Extraction of Neologisms from Japanese Corpora

A thesis presented
by
James Breen
to
The School of Computing and Information Systems
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

University of Melbourne
Melbourne, Australia
December 2017
Extraction of Neologisms
from Japanese Corpora

Abstract

In this thesis an exploration of the application of natural-language processing techniques to the extraction of neologisms from Japanese corpora is described. The research aim was to establish techniques which can be developed and exploited to assist significantly in neologism extraction for compiling Japanese monolingual and bilingual dictionaries. The particular challenge of the task is presented by the lack of word boundaries in Japanese text which creates a problem in the identification of unrecorded words.

Three broad approaches have been explored, using a variety of language processing and artificial intelligence techniques, and drawing on large-scale Japanese corpora and reference lexicons: synthesis of possible Japanese words by mimicking Japanese morphological processes, followed by testing for the presence of candidate words in Japanese corpora; analysis of morpheme sequences in Japanese texts to determine the presence of potential new or unrecorded terms; and analysis of language patterns which are often used in Japanese in association with new and emerging terms.

The research described in this thesis has identified a number of processes which
Abstract

can be used to assist lexicographers in the identification of unrecorded lexical items in Japanese texts.
# Contents

Title Page ................................................................. i  
Abstract ................................................................. iii  
Table of Contents ......................................................... v  
Abbreviations used in this Thesis ...................................... x  
Citations to Previously Published Work ............................... xii  
Acknowledgments ........................................................... xiii  

1 Introduction .......................... 1  
1.1 Thesis Overview ...................... 1  
1.2 Background, Relevance and Importance of the Project .......... 3  
  1.2.1 Japanese Orthography .................. 3  
  1.2.2 Neologism Formation in Japanese .......... 4  
  1.2.3 Project Goals ............................ 8  
1.3 Conceptual Framework of the Study, Summary of Experimental Methods 9  
1.4 Lexicographic Background .................. 12  
1.5 Thesis Structure and Timeline .................. 14  

2 Neologisms: Lexicographic Issues and Terminology .......... 16  
2.1 Introduction .............................. 16  
2.2 Nomenclature ............................. 19  
2.3 What Makes Up A Lexical Item? ............. 23  
2.4 The Japanese Perspective .................. 26  
2.5 Single Items .............................. 26  
2.6 Multiword Expressions ...................... 29  
2.7 Potential Lexical Items ...................... 33  
2.8 Summary ................................. 40  

3 Prior Work and Literature Review .......................... 41  
3.1 Introduction .............................. 41  
3.2 General Lexicography ..................... 41  
3.3 Japanese NLP - Morphological Analyzers and Parsers .................. 43
### Contents

#### 3.4 Alternative Segmentation Approach ........................................ 43
#### 3.5 Identification of Unknown Words ............................................. 44
#### 3.6 Pre-candidature Work ............................................................ 46

#### 4 Resources ................................................................. 49
  4.1 Introduction .............................................................................. 49
  4.2 Dictionaries and General Lexicons ............................................ 50
  4.3 Text Corpora ........................................................................... 53
  4.4 \(n\)-gram Corpora ................................................................. 55
  4.5 Software .................................................................................. 58
    4.5.1 Morphological Analyzers ...................................................... 58
    4.5.2 Morpheme Lexicons .............................................................. 60
    4.5.3 Machine-Learning Systems .................................................... 62

#### 5 Lexical Item Identification From Morpheme Analysis .................. 64
  5.1 Introduction .............................................................................. 64
  5.2 Morphological Analysis of Japanese Text ....................................... 65
  5.3 Prior Work ................................................................................ 69
  5.4 Using Unknown Word Functions in Morphological Analyzers ............ 71
  5.5 Rule-Based Lexical Item Identification ............................................ 75
  5.6 Application of Machine Learning to Lexical Item Extraction ............ 79
    5.6.1 Basic Model ....................................................................... 79
    5.6.2 Labelling ........................................................................... 80
    5.6.3 Feature Development ............................................................. 82
  5.7 Evaluation .................................................................................. 86
    5.7.1 Overview ............................................................................ 86
    5.7.2 Testing with Automatically Marked-up Texts ......................... 87
    5.7.3 Initial Testing ...................................................................... 87
    5.7.4 Testing with Hand-Annotated Texts ........................................ 91
    5.7.5 Precision Issues ................................................................... 99
    5.7.6 Impact of Training Texts ......................................................... 100
    5.7.7 Impact of Training Size ........................................................ 101
  5.8 Testing for Potential Lexical Items in Unseen Texts ....................... 103
  5.9 Summary, Discussion and Conclusions ......................................... 107
  5.10 Postscript: Alternative Chunking Approach .................................... 109

