #

করিত । verb । করিত /VM/VGF/(কর্^v^*^*^5^*^কর্^কর্)

But, the extraction of conjunct verbs (*ConjVs*) that have a "*noun+verb*" construction is descriptively and theoretically puzzling (Das 2009). The identification of lexical patterns is not sufficient to recognize the compound verbs (*CompVs*). For example, বই দেওয়া *boi deoya* ‘give book’ and ভরসা দেওয়া *bharsa deyoa* ‘to assure’ both contain similar lexical pattern (noun+verb) and same *Light Verb* দেওয়া *deyoa*. But, ভরসা দেওয়া *bharsa deyoa* ‘to assure’ is a conjunct verb (*ConjV*) where as বই দেওয়া *boi deoya* ‘give book’ is not a conjunct verb (*ConjV*). Linguistic observation shows that the inclusion of this typical category into conjunct verbs (*ConjVs*) requires the additional knowledge of syntax and semantics. In connection to conjunct verbs

(*ConjVs*), (Mohanty 2010) defines two types of conjunct verbs (*ConjVs*), synthetic and analytic. A synthetic conjunct verb is one in which both the constituents form an inseparable whole from the semantic point of view or semantically non-compositional in nature. On the other hand, an analytic conjunct verb is semantically compositional. Hence, the identification of conjunct verbs requires knowledge of semantics rather than only the lexical patterns. It is to be mentioned that sometimes, the negative markers (না *no,* নাই *nai*) are attached with the *Light Verbs* উঠনা *uthona* 'do not get up' ফেলনা *phelona* 'do not throw". Negative attachments are also considered in the present task while checking the suffixes of *Light Verbs* (*LVs*).

7.6.3.3 Identification of Lexical Scope for Complex Predicates (*CPs*)

The identification of lexical scopes of the Complex Predicates (*CPs*) from their successive sequences shows that multiple Complex Predicates (*CPs*) can occur in a long sequence. An automatic method is employed to identify the Complex Predicates (*CPs*) along with their lexical scopes. The lexical category or basic POS tags are obtained from the parsed sentences. If the compound and conjunct verbs occur successively in a sequence, the left most two successive tokens are chosen to construct the Complex Predicate (*CP*). If successive verbs are present in a sequence and the dictionary form of the second verb reveals that the verb is present in the lists of compound *Light Verbs* (*LV*), then that *Light Verb* (*LV*) may be a part of a compound verb (*CompV*). For that reason, the immediate previous word token is chosen and tested for its basic POS in the parsed result. If the basic POS of the previous word is "*verb* (*v*)" and any suffixes of either conjunctive participial form –এ/-ইয়া *-e/iya* or the infinitive form –তে/-ইতে *–te/ite* is attached to the previous verb, the two successive verbs are grouped together to form a compound verb (*CompV*) and the lexical scope is fixed for the Complex Predicate (*CP*).

If the previous verb does not contain –এ/-ইয়া *-e/iya* or –তে/-ইতে *–te/ite* inflections, no compound verb (*CompV*) is framed with these two verbs. But, the second *Light Verb* (*LV*) may be a part of another Complex Predicate (*CP*). This *Light Verb* (*LV*) is now considered as the *Full Verb* (*FV*) and its immediate next verb is searched in the list of compound *Light Verbs* (*LVs*) and the formation of compound verbs (*CompVs*) progresses similarly. If the verb is not in the list of compound *Light Verbs*, the search begins by considering the present verb as *Full Verb* (*FV*) and the search goes in a similar way.

The following examples are given to illustrate the formation of compound verbs (*CompVs*) and find the lexical scopes of the compound verbs (*CompVs*).

আমি চলতে গিয়ে পড়ে গেলাম ।

(*ami*) (*chalte*) (*giye*) (*pore*) (*gelam*).

I <fell down while walking>.

Here, "*chalte giye pore gelam*" is a verb group. The two left most verbs চলতে গিয়ে *chalet giye* are picked and the dictionary form of the second verb is searched in the list of compound *Light Verbs*. As the dictionary form (*jaoya* 'go') of the verb গিয়ে *giye* is present in the list of compound *Light Verbs* (as shown in Table 7.9), the immediate previous verb চলতে *chalte* is checked for inflections –এ/–ইয়া-*e/iya* or তে/ইতে *–te/ite*. As the verb চলতে *chalte* contains the inflection -তে -*te* , the verb group চলতে গিয়ে *chalte giye* is a compound verb (*CompV*) where গিয়ে *giye* is a *Light Verb* and চলতে *chalet* is the *Full Verb* with inflection (-তে -*te*). Next verb group, পড়ে গেলাম *pore gelam* is identified as compound verb (*CompV*) in a similar way (পড়্+ (-e) *por+ (-e)* + গেলাম*gelam* (*jaoya* 'go')). Another example is given as follows.

আমি উঠে পড়ে দেখলাম যে সে নেই ।

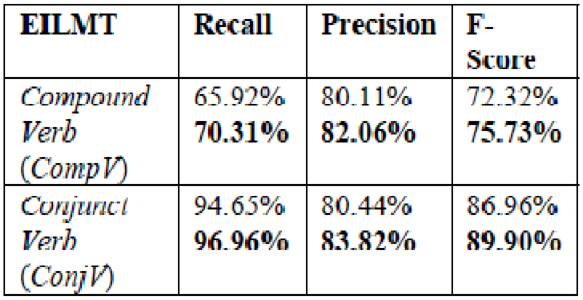
(*ami*) (*uthe*) (*pore*) (*dekhlam*) (*je*) (*tumi*) (*ekhane*) (*nei*)

I <get up and saw> that you are not here

Here, উঠে পড়ে দেখলাম *uthe pore dekhlam* is another verb group. The immediate next verb of উঠে *uthe* is পড়ে *pore* that is chosen and its dictionary form is searched in the list of compound *Light Verbs* (*LV*) similarly. As the dictionary form পড়া(pOra) of the verb পড়ে *pore* is present in the list of *Light Verbs* and the verb উঠে *uthe* contains the inflection -e *–e,* the consecutive verbs frame a compound verb (*CompV*) উঠে পড়ে where উঠে *uthe* is a *Full Verb* with inflection --এ *–e* and পড়ে *pore* is a *Light Verb*. The final verb দেখলাম *dekhlam* is chosen and as there is no other verb present, the verb দেখলাম *dekhlam* is excluded from any formation of compound verb (*CompV*) by considering it as a simple verb. Similar technique is adopted for identifying the lexical scopes of conjunct verbs (*ConjVs*). The method seems to be a simple pattern matching technique in a left-to-right fashion but it helps in case of conjunct verbs (*ConjVs*). As the noun or adjective occur in the first slot of conjunct verbs (*ConjVs*) construction, the search starts from the point of noun or adjective. If the basic POS of a current token is either "*noun*" or "*adjective*" and the dictionary form of the next token with the basic POS "*verb* (*v*)" is in the list of conjunct *Light Verbs* (*LVs*), then the two consecutive tokens are combined to frame the pattern of a conjunct verb (*ConjV*). For example, the identification of lexical scope of a conjunct verb (*ConjV*) from a sequence such as *uparjon korte gelam* 'earn-do-go' (went to earn) identifies the conjunct verb (*ConjV*) *uparjon korte*. There is another verb group *korte gelam* that seems to be a compound verb (*CompV*) but is excluded by considering *gelam* as a simple verb.

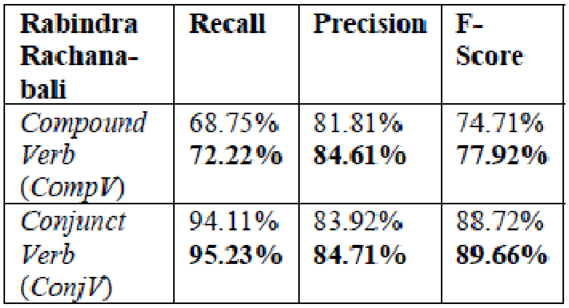
7.6.4 System Evaluation

The system is tested on 800 development sentences and finally applied on a collection of 500 sentences from each of the two Bengali corpora. As there is no annotated corpus available for evaluating Complex Predicates (*CPs*), the manual evaluation of total 1000 sentences from the two corpora is carried out in the present task. The *recall*, *precision* and *F-Score* are considered as the standard metrics for the present evaluation. The extracted Complex Predicates (*CPs*) contain compound verb (*CompV*) and conjunct verbs (*ConjVs*). Hence, the metrics are measured for both types of verbs individually. The separate results for two separate corpora are shown in Table 7.11 and Table 7.12 respectively.



EILMT	**Recall**	**Precision**	**F-Score**
Compound Verb (CompV)	65.92% **70.31%**	80.11% **82.06%**	72.32% **75.73%**
Conjunct Verb (ConjV)	94.65% **96.96%**	80.44% **83.82%**	86.96% **89.90%**

Table 7.11. *Recall*, *Precision* and *F-Score* of the system for acquiring the *CompVs* and *ConjVs* from EILMT Travel and Tourism Corpus.



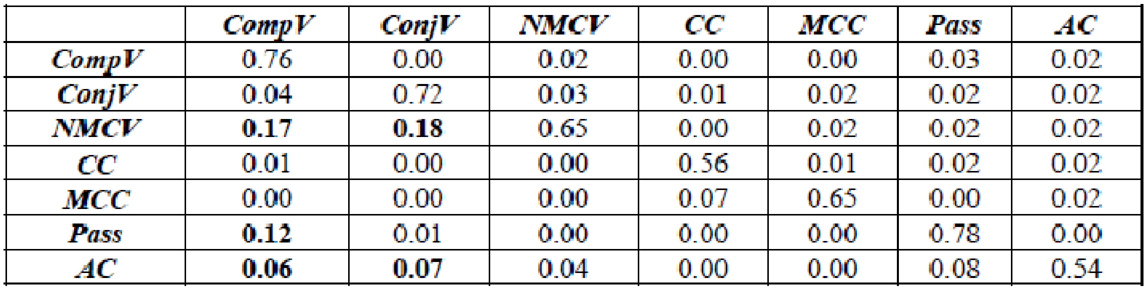
Rabindra Rachana-bali	**Recall**	**Precision**	**F-Score**
Compound Verb (CompV)	68.75% **72.22%**	81.81% **84.61%**	74.71% **77.92%**
Conjunct Verb (ConjV)	94.11% **95.23%**	83.92% **84.71%**	88.72% **89.66%**

Table 7.12 *Recall*, *Precision* and *F-Score* of the system for acquiring the *CompVs* and *ConjV*s from Rabindra Rachanabali corpus

The results show that the system identifies the Complex Predicates (*CPs*) satisfactorily from both of the corpus. In case of Compound Verbs (*CompVs*), the precision value is higher than the recall. The lower recall value of Compound Verbs (*CompVs*) signifies that the system fails to capture the other instances from overlapping sequences as well as non-Complex predicates (non-*CPs*). But, it is observed that the identification of lexical scopes of compound verbs (*CompVs*) and conjunct verbs (*ConjVs*) from long sequence of successive Complex Predicates (*CPs*) increases the number of Complex Predicates (*CPs*) entries along with compound verbs (*CompVs*) and conjunct verbs (*ConjVs*). The figures shown in bold face in Table 7.11 and Table 7.12 for the Travel and Tourism corpus and Short Story corpus of Rabindranath Tagore indicates the improvement of identifying lexical scopes of the Complex Predicates (*CPs*). In comparison to other similar language such as Hindi (Mukerjee *et al*. 2006) (the reported precision and recall are 83% and 46% respectively), our results (84.66% precision and 83.67% recall) are higher in case of extracting Complex Predicates (*CPs*). The reason may be of resolving the lexical scope and handling the morpho-syntactic features using shallow parser. In addition to *Non-MonoClausal Verb* (*NMCV*) or Serial Verb, the other criteria (Butt 1993; Paul 2004) are used in our present diagnostic tests to identify the complex predicates (*CPs*). The frequencies of *Compound Verb* (*CompV*), *Conjunct Verb* (*ConjV*) and the instances of other constraints of non Complex Predicates (non-*CPs*) are shown in Figure 2. It is observed that the numbers of instances of *Conjunct Verb* (*ConjV*), *Passives* (*Pass*), *Auxiliary Construction* (*AC*) and *Non-MonoClausal Verb* (*NMCV*) or Serial Verb are comparatively high than other instances in both of the corpus.

The error analysis is conducted on both of the corpus. Considering both corpora as a whole single corpus, the confusion matrix is developed and shown in Table 7.13. The bold face figures in Table 7.13 indicate that the percentages of non-Complex Predicates (non-*CPs*) such as

Non-MonoClausal Verbs (*NMCV*)*, Passive*s (*Pass*) and *Auxiliary Construction* (*AC*) that are identified as compound verbs (*CompVs*). The reason is the frequencies of the non-Complex Predicates (non-*CPs*) that are reasonably higher in the corpus. In case of conjunct verbs (*ConjVs*), the *Non-MonoClausal Verbs* (*NMCV*) and *Auxiliary Construction* (*AC*) occur as conjunct verbs (*ConjVs*). The system also suffers from clausal detection that is not attempted in the present task. The *Passive*s (*Pass*) and *Auxiliary Construction* (*AC*) requires the knowledge of semantics with argument structure knowledge.



	CompV	***ConjV***	***NMCV***	***CC***	***MCC***	***Pass***	***AC***
CompV	0.76	0.00	0.02	0.00	0.00	0.03	0.02
ConjV	0.04	0.72	0.03	0.01	0.02	0.02	0.02
NMCV	**0.17**	**0.18**	0.65	0.00	0.02	0.02	0.02
CC	0.01	0.00	0.00	0.56	0.01	0.02	0.02
MCC	0.00	0.00	0.00	0.07	0.65	0.00	0.02
Pass	**0.12**	0.01	0.00	0.00	0.00	0.78	0.00
AC	**0.06**	**0.07**	0.04	0.00	0.00	0.08	0.54

Table 7.13 Confusion Matrix for *CPs* and constraints of non-*CPs* (in %).

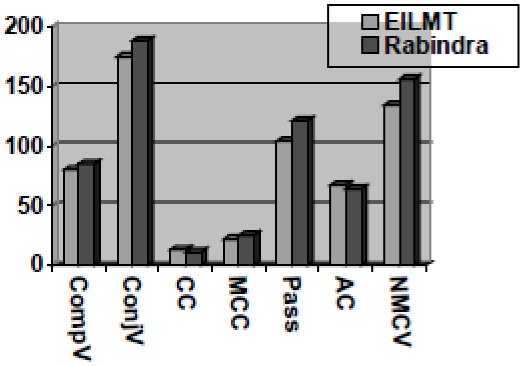


Figure 7.10 The frequencies of Complex Predicates (*CPs*) and different constrains of non-Complex Predicates (*non-CPs*).

7.6.5 Conclusion

In this paper, we have presented a study of Bengali Complex Predicates (*CPs*) with a special focus on compound verbs, proposed automatic methods for their extraction from a corpus and diagnostic tests for their evaluation. The problem arises in case of distinguishing Complex Predicates (*CPs*) from Non-Mono-Clausal verbs, as only the lexical patterns are insufficient to identify the verbs. In future task, the subcategorization frames or argument structures of the sentences are to be identified for solving the issues related to the errors of the present system.

Chapter 8

Measuring the Compositionality of Bigrams in English

8.1 Introduction

The measurement of relative compositionality of bigrams is crucial to identify Multi-word Expressions (MWEs) in Natural Language Processing (NLP) tasks. The article presents the experiments carried out as part of the participation in the shared task '*Distributional Semantics and Compositionality (DiSCo)*' organized as part of the *DiSCo* workshop in ACL-HLT 2011. The experiments deal with various collocation based statistical approaches to compute the relative compositionality of three types of bigram phrases (Adjective-Noun, Verb-subject and Verb-object combinations). The experimental results in terms of both fine-grained and coarse-grained compositionality scores have been evaluated with the human annotated gold standard data. Reasonable results have been obtained in terms of average point difference and coarse precision.