#### 6 Japanese Loanword Multi-Word Expressions: Extraction, Segmen-
									
tation and Translation ................................................................ 111
  6.1 Introduction .............................................................................. 111
  6.2 Orthographical Aspects of Loanwords in Japanese ....................... 112
  6.3 Assimilation of Loanwords in Japanese ....................................... 114
  6.4 Loanword Multi-Word Expressions ............................................. 115
7.9.6 Discussion of 4-kanji Synthesis and Evaluation 182
7.10 Summary, Discussion and Conclusions 185
7.10.1 Future Work 186

8 Generation and Extraction of Compound Verbs 188
8.1 Introduction 188
8.2 Overview of Japanese Compound Verbs 189
8.3 Prior Work 191
8.4 General Approach and Resources 191
8.4.1 Approaches 191
8.4.2 Resources Used 193
8.4.3 Synthesis of Compound Verbs 193
8.4.4 Direct n-gram Search 195
8.5 Analysis of the Potential Compound Verbs 198
8.6 Productivity Measures of V1 and V2 Components 200
8.7 Summary, Conclusions and Future Work 206

9 Neologism Identification through Language Contexts 208
9.1 Introduction 208
9.2 Prior Work 209
9.3 Text Corpora 211
9.4 Initial Exploration 212
9.4.1 Pattern Frequencies 212
9.4.2 Testing Contexts of Known New Terms 213
9.4.3 Explicit Neologism Labelling 216
9.4.4 Parenthesized kana 217
9.5 Further Investigation 219
9.5.1 Expansion of Linguistic Patterns 219
9.5.2 Comparative Evaluation of the Language Patterns 221
9.5.3 Discussion 222
9.6 Locating Target Terms 226
9.7 Precision and Recall Issues 228
9.8 Detailed Investigation 229
9.8.1 Pattern Selection 229
9.8.2 Text Scanning and Target Term Extraction 231
9.9 Evaluation of Extracted Target Terms 235
9.9.1 WWW Corpus 238
9.9.2 Significance of Multiple Occurrences 243
9.9.3 Twitter Data 244
9.9.4 Retweets 244
9.9.5 Analysis of Retweets 246
9.9.6 Multiple Occurrence: Not Retweets 248
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.9.7 Classification of Names</td>
<td>249</td>
</tr>
<tr>
<td>9.9.8 Issue of Parenthesized Terms</td>
<td>250</td>
</tr>
<tr>
<td>9.9.9 Burstiness</td>
<td>251</td>
</tr>
<tr>
<td>9.9.10 Terms In and Out of Patterns</td>
<td>252</td>
</tr>
<tr>
<td>9.10 Summary, Discussion and Conclusions</td>
<td>254</td>
</tr>
<tr>
<td>10 Summary, Conclusions and Future Work</td>
<td>257</td>
</tr>
<tr>
<td>A Rule-Based Chunker - Encapsulations</td>
<td>279</td>
</tr>
<tr>
<td>B Lexical Item Tagging System</td>
<td>292</td>
</tr>
<tr>
<td>C Neologism Synthesis - Data and Examples</td>
<td>303</td>
</tr>
</tbody>
</table>
Abbreviations used in this Thesis

API: application programming interface
ATR: automatic term recognition
CRF: conditional random field
HMM: hidden Markov model
JCV: Japanese compound verb
ML: machine learning
MWE: multiword expression
NLP: natural language processing
NP: noun phrase
OOV: out-of-vocabulary
POS: part of speech
RNN: recurrent neural network
RPG: role-playing game
SNR: signal-to-noise ratio
SVM: support vector machine
Citations to Previously Published Work

Large portions of Chapter 1, Chapter 6, Chapter 8 and Chapter 9 have appeared or will shortly appear in the following papers and book chapters:


Plans for additional publications were unfortunately disrupted by prolonged periods of medical treatment and recovery during the latter part of the candidature.
Acknowledgments

The research project reported in this thesis was carried out under the supervision of Professor Timothy Baldwin of the School of Computing and Information Systems, The University of Melbourne, and Associate Professor Francis Bond of the Division of Linguistics and Multilingual Studies, Nanyang Technological University, Singapore.

The author greatly appreciates the support, guidance and counsel of the supervisors, both of whom are friends of long standing. Indeed without their important roles this project would never have been possible.

Many other people have assisted with advice, suggestions and commentary, in particular the people, too numerous to name, who contributed to the annotation exercise described in Chapter 5.

The project was supported financially by the Australian Government through the Research Training Scheme (RTS) and a Research Training Program (RTP) Fee-Offset Scholarship.

Finally, I must acknowledge the support and forbearance of my wife Julia during this extended project. In fact it was her engagement with the Suzuki Method of music education and the move of our family to Japan in 1981 to enable her to undertake study in that method that has led, by many tortuous paths, to this project.
Chapter 1

Introduction

1.1 Thesis Overview

This thesis reports on an exploration of the application of natural-language processing techniques to the extraction of neologisms from Japanese corpora. The research aim is to establish techniques which can be developed and exploited to assist significantly in neologism extraction for compiling monolingual and bilingual dictionaries.

For the purposes of the project, the term “neologism” has been extended to include all words, phrases, expressions, etc. which could reasonably be expected to appear in a bilingual lexicon, but which are not so recorded. The title of the thesis could have referred to “potential lexical items” but the focus of the project has largely been on new terms and in fact the majority of terms extracted during the investigation have been relatively new. The term “extraction” is used to encompass the identification of neologisms as well as, where possible, the establishment of their pronunciation and
meanings.\textsuperscript{1}

As an illustrative example of some of the challenges and issues associated with neologism identification and extraction in Japanese, consider the following relatively simple and topical sentence:\textsuperscript{2}

オバマの成功が生み出した脱真実の米国、トランプの時代がやってくる。\textsuperscript{3}

Some of the issues, such as the lack of clear word boundaries and use of a mixture of three different scripts are discussed below, but the focus of this example is on the sequence of three \textit{kanji} characters 脱真実. Their location in the sentence indicates that they probably, but not inevitably, form a single term. Closer investigation indicates that they are likely to be the word 真実 shinjitsu “truth” in association with the \textit{kanji} 脱 which is read datsu and usually operates as a prefix meaning “de-”, “ex-”, “un-”, etc., and in fact 脱真実 is indeed a newly-coined term equating to the English “post-truth”. Note that this may have not have been the case, and that such a 3-\textit{kanji} sequence may have straddled a boundary between adjacent clauses.

\footnotesize
\textsuperscript{1}In general the term “reading” will be used when labelling the spoken syllables or morae associated with a Japanese term, especially one written using \textit{kanji}. This is the more usual practice with Japanese as the syllabic transcription is not always the same as the actual pronunciation, which may differ slightly due to such things as devoicing and moraic nasals.

\textsuperscript{2}In general and where appropriate Japanese terms and text passages will be accompanied by a transliteration using the Modified Hepburn variety of Japanese romanization (in italics) and a translation into English.

\textsuperscript{3}obama no seikō ga umidashita datsushinjitsu no beikoku, toranpu no jidai ga yatte kuru “[In the] post-truth America created by Obama’s success, the Trump era has arrived.”
1.2 Background, Relevance and Importance of the Project

1.2.1 Japanese Orthography

At the heart of this project is the nature of Japanese orthography, with its use of multiple scripts and in particular the absence of any indication of the boundaries between the syntactic elements in texts (Backhouse 1993; Seeley 2000; Bond et al. 2016). Therefore it is appropriate before proceeding any further to provide a very brief overview of Japanese orthography.\(^4\)

Modern Japanese is written in a mixture of four scripts:

a. *kanji* (Chinese characters), e.g. 猫, 犬, 鳥, 牛, etc. which are used mainly for nouns and the roots of verbs, adjectives, etc. Approximately 2,500 *kanji* are in common use, although the full set available is estimated to be around 80,000. Most nouns written with *kanji* use two or more characters, whereas verbs typically use a single *kanji* as the root component;

b. the *hiragana* syllabary (46 symbols plus diacritics: あいうえおかきくけこ, etc.)

In modern Japanese *hiragana* is used mainly for particles, verb and adjective inflections, conjunctions, etc.;

c. the *katakana* syllabary (also 46 symbols plus diacritics: アイウエオカキクケコ, etc.) *Katakana* are currently used for loanwords, scientific names, transcriptions

---

Chapter 1: Introduction

of foreign names, etc.;

d. the letters of the Latin alphabet. While it is possible to write Japanese using this alphabet (in Japanese this is known as ローマ字 or rōmaji), in practice the use of the Latin alphabet in Japanese text is confined to such things as initials and acronyms (PC, NEC, USB, bps, etc.) plus occasional words such as product names (iPhone, Windows, etc.).

An illustration of the use of the main scripts can be seen in the sentence スーパーで食品を買いました:

スーパー で 食品 を 買いました
sūpā de shokuhin wo kaimashita
supermarket at food bought
“[I] bought some food at [a/the] supermarket”

Here kanji are used for the noun 食品 shokuhin “foodstuffs” and the root of the verb 買う kau “to buy”, hiragana are used for the particles で de: (location of action) and を wo (object marker) and the polite past-tense inflection of the verb (いました imashita), and katakana is used for the abbreviated form of the loanword スーパーマーケット sūpāmāketto “supermarket”.