The present work examines the relative compositionality of Adjective-Noun (ADJ-NN; e.g., *blue chip*), Verb-subject (V-SUBJ; where noun acting as a subject of a verb, e.g., *name imply*) and Verb-object (V-OBJ; where noun acting as an object of a verb, e.g., *beg question*) combinations using collocation based statistical approaches. Measuring the relative compositionality is useful in applications such as machine translation where the highly non-compositional collocations can be handled in a special way (Hwang and Sasaki 2005).

Multi-word expressions (MWEs) are sequences of words that tend to co-occur more frequently than chance and are either idiosyncratic or decomposable into multiple simple words (Baldwin 2006). Deciding idiomaticity of MWEs is highly important for machine translation, information retrieval, question answering, lexical acquisition, parsing and language generation. *Compositionality* refers to the degree to which the meaning of a MWE can be predicted by combining the meanings of its components. Unlike *syntactic compositionality* (e.g. *by and large*), *semantic compositionality* is continuous (Baldwin 2006).

Several studies have been carried out for detecting compositionality of noun-noun MWEs using WordNet hypothesis (Baldwin et al. 2003), verb-particle constructions using statistical similarities (Bannard et al. 2003; McCarthy et al. 2003) and verb-noun pairs using Latent Semantic Analysis (Katz and Giesbrecht 2006).

Our contributions are two-fold: firstly, we experimentally show that collocation based statistical compositionality measurement can assist in identifying the continuum of compositionality of MWEs. Secondly, we show that supervised weighted parameter tuning results in accuracy that is comparable to the best manually selected combination of parameters.

8.2 Proposed Methodologies

The present task was to identify the numerical judgment of compositionality of individual phrase. The statistical co-occurrence features used in this experiment are described.

Frequency: If two words occur together quite frequently, the lexical meaning of the composition may be different from the combination of their individual meanings. The frequency of an individual phrase is directly used in the following methods.

Point-wise Information (PMI): An information-theoretic motivated measure for discovering interesting collocations is *point-wise mutual information* (Church and Hanks 1990). It is originally defined as the mutual information between particular events *X* and *Y* and in our case the occurrence of particular words, as follows:

$$PMI(x\ y) = \log_2 \frac{P(x,y)}{P(x).P(y)} \approx \log_2 \frac{NC(x,y)}{C(x).C(y)} \tag{8.1}$$

PMI represents the amount of information provided by the occurrence of the event represented by *X* about the occurrence of the event represented by *Y*.

T-test: T-test has been widely used for collocation discovery. This statistical test tells us the probability of a certain constellation (Nugues, 2006). It looks at the mean and variance of a sample of measurements. The null hypothesis is that the sample is drawn from a distribution with mean. T-score is computed using the equation (8.2):

$$t(x,y) = \frac{mean\big(P(X,Y)\big) - mean\ (P(X))mean(P(Y))}{\sqrt{\big(\sigma^2 P(X,Y)\big) + \sigma^2\big(P(X)\big)\sigma^2\ (P(Y))}}$$

$$\approx \frac{C(X,Y) - \frac{C(X)C(Y)}{N}}{\sqrt{C(X,Y)}} \tag{8.2}$$

In both the equations (1) and(2), *C(x)* and *C(y)* are respectively the frequencies of word *X* and word *Y* in the corpus, *C(X,Y)* is the combined frequency of the bigrams <X Y> and N is the total number of tokens in the corpus. Mean value of *P(X,Y)* represents the average probability of the bigrams <X Y>. The bigram count can be extended to the frequency of word X when it is followed or preceded by Y in the window of K words (here K=1).

Perplexity: Perplexity is defined as $2^{H(X)}$

$$2^{H(X)} = 2^{-\sum_x P(x)\log_2 P(x)} \tag{8.3}$$

where *H(X)* is the cross-entropy of *X*. Here, *X* is the candidate bigram whose value is measured throughout the corpus. Perplexity is interpreted as the average "branching factor" of a word: the statistically weighted number of words that follow a given word. As we see from equation (4), Perplexity is equivalent to entropy. The only advantage of perplexity is that it results in numbers more comprehensible for human beings. Here, perplexity is measured at both root level and surface level.

Chi-square test: The t-test assumes that probabilities are approximately normally distributed, which may not be true in general (Manning and Schütze, 2003). An alternative test for dependence which does not assume normally distributed probabilities is the χ^2-test (pronounced "chi-square test"). In the simplest case, this 2 test is applied to a 2-by-2 table as shown below:

	X = *new*	X ≠ *new*
Y= *companies*	n_{11} (*new companies*)	n_{12} (e.g.,*old companies*)
Y ≠ *companies*	n_{21} (e.g., *new machines*)	n_{22} (e.g., *old machines*)

Table 8.1: A 2-by-2 table showing the dependence of occurrences of *new* and *companies*

Each variable in the above table depicts its individual frequency, e.g., n_{11} denotes the frequency of the phrase "new companies".

The idea is to compare the observed frequencies in the table with the expected frequencies when the words occur independently. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence. The equation for this test is defined below:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{8.4}$$

$$\text{where } O_{ij} = \frac{\sum_k n_{ik}}{N} \times \frac{\sum_k n_{kj}}{N} \times N$$

N is the number of tokens in the corpus.

8.3 Used Corpora and Dataset

The system has used the WaCkypedia_EN[1] corpora (discussed in Section 4.1.2) which are a 2009 dump of the English Wikipedia (about 800 million tokens). The corpus was POS-tagged and lemmatized followed by full dependency parsing. The total number of candidate items for each relation type extracted from the corpora is: ADJ-NN (144, 102), V-SUBJ (74, 56), V-OBJ (133, 96). The first number within brackets is the number of items with fine-grained score, while the second number refers to the number of items with coarse grained score. These candidate phrases are split into 40% training, 10% validation and 50% test sets. The training data set consists of three columns: relation (e.g., EN_V_OBJ), phrase (e.g., *provide evidence*) and judgment score (e.g. "38" or "high"). Scores were averaged over valid judgments

[1] http://wacky.sslmit.unibo.it/

per phrase and normalized between 0 and 100. These numerical scores are used for the Average Point Difference score. For coarse-grained score, phrases with numerical judgments between 0 and 33 as “low”, 34 to 66 as “medium” and 66 and over got the label "high".

8.4 System Architecture

The candidate items for each relation type are put in a database. For each candidate, all the statistical co-occurrence feature values like frequency, PMI, T-test, Perplexity (root and surface levels) and Chi-square tests are calculated. The final fine-grained scores are computed as the simple average and weighted average of the individual statistical co-occurrence scores. Another fine-grained score is based on the T-test score that performed best on the training data. Coarse-grained scores are obtained for all the three fine-grained scores.

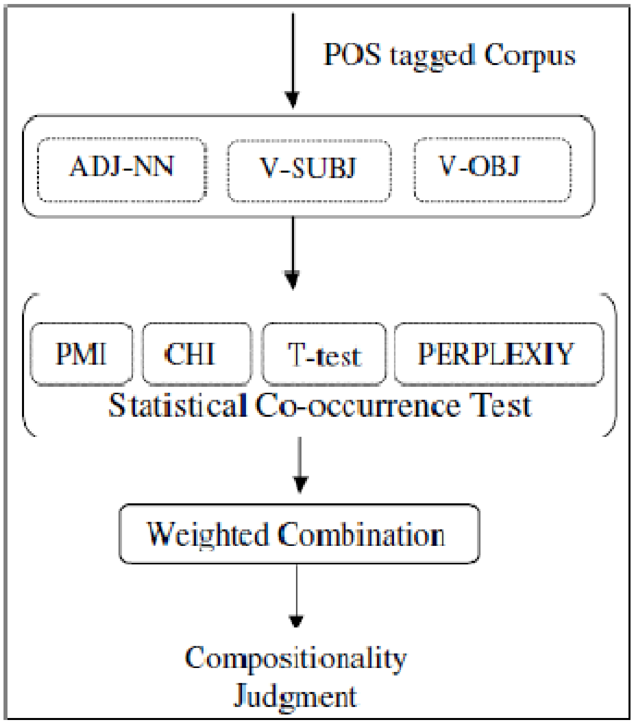


Figure 8.1: System Architecture

8.5 Weighted Combination

The validation data is used as the development data set for our system. The weighted average of the individual statistical co-occurrence scores is calculated by assigning different weights to each co-occurrence feature score. The weights are calculated from the training data using the average point difference error associated with the co-occurrence feature. The feature which gives minimum error score is assigned the higher weight. For each co-occurrence feature score i, if the error on the training data is e_i, the weight W_i assigned to the co-occurrence feature score i is defined as:

$$W_i = \frac{100 - e_i}{\sum_i (100 - e_i)} \tag{8.5}$$

The individual co-occurrence feature scores are normalized to be in the range of 0 to 1 before calculating the weighted sum.

Note that, when measuring coarse-precision, the fine-grained scores are bucketed into three bins as explained in Section 8.3.

Errors	**PMI**	**T test**	**Perx-Root**	**Perx-Surface**	**chi square**	**Average**	**Weighted Average**
APD	29.35	**24.25**	35.23	31.4	36.57	**21.22**	**21.20**
CP	0.31	**0.60**	0.48	0.42	0.45	**0.57**	**0.62**

Table 8.2 Evaluation results on different approaches on validation data

8.6 Evaluation Metrics

The system output is evaluated using the following evaluation metrics:

- **Average Point Difference (APD):** the mean error (0 to 100) is measured by computing the average difference of system score and test data score. The minimum value implies the minimum error and the maximum accuracy of the system.
- **Coarse Precision (CP):** the test data scores are binned into three grades of compositionality (non-compositional, somewhat compositional, and fully-compositional), ordering the output by score and optimally mapping the system output to the three bins.
- **Spearman's rho coefficient:** it is used to estimate strength and direction of association between two ordinal level variables (i.e., gold standard results and system results). It can range from -1.00 to 1.00.
- **Kendall’s tau rank coefficient:** it is a measure of rank correlation, i.e., the similarity of the orderings of the gold standard results and the system results. This coefficient must be in the range from -1 (complete disagreement) to 1 (complete agreement).

System	Spearman rho	Kendall's Tau	Average Point Difference (APD)				Coarse Precision (CP)			
			ALL	ADJ-NN	V-SUBJ	V-OBJ	ALL	ADJ-NN	V-SUBJ	V-OBJ
Baseline	0.20	0.20	32.82	34.57	29.83	32.34	0.297	0.288	0.300	0.308
RUN-1	**0.33**	**0.23**	**22.67**	**25.32**	**17.71**	**22.16**	0.441	0.442	0.462	0.425
RUN-2	0.32	0.22	22.94	25.69	17.51	22.60	0.458	0.481	0.462	0.425
RUN-3	-0.04	-0.03	25.75	30.03	26.91	19.77	**0.475**	**0.442**	**0.346**	**0.600**

Table 8.3: Overall System results on test set

8.7 Experimental Results

The system has been trained using the training data set with their fine-grained score. The evaluation results on the validation set are shown in Table 8.2. It is observed that T-test gives the best results on the validation data set in terms of precision. Based on the validation set results, three procedural approaches are run and three results are reported on the test data.

RUN-1 (Weighted Combination): These results are obtained from the weighted combination of individual scores. Both the perplexity measures are not useful to make significant gain over the compositionality measure. For the rank combination experiments, the best co-occurrence measures, i.e., PMI, Chi-square and T-test are considered. For the weighted combination, the results are reported for the weight triple (0.329, 0.309, 0.364) for PMI, Chi-square and T-test respectively.

RUN-2 (Average Combination): These results are reported by simply averaging the values obtained from the five measures.

RUN-3 (Best Scoring Measure: T-test): The T-test results are observed as the best scoring measure used in this experiment.

When calculating the coarse-grained score the compositionality of each phrase is tagged as *'high'*, *'medium'* or *'low'* discussed in Section 8.3.

The final test data set has been evaluated on the gold standard data developed by the organizers and the results on the three submitted runs are described in Table 8.3. The positive value of Spearman's rho coefficient implies that the system results are in the same direction with the gold standard results; while the Kandell's tau indicates the independence of the system value with the gold standard data. As expected, Table 8.3 shows that the weighted average score (Run 1) gives better accuracy for all phrases based on the APD scores. On the other hand, the T-test results (Run 3) give high accuracy for the coarse precision calculation while it is in the last position for ADP scores.

8.8 Conclusion

We have demonstrated that the statistical evidences can be useful to indicate the continuum of compositionality of the bigrams i.e. adjectivenoun, verb-subject and verb-object combination. We are extremely confident with these empirical approaches to a semantic measure as compositionality directly relates to the semantics of a phrase. The coarse precision can be improved if three ranges of numerical values can be tuned properly and the size of the three bins can be varied significantly. As our future task, we can use other statistical collocation-based methods (e.g. Log-likelihood ratio, Relative frequency ratios etc.). Furthermore, we will plan to incorporate standard lexical resources like WordNet, VerbNet and use their lexical ontology to enhance the compositionality judgment of the collocations.

Chapter 9

Applications of Multiword Expressions

9.1 Authorship Identification and Stylometry Analysis

9.1.1 Introduction

Stylometry, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and belongs to the core task of Text categorization that involves authorship identification, plagiarism detection, forensic investigation, computer security, copyright and estate disputes etc. In this work, we present a strategy for Stylometry detection of documents written in Bengali. We adopt a set of fine-grained attribute features with a set of lexical markers for the analysis of the text and use three semi-supervised measures for making decisions. Finally, a majority voting approach has been taken for final classification. We also try our experiment using Conditional Random Filed (CRF). The system is fully automatic and language-independent. Evaluation results of our attempt for

Bengali author's Stylometry detection show reasonably promising accuracy in comparison to the baseline model.

9.1.2 Stylometry Analysis

Stylometry is an approach that analyses text in text mining e.g. novels, stories, dramas written by authors, trying to measure the author's style, rhythm of his pen, subjection of his desire, prosody of his mind by choosing some attributes that are consistent throughout his writing and play the linguistic fingerprint of that author. In other words, stylometry is the application of the study of linguistic style, usually with reference to written text that concerns the way of writing rather than its contents. Computational Stylometry is focused on subconscious elements of style less easy to imitate or falsify.

Stylistic analysis that has been done by Croft (1981) claimed that for a given author, the habits "*of style*" are not affected "*by passage of time, change of subject matter or literary form. They are thus stable within an authors writing, but they have been found to vary from one author to another*" (Mustafa, Mustapha, Azmi and Sulaiman, 2010). However, stylometric authorship attribution can be considered as a typical clustering, classification and association rule problem, where a set of documents with known authorships are used for training and the aim is to automatically determine the corresponding author of an anonymous text, but the way of selecting the appropriate features is not focused in that sense and vary from one research to other.

Most of the authorship identification studies are better at dealing with some closed questions like (i) who wrote this, A or B, (ii) if A wrote these, did he also writes this, (iii) how likely is it that A wrote this etc. The main target in this study is to build a decision making system that enables users to predict and to choose the right author from an anonymous author's novel under consideration, by choosing various lexical, syntactic, analytical features known as ***stylistic markers***. The system uses three semi-supervised, reference based measurements (Cosine-similarity, Chi-square measurement and Euclidean distance) which behave as an expert opinion to map the testing documents to the appropriate authors. Without focusing much on the distributional lexical measures like vocabulary richness or frequency of individual word counts, we mainly focus on some low-level measures (sentence count, word count, punctuation count, length of words and sentences etc.), phrase level measures (noun chunk, verb chunk, etc.) and context level measures (number of dialog, length of dialog, sentence structure analysis etc.). Additionally, we propose a baseline system for Bengali Stylometry analysis using ***vocabulary richness function***. The present attempt basically deals with the microscopic observation for the stylistic behaviours of the articles written by the famous novel laureate Rabindranath Tagore long years back and tries to disambiguate them from the anonymous articles written by some other authors in that period.