1.2.2 Neologism Formation in Japanese

Despite having a rich lexicon, the Japanese language has a noted tendency to adopt and create new words (Kageyama 2014). While the reasons for adopting new words are varied, there are a number of processes associated with the Japanese language which tend to encourage neologism creation:

5The hiragana and katakana syllabaries are often referred to collectively as kana.
a. the readiness to accept loanwords. Unlike some countries, which attempt to restrict loanword usage, Japan has placed no formal restriction on their use. Estimates of the number of loanwords used in Japanese range as high as 100,000. Most of these words have been borrowed directly from English, however many are from other languages such as Chinese and Korean and a significant number, known as 和製英語 waseieigo “Japanese-made English”, have been assembled from English words or word fragments.

Further description of Japanese loanwords can be found in Section 6.

b. the accepted morphological process of creating words by combining two or more kanji chosen for their semantic properties. This process was used extensively in the mid-19th century when Japan re-engaged with the rest of the world and needed an expanded lexicon to handle the technological and cultural information flowing into the country. (For example, the Japanese word for the sound-post in a string instrument is 魂柱 konchū, which is formed from the kanji meaning “soul, spirit” and “pillar, post”. This has probably been created from the Italian term for sound-post which is “anima” or the French term “âme”, both meaning “soul”.) This process has continued, although creation of such basic lexical items is not as common, and more use is made of loanwords for new objects, concepts, etc.

A broadly similar process is used to create compound verbs (see Section 8.2).

c. the tendency to create abbreviations, particularly from compound nouns and long loanwords. For example, the formal term for “student discount” in Japanese
is 学生割引 gakusei waribiki, however the common term is 学割 gakuwari formed from the first kanji in each of the two constituent nouns. A similar process is applied to loanwords, resulting in words such as セクハラ sekuhara for “sexual harassment”, a contraction of セクシュアルハラ スメント sekushuaru harasumento (Itō 1990; Tsujimura 2006).

Many neologisms find their way eventually into published dictionaries, and there are several special neologism dictionaries known as 新語辞典 shingojiten “new word dictionary” or 現代用語辞典 gendaiyōgojiten “neologism dictionary”, however many abbreviations, compound verbs and loanwords are less well lexicalized as native speakers can usually recognize them as such and derive the reading and meaning.

Traditional techniques for identifying neologisms involve extracting lexical items from texts and comparing them with a lexical database. European languages in recent centuries have established the practice of marking word boundaries with a space or punctuation character, thus making the extraction phase of the identification of neologisms a relatively straightforward process. (“It is easy for computer programs to spot completely new words” (Atkins and Rundell 2008).) That ease of identification of neologisms vanishes in languages such as Japanese and Chinese where no explicit marking of word boundaries takes place. As a result, automated parsing and morphological analysis of such languages is a much more significant challenge. With Japanese the process is somewhat aided by the use of four scripts, as the script changes are valuable clues when parsing texts.

Much of the attention given to unlexicalized Japanese words and expressions in the Japanese natural language processing (NLP) literature has been in the context
of the development and use of parsers and morphological analysis systems. In this context, the detection of out-of-lexicon words has been usually been characterized as a problem to be overcome, rather than a goal of the analysis, and often as an adjunct to part-of-speech (POS) assignment. Accurate parsing and word segmentation of Japanese (and Chinese) was initially regarded as a task too hard for computers, however from the early 1990s a series of research and commercial analyzers have been developed with significantly improving accuracy and effectiveness. All these systems use a combination of large lexicons and a range of artificial intelligence techniques. The field is now generally regarded as mature (Den et al. 2008), however given the role of lexicons in the parsing process, neologisms and other out-of-vocabulary (OOV) terms may pose problems, and typically will produce sequences of isolated morphemes until they can resynchronize.

As the lexicographer wishes to discover unrecorded terms in the context of their usage, it is important that techniques be explored and developed that enable them to be detected in texts, despite the problems they cause to analysis systems. In addition, lexicographers wish to extract not just words, but common collocations of words in multiword expressions, compound nouns and verbs, idiomatic expressions, etc. In Japanese there is also the need to establish the reading of unrecorded words. When these words are in hiragana or katakana this is not a problem, but for words written using kanji a problem arises as kanji usually have several readings, and there are also euphonic changes associated with some character combinations (連濁 rendaku) which are not reliably predictable (Kubozono 2005).
1.2.3 Project Goals

The aim of this study is to explore techniques that will enable neologisms to be identified and extracted from texts, and for common groupings of words to be also identified and extracted. As mentioned above, the term “neologism” has been extended to include all words, phrases, expressions, etc. which could reasonably be expected to appear in a bilingual lexicon, but which are not so recorded. This bilingual context is important, as a strictly monolingual focus will tend to overlook constructions which arise from regular Japanese morphological processes but which are not clear to non-native speakers. For example the word凹状 ōjō does not appear in Japanese monolingual dictionaries or in Japanese-English dictionaries, but appears in several English-Japanese dictionaries as a Japanese gloss for the word “concavity”. The reason for its absence is that it is a construction from two kanji, a prefix凹 meaning “concave” and a noun状 meaning “shape”, something that may be obvious to a literate Japanese, but not necessarily to others. This means that to address the requirements of bilingual dictionaries a somewhat broader scope will need to be applied than would normally be found in Japanese lexicography.