9.1.3 Related Work on Stylometry Analysis

Stylometry, which may be considered as an investigation of "Who was behind the keyboard when the document was produced?" or "Did Mr. X wrote the document or not?" is a long term study mainly in forensic investigation department that started from late Nineties. In the past, where Stylometry emphasized the rarest or most striking elements of a text, contemporary techniques can isolate identifying patterns even in common parts of speech. The pioneering study on authorship attributes identification using word-length histograms appeared at the very end of nineteen century (Malyutov 2006). After that, a number of studies based on content analysis (Krippendorf 2003), computational stylistic approach (Stamatatos, Fakotakis and Kokkinakis, 1999), exponential gradient learn algorithm (Argamon, Saric and Stien, 2003), Winnow regularized algorithm (Zhang, T., Damerau, F., Johnson, 2002), Support Vector Machine based approach (Pavelec, Justino and Oliveira, 2007) etc have been proposed for various languages like English and Portuguese. Recently, research has started to focus on authorship attribution on larger sets of authors: 8 (Halteren 2005), 20 (Argamon, Saric and Stien, 2005), 114 (Madigan, David, Alexander Genkin, Lewis, Argamon, Fradkin and Ye, 2005), or up to thousands of authors (Moshe, Schler, and Elisheva, 2007). The use of computers regarding the extraction of Stylometrics has been limited to auxiliary tools (i.e. simple program for counting user-defined features fast and reliably). Hence, authorship attribution studies so far may be looked like *computer-assisted*, not *compute-based*. As a beginning of Indian language Stylometry analysis, our research does not consider any manual intervention for extracting features (like identification of some high frequent start-up words), moreover we have dealt with a number of large-size non-homogeneous texts since they are composed of dialogues, narrative parts etc and try to build a language and text-length independent system for attribute analysis.

9.1.4 Inference of Fine-grained Attributes for Stylometry Detection

The methodology used in this work generally depends on the combination of 76 fine-grained style-markers for feature engineering and three semi-supervised approaches for decision making. As an initial attempt, we have decided to work with the simple approach like statistical measurement, analyze the drawbacks and further go beyond for working with other machine learning or hybrid approaches. Furthermore, the reasons for not attempting with the methods described in the related work section are as follows: the content analysis is one of the earliest types of computations, also for exponential and Winnow algorithms as both are purely mathematical models and the SVM based method has a strong affinity to the language for which the system is designed. Currently, authorship attribute studies are dominated by the use of lexical measures. In a review paper (Holmes 1994), the author asserted that:

" ..... yet, to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than based on lexical items."

For this reason, in order to set a baseline for the evaluation of the proposed method, we have decided to implement a lexical based approach called *vocabulary richness*. Detailed discussion about the baseline system and our approach are mentioned in the next section.

9.1.4.1 Proposed Methodology Design

As mentioned, the proposed stylistic markers used in this study take full advantage of the analysis of the distributed contextual clues as well as full analysis by Natural Language Processing tools. The system architecture of the proposed Stylometry Detection system is shown in Figure 9.1. In this section, we first describe brief properties of different components of the system architecture and then the set of stylistic features is analytically presented. Finally the classification methods are elaborated with brief description of their functionalities.

1. Textual Analysis: Basic pre-processing before actual textual analysis has been done so that stylistic markers are clearly viewed for further analysis by the system. Token-level markers discussed in the next Section, are extracted from the preprocessed corpus. Then parsing using Shallow parser1 has been done to separate the sentence and the chunk boundaries and parts-of-speech. From this parsed text, chunk-level and context-level markers are identified.

2. Stylistic Features Extraction: Stylistic features have been proposed as more reliable style markers than for example, word-level features since they are not under the conscious control of the author. To allow the selection of the linguistic features rather than n-gram terms, robust and accurate text analysis tools such as lemmatizers, part-of-speech (POS) taggers, chunkers etc. are needed. We have used the Shallow parser, which gives a parsed output of a raw input corpus. It tokenizes the input, performs a part-of-speech analysis, looks for chunks and inflections and a number of other grammatical relations. The stylistic markers which have been selected in this experiment are coarsely classified into three categories and discussed in the Table 9.1. Sentences are detected using the sentence boundary markers mainly *‘dari’* or *‘viram’* (‘_ ’), question marks (‘?’) or exclamation notation (‘!’) in Bengali. Sentence length and word count are the traditional and well-defined measures in authorship attribute studies and punctuation count is another interesting characteristics of the personal style of a writer. Chunk or phrase level markers are indications of various stylistic aspects, e.g., syntactic complexity, formality etc. Out of all detected chunk sets, mainly nine chunk types have been considered in this experiment. They are noun chunk (NP), verb-finite chunk (VGF), verb-non-finite chunk (VGNF), gerunds (VGNN), adjective chunk (JJP), adverb chunk (RBP), conjunct phrase (CCP), chunk fragment (FRAGP) and others (OTHERS). Shallow parser identifies 25 Part Of Speech (POS) categories. Among them, 24 POSs have been taken into consideration except UNK. Words tagged with UNK are unknown words and are verified by Bengali monolingual dictionary. Since Shallow parser is an automated text-processing tool, the style markers of the above levels are measured approximately. Depending on the complexity of the text, the provided measures may vary from real values which can only be measured using manual intervention. Making the system fully automated, the system performance depends on the performance of the parser. As we can see in the Table 9.1 that each marker is defined as a percentage measure of the ratio of two relevant measures, this approach was followed in order to work with text-length independent style markers as possible. However, it is worth noting that we do not claim that the proposed set of 76 markers is the final one. It could be possible to split them into more fine-grained measures e.g. F21 can be split into separate measures i.e. individual occurrence of the punctuation symbols (comma per word, colon per word, dari per word etc.). Here, our goal is to make an attempt towards the investigation of Bengali author’s writing style and to prove that an appropriately defined set of such style markers performs better than the traditional lexical based approaches.

Coarse-grained Classification	Stylistic Markers	Description	Total
Token-level	F1 to F10	Word length (1 to 9 and above) in %	10
	F11 to F20	Words per sentence (0-10, 10-20 and so on, up to 80-90 and above) in %	10
	F21	Punctuations per word in %	1
	F22 to F31	Individual punctuations in % (10 punctuations)	10
Chunk-level	F32 to F40	Detected NP, VGF, VGNF, VGNN, JJP, RBP, CCP, FRAGP, OTHER out of total chunks in %	9
	F41 to F49	Average words included in all above mentioned chunks in %	9
	F50 to F73	Individual percentage of detected POS (24) by Shallow parser	24
Context-level	F74	Average words per dialog in %	1
	F75	Words not included in the dictionary including Named-Entities in %	1
	F76	Hapax-legomena count out of all words in %	1

Table 9.1 Fine-grained stylistic features

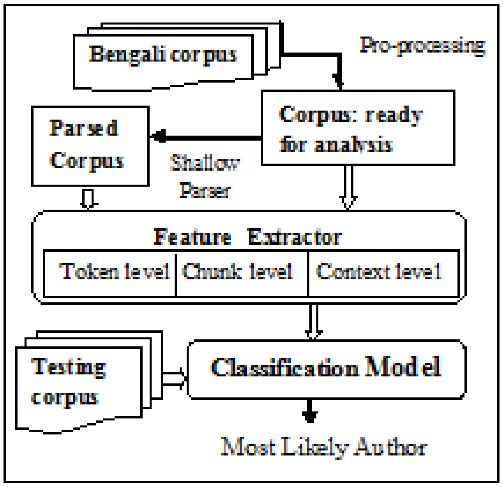


Figure 9.1 System Architecture of the Stylometry Detection System

3. Classification Model: A number of discriminative models based on statistical and machine learning measures, such as Bayesian Network, decision trees, neural networks, support vector machines, K-nearest neighbor approach etc. are available for text categorization. In this experiment, three semi-supervised, reference-based classification models have been used: (1) Cosine-similarity measurement, (2) Chi-square measure and (3) Euclidean distance. These are briefly discussed below.

Cosine-similarity measurement: Cosine-similarity is a measure of similarity between two vectors of $\boldsymbol{n}$ dimensions by finding the cosine of the angle between them, often used to compare documents in text mining. Given two vectors of attributes, $\boldsymbol{R}$ and $\boldsymbol{T}$, the cosine similarity, $\boldsymbol{\theta}$ is represented using a dot product and magnitude as:

$$Similarity = \cos(\theta) = \frac{R.T}{|R|.|T|} = \frac{\sum_{i=1}^{n} r_i.t_i}{\sqrt{\sum_{i=1}^{n} r_i^2} * \sqrt{\sum_{i=1}^{n} t_i^2}} \tag{9.1}$$

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence and in-between values indicating intermediate similarity or dissimilarity. Here, n is the number of features (i.e., 76) that act as dimensions of the vectors and r_i and t_i are the features of reference and test vectors respectively.

Chi-square measure*:** Chi-square is a statistical test commonly used to compare observed data with the expected data according to a specific hypothesis. That is, chi-square (χ^2) is the sum of the squared differences between observed (O***) and the expected (***E***) data (or the deviation, ***d***), divided by the expected data in all possible categories.

$$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i} \tag{9.2}$$

Here, the mean of each cluster is used as the observation data for that cluster and used as reference ***O***. ***n*** is the number of features and $\boldsymbol{O_i}$ is the observed value of the i^th^ feature. The Chi Square test gives a value of χ^2 that can be converted to Chi Square ($\boldsymbol{c^2}$) using chi-square table which is an n×n matrix with row representing the degree of freedom (i.e., difference between number of rows and columns of the contingency matrix) and column representing the probability we expect. This can be used to determine whether there is a significant difference from the null hypothesis or whether the results support the null hypothesis. After comparing the chi-squared value in the cell with our calculated χ^2 value, if the χ^2 value is greater than the 0.05, 0.01 or 0.001 column, then the goodness-of-fit null hypothesis can be rejected, otherwise accepted.

Euclidean distance: The Euclidean distance between two points, ***p*** and ***q*** is the length of the line segment. In Cartesian coordinates, if $\boldsymbol{p} = (p_1, p_2... p_n)$ and $\boldsymbol{q} = (q_1, q_2,..., q_n)$ are two points in Euclidean n-space, then the distance from ***p*** to ***q*** is given by:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{9.3}$$

where, ***n*** is the number of features or dimension of a point, ***p*** is the reference point (i.e. mean vector) of each cluster and ***q*** is the testing vector. For every test vector, three distances from three reference points have been calculated and smallest distance defines the probable cluster.

9.1.4.2 Experimental Details

• Corpus Acquisition

Aforementioned, Resource acquisition is one of the most important challenge to work with resource constrained languages like Bengali. The system has used thirty stories in Bengali written by the noted Indian Nobel laureate Rabindranath Tagore (discussed in section 4.1.1). Among them, we have selected twenty stories for training purpose and rest for testing. We choose this domain for the reason that in such writings the idiosyncratic style of the author is not likely to be overshadowed by the characteristics of the corresponding text genre. To differentiate them from other author's articles, we have selected 30 articles from author A and 30 articles of other authors[1]. In this way, we have three clustered set of documents identified as articles of Author R (Tagore's articles), Author A and others (O). This paper focuses on two topics: (i) the effort of earlier works on feature selection and learning and (ii) the effort of limited data in authorship detection.

• Baseline System

In order to set up a baseline system, we proposed traditional lexical based methodology called ***vocabulary richness***. Among the various measures like Yule's K measure, Honore's R measure, we have taken most typical one as the type-token ratio ***(V/N)*** where ***V*** is the size of the

[1] http://banglalibrary.evergreenbangla.com

vocabulary of the sample text and ***N*** is the number of tokens which forms the simple text. We have gathered dimensional features of the articles of each cluster and averaged them to make a mean vector for every cluster. So these three mean vectors indicate the references of three clusters respectively. Now, for every testing document, similar features have been extracted and a test vector has been developed. Now, using Nearest-neighbour algorithm, we have tried to identify the author of the test documents. The results of the baseline system are shown using confusion matrix in Table 9.2. Each row shows classification of the ten texts of the corresponding authors. The diagonal contains the correct classification. The baseline system achieves 37% average accuracy. Approximately 60% of average accuracy error (for author A and O) is due to the wrong identification of the author as Author R.

Baseline System				
	R	A	O	e (Error)
R	6	0	4	0.40
A	7	2	1	0.80
O	5	2	3	0.70
Average error				0.63

Table 9.2 Confusion matrix of our system

9.1.4.3 Performance of Our System

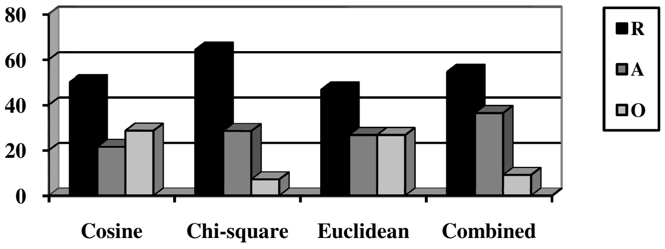
We have discussed earlier that our classification model is based on three statistical techniques. A voting approach combining the decision of the three models for each test document have also been measured for expecting better results. The confusion matrix is shown in Table 9.3. This table shows that Chi-square measure has relatively less error (46%) rate compared to other measures. A majority voting technique has an accuracy rate of 63% which is relatively better than others. In the case when the three statistical techniques produce different results, the result of Chi-square measure has been taken as correct result because it has given more accuracy compared to the others when measured individually.

Our System																
	Cosine-similarity				**Chi-square measure**				**Euclidean distance**				**Combined voting**			
	R	A	O	e	R	A	O	e	R	A	O	e	R	A	O	e
R	5	2	3	0.5	7	3	0	0.3	6	2	2	0.4	8	2	0	0.2
A	3	6	1	0.4	5	4	1	0.6	4	4	2	0.6	4	5	1	0.5
O	4	1	5	0.5	4	1	5	0.5	3	2	5	0.5	2	2	6	0.4
Average error				0.46	Average error			0.46	Average error			0.5	Average error			0.37

Table 9.3 Confusion matrix of our system

9.1.4.3 Discussion

Form the experimental results, it is clear that statistical approaches show nearly similar performance and accuracies of all of them are around 50%. Also the major sources of the errors are for the inappropriate identification of author as Author R. From the figure 9.2, we can see that the system looks little bit biased towards the identification of Rabindranath Tagore as author of the test documents. In all cases, the bar graphs for Author R are higher than others. The reason behind this is the acquisition of resources. Developing appropriate corpus for this study is itself a separate research area and takes huge amount of time. Furthermore, the collected articles from other authors are heterogeneous and not domain constrained. Our studies will be planned to focus on the identification of the unpublished articles of Rabindranath Tagore. For this, more microscopic observation in various fields of his writings will be needed. Here we only try our experiments on the stories of the writer. The success of the system lies not on the correct mapping of the articles



Figur 9.2 Error analysis of different approachs

to their corresponding three authors but to filter all the inventions of Rabindranath Tagore from a bag of documents and the more the accuracy of the filtering, the more the accuracy of the system. Apart from being the first work of its kind for Bengali language, the contributions of this experiment can be identified as: (i) application of statistical approach in the field of Stylometry, (ii) development of classification algorithm in n-dimensional vector space, (iii) developing a baseline system in this field and (iv) more importantly, working with the great writings of Rabindranath Tagore to reveal his swinging of thought and dexterity of pen when writing articles.

9.1.4.5 Conclusion

This paper introduced the use of a large number of fine-grained features for Stylometry detection. The presented methodology can also be used in author verification task i.e. the verification of the hypothesis whether or not a given person is the author of the text under study. The methodology can be adopted for other languages since maximum of the features are language independent. The classification is very fast since it is based on the calculation of some simple statistical measurements. Particularly, it appears from our experiments that texts with less word are less likely to be classified correctly. For that, our system is little biased towards the stylometry of Rabindranath Tagore. It is due to the lack of the large number of resources of other authors under study. However from this preliminary study, future works are planned to increase the database with more fine grained features and to identify more context dependent attributes for further improvement.

9.1.5 Authorship Identification Using CRF

In this experiment, we have tried to adapt a system beyond the conventional approaches of Stylometry detection. For this, Conditional Random Field (CRF) model has been introduced very first time in this field of application and features have been selected after microscopic investigation of the contextual information of the documents. We have also proposed a statistical approach for comparing the results with the previous one. We have used the articles which were written by the famous novel laureate Rabindranath Tagore long years back and try to dissimilate them from the anonymous articles written by other authors at that period of time. In the rest of the article, we use the expressions *Stylometry detection* and *authorship identification* interchangeably to express their anonymous senses. Experimental results indicate that the CRF model can enhance the task of identifying the authors.

9.1.5.1 System Architecture

Figure 9.3 shows the process of the CRF-based authorship identification. The implementation carries out in the following steps.

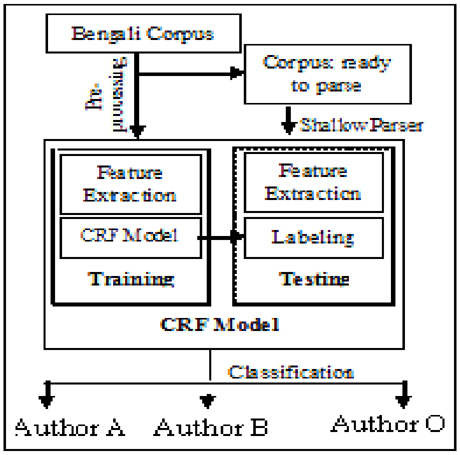


Figure 9.3 Proposed system architecture

- **Preprocessing and Parsing**

The documents are row and so unformatted that an initial cleaning is required before CRF model training. We must transfer the document into the formatted sequences, i.e. a bag of words or phrases of the document. From the pre-processed document, token-level and some of the context-level features are extracted. For a new document, we conduct the sentence segment, POS tagging using Shallow Parser so that the stylistic features are easily viewed to the system. From this, chunk-level and context-level markers are identified.

- **CRF Model:**

The input is the feature vector discussed in Table 9.4. There are three kinds of features, i.e. (1) token-level features, (2) phrase-level features and (3) context-level features. Token-level features include length of the word, number of keywords, starting word of a dialog maximum time present, count of hapax legomena. Phrase-level include count of POSs, chunks those we have considered here (not all POSs or chunks, the Shallow parser generally gives, are considered). Average length of the paragraph and length of the dialog are included in context-level features. Detected sentences are the sentence boundary ended mainly with *'dari'* ('।'), question marks ('?') or exclamation notation ('!') in Bengali. Sentence-length, word-count are the traditional and well-defined measures in authorship attribute studies and punctuation count is the very interesting characteristics of the personal style of a writer. Problem occurs to identify keywords as there is no standard tool to extract keywords for Bengali documents. For this, we have identified top ten high frequent words (excluding stop-words in Bengali) for every cluster using TF*IDF method which act as the list of keywords of that cluster corresponding to that author. Now, similarly, we have extracted a list of top ten high frequent words from every testing document and intersect them with the keywords of cluster1, cluster 2 and cluster 3 which are the count of the features KW1, KW2, KW3 respectively. Since Shallow parser is an automated text-processing tool, the style markers of the above levels are measured approximately. Depending on the complexity of the text, the provided measures may vary from real values which can only be measured using manual intervention. Making the system fully automated, the system fully believes on the performance of the parser for the extraction of all POS and chunk level features. The last column of the training feature file is labeled as R, A or O which is the indication of the three authors and for testing, all are labeled as X which is an arbitrary word indicating unknown author. CRF adds an extra column at the last position which indicates the label of the author for that document (R, A or O). As we can see that maximum of these features are the ratio of two relevant measures, this approach was followed in order to achieve as text-length independent style markers as possible. However, it is worth noting that we do not claim that the proposed set of features is the final one. It could be possible to split them into more fine-grained measures. Here, our goal is to make a pioneer approach towards the investigation of Bengali author's writing style and to prove that an appropriately defined set of such style markers performs better than the traditional lexically-based approaches.

No.	Features	Explanations	Normalization
1	Doc	Name of the document	-
2	Len_w	Average length of the word	Avg. len (word)/ Max_len of word
3	Len_d	Average length of the dialog	Avg. words per dialog / no. of sentences
4	Len_p	Average length of the paragraph	Avg. sentences per paragraph / no. of sentences
5	Punc	No. of punctuations	count (punc.) / no. of word
6	Chunk_N	Detected Noun phrase	count (NP) / no. of all chunk count
7	Chunk_V	Detected Verb phrase	count (VP) / no. of all chunk count
8	Chunk_CCP	Detected conjunct phases	count (CCP) / no. of all chunk count
9	POS_U	Detected unknown word	count (unknown)/ count (word)
10	POS_RE	Detected reduplication and echo-word	count (RDP+ECH) / count (word)
11	KW1	Intersection of the keywords of cluster 1 and the tested document	| KW (doc) ∩ KW (cluster 1)| / no. of KW in cluster 1
12	KW2	Intersection of the keywords of cluster 2 and the tested document	| KW (doc) ∩ KW (cluster 2)| / no. of KW in cluster 2
13	KW3	Intersection of the keywords of cluster 3 and the tested document	| KW (doc) ∩ KW (cluster 3)| / no. of KW in cluster 3
14	Start	Starting word (stemming form) of the dialog which is present in maximum time	_
15	Hapax Legomena	No. of words with frequency = 1, including named-entities	count (Hapax legomena)/count (word)

Table 9.4 Features in the CRF model

- **Used Corpus**

The corpus used in this experiment is discussed in Section 9.4.1.2 of the ‘Corpus acquisition’ subsection.

- **Baseline System**

Same baseline system as discussed in Section 9.4.1.2 of the 'Baseline System' subsection is also used hare.

9.1.5.2 Performance of Statistical Model

As aforementioned, we are dealing with two ways of classification model in Natural Language Processing i.e. statistical and machine learning approaches to make a comparative study among them to show which is more accurately performed in Stylometry detection task. Same features are used in this methodology and they act as the dimensions of the vector. After grouping the documents into three clusters, we have made a reference vector (mean vector) individually for every cluster which performs as the representatives for that cluster. For every test document, the cosine-similarity measures are performed with the three reference vectors and the document is assigned to that cluster with which, the similarity is higher. The first half of the confusion matrix named as "Cosine-similarity" in Table 9.4 depicts the results of this measure. As we can notice from the table that the accuracy of the statistical measure is 54% which is far better than the traditional baseline system and the articles of Rabindranath Tagore are identified more perfectly than others. It may be because of the resource acquisition of the corpus of Rabindranath Tagore is homogeneous in nature, where as for other authors, it has not been possible to collect same length corpus and sometime the collected corpus are of different domains.

	Cosine-similarity				**CRF**			
	R	***A***	***O***	***e***	***R***	***A***	***O***	***e***
R	6	2	2	0.4	7	1	2	0.3
A	2	5	3	0.5	3	5	2	0.5
O	4	2	5	0.5	3	1	6	0.4
Average error				0.46	Average error			0.40

Table 9.5 Confusion matrix for both measures

9.1.5.3 Performance of CRF based Modal

Performance of the CRF model for authorship identification is shown in the second half of the Table 9.4 named as "CRF". The average accuracy of this system is 60% which shows a tremendous improvement in comparison with the baseline system. The identification of the documents of author A is more or less same with the previous statistical approach and 30% of the error for Author A and Author O have been occurred for wrong identification of the author as Author R. This shows a little biasness of the system to the Stylometry of Tagore's writing.

9.1.5.3 Conclusion

Conditional Random Field is a state-of-the-art sequence modeling approach, which can use the features of the documents more sufficiently and effectively. In this experiment, we have studied in Bengali corpus to detect the stylistic features of the anonymous writings and try to map them with their possible authors. The presented methodology can also be used in author verification task i.e. the verification of the hypothesis whether or not a given person is the author of the text under study even if in other languages since maximum of the features are language independent. Particularly, it seems from our experiments that texts with less word are less likely to be correctly classified. However, for our future study, we would like to apply this system for other languages. Furthermore, we plan for a hybrid approach that can takes into account the advantage of both the unsupervised as well as machine learning approaches and look for the improvement of the performance. For this, more textual analysis and relevant corpus collection will be needed. Above all, we would implement this system on the other fields of Text mining i.e. e-mail identification, forensic investigation, copyright and estate disputes etc. to make it more robust and general.

9.2 Handling MWES in Phrase-Based Statistical Machine Translation

9.2.1 Introduction

Preprocessing of the parallel corpus plays an important role in improving the performance of a phrase-based statistical machine translation (PB-SMT). In this experiment, we propose a frame work in which predefined information of Multiword Expressions (MWEs) can boost the performance of PB-SMT. Here, we preprocess the parallel corpus to identify Noun-noun MWEs, reduplicated phrases, complex predicates and phrasal prepositions. Single-tokenization of Noun-noun MWEs, phrasal preposition (source side only) and reduplicated phrases (target side only) provide significant gains over our previous best PB-SMT model. Automatic alignment of complex predicates substantially improves the overall MT performance and the word alignment quality as well. For establishing NE alignments, we transliterate source NEs into the target language and then compare them with the target NEs. Target language NEs are first converted into a canonical form before the comparison takes place. The proposed system achieves significant improvements (6.38 BLEU points absolute, 73% relative improvement) over the baseline system on an English- Bengali translation task.

9.2.2 MWEs and Machine Translation

Performance of a Statistical machine translation (SMT) system depends mainly upon the good quality word and phrase alignment tables that constitute the translation knowledge acquired from a parallel corpus. In this experiment, we show that handling the Multiword Expressions (MWE) can improve the performance of a SMT system. The structure and meaning of MWEs cannot be derived from their component words, as they occur independently. Examples include conjunctions (*'as well as'*), idioms (*'kick the bucket'* means 'to die'), phrasal verbs (*'find out'*), compound noun (*'village community'*) etc. Briefly, MWE can be roughly defined as idiosyncratic interpretations that cross word boundaries (Sag et al. 2002). Traditional approaches to word alignment follow the IBM Models (Brown et al. 1993). These approaches are unable to handle many-to-many alignments and hence do not work well with multi-word expressions, especially with NEs, reduplications and complex predicates. The IBM models allow only one-to-one mapping to make a correspondence between each word in the source side to one word in the target side (Marcu 2001, Koehn et al. 2003). The alignment probabilities in the well-known Hidden Markov Model (HMM: Vogel et al. 1996) depend on the alignment position of the previous word. The HMM model does not explicitly consider many-to-many alignments. In this experiment, we address this many-to-many alignment problem indirectly. Our objective is to see how the identification of MWEs enhances the performance of the SMT system. In this experiment, several types of MWEs like phrasal prepositions and Verb-object combinations are automatically identified on the source side while named-entities and complex predicates are identified on both sides of the parallel corpus. In the target side, identification of the Noun-noun MWEs and reduplicated phrases are carried out. We use simple rule-based and statistical approaches to identify these MWEs. Source and target language NEs are aligned using a statistical transliteration technique. We rely on these automatically aligned NEs and treat them as translation examples (Pal.et.al 2010). Adding bilingual dictionaries, which in effect are instances of atomic translation pairs, to the parallel corpus is a well-known practice in domain adaptation in SMT (Eck et al. 2004; Wu et al. 2008). We modify the parallel corpus by converting the MWEs into single tokens and adding the aligned NEs and complex predicates in the parallel corpus to improve the word alignment and hence the phrase alignment quality. The preprocessing of the parallel corpus results in improved MT quality in terms of automatic MT evaluation metrics.

9.2.3 Related Work

Moore (2003) used capitalization cues for identifying NEs on the English side and then applied statistical techniques to decide which portion of the target language corresponds to the specified English NE. A Maximum Entropy model based approach for English—Chinese NE alignment has been proposed in Feng et al. (2004) which significantly outperforms IBM Model4 and HMM. The following features were used during the alignment: translation score, transliteration score, source NE and target NE's co-occurrence score and finally distortion score for distinguishing identical NEs in the same sentence. A method for automatically extracting NE translingual equivalences between Chinese and English based on multi-feature cost minimization has been proposed in Huang et al. (2003). The following costs were considered: transliteration cost, word-based translation cost and NE tagging cost.

Venkatapathy and Joshi (2006) reported a discriminative approach to use the verb-based multi-word expressions compositionality information in order to improve the word alignment quality. Ren et al. 2009 presented log likelihood ratio-based hierarchical reducing algorithm to automatically extract bilingual MWEs. They investigated the usefulness of these bilingual MWEs in SMT by integrating bilingual MWEs into the Moses decoder (Koehn et al. 2007). They observed the highest improvement with an additional feature that identifies whether or not a bilingual phrase contains bilingual MWEs. This approach was generalized in Carpuat and Diab (2010) who replaced the binary feature by a count feature representing the number of MWEs in the source language phrase. Intuitively, MWEs on the source and the target sides should be both aligned in the parallel corpus and translated as a whole. However, in the state-of-the-art PB-SMT systems, the constituents of an MWE are marked and aligned as parts of consecutive phrases, since PB-SMT (or any other approaches to SMT) does not generally treat MWEs as special tokens. Another problem with SMT systems is the wrong translation of verb phrases. Sometimes verb phrases are deleted in the output sentence. Moreover, the words inside verb phrases are generally not aligned one-to-one; the alignments of the words inside source and target verb phrases are mostly many-to-many, particularly so for the English-Bengali language pair. These are the motivations behind considering MWEs like NEs, reduplicated phrases, prepositional phrase and compound verbs for special treatment in this work. By converting the MWEs into single tokens, we make sure that PB-SMT also treats them as a whole. The first objective of the present work is to see how single tokenization and alignment of NEs on both the sides, single tokenization of phrasal verbs and phrasal prepositions on them source side and single tokenization of reduplicated phrases and noun-noun compounds on the target side affects the overall MT quality. The second objective is to see whether prior automatic alignment of complex predicates and single tokenized MWEs can bring any further improvement in the overall performance of the MT system. We carried out the experiments on English-Bengali translation task. Bengali shows high morphological richness at lexical level. Language resources in Bengali are not widely available. Furthermore, this is the first time when the identification of MWEs in Bengali language is used to enhance the performance of an English-Bengali Machine Translation System.

9.2.4 System Description

9.2.4.1 PB-SMT

Translation is modeled in SMT as a decision process, in which the translation $e_1^I = e_1 \ldots e_i \ldots e_I$ of a source sentence $f_1^J = f_1 \ldots f_j \ldots f_J$ is chosen to maximize the following equation (9.4):

$$\arg\max_{I,e_1^I} P(e_1^I \mid f_1^J) = \arg\max_{I,e_1^I} P(f_1^J \mid e_1^I).P(e_1^I) \qquad (9.4)$$

where $P(f_1^J \mid e_1^I)$ and $P(e_1^I)$ denote respectively the translation model and the target language model (Brown et al. 1993). In log-linear phrase-based SMT, the posterior probability $P(e_1^I \mid f_1^J)$ is directly modeled as a log-linear combination of features (Och and Ney 2002), that usually comprise *M* translational features, and the language model, as in equation (9.5):

$$\log P(e_1^I \mid f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) + \lambda_{LM} \log P(e_1^I) \tag{9.5}$$

where $s_1^k = s_1 ... s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases $(\hat{e}_1,...,\hat{e}_k)$ and $(\hat{f}_1,...,\hat{f}_k)$ such that (we set i_0 = 0) (9.6):

$$\begin{aligned} \forall 1 \le k \le K,\ & s_k = (i_k, b_k, j_k), \\ & \hat{e}_k = e_{i_{k-1}+1} ... e_{i_k}, \\ & \hat{f}_k = f_{b_k} ... f_{j_k} \end{aligned} \tag{9.6}$$

and each feature $\hat{h}_m$ in equation (9.5) can be rewritten as in equation (9.7):

$$h_m(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \tag{9.7}$$

where $\hat{h}_m$ is a feature that applies to a single phrase-pair. It thus follows:

$$\sum_{m=1}^{M} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) = \sum_{k=1}^{K} \hat{h}(\hat{f}_k, \hat{e}_k, s_k) \tag{9.8}$$

where $\hat{h} = \sum_{m=1}^{M} \lambda_m \hat{h}_m$.