As will be explained further below, the purpose of this study is to carry out research into the appropriate NLP and artificial intelligence techniques which will facilitate such neologism extraction, examining such alternatives as an extension of morphological analysis, synthesis of potential neologisms, and text classification, with the long-term goal of enabling the development of tools which assist lexicographers.
1.3 Conceptual Framework of the Study, Summary of Experimental Methods

The investigation into neologism extraction has been broken into three main approaches:

A. analysis of morpheme sequences in Japanese texts to determine the presence of potential new or unrecorded terms. This analysis is broken into two groups:

   a. analysis of native Japanese words using morphological analysis software.

      This analysis is in several stages:

      i. processing the texts with a morphological analyzer to produce sequences of tagged morphemes;

      ii. preliminary analysis of the output to detect possible OOV occurrences based on the analyzer’s assessment of the morphemes;

      iii. additional tagging of the morphemes with features derived from combinatorial data derived from large lexicons and corpora;

      iv. processing the tagged morphemes with an ML-based chunker with the aim of (re)assembling candidate words and expressions;

      v. processing the morpheme sequences using alternative techniques such as rule-based chunking with the same aim.

   b. analysis of loanwords, in particular, compounds made from several source words. Since loanwords are written in the katakana script, they can be readily separated from other Japanese text and analyzed using traditional
lexicon-checking techniques. The particular focus will be on appropriate segmentation and translation of these compounds.

B. synthesis of possible Japanese words by mimicking Japanese morphological processes, followed by testing for the presence of candidate words in Japanese corpora. Possible candidates for this approach include:

a. abbreviations of multiword expressions and compound nouns. This form of ellipsis is very common in Japanese, as in the example of 学生割引 quoted above (Section 1.2.2).

b. single/double-kanji word plus affix. Affixation using a single kanji is a very common morphological process in Japanese. Around 100 kanji are commonly used for this role, e.g. 超 chō “super-”, “ultra”, 各 kaku “each”, “every”, 新 shin “new”, 全 zen “all”, “whole”, 化 ka “-ization”, “-ification”, 流 ryū “-style”, “-like”, etc. however many more can play such a role. It can be debated whether these really lead to neologisms. As in the case of 凹状 quoted above, they may lead to terms which are useful in the context of a bilingual dictionary. (It is worth noting that the UniDic morpheme lexicon (see Section 4.5.2) includes a large number of constructions arising from affixation, among them 凹状.)

c. mass compound generation. With approximately 2,500 kanji used in modern Japanese, the number of possible 2-kanji compounds is “only” 6.25 million. The synthesis of 4-kanji compounds from pairs of 2-kanji compounds is on a larger scale, as for example the JMdict dictionary contains about 48,000 such 2-kanji compounds. The availability of a Japanese n-gram
corpus makes at least the initial testing of such combinations a relatively tractable task.

d. compound verbs. These are formed from two (or more) verbs, and are very common and highly productive, e.g. 歌い始める utaihajimeru “to start singing” from 歌う utau “to sing” and 始める hajimeru “to start” or “to begin”.

C. analysis of language patterns which are often used in Japanese in association with new and emerging terms. These patterns are usually associated with discussions or explanations of new terms, and can often be found in newspaper articles, WWW pages, blogs, etc. For example, in the passage オタ芸(オタげい・ヲタげい)とは、アイドルや声優などのコンサートや... the parenthesized readings in hiragana/katakana and the following と is particle are typical of an explication of a term.

All the experimental approaches outlined above are complementary to a degree. The synthesis method, which is relatively mechanical and largely context-free, results in quite high levels of recall and relatively low precision whereas the goal of the language context method is to see if very specific contexts aid precision and relevance at the price of a reduced recall.

In addition to the above, the generation of possible readings for neologisms in kanji and English glosses will be investigated.

As far as possible, candidate neologisms will be tested for diachronic behaviour by checking historical data holdings (newspapers, WWW archives, etc.) both to examine

---

6 otagei(otagei,wotagei)towa, aidoru ya seiyū nado no konsāto ya ... “as for the term ‘dancing and shouting by fans’ (pronunciations), for young stars and voice actors, etc. in concerts ...”
the development of the words and to determine whether they are ephemeral.

1.4 Lexicographic Background

The stated research aims of the project (above) included: “to establish techniques which can be developed and exploited to assist significantly in neologism extraction for compiling monolingual and bilingual dictionaries.” The particular lexicographic context of the research project is the candidate’s engagement since 1991 in the ongoing compilation of a major open-source and largely crowd-sourced Japanese dictionary. This extensive experience and engagement with Japanese lexicography has provided the impetus for the research program reported in this thesis, and has informed the program in terms of bringing to bear the past experience and the requirements of practical lexicography. In turn the ongoing lexicographic work has benefited from the developments in the research program and the insights it has provided.