9.2.4.2 Preprocessing of the Parallel Corpus

The initial English-Bengali parallel corpus is cleaned and filtered using a semi-automatic process. We employed several kinds of multi-word information: phrasal preposition, phrasal verb, reduplication, noun-noun MWEs, complex predicates and NEs. Compound verbs are first identified on both sides of the parallel corpus. Das et al. (2010) analyzed and identified a category of compound verbs (*Verb* + *Verb*) and conjunct verbs (*Noun /Adjective* + *Verb*) for Bengali. We adapted their strategy for identification of compound verbs as well as serial verb (*Verb* + *Verb* + *Verb*) in Bengali.

For the identification of Named-Entities and their technique of alignment, we have adopted similar technique discussed in Pal et al. (2010). Reduplicated phrases are not quite often in English side corpus; some of them (like correlative, semantic reduplications) are not at all a good habit of being used in English (Chakraborty and Bandyopadhyay 2010). But reduplication plays a crucial role in target side language and their frequencies are also quite high so for this, there is a chance to make these reduplicated words as a single-token of the issue for many to one alignment problem because these kinds of reduplicated words should have mapped with the single word of the source side. Phrasal preposition and phrasal verb may have carried different meaning for the target side , So we treat these kind of word as single token to inform the translation model that this word have carried different meaning instead of single occurrence of the words. Once the compound verbs and the NEs are identified on both sides of the parallel corpus, they are converted into and replaced by single tokens. When converting these MWEs into single tokens, we replace the spaces with underscores ('_'). Since there are already some hyphenated words in the corpus, we do not use hyphenation for this purpose; besides, the use of a special word separator (underscore in our case) facilitates the job of deciding which single-token (target language) MWEs to detokenize into words comprising them, before evaluation.

9.2.4.3 MWE Identification in Source Side

We adopt the tool named as UCREL[2] Semantic analysis System developed by Lancaster University (Rayson et al. 2004). The UCREL semantic analysis system (USAS) is a software tool for undertaking the automatic semantic analysis of English spoken and written data. It contains hierarchical semantic tag set containing 21 major discourse fields and 232 fine-grained semantic field tags. The semantic tags show semantic fields which group together word senses that are related by virtue of their being connected at some level of generality with the same mental concept. The groups include not only synonyms and antonyms but also hypernyms and hyponyms. Currently, the lexicon contains nearly 37,000 words and the template list contains over 16,000 multi-word units. Each template consists of a pattern of words and syntactic tags, some using wildcards to enable tagging with inflectional variants and less strictly defined patterns. The semantic tags for each template are arranged in rank frequency order in the same way as the lexicon. Various types of MWUs are included: phrasal verbs (e.g. stubbed out), noun phrases (e.g. riding boots), proper names (e.g. United States of America), true idioms (e.g. living the life of Riley) etc.

Currently, the USAS system consists of the CLAWS POS tagger (Garside and Smith 1997), a lemmatiser, a semantic tagger and some auxiliary format manipulating components. For POS

[2] http://www.comp.lancs.ac.uk/ucrel

tagging, this system employs the C7 tagset[3]. Subsequent semantic disambiguation, to a large extent, depends on POS information encoded in this tagset. They report an evaluation of the accuracy of the system compared to a manually tagged test corpus on which the USAS software obtained a precision value of 91% after testing it in a corpus containing about 124,900 words.

9.2.4.4 MWE Identification in Target Side

- **Identification of Reduplication**

In all languages, the repetition of noun, pronoun, adjective and verb are broadly classified under two coarse-grained categories: repetition at the (a) ***expression level***, and at the (b) ***contents** or **semantic level.*** The repetition at both the levels is mainly used for emphasis, generality, intensity or to show continuation of an act. The works on MWE identification and extraction have been continuing in English (Fillmore 2003). In this experiment, we have used simple rule-based approach (Chakraborty and Bandyopadhyay 2010) (discussed in Section 6) to identify reduplication in Bengali-side corpus. In that paper, the author classified expression-level Bengali reduplication into five fine-grained subcategories. They are (i) Onomatopoeic expressions (**khat khat***, knock knock*), (ii) Complete Reduplication (**bara-bara,** *big big*), (iii) Partial Reduplication (**thakur-thukur,** *God*), (iv) Semantic Reduplication (**matha-mundu***, head*) and (v) Correlative Reduplication (**maramari**, *fighting*). We have tried to cover almost all above mentioned types. We have used simple rules and morphological properties in lexical level and Bengali-monolingual dictionary for semantic reduplications.

- **Noun-Noun MWE Identification**

In the past few years, noun compounds have received increasing attention as researchers work towards the goal of full text understanding. Compound nouns are nominal compound where two or more nouns are combined to form a single phrase such as *'golf club'* or '*computer science department'* (Baldwin and Kim 2010). Compound noun MWEs can be defined as a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the others(s) in other contexts and which shows some phonological and/or grammatical isolation from normal syntactic usage. In English, Noun-Noun (NN) compounds

[3] See http://www.comp.lancs.ac.uk/ucrel/claws7tags.html

occur with high frequency and high lexical and semantic variability (Tanaka and Baldwin 2003). In this experiment, we have used simple statistical methodology for identifying Noun-noun MWEs. For that, the system uses Point-wise Mutual Information (PMI), Log-likelihood Ratio (LLR) and Phi-coefficient, Co-occurrence measurement and Significance function ((Agarwal et al. 2004). Final evaluation has been carried out by combining all the above mentioned features. A predefined cut-off has been taken out and the candidates having above threshold value have been considered as MWEs.

9.2.4.5 Automatic Alignments of NEs and complex predicates

We first create an NE parallel corpus by extracting the source and target (single token) NEs from the NE-tagged parallel translations and align those as the strategies applied by Pal.et.al, 2010.

- **Complex predicate Extractor:**

For the extraction of Complex Predicates (CPs) in Bengali, specially focused on compound verbs (CPs) (Verb + Verb) and conjunct verbs (Noun /Adjective + Verb) we have adopted the method applied by Das et al. (2010). But here we have also considered serial verbs (SVs) (Verb + Verb or the patterns like Verb + Verb + Verb). To extract serial verb, we have taken those pattern, which occur serially in a sentence and do not have to be considered further in Compound verb extraction. . Below we have given some example of complex predicates and serial verb in Bengali side which is associated with their extracted form in source English side.

294	দেখা_যায়/NCP(Serial Verb)
294	can be viewed
270	নিয়ে_যেতে_পারেন/NCP(Serial Verb)
270	can carry
2958	অবরোধ_করতে_পারত/CP(Conjunct Verb)
2958	would have blocked
1313	পাড়ি_দেয়/CP(Compound Verb)
1313	arrived
3541	চেখে_দেখুন/CP(Compound Verb)
3541	test

(Here left column indicates sentence ids, NCP: = Not complex Predicate, CP: = Complex Predicate)

Analysis and Extraction procedure mainly follows on the target Bengali side. At first, we have extracted and listed all serial verbs and complex predicates with their sentence id from the target side. By using those sentence ids from the target list, we have extracted and listed the entire verb chunk associated with them in the source English side.

- **Verb Chunk Aligner**

(i) Initial source and Target chunk aligner:

Form the extracted list we have aligned both side as all possible combination of complex predicates and produced a roughly unaligned list with sentence id as follows Example

1069 ||| designed ||| তৈরি_হয়/NCP

1069 ||| designed ||| নকশা_করা/CP(Conjunct Verb)

1069 ||| is ||| তৈরি_হয়/NCP

1069 ||| is ||| নকশা_করা/CP(Conjunct Verb)

1069 ||| was built ||| তৈরি_হয়/NCP

1069 ||| was built ||| নকশা_করা/CP(Conjunct Verb)

(ii) Statistical Aligner

This module produces an alignment list from the unaligned list using statistical method. From a single English verb chunk, we have made more corresponding verb chunk by extracting synset from wordnet for the main verb of that particular verb chunk. Using this synset we produce more verb chunk from the one verb chunk. When we check source target combination frequency in the entire unaligned list, we also check same for the produced synset chunk. If more than one combination occurs so frequent in the unaligned list then we have consider this should be aligned. With this strategy, we have prepared an alignment list of source – target complex predicates and serial verb list. After getting an alignment list then remove these entries from the initial unaligned list and proceed to the next steps. For finding frequency of occurrence of words, we analyze the morphology of the word on both side and matching root word only.

(iii) Iterative decision Maker and Iteration:

(a) Both side single chunk aligner

From the modified unaligned list we find unique occurrence sentence id on the both source and target side and align both side for the specified sentence id. We run this module iteratively because if there may be a situation occurs that a sentence may have two or more verbs on both side to aligned if all verbs are aligned by statistical aligner and other module except a single one then the remaining single verb can be aligned together using this module. After aligning, we modify the alignment list which has given by statistical aligner.

(b) Pattern generator and Aligner:

The pattern Generator extracts patterns for both source and target from the alignment list and produces a source target pattern list. The extracted patterns are as follows:

Root form of main verb of source side_ suffix target side pattern

MV_ed MV_করে/কর্/এ (pattern)

2405 started |start চালু/চালু_করে/কর্/এ (pattern alignment)

was_MV_ed MV_হয়েছিল/হ/ছিল

2508 was |be completed |complete শেষ/শেষ/poslcat="NM"_হয়েছিল/হ/ছিল

In target side pattern we have consider root word, and inflection only. We generate pattern for the each verb chunk from the unaligned list and produce a list of pattern for the unaligned list. Now match both list if both source and target pattern are match conjugally in the unaligned pattern list then we make align those chunk together. After getting list with this module we increase the alignment list.

(c) Iterative decision maker:

If module I and II increases the size of alignment list then this module will take decision that the process will start again otherwise it will stop the iteration.

9.2.5 Tools and Resources used

A sentence-aligned English-Bengali parallel corpus containing 14,187 parallel sentences from a travel and tourism domain was used in the present work. The corpus was obtained from the consortium-mode project "Development of English to Indian Languages Machine Translation (EILMT) System[4]". The Stanford Parser[5] and the CRF chunker[6] have been used for identifying compound verbs in the source side of the parallel corpus. The Stanford NER[7] was used to identify NEs on the source side (English) of the parallel corpus.

The sentences on the target side (Bengali) were POS-tagged by using the tools obtained from the consortium mode project "Development of Indian Languages to Indian Languages Machine Translation (IL-ILMT) System". NEs in Bangla are identified using the NER system of Ekbal and Bandyopadhyay (2008). We use the Stanford Parser, Stanford NER and the NER for Bangla along with the default model files provided, i.e., with no additional training.

The effectiveness of the MWE-aligned parallel corpus developed in the work is demonstrated by using the standard log-linear PB-SMT model as our baseline system: GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics described in (Koehn et al. 2003), minimum-error-rate training (Och 2003) on a held-out development set, target language model with Kneser-Ney smoothing (Kneser and Ney 1995) trained with SRILM (Stolcke 2002), and Moses decoder (Koehn et al. 2007).

9.2.6 Experiments and Results

We randomly extracted 500 sentences each for the development set and testset from the initial parallel corpus, and treated the rest as the training corpus. After filtering on maximum allowable sentence length of 100 and sentence length ratio of 1:2 (either way), the training corpus contained 13,176 sentences. In addition to the target side of the parallel corpus, a monolingual Bangla corpus containing 293,207 words from the tourism domain was used for the target language model. We experimented with different n-gram settings for the language model and the maximum phrase length, and found that a 4-gram language model and a maximum

[4] The EILMT and ILILMT projects are funded by the Department of Information Technology (DIT), Ministry of Communications and Information Technology (MCIT), Government of India.

[5] http://nlp.stanford.edu/software/lex-parser.shtml

[6] http://crfchunker.sourceforge.net/

[7] http://nlp.stanford.edu/software/CRF-NER.shtml

phrase length of 4 produced the optimum baseline result. We therefore carried out the rest of the experiments using these settings.

In training set	English		Bengali	
	T	U	T	U
CPs	4874	2289	14174	7154
reduplicated word	-	-	85	50
Noun-noun compound	892	711	489	300
Phrasal preposition	982	779	-	-
Phrasal verb	549	532	-	-
Total NE words	22931	8273	17107	9106

Table 9.6 MWE statistics. (T - Total occurrence, U – Unique, CP – complex predicates)

This system continues with the various preprocessing of the corpus and going on observing the improvement achieved by the identification of MWEs in phrase level. In this experiment, our intuition is that the more the MWEs are identified and aligned properly, the more the system shows the improvement in the translation procedure. In the source side, the system treats the phrasal prepositions and noun-noun compounds as a single token which shows n:m alignment in the bilingual context. After identifying them as single token and align them using GIZA++, the system has achieved an accuracy of 13.99 BLEU score. But when noun-noun compounds are identified separately, the system shows relatively degradable results with respect to the other identification. The reason behind these results is manifold. Firstly, the accuracy of the UCREL semantic toolkit is not satisfactory especially for the tourism domain. Secondly, it has been observed that noun-noun compounds are translated in target side with n:n alignment basis. For them, the single tokenization is not desirable at all. However, overall combined result infers our actual intuition.

In that target side, reduplication has been identified and aligned it with the source side. The system draws an estimating result after aligning reduplication with the improvement of 0.51 BLEU score as reduplications in the target side may not show any significant existence in the source side. In the target side, reduplications, noun-noun compound as well as both have given the satisfactory results with the improvement of 0.50 BLEU score which again proves our intuition.

Experiments		Exp	BLEU	NIST
Baseline Best System (Alignment of NEs of any length)		1	**13.33**	**4.44**
Source Side Treatment + NEA	Phrasal preposition as single-token (SPPaST)	2	13.76	4.39
	Verb-object combination as a single-token (SVOaST)	3	13.61	4.40
	Verb-object combination and phrasal preposition as a single-token (SPPaST+SVOaST)	4	13.99	4.41
	Noun-noun compound as Single token (SNNaST)	5	13.61	4.40
	(SPPaST+SNNaST)	6	13.71	4.41
	(SPPaST+SNNaST+SPVaST)	7	13.89	4.42
Target Side Treatment+ NEA	Reduplicated word as single-token (TRWaST)	8	13.84	4.42
	Noun-noun compound as Single token (TNNaST)	9	13.75	4.42
	Reduplicated word and Noun noun compound as single-token (TRWaST + TNNaST)	10	13.83	4.42
Both Side Treatment+ NEA	SPPaST+TRWaST	11	14.07	4.41
	SPPaST+TRWaST+TNNaST	12	14.38	4.43
	SPPaST+SNNaST+TRWaST+TNNaST	13	14.20	4.43
	SPPaST+SVOaST+TRWaST+TNNaST†	**14**	**14.58**	**4.44**
	SPPaST+SNNaST+SVOaST+TRWaST+TNNaST	15	14.51	4.43
Best System + NEA	Complex predicates alignment (CPA)	16	14.14	4.43
Best System + NEA	CPA+(Best combination) SPPaST+SVOaST+TRWaST+TNNaST	17	15.12	4.48

Table 9.7 Evaluation results for different experimental setups. (The '†' marked systems produce statistically significant improvements on BLEU over the baseline system)

Finally, we have treated both the source and destination side corpus by combining the previous identified phrases. Table 9.7 shows that when we combined the prepositional phrase, verb-object combination, reduplicated word and Noun-noun compound as single token, the alignment system achieves the best results with 14.58 BLEU score. The table also reflects the results for the other combination which also proves our intuition with respect to the baseline system.

Table 9.6 shows the MWE statistics of the parallel corpus as identified by the NERs. The average NE length in the training corpus is 2.16 for English and 1.61 for Bangla. As can be seen from Table 9.5, 44.5% and 47.8% of the NEs are single-word NEs in English and Bangla respectively, which suggests that prior alignment of the single-word NEs, in addition to multi-word NE alignment, should also be beneficial to word and phrase alignment.

Of all the NEs in the training and development sets, the transliteration-based alignment process was able to establish alignments of 4,711 single-word NEs, 4,669 two-word NEs and 1,745 NEs having length more than two. It is to be noted that, some of the single-word NE alignments, as well as two-word NE alignments, result from multi-word NE alignment.

We analyzed the output of the NE alignment module and observed that longer NEs were aligned better than the shorter ones, which is quite intuitive, as longer NEs have more tokens to be considered for intra-NE alignment. Since the NE alignment process is based on transliteration, the alignment method does not work where NEs involve translation or acronyms. We also observed that English MW NEs are sometimes fused together into single-word NEs.

We performed three sets of experiments: treating compound verbs as single tokens, treating NEs as single tokens, and the combination thereof. Again for NEs, we carried out three types of preprocessing: single-tokenization of (i) two-word NEs, (ii) more than two-word NEs, and (iii) NEs of any length. We make distinctions among these three to see their relative effects. The development and test sets, as well as the target language monolingual corpus (for language modeling), are also subjected to the same preprocessing of single-tokenizing the MWEs. For NE alignment, we performed experiments using 4 different settings: alignment of (i) NEs of length up to two, (ii) NEs of length two, (iii) NEs of length greater than two, and (iv) NEs of any length. Before evaluation, the single-token (target language) underscored MWEs are expanded back to words comprising the MWEs.

Since we did not have the gold-standard word alignment, we could not perform intrinsic evaluation of the word alignment. Instead we carry out extrinsic evaluation on the MT quality using the well known automatic MT evaluation metrics: BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), NIST (Doddington 2002), WER, PER and TER (Snover et al. 2006). As can be seen from the evaluation results reported in Table 9.6, baseline Moses without any preprocessing of the dataset produces a BLEU score of 8.74. The low score can be attributed to the fact that Bangla, a morphologically rich language, is hard to translate into. Moreover, Bangla being a relatively free phrase order language (Ekbal and Bandyopadhyay 2009) ideally requires multiple set of references for proper evaluation. Hence using a single reference set does not justify evaluating translations in Bangla. Also the training set was not sufficiently large enough for SMT. Treating only longer than 2-word NEs as single tokens does not help improve the overall performance much, while single tokenization of two-word NEs as single tokens produces some improvements (.39 BLEU points absolute, 4.5% relative). Considering compound verbs as single tokens (CVaST) produces a .82 BLEU point improvement (9.4% relative) over the baseline. Strangely, when both compound verbs and NEs together are counted as single tokens, there is hardly any improvement. By contrast, automatic NE alignment (NEA) gives a huge impetus to system performance, the best of them (4.59 BLEU points absolute, 52.5% relative improvement) being the alignment of NEs of any length that produces the best scores across all metrics. When NEA is combined with CVaST, the improvements are substantial, but it can not beat the individual improvement on NEA. The (†) marked systems produce statistically significant improvements as measured by bootstrap resampling method (Koehn 2004) on BLEU over the baseline system. Metric-wise individual best scores are shown in bold in Table 9.6.

9.2.7 Conclusion and Future Work

In this experiment, we have successfully shown how the simple yet effective preprocessing of treating various types of MWEs, namely NEs, reduplications and compound verbs, as single-tokens, and conjunction with prior NE alignment can boost the performance of PB-SMT system on an English-Bengali translation task. Treating compound verbs as single-tokens provides significant gains over the baseline PB-SMT system. Amongst the MWEs, NEs perhaps play the most important role in MT, as we have clearly demonstrated through experiments that automatic alignment of NEs by means of transliteration improves the overall MT performance substantially across all automatic MT evaluation metrics. Our best system yields 4.59 BLEU points improvement over the baseline, a 52.5% relative increase. We compared a subset of the output of our best system with that of the baseline system, and the output of our best system almost always looks better in terms of either lexical choice or word ordering. The fact that only 28.5% of the test set NEs appear in the training set, yet prior automatic alignment of the NEs brings about so much improvement in terms of MT quality, suggests that it not only improves the NE alignment quality in the phrase table, but word alignment and phrase alignment quality must have also been improved significantly. At the same time, single-tokenization of MWEs makes the dataset sparser, but yet improves the quality of MT output to some extent. Data-driven approaches to MT, specifically for scarce-resource language pairs for which very little parallel texts are available, should benefit from these preprocessing methods. Data sparseness is perhaps the reason why single-tokenization of NEs and compound verbs, both individually and in collaboration, did not add significantly to the scores. However, a significantly large parallel corpus can take care of the data sparseness problem introduced by the single-tokenization of MWEs.

The present work offers several avenues for further work. In future, we will investigate how these automatically aligned NEs can be used as anchor words to directly influence the word alignment process. We will look into whether similar kinds of improvements can be achieved for larger datasets, corpora from different domains and for other language pairs. We will also investigate how NE alignment quality can be improved, especially where NEs involve translation and acronyms. We will also try to perform morphological analysis or stemming on the Bangla side before NE alignment.

Chapter 10

Conclusion

10.1 Summary and Findings of this Thesis

An attempt has been made in the thesis to model the syntax and semantics of Bengali Multi Word Expressions (MWEs) based on the following statistical approaches: substitutability, co-occurrence properties, semantic clustering and linguistic properties. A detailed discussion of the experiments to automatically acquire the syntax and semantics of MWEs has been presented in the thesis. We have experimented with different approaches to idetify the statistical approaches that are suited to specific MWE types and tasks.

The findings of the various experiments carried out in the thesis are summarized below:

1. The experiments mainly focus on the Bengali Multiword Expressions; though an experiment on identification of compositionality for English bigram MWEs has been carried out.
2. Identification and Extraction of MWEs have been done with various statistical methodologies.
3. All types of Bengali bigram Noun compounds and complex predicates are handled in these experiments.

4. The thesis incorporates a graph-based semantic clustering approach to identify MWEs from limited size corpus in Bengali.
5. Most importantly, working on resource constraint language line Bengali is itself a challenging task where lack of corpus and lexical resources put obstracles during the experiment.
6. Most importantly, this thesis is a pioneer to handle the stylometric features in Bengali. Our experiment focuses on making various statistics of the writers and tests them on their other writings. We also experimented with machine learning approaches to show the improvement over statistical approaches.
7. We also applied Bengali MWEs in Statistical Machine translation system to improve the translation quality by enhacing the alignment of Bengali-English parallel corpus.

10.2 Future Road Map

10.2.1 Future Researches on MWEs

Future research directions on the core part of MWEs are:

1. Expanding the windows when identifying the MWEs and focusing on the n-gram phrases (n>2)
2. Working on other types of MWEs in Bengali like adjective-noun compounds
3. Building standard MWE lexicon in Bengali and adapt supervised approaches in the identification task
4. Developing more unsupervised methods
5. Incorporating Named-Entity recognizer with every MWE identification system
6. Working on the other research issues of MWEs like semantic interpretation, internal semantic disambiguation
7. Experimenting with these developed systems on other languages like Hindi and other Indian languages
8. Augmenting the effort of building Bengali WordNet

10.2.2 Future Researches on the Applications of MWEs

Future researches on the real-world NLP applications of MWEs are:

1. Direct application of MWEs in Stylometry analysis
2. Adapting more statistical approaches when building the statistics of the writer
3. Applying the MWEs in oher NLP applications line Textual Entailment, Sentiment analysis, Summarization and Information Rtrieval tasks
4. Above all, collecting more and more Bangali corpus

In the concluding part, we agree that complete identification of Bengali MWEs is not yet done. The systems developed start a new direction on working with MWEs in Bengali language.

Appendix

Research Publications

1. **Tanmoy Chakraborty** and Sivaji Bandyopadhyay, Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. In *Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010)*, *The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August 28, 2010, pp. 73-76.

2. Dipankar Das, Santanu Pal, Tapabrata Mondal, **Tanmoy Chakraborty** and Sivaji Bandyopadhyay, Automatic Extraction of Complex Predicates in Bengali. In *Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010)*, *The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, August 28, 2010, pp. 37- 45.

3. **Tanmoy Chakraborty** and Sivaji Bandyopadhyay, Authorship Identification Using Stylometry Analysis: A CRF Based Approach, In *Proceedings of IEEE Cascom Postgraduate Student paper Conference*, Jadavpur University, Kolkata, November 27, 2010, pp. 66-69.

4. **Tanmoy Chakraborty**, Identification of Noun-Noun (N-N) Collocations as Multi-Word Expressions in Bengali Corpus, *The 8th International Conference on Natural Language Processing (ICON 2010)*, IIT Kharagpur, India, December 8-11, 2010.

5. **Tanmoy Chakraborty** and Sivaji Bandyopadhyay, Inference of Fine-grained Attributes of Bengali Corpus for Stylometry Detection, *The 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2011)*, Tokyo, Japan, February 20-26, 2011.

6. **Tanmoy Chakraborty**, Santanu Pal, Tapabrata Mondal, Tanik Saikh and Sivaji Bandyopadhyay, Shared task system description: Measuring the Compositionality of Bigrams using Statistical Methodologies, In *Proceedings of Distributional Semantics and Compositionally (DiSCo)*, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, Portland, Oregon, USA, June 24, 2011. (Accepted)

7. **Tanmoy Chakraborty**, Dipankar Das, Sivaji Bandyopadhyay, Semantic Clustering: an Attempt to Extract Multiword Expressions in Bengali, In *Proceedings of Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)* Portland, Oregon, USA, June 23, 2011. (Accepted).

Bibliography

Argamon, S., Saric, M., Stien, S.S. 2003. Style mining of electronic messages for multiple authorship discrimination: First results, In proceedings of 9th ACM SIGKDD, pp. 475-480.

Abeillè, Anne. 1988. Light verb constructions and extraction out of NP in a tree adjoining grammar. In *Papers of the 24th Regional Meeting of the Chicago Linguistics Society*.

Abbi, Anvita. 1991. Semantics of Explicator Compound Verbs. *In South Asian Languages, Language Sciences*, 13(2): pp. 161-180.

Alsina, Alex. 1996. Complex Predicates: Structure and Theory. *Center for the Study of Language and Information Publications*, Stanford, CA.

Agirre, Eneko, and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht, Netherlands: Springer.

Agirre, Eneko, and David Martinez. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of COLING workshop on Semantic Annotation and Intelligent Content*, 11–19, Saarbrucken, Germany.

Agarwal, Aswini, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. In *Proceedings of International Conference on Natural Language Processing (ICON)*, pp. 165-174.

Argamon, Shlomo, Marin Saric, and Sterling S. Stein.. 2003. Style mining of electronic messages for multiple authorship discrimination: First results, In: Proceedings of the 2003 Association for Computing Machinery Conference on Knowledge Discovery and Data Mining (ACM SIGKDD), pp 475—480.

Alba-Salas, Josep, 2002. *Light Verb Constructions in Romance. A Syntactic Analysis*. Cornell University dissertation.

Apte, M.L. 1968. *Reduplication, Echo Formation and Onomatopoeic in Marathi*. Deccan College Post-Graduate and Research Institute, Pune, India.

Bhaskararao and Peri. 1977. *Reduplication and Onomatopoeia in Telugu*. Deccan College Post-Graduate and research Institute, Pune, India.

Brown Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263– 311.

Bashir, Elena. 1993. Causal chains and compound verbs. *In M. K. Verma ed. (1993) Complex Predicates in South Asian Languages,* Manohar Publishers and Distributors, New Delhi.

Burton-Page, John. 1957. Compound and conjunct verbs in Hindi. *Bulletin of the School of Oriental and African Studies,* 19: 469-78.

Baldwin, Timothy. 2005a. The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions* 19.398–414.

Baldwin, Timothy. 2005b. Looking for prepositional verbs in corpus data. In *Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 115–126, Colchester, UK.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003a. An empirical model of multiword expression decomposability. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, pp. 89–96, Sapporo, Japan.

Baldwin, Timothy, John Beavers, Leonoor Vander Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. 2003b. In search of a systematic treatment of determinerless PPs. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 145–156, Toulouse, France.

Baldwin, Timothy, John Beavers, Leonor Van Der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. 2006. In search of a systematic treatment of determinerless PPs. In *Syntax and Semantics of Prepositions*, ed. by Patrick Saint-Dizier. Springer.

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, 2047–2050, Lisbon, Portugal.

Baldwin, Timothy, and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, 24–31, Barcelona, Spain.

Baldwin, Timothy, and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, 98–104, Taipei, Taiwan.

Bame, Ken, 1999. Aspectual and resultative verb-particle constructions with up. Handout for talk presented at the Ohio State University Linguistics Graduate Student Colloquium.

Banerjee, Satanjeev, and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003)*, 805–810, Acapulco, Mexico.

Bannard, Colin, 2003. Statistical techniques for automatically inferring the semantics of verb-particle constructions. Master's thesis, University of Edinburgh.

Bannard, Colin, 2006. *Acquiring Phrasal Lexicons from Corpora*. University of Edinburgh, UK dissertation.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 65–72, Sapporo, Japan.

Barker, Ken, and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, 96–102, Montreal, Canada.

Becker J. D. 1975. The phrasal lexicon*. In Theoretical Issues of NLP, Workshop in Computational Linguistics, Psychology and AI, Cambridge*, MA.

Bauer, Laurie. 1983. *English Word-formation*. Cambridge, UK: Cambridge University Press.

Ben Taskar, Pieter Abbeel, and Daphne Koller. 2002, Discriminative probabilistic models for relational data. In Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02).

Bolinger, Dwight, 1976a. Meaning and memory.

Bolinger, Dwight. 1976b. *The Phrasal Verb in English*. Boston, USA: Harvard University Press.

Bond, Francis, 2001. *Determiners and number in English, contrasted with Japanese, as exemplified in machine translation*. Brisbane, Australia: University of Queensland dissertation.

Borthen, Kaja, 2003. *Norwegian bare singulars*. Norwegian University of Science and Technology dissertation.

Brinton, Laurel. 1985. Verb particles in English: Aspect or aktionsart. *Studia Linguistica* 39. pp. 157–168.

Briscoe, Ted, and John Carroll. 2002. Accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 1499–1504, Las Palmas, Canary Islands.

Butt, Miriam. 1995. The Structure of Complex Predicates in Urdu. Doctoral Dissertation, Stanford University.

Buckeridge, Alan M., and Richard F. E. Sutcliffe. 2002. Disambiguating noun compounds with Latent Semantic Indexing. In *Proceedings of the 2nd International Workshop on Computational Terminology*, Patras, Greece.

Buitelaar, Paul, 1989. *CoreLex: Systematic Polysemy and Underspecification*. Brandeis University dissertation.

Burnard, Lou, 1995. User guide for the British National Corpus. Butt, Miriam. 2003. The light verb jungle. In *Proceedings of the Workshop on Multi-verb Constructions*, 1–49, Trondheim, Norway.

Croft, D.J.: Book of Mormon word prints reexamined. Sun Stone Publish., 6, 15--22 (1981)

Calzolari, Nicoletta, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, 1934–1940, Las Palmas, Canary Islands.

Cao, Yunbo, and Hang Li. 2002. Base noun phrase translation using web data and the EM algorithm. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, 37–40, Taipei, Taiwan.

Carpuat, Marine, and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics*, 387–394, Ann Arbor, USA.