The dictionary compilation project began with work on the expansion of the small EDICT (Electronic DICTionary) Japanese-English dictionary file (Breen 1995). This dictionary had a very limited microstructure, and in 1999 the information was converted into an expanded XML structure in order to provide a better basis for recording Japanese terms and to reflect the complexities associated with the orthography and lexical items. The resulting dictionary, which has a significantly expanded microstructure and allows for multiple surface forms and readings within entries, and for glosses in several languages, was titled the Japanese-Multilingual Dictionary (JMdict) (Breen 2004b). The EDICT format of the data, which was in use by a number of applications, was retained as a legacy system and generated regularly from the
expanding JMdict file. Further details about JMdict are provided in Section 4.2.

Initially the dictionary file was maintained by the candidate with occasional contributions from others. In 2003 a WWW-based system was developed to enable online submission of amendments and new entries, and in 2010 the data was incorporated into a specially-developed online database system which allows for multiple editors to evaluate proposed additions and amendments (Breen and McGraw 2013), and provides full and open recording of the amendments and editorial process. At present the dictionary database is handling an average of 10 new entries and 50 amendments per day.

While much of the work associated with the compilation of a bilingual dictionary consists of extracting and translating material from source language lexicons, e.g. the excellent Japanese dictionaries mentioned in Section 4.2, a particular challenge is the identification and translation of neologisms. In many ways the investigation of neologisms is one of the more interesting and challenging tasks in lexicography. The usual path for the inclusion of an entry for a neologism in JMdict has been for it to be detected and submitted by a contributor, but there has also been an interest in anticipating such submissions by searching contemporary text sources for neologisms and other unrecorded terms. It is this interest that has provided the impetus and focus for the present project.

\footnote{http://www.edrdg.org/wiki/index.php/Main_Page#The_JMdict.2FEDICT_Project}
1.5 Thesis Structure and Timeline

The structure of the thesis is as described below. As the candidature was on a part-time basis, and was severely impacted by a major medical issue which involved extensive periods of suspension, it is appropriate to indicate when the various components of the project were conducted. These are included below.

a. Chapter 1 (Introduction) covers the background and broad approach taken in the research, and includes an overview of Japanese orthography, which lies at the heart of the research challenge.

b. Chapter 2 (Neologisms: Lexicographic Issues and Terminology) examines the lexicographic issues associated with the identification of terms for inclusion in dictionaries. (2013)

c. Chapter 3 (Prior Work and Literature Review) explores the prior work in the field, including the development of morphological analysis systems which are an essential element in Japanese NLP.

d. Chapter 4 (Resources) describes the lexical, corpora and software resources that have been assembled and used in the project.

e. Chapter 5 (Lexical Item Identification From Morpheme Analysis) describes a major investigation into the broad evaluation of Japanese text at the morpheme level, with the goal of identifying suitable approaches for assembling and evaluating candidate lexical items. (2010–2011)
f. Chapter 6 (Japanese Loanword Multi-Word Expressions: Extraction, Segmentation and Translation) continues the morpheme-level investigation, with a focus on the segmentation and ultimate translation of the very common loanword multiword expressions in Japanese. (2012)

g. Chapter 7 (Neologism Synthesis) explores the approach of mimicking Japanese morphological processes with a view to identifying unlexicalized Japanese words, and evaluates the use of machine-learning and other approaches to assist in the identification of candidates for further evaluation. (2013–2015)

h. Chapter 8 (Generation and Extraction of Compound Verbs) continues the approach of imitation of Japanese morphological processes with a detailed investigation into synthesizing the common and productive compound verb form. (2009)

i. Chapter 9 (Neologism Identification through Language Contexts) investigates the use of language patterns commonly associated with new or rare terms to identify possible neologisms for further analysis. (2016)

j. Chapter 10 (Summary, Conclusions and Future Work) provides an overview of the research outcomes and indicates possible future work in the field.
Chapter 2

Neologisms: Lexicographic Issues
and Terminology

2.1 Introduction

In the statement of the research goals of this project, i.e. the exploration of the application of natural-language processing techniques to the extraction of neologisms from Japanese corpora, the term “neologism” has been interpreted as meaning all words, phrases and expressions which could reasonably be expected to appear in a bilingual lexicon, but which are not so recorded. Thus our target is not confined to new formations, and may include items which either have hitherto escaped the attention of lexicographers, or which have been used infrequently but have now come to be used frequently enough to warrant inclusion.

In this chapter the important issue of the criteria for determining whether or not a word, phrase or expression should be included in a dictionary is discussed, both in
general terms, and in the context of the Japanese language and orthography. The secondary but important issue of what to call these “things that go in dictionaries” is also examined. Neologism detection in general is discussed elsewhere (see Chapter 3).