Charles Sutton. Conditional probabilistic context-free grammars. Master's thesis, University of Massachusetts, 2004.

Chafe, Wallace L. 1968. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language* . pp. 109–127.

Chander, Ishwar, 1998. *Automated postediting of documents*. University of Southern.

California dissertation. Charniak, Eugene. 2000. A maximum entropy-based parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics*, Seattle, USA.

Chaudhuri, B.V, C. Kar and S. Ghosh. 2005. Computer-Based Multi-Word Analysis. *In Proceedings of International Conference on Natural Language Processing-2005*, pp. 96-98.

Choueka, Yaacov. 1988. Lookin for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of RIAO*, pp. 43–38.

Choueka, Yaacov, Shmuel T. Klein, and E. Neuwitz. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic* 4.34–38.

Church, Kenneth W., and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics (ACL-1989)*, 76–83, Vancouver, Canada.

Carpuat, Marine, and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In Proceedings of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics conference (HLTNAACL 2010), Los Angeles, CA.

Chakrabarti, Debasri, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma, and Pushpak Bhattacharyya. 2008. Hindi compound verbs and their automatic extraction. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Posters and demonstrations, Manchester, UK, pp. 27-30.

Chakraborty, Tanmoy and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. In *proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Beijing, China.

Cook, Paul, and Suzanne Stevenson. 2006. Classifying particle semantics in English verb-particle constructions. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 45–53, Sydney, Australia.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 20, pp. 37–46.

Copestake, Ann, and Alex Lascarides. 1997. Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association of Coomputational Linguistics and 8th Conference of the European Chapter of Association of Computational Linguistics (ACL/EACL-1997)*, 136–143, Madrid, Spain.

Cruse, Alan D. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, 2004. TiMBL: Tilburg memory based learner, version 5.1, reference guide.

Das, Pradeep Kumar. 2009. The form and function of Conjunct verb construction in Hindi. *Global Association of Indo-ASEAN Studies*, Daejeon, South Korea.

Das Dipankar, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty, Sivaji Bandyopadhyay .2010. Automatic Extraction of Complex Predicates in Bengali In proceedings of the workshop on Multiword expression: from theory to application (MWE-2010), The 23rd International conference of computational linguistics (Coling 2010),Beijing, Chaina, pp. 37-46.

Dasgupta, Sajib, Naira Khan, Asif Iqbal sarkar, Dewan Shahriar Hossain Pavel and Mumit Khan. Morphological Analysis of Inflecting Compound Words in Bengali. In *Proceedings of the 8th International Conference on Computer and Information Technology (ICCIT),* Bangladesh, 2005.

Dehe, Nicole. 2002. *Particle Verbs in English: syntax, information structure and intonation*. Amsterdam/Philadelphia: John Benjamins Publishing.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B (Methodological) 39 (1): 1–38.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co occurrence statistics. In Proceedings of the Second International Conference on Human Language Technology Research (HLT-2002), San Diego, CA, pp. 128-132.

Dias, Gaël Harry. 2003. Multiword Unit Hybrid Extraction. In *proceedings of the First Association for Computational Linguistics, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 41-48.

Dias, Gaël Harry, Ray Jackendoff, Andrew McIntyre, and Silke Urban (eds.) 2001. *Verb-Particle Explorations*. Berlin/New York: Mounton de Gruyter.

Dias, Gaël, S. Guillor´e, and J. G. Pereira Lopes. 1999. Multilingual aspects of multiword lexical units. In *Workshop on Language Technologies in the Framework of the 32rd Annual Meeting of the* Societas Linguistica Europaea, Ljubljana, Slovenia.

Diab M. T, and P. Bhutada. 2009. Verb Noun Construction MWE Token Supervised lassification, In Proceedings of the Joint conference of Association for Computational Linguistics and International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing 2009 (ACL-IJCNLP 2009), Workshop on Multiword Expression., Singapore, pp.17-22.

Dirven, Ren´e. 2001. The metaphoric in recent cognitive approaches to English phrasal verbs. *metaphorik.de*.pp. 39–54.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language* 53.pp. 810–842.

Dras, Mark, and Mike Johnson. 1996. Death and lightness: Using a demographic model to find support verbs. In *Proceedings of the 5th International Conference on the Cognitive Science of Natural Language Processing*, 165–172, Dublin, Ireland.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19.pp. 61–74.

Enivre J, Nilson. 2004. Multiword Units in Syntactic Parsing. *In Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MAMURA) 2004 Workshop*, Lisbon, pp.39-46.

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the Unified Medical Language System. In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, pp. 792-798.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2008. Maximum Entropy Approach for Named Entity Recognition in Indian Languages. International Journal for Computer Processing of Languages (IJCPOL), Vol. 21 (3), pp. 205-237.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2009. Voted NER system using appropriate unlabeled data. In proceedings of the ACL-IJCNLP-2009 Named Entities Workshop (NEWS 2009), Suntec, Singapore, pp.202-210.

Evert, Stephen, 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. University of Stuttgart dissertation.

Ekbal, Asif, Rejwanul Haque and Sivaji Bandyopadhyay. (2008). Maximum Entropy Based Bengali Part of Speech Tagging. In *proceedings of Advances in Natural Language Processing and Applications Research in Computing Science,* 2008, pp. 67-78.

Feng, Donghui, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), Barcelona, Spain, pp. 372-379.

Fillmore C. 2003. *An extremist approach to multi-word expressions*. In a talk given at IRCS, University of Pennsylvania, 2003. 2003.

Fan, James, Ken Barker, and Bruce W. Porter. 2003. The knowledge required to interpret noun compounds. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1483–1485, Acapulco, Mexico.

Fazly, Afsaneh, Ryan North, and Suzanne Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, 38–47, Ann Arbor, USA. Association for Computational Linguistics.

Fazly, Afsaneh, and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 9–16, Prague, Czech Republic.

Fellbaum, Christiane (ed.) 1998. *WordNet, An Electronic Lexical Database*. Cambridge, Massachusetts, USA: MIT Press.

Fernando, Chitra, and Roger Flavell. 1981. *On idioms*. Exeter: University of Exeter.

Fillmore, Charles, Paul Kay, and Mary C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions. *Language* 64.501–538.

Fillmore C. 2003. *An extremist approach to multi-word expressions*. In a talk given at RCS, University of Pennsylvania, 2003.

Finin, Timothy Wilking, 1980. *The semantic interpretation of compound nominals*. University of Illinois, Urbana Champaign dissertation.

Firth, John Rupert. 1957. A Synopsis of Linguistic Theory, 1933-1955. In *Studies in Linguistic Analysis*, ed. by J. R. Firth, 1–32. Oxford: Blackwell.

Folli, Raffaella, Heidi Harley, and Simin Karim. 2003. *Determinants of even type in Persian complex predicates*. Cambridge Working Papers in Linguistics.

Fraser, Bruce. 1976. *The Verb-Particle Combination in English*. The Hague: Mouton. Gates, Edward. 1988. *The treatment of multiword lexemes in some current dictionaries of English*. Snell-Hornby.

Gibbs, Raymond W. 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition* 8.149–156.

Girju, Roxana. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 568–575, Prague, Czech Republic.

Girju, Roxana, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language* 19.479–496.

Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th Semantic Evaluation Workshop(SemEval-2007)*, 13–18, Prague, Czech Republic.

Grefenstette, Gregory, and Simual Teufel. 1994. What is a word, what is a sentence? problems of tokenization. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*, 79–87, Budapest, Hungary.

Grefenstette, Gregory, and Simual Teufel. 1995. A corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings of the 7th European Chapter of Association of Computational Linguistics (EACL-1995)*, pp. 98–103, Dublin, Ireland.

Gries, Stefan T. 1999. Particle movement: A cognitive and functional approach. *Cognitive Linguistics* 10. pp. 105–145.

Grishman, Ralph, Catherine Macleod, and Adam Myers, 1998. Complex syntax reference manual.

Grover, Claire, Maria Lapata, and Alex Lascarides. 2004. A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering* 1. Pp. 1–38.

Huang, Fei, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In *Proceedings of the ACL-2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 2003*, Sapporo, Japan, pp. 9-16.

Haspelmath, Martin. 1997. *From Space to Time in The World's Languages*. Munich, Germany: Lincorn Europa.

Hearst, Marti. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING '92)*, Nantes, France.

Hermjakob, Ulf, Eduard Hovy, and Chin-Yew Lin. 2002. Automated question answering in Webclopedia. In *Proceedings of the ACL-02 Demonstrations Session*, 98–99, Philadelphia, USA.

Hirst, Graeme, and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In (Fellbaum 1998), pp. 305–332.

Hook, Peter. 1974. The Compound Verbs in Hindi. *The Michigan Series in South and South-east Asian Language and Linguistics*. The University of Michigan.

Halteren, V. H.: Linguistic profiling for author recognition and verification, *In: Proceedings of the 2005 Meeting of the Association for Computational Linguistics (ACL) (2005).*

Holmes, D.: 1994. Authorship Attribution, Computers and the Humanities, 28, pp. 87—106.

Hoshi, H., 1994. *Passive, Causive, and Light Verbs: A Study of Theta Role Assignment*. University of Connecticut dissertation.

Huddleston, Rodney, and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Hull, Richard D., and Fernando Gomez. 1996. Semantic interpretation of nominalizations. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-1996)*, 1062–1068, Portland, Oregon.

Humphreys, L., D. Lindberg, H. Schoolmand, and G.O. Barnett. 1998. The unified medical language system: An informatics research collabration. *Journal of the American Medical informatics Assocation* 5.1–13.

Ide, Nancy, and Jean Veronis. 1998. Word sense disambiguation: The state of the art. *Computational Linguistics* 24.pp. 1–40.

Isabelle, Pierre. 1984. Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING-1984)*, pp. 509–516, San Francisco, USA.

Ishikawa, K. 1999. Enlish verb-particle constructions and V-internal structure. *English Linguistics* 16.329–352.

Jackendoff, Ray. 1973. The base rules for prepositional phrases. In *A Festschrift for Morris Halles*, 345–356. New York: Halt: Rinehart and Winston.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. Cambridge, USA: MIT Press.

Jackendoff, Ray. 2002. *Foundation of Language*. Oxford, UK: Oxford University Press. Jespersen, Otto. 1965. *A Modern English Grammar on Historical Principles, Part VI, Morphology*. London, UK: George Allen and Unwin Ltd.

Jiang, Jay, and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, 19–33, Taipai, Taiwan.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. 18th International Conference on Machine Learning.

Johnston, Michael, and Frederica Busa. 1996. Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons*, 77–88, Santa Cruz, USA.

Kan, Yee Fan Tan Min-Yen, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context (MWEmc)*, Trento, Italy.

Kilgarriff, Adam and Joseph Rosenzweig. 2000. Framework and results for english SENSEVAL. *Computers and the Humanities*. Senseval Special Issue, 34(1-2). pp.15-48.

Kastovsky, Dieter. 1982. *Wortbildung und Semantik*. Dusseldorf: Bagel/Francke.

Katz, Graham, and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedingsof the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 28–35, Sydney, Australia.

Katz, Jerrold J., and Paul M. Postal. 2004. Semantic interpretation of idioms and sentences containing them. In *Quarterly Progress Report (70), MIT Research Laboratory of Electronics*, 275–282. MIT Press.

Keane E. 2001. *Echo Words in Tamil*. PhD thesis, Meriton College, Oxford.

Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 181–184. Detroit, MI.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series,* Edmonton, Canada, pp. 48-54.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL 2007):* Proceedings of demo and poster sessions, Prague, Czech Republic, pp. 177-180.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In EMNLP-2004: *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 25-26 July 2004, Barcelona, Spain, pp 388-395.

Kaul, Vijay Kumar. 1985. The Compound Verb in Kashmiri. Unpublished Ph.D. dissertation. Kurukshetra University.

Kayne, Richard S. 1985. Principles of particle constructions. In *Jacqueline Gueron*, 101–140. Dordrecht:Foris: Grammatical Representation.

Kearns, Kate, 2002. Light verbs in English. Kengo Sato and Yasubumi Sakakibara. RNA secondary structural alignment with conditional random fields. Bioinformatics, 21:pp. 237–242, 2005.

Kunchukuttan F A and Om P. Damani, 2008. A System for Compound Noun Multiword Expression Extraction for Hindi. *6th International. Conference on Natural Language Processing (ICON).*

Krippendorf, K. H.: Content Analysis-An Introduction to its Methodology, 2nd Edition, Sage Publications Inc., ISBN: 13: 978- 0761915454, pp. 440 (2003)

Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research, 8:1261–1276 (2007)

Kipper-Schuler, Karin. 2005. VerbNet: A broadcoverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Department, University of Pennsylvania, Philadelphia,PA.

Kim, Su Nam. 2008. Statistical Modeling of Multiword Expressions. PhD thesis. University of Melbourne, Melbourne, Australia.

Kim, Su Nam, and Timothy Baldwin. 2005. Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of 2nd International Joint Conference on Natual Language Processing (IJCNLP-2005)*, 945–956, Jeju, Korea.

Kim, Su Nam, and Timothy Baldwin. 2006a. Automatic extraction of verb-particles using linguistic features. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, 65–72, Trento, Italy.

Kim, Su Nam and Timothy Baldwin. 2006b. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, 491–498, Sydney, Australia.

Kim, Su Nam, and Timothy Baldwin. 2007a. Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of Conference of the Pacific Association for Computational Linguistics*, 40–48, Melbourne, Australia.

Kim, Su Nam, and Timothy Baldwin. 2007b. Disambiguating noun compounds. In *Proceedings of 22nd AAAI Conference on Artificial Intelligenc*, 901–906, Vancouver, Canada.

Kim, Su Nam, and Timothy Baldwin. 2007c. Interpreting noun compounds using bootstrapping and sense collocation. In *Proceedings of Conference of the Pacific Association for Computational Linguistics*, 129–136, Melbourne, Australia.

Kim, Su Nam, and Timothy Baldwin. 2007d. Melb-kb: Nominal classification as noun compound interpretation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 231–236, Prague, Czech Republic.

Kim, Su Nam, and Timothy Baldwin. 2008. Benchmarking noun compound interpretation. In *Proceedings of 3rd International Joint Conference on Natual Language Processing (IJCNLP- 2008)*, 569–576, Hyderabad, India.

Kim, Su Nam, Meladel Mistica, and Timothy Baldwin. 2007. Extending sense collocation on interpreting noun compounds. In *Proceedings of Australasian Language Technology Workshop*, 49–56, Melbourne, Australia.

Landauer, Thomas K., Peter W. Faltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*. pp. 259–284.

Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics* 28.357–388.

Lapata, Mirella, and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Langauge Techinology Conference and Conference on Empirical Methods in National Language Processing (HLT/NAACL-2004)*, pp. 121–128, Boston, USA.

Lauer, Mark, 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Macquarie University dissertation.

Leacock, Claudia, and Nancy Chodorow. 1998. *Combining local context and WordNet similarity for word sense identification*. Cambridge, USA: MIT Press.

Levi, Judith. 1978. *The Syntax and Semantics of Complex Nominals*. New York, New York, USA: Academic Press.

Li, Wei, Xiuhong Zhang, Cheng Niu, Yuankai Jiang, and Rohini K. Srihari. 2003. An expert lexicon approach to identifying English phrasal verbs. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 513–520, Sapporo, Japan.

Liberman, Mark, and Richard Sproat. 1992. The stress and structure of modified noun phrases in English. In *Lexical Matters – CSLI Lecture Notes No. 24*, ed. by Ivan A. Sag and A. Szabolcsi. Stanford, USA: CSLI Publications.

Lidner, Sue., 1983. *A lexico-semantic analysis of English verb particle constructions with OUT and UP*. University of Indiana at Bloomington dissertation.

Lin, Dekang. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31th Association of Computational Linguistics (ACL-1993)*, 112–120, Columbus, Ohio, USA.