To illustrate the issue of scope of inclusion in dictionaries, consider the entry for the Japanese word 協商 kyōshō as it occurs the 5th edition of the Kenkyūsha’s New Japanese-English Dictionary (Toshiro et al. 2003). A summary of the entry is:

きょうしょう 【協商】
an entente; an understanding; an agreement.
→ 協商の上定める decide upon sth in consultation <with>
→ 協商を結ぶ conclude an entente <with>
協商国 a party to an entente; an entente.

The key elements of the entry are:

a. the basic entry information consisting of the word itself (協商), its reading (きょうしょう) and its meaning (an entente; an understanding; an agreement)

b. some phrases showing typical usage of the word;

c. a compound noun (協商国) formed from 協商, along with its translation.

The entry can thus be seen to be covering two words: 協商 and 協商国, with the latter presented as a subentry of the former. In the printed edition the entries are ordered by reading, in this case きょうしょう, and in the electronic and online editions the entry is indexed on 協商, きょうしょう and 協商国. Thus the coverage of the dictionary can be seen to include both 協商 and 協商国, with the presentation of them within the one entry being a matter for the microstructure of the individual dictionary.

In the illustration above, we showed an example of the inclusion of two terms, 協商 and 協商国, in a major dictionary. A question arises as to what to call these
terms within the context of a dictionary and within the terminology of the field of lexicography in general. There are several terms commonly used for such terms: e.g. headwords, lemmas, lexemes, lexical items, and it is useful to have a consistent and accepted term to use throughout this thesis.

It is also appropriate to examine what sorts of terms are appropriate for inclusion in dictionaries, and the criteria for assessing them. Some, such as basic nouns, verbs, etc. in common use, are clear candidates for inclusion, but in the cases of longer formations such as expressions and compound nouns the situation is not as clear. In regular dictionary compilation a degree of looseness is not usually a problem; most dictionary compilers have sets of rules as to what is and is not suitable for inclusion, and these can evolve over time, be rather fuzzy in edge cases, be subject to intuition or “gut feel” without causing major issues. It can get difficult, however, when a term-extraction methodology is under test and human annotators are attempting to assess the acceptability or otherwise of candidates. In order to present valid research findings with meaningful levels of inter-annotator agreement the actual definition of what is being sought and criteria for assessing candidates becomes pressing, and it is important to place it on a firm and objective footing.

As this study is not one of lexicographic theory and practice, it is appropriate that we draw guidance from the extensive literature in the field, and in particular from the material published by professional lexicographers in the form of guidelines for dictionary compilation. Accordingly we have concentrated on the structures and advice found in three major recent publications in this area:

(Svensén 2009)

c. Dictionaries; The Art and Craft of Lexicography (Landau 2001)

2.2 Nomenclature

Perhaps the most common label for something that could go in a dictionary is “headword”, which the Shorter Oxford English Dictionary (SOED) (Little et al. 2007) defines as “a word forming a heading (of an entry in a dictionary, etc.)”. The SOED uses that label in its front-matter, where it says “(e)very entry opens with a headword [...] The headword is the word whose meaning, etymology, history, reading, etc. are the subject of the entry.” The clear implication is that the term is only used for single words, which is indeed the case with the vast majority of entries in the SOED, exceptions being a small number of foreign phrases such as au revoir. Where the SOED does record English phrases they appear in what it refers to as “sub-entries”; thus the phrase angry young men appears within the entry for angry.

The use of “headword” as a label then is slightly problematical in that it carries the implication of referring mainly to single words, and also as something that “heads” an entry in some way, whereas our targets would include sub-entries such as angry young men and 協商国 (in the Kenkyūsha example above).

An early candidate in our search for an appropriate label for something that could go in a dictionary was “lexeme”, which seemed suitably general and which intuitively carries the meaning of something in a lexicon. The SOED defines it as “a lexical unit in the vocabulary of a language; a morpheme representing such a unit”, and
for “lexical” it has: “(1) of or pertaining to the words or vocabulary of a language”. Between them they are saying that a lexeme is a word in a language, which again is a rather tight definition, and also does not seem to lend itself to being applied to phrases or expressions. A further problem with “lexeme” is that it carries some baggage from its use in fields such as computer science, where in lexical analysis it is often a synonym for token.

Atkins and Rundell avoid lexeme completely, and use the phrase “lexical item” to describe the subject of an entry. It is useful to quote their definition:

“A ‘lexical item’ is any word, abbreviation, partial word, or phrase which can figure in a dictionary (often described as the headword of an entry) as the ‘target’ of some form of lexicographic description, most commonly a definition or a translation. [...] What constitutes a lexical item is to a certain extent language-specific, ...”

That appears to be elegant and to the point, however they muddy the water somewhat in the text that precedes this definition where they discuss the concept of lemmas, and remark:

“some types of multiword lemma [...] regularly appear as headwords in dictionaries”, but then go on to say “(f)or clarity in this chapter we will use the term headword to denote a lemma when it is the headword of an entry in a dictionary, and keep lemma for discussions of meaning and grammar.”

So is their preferred term “headword” or “lexical item”, or both? They appear to treat them as interchangeable in their Guide, but it is apparent that they use “headword” more often.

Landau does not really provide a tight definition; instead beginning his chapter on the key elements in a dictionary by briefly mentioning,
“the elements that make up a dictionary entry, beginning with the alphabetized headword, or main entry, by which the word or expression being defined (the lexical unit) is identified. The canonical form, sometimes called the *lemma*, is the form chosen to represent a paradigm; most headwords, with the exception of cross-references and names, are canonical forms.” (p. 98)

From there he goes into several pages on canonical forms; obviously a favourite topic of his. He occasionally uses “lexical unit”, although it seems to be a headword too, and is sometimes, but not always, a “lemma”. He also agrees with Atkins and Rundell in regarding a lemma as really being the canonical form of something that inflects.

Svensén is in no doubt about what to call it; for him it is a “lemma”; in fact he devotes a whole chapter to “The Lemma”,¹ which he begins with the comment:

“(a) common synonym of ‘lemma’ is ‘headword’. In this book the term ‘lemma’ is preferred mainly because ‘headword’ will be somewhat problematic when the lemma consists of a multiword lexical item”.

There is some support for this use of “lemma” from the SOED which carries the following gloss: “3. A word or phrase defined in a dictionary, glossed in a dictionary, entered in a word-list, etc.; the form of a word or phrase chosen to represent all inflectional or spelling variants in a dictionary entry, etc.” (The SOED lemma entry contains two other senses relating to axioms and arguments.)

Svensén does not seem to regard the common reservation of “lemma” to mean the canonical form of an inflecting word or expression as a matter of importance, which is a pity as a useful distinction is lost. The anonymous author(s) of the Wikipedia article on “lemma” make this point:

¹He heads that chapter with a witty quotation from the French lexicographer Alain Rey: “Mais qu’est-ce qu’un mot? Question simple en apparence, sauf pour les linguistes.” (So what is a word? An apparently simple question, except for linguists.)
“(i)n morphology and lexicography, a lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words (headword). In English, for example, run, runs, ran and running are forms of the same lexeme, with run as the lemma. Lexeme, in this context, refers to the set of all the forms that have the same meaning, and lemma refers to the particular form that is chosen by convention to represent the lexeme. In lexicography, this unit is usually also the citation form or headword by which it is indexed.”

We note that the term lemma is also widely used in computational linguistics in the context of canonical forms, with the well-recognized process of “lemmatization” playing a major role in many applications.

Thus, although there is some support for the use of “lemma” as the label for dictionary candidates, on balance its use is not really appropriate, except in the context of canonical forms where it plays a useful role.

That leaves a choice between “headword” and “lexical item”. Certainly in practical lexicography “headword” is a very commonly used term, but one can share Svensén’s discomfort at it being used for multiword terms. Also it is useful to distinguish between the words/terms being lexicalized in the construction of an electronic dictionary database, and the words/terms in the macrostructure of an electronic dictionary application by which the entry is accessed.

To illustrate this, consider a typical dictionary entry for the Japanese verb 掛ける kakeru, which has an alternative surface form of 懸ける and the reading かける. This can be regarded as one lexical item (< 掛ける | 懸ける, かける >) with three headwords (掛ける, 懸ける and かける). After all, all three are “the ‘target’ of some form of lexicographic description” (Atkins and Rundell).

---

From this, it is proposed that the following terminology be used throughout this thesis:

a. lexical item: the collection of Japanese words, terms and MWEs, including readings and alternative surface forms, which are the basis of a dictionary entry and which form the target for lexicographic description;

b. headword: the component or components of the lexical item which are used to index and access an entry in an electronic dictionary application;

c. lemma: the canonical form of an inflecting word or term within a lexical item.

### 2.3 What Makes Up A Lexical Item?

A clear message in the literature of lexicography is that the overarching principle on what could or should go in a dictionary is something that a user or reader could reasonably be expected to look up. As Landau puts it: “(a) dictionary beyond all else must be useful” and “(t)he selection of entries depends very much on the compiler’s sense of what is wanted” (p. 359).

As mentioned earlier, our target is for lexical items which “could reasonably be expected to appear in a bilingual lexicon”. The bilingual aspect is important — bilingual dictionaries have to address two target sets of users:

a. those who speak the source language and who wish to know how a lexical item is expressed in the target language (Atkins and Rundell call this an encoding dictionary and Svensén