Lin, Dekang. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING/ACL-1998)*, pp. 768–774, Montreal, Canada.

Lin, Dekang. 1998b. Extracting collocations from text corpora. In *Proceedings of the 1st Workshop on Computational Terminology*, Montreal, Canada.

Lin, Dekang. 1998c. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*, 296–304, Madison, Wisconsin, USA.

Lin, Dekang. 1998d. Using collocation statistics in information extraction. In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, USA.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Association of Computational Linguistics (ACL-1999)*, 317–324, College Park, Maryland, USA.

Lohse, Barbara, John A. Hawkins, and Thomas Wasow. 2004. Domain minimization in English verb-particle constructions. *Language* 80. pp. 238–261.

Lynott, Dermot, and Mark T. Keane. 2004. A model of novel compound production. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Chicago, Illinois, USA.

Mustafa, T. K.., Mustapha, N., Azmi, M. A., Sulaiman, N. B. 2010. Dropping down the Maximum Item Set: Improving the Stylometric Authorship Attribution Algorithm in the Text Mining for Authorship Investigation, *Journal of Computer Science 6 (3),* pp. 235—243.

Malyutov, M. B. 2006. Authorship attribution of texts: A review. Lecture Notes in Computer Science, Volume 4123, DOI: 10.1007/11889342_20, pp. 362-380.

Marcu, Daniel. 2001. Towards a Unified Approach to Memory and Statistical-Based Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France, pp 386-393.

Moore, Robert C. 2003. Learning translations of named entity phrases from parallel corpora. In *Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003),* Budapest, Hungary; pp. 259-266.

MacQueen, James B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistcs and Probability*, 281–297, Berkeley, USA. University of California at Berkeley Press.

M. Baroni, S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3): 209-226.

Marcus, Mitchell. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, USA: MIT Press.

Matthew Richardson and Pedro Domingos. Markov logic networks. Machine Learning, 2005.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics* 19.pp. 313–330.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 73–80, Sapporo, Japan.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*, 280–287, Barcelona, Spain.

McCarthy, Diana, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 200 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 369–379.

Mendes, A., Antunes, S., Nascimento, M., Casteleiro, J., Pereira, L. and Sá, T. 2006. COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 24-26.5. 2006, Genova, pp. 1900-05.

Mihalcea, Rada, and Ehsanul Faruque. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of the ACL/SIGLEX Senseval-3* , 155–158, Barcelona, Spain.

McCarthy, Diana, and Dan Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of the 16th Conference of the American Association of Aritificial Intelligence (AAAI-1999)*, pp. 461–466, Orlando, USA.

Mimmelmann, Nikolaus P. 1998. Regularity in irregularity: Article use in appositional phrases. *Linguistic Typology*. Pp. 315–353.

Minnen, Guido, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*. Pp. 207–223.

Miyagawa, Shigeru. 1989. Light verbs and the ergative hypothesis. *Linguistic Inquiry* 20.659–668. Mohanty, Gopabandhu. 1992. The Compound Verbs in Oriya. Ph. D. dissertation, Deccan College Post-Graduate and Research Institute, Pune.

Mohanty, Panchanan. 2010. WordNets for Indian Languages: Some Issues. *Global WordNet Conference-2010*, pp. 57-64.

Madigan, David, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye.: Author identification on the large scale, In: Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA) (2005)

Moldovan, Dan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, 60–67, Boston, USA.

Mukherjee, Amitabha, Soni Ankit and Raina Achla M. 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. *Multiword Expressions: Identifying and Exploiting Underlying Properties. Association for Computational Linguistics.*

Nakov, Preslav, and Marti Hearst. 2005. Search engine statistics beyond the *n*-gram: Application to noun compound bracketting. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 17–24, Ann Arbor, USA.

Nakov, Preslav, and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA)*, 233–244, Bularia.

Nongmeikapam K and S. Bandyopadhyay. 2010. Identification of Reduplication MWEs in Manipuri, a rule based approach, *In Proceedings of the 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL).*

Nastase, Vivi, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 781–787, Boston, USA.

Ney, H. and Popovic, M. 2004. Improving Word Alignment Quality using Morphosyntactic Information. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva 23-27.8. 2004, pp. 310-314.

Ngai, Grace, and Radu Florian. 2001. Transformation-based learning in the fast lane. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL)*, 40–47, Pittsburgh, USA.

Nicholson, Jeremey, and Timothy Baldwin. 2005. Statistical interpretation of compound nominalisations. In *Proceedings of the Australian Language Technology Workshop*, 152–159, Sydney, Australia.

Nulty, Paul. 2007. Semantic classification of noun phrases using web counts and learning algorithms. In *Proceedings of the Association of Computational Linguistics 2007 Student Research Workshop*, 79–84, Prague, Czech Republic.

Nunberg, Geoffrey, Ivan A. Sag, and Tom Wasow. 1994. Idioms. *Language*. Pp. 491–538.

Odijik J. 2004. Reusable Lexical Representation for Idioms. *In Proceedings of Language Resources and Evaluation Confrrence (LREC)*, Lisbon, pp. 903-906.

ŐDowd, Elizabeth M. 1998. *Prepositions and Particles in English*. Oxford University Press.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Sapporo, Japan, pp. 160-167.

ŐHara, Tom, and Janyce Wiebe. 2003. Preposition semantic classification via Treebank and framenet. In *Proceedings of the 7th Conference on Natural Language Learning*, pp. 79–86, Edmonton, Canada.

Olsen, Susan. 2000. Against incorporation. In *Linguistische Arbeitsbertichte 74* , pp. 149–172. University of Leipzig.

ŐS´eaghdha, Diarmuid. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*.

ŐS´eaghdha, Diarmuid, and Ann Copestake. 2007. Co-occurrence contexts for noun compound interpretation. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, 57–64, Prague, Czech Republic.

P. Singla and P. Domingos. Discriminative training of Markov logic networks. In Proceedings of the Twentieth National Conference on Artificial Intelligence, pp. 868–873, Pittsburgh, PA, 2005. AAAI Press.

Paul, Soma. 2010. Representing Compound Verbs in Indo WordNet. *Golbal Wordnet Conference-2010*, pp. 84-91.

Paul, Soma. 2004. An HPSG Account of Bangla Compound Verbs with LKB Implementation. Ph.D dissertation, University of Hyderabad, Hyderabad.

Paul, Soma. 2003. Composition of Compound Verbs in Bangla. *Multi-Verb constructions*. Trondheim Summer School.

Pavelec, D., Justino, E., Oliveira, L. S. 2007. Author Identification using Stylometric features, Inteligencia Artificial, Revista Ideroamericana de Inteligencia Artifical. Valencia, Espana, Vol 11, pp. 59—65.

Patrick, Jon, and Jeremy Fletcher. 2004. Differentiating types of verb particle constructions. In *Proceedings of Australian Language Technology Workshop*, 163–170, Sydney, Australia.

Patwardhan, Siddharth, 2003. Incorporating dictionary and corpus information into a context vector. Master's thesis, University of Minnesota, USA.

Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, 17–21, Mexico City, Mexico.

Pawley, Andrew, and Frances Hodgetts Syder, 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency.

Peabody, K.W., 1981. Constraints on the productivity of verb-particle combinations. Master's thesis, Ohio State University.

Pearce, Darren. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, 41–46, Pittsburgh, Pennsylvania, USA.

Pecina, Pavel. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pp. 13–18, Ann Arbor, USA. Association for Computational Linguistics.

Piao, Scott, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions wth a semantic tagger. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 49–56, Sapporo, Japan.

Piao, Scott, Paul Rayson, Olga Mudraya, Andrew Wilson, and Roger Garside. 2006. Measuring mwe compositionality using semantic annotation. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 2–11, Sydney, Australia.

Potter, Elizabeth, Jenny Watson, Michael Lax, and Miranda Timewell. 2000. *Collins Cobuild Dictionary of Idioms*. Cambridge, UK: Harper Collins Publishers.

Pal Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay and Andy Way.2010. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation, In *proceedings of the workshop on Multiword expression: from theory to application (MWE-2010), The 23rd International conference of computational linguistics (Coling 2010),* Beijing, Chaina, pp. 46-54.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002),* Philadelphia, PA, pp. 311-318.

Prager, J., and J. Chu-Carroll, 2001. Use of WordNet hypernyms for answering what-is questions.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, USA: MIT press.

Quirk, Randolph, Sydney GreenBaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London, UK: Longman.

Ramchand, Gillian, and Peer Svenonius. 2002. The lexical syntax and lexical semantics of the verb-particle construction. In *Proceedings of WCCFL*, pp. 387–400, Somerville, USA.

Resnik, Philip. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the 3rd Workshop on Very Large Corpus*, 77–98, Cambridge, USA.

Rayson, Paul, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL Semantic Analysis System. In *proc. Of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks*, pages 7-12, Lisbon, Porugal.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, Suntec, Singapore, pp. 47-54.

Riehemann, Susanne, 2001. *A Constructional Approach to Idioms and Word Formation*. Stanford University dissertation.

Rosario, Barbara, and Hearst Marti. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, 82–90, Pittsburgh, Pennsylvania, USA.

Ross, Haj. 1995. Defective noun phrases. In *In Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, 398–440, Chicago, Illinois, USA.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, 1–15, Mexico City, Mexico.

Salton, Gerard, Allan Wong, and C.S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18.613–620.

Sanderson, Mark, 1996. *Word sense disambiguation and Information retrieval* . University of Glasgow dissertation.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation.

Sinha, R. Mahesh, K. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. *Multiword Expression Workshop, Association of Computational Linguistics - International Joint Conference on Natural Language Processing-2009*, pp. 40-46, Singapore.

Sanjiv Kumar and Martial Hebert. Discriminative fields for modeling spatial dependencies in natural images. In Sebastian Thrun, Lawrence Saul, and Bernhard Schȍlkopf, editors, Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, 2003.

Schone, Patrick, and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, pp. 100–108, Hong Kong.

Schutze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics*. pp. 97–123.

Side, Richard. 1990. Phrasal verbs: sorting them out. *ELT Journal*. pp144–52.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*. pp. 143–77.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, pp. 1297–1304, Vancouver, Canada.

Sarkar, Pabitra. 1975. Aspects of Compound Verbs in Bengali. Unpublished M.A. dissertation, Chicago University.

Stamatatos, E., Fakotakis N., Kokkinakis, G.: Automatic authorship attribution, In: proc. of the 9th Conference on European Chapter of the ACL, pp. 158--165, June 8-12, (1999) Sparck Jones, Karen. 1983. *Compound noun interpretation problems*. Englewood Cliffes, USA: Prentice-Hall.

Stevenson, Suzanne, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 1–8, Barcelona, Spain.

Su Nam Kim and Min-Yen Kan. 2009. *Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles*. In: Proc. of the 2009 Workshop on multiword Expressions, ACL-IJCNLP 2009. Suntec, Singapore. pp. 9-16.

Stvan, Laurel Smith, 1998. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. Northwestern University dissertation.

Svenonius, Peter, 1994. *Dependent nexus. Subordinate predication structures in English and the Scandinavian languages*. University of California at Santa Cruz dissertation.

S. Stevenson, A. Fazly and R. North. Statistical Measures of the Semi-Productivity of Light Verb Constructions, In *Proc. of the Second ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 1-8, 2004

Thoudam D.S and S. Bandyopadhyay. 2008. Morphology Driven Manipuri POS Tagger. *In workshop on NLP fo Less Privileged Languages*, Hyderabad, pp. 91-98

Thanopoulos, Aristomenis, Nikos Fakotakis, and George Kokkinakis. 2003. Identification of multiwords as preprocessing for automatic extraction of lexical similarities. In *Proceedings of 6th International Conference on Text, Speech and Dialogue*, pp. 8–11.

Ceske Budejovice, Czech Republic. Tschichold, Cornelia, 1998. *Multi-word Units in Natural Language Processing*. University of Basel dissertation.

Turney, Peter D. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of 9th International Joint Conference on Aritificial Intelligence (IJCAI-2005)*, 1136–1141, Edinburgh, Scotland.

Uchiyama, Kiyoko, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions* 19.497–512.

Utsuro, Takehito, Takao Shime, Masatoshi Tsuchiya, Suguru Matsuyoshi, and Satoshi Sato. 2007. Learning dependency relations of Japanese compound functional expressions. In *Proceedings of the ACL-2007 Workshop on a Broader Perspective on Multiword Expressions*, 65–72, Prague, Czech Republic.

Van de Curys, Tim, and Begona Villada Moiron. 2007. Semantics-based multiword expression extraction. In *Proceedings of the ACL-2007 Workshop on a Broader Perspective on Multiword Expressions*, 25–32, Prague, Czech Republic.

Van Der Beek, Leonoor, 2005. *Topics in Corpus-Based Dutch Syntax* . University of Rijksuniversiteit Groningen dissertation.

Vanderwende, Lucy. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, pp. 782–788, Kyoto, Japan.

Venkatapathy, Sriram, and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *the Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pp. 899–906, Vancouver, Canada.

Venkatapathy, Sriram, and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of Coling-ACL 2006: Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties,* Sydney, pp. 20-27.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In Pr*oceedings of the 16th International Conference on Computational Linguistics (COLING 1996),* Copenhagen, pp. 836-841.

Word-sense disambiguation for machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, 771–778, Vancouver, Canada.

Verma, Manindra K.1993. Complex Predicates in South Asian Languages. Manohar Publishers and Distributors, New Delhi.

Villada Moiron, Begona, 2005. *Data-driven Identification of Fixed Expressions and Their Modifiability*. University of Rijksuniversiteit Groningen dissertation.

Villavicencio, Aline. 2003a. Verb-particle constructions and lexical resources. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, 57–64, Sapporo, Japan.

Villavicencio, Aline. 2003b. Verb-particle constructions in world wide web. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in Computational Linguistics Formalisms and Applications*, Toulouse, France.

Villavicencio, Aline. 2005. The availability of verb-particle constructions in lexical resources:how much is enough? *Computer Speech and Language, Special Issue on Multiword Expressions* 19.415–432.

Villavicencio, Aline, Timothy Baldwin, and Benjamin Waldron. 2004. A multilingual database of idioms. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, 1127–1130, Lisbon, Portugal.

Wacholder, Nina, and Peng Song. 2003. Toward a task-based gold standard for evaluation of NP chunks and technical terms. In *Proceedings of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT/NAACL-2003)*, 189–196, Edmonton, Canada.

Warren, Beatrice, 1978. *Semantic Patterns of Noun-Noun Compounds*. Actr Universitatis Gothoburgensis dissertation.

Widdows, Dominic. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, 276–83, Edmonton, Canada.

Widdows, Dominic, and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of ACL2005 Workshop on Deep Lexical Axquisition*, 48–56, Ann Arbor, USA.

Wierzbicka, Anna. 1982. Why can you have a drink when you can't *have an eat? *Language* 58.753–799.

Wood, Frederick T. 1964. *English Verbal Idioms*. London, UK: Macmillan.

Wu, Zhibiao, and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-1994)*, 133–138, Las Cruces, New Mexico, USA.

Yarowsky, David. 1993. One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, 266–271, Plainsboro, New Jerey, USA.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics (ACL-1995)*, 189–196, Cambridge, USA.

Yoon, Juntae, Key-Sun Choi, and Mansuk Song. 2001. A corpus-based approach for Korean nominal compound analysis based on linguistic and statistical information. *Natural Language Engineering* 7.251–270.

Zhao, Jinglei, Hui Liu, and Ruzhan Lu. 2007. Semantic labeling of compound nominalization in Chinese. In *Proceedings of the ACL-2007 Workshop on a Broader Perspective on Multiword Expressions*, 73–80, Prague, Czech Republic.

Zhang, T., Damerau, F., Johnson, D. 2002. Text chunking using regularized winnow, In: proc. 39th Annual Meeting on ACL, pp. 539--546, July 6-11, 2001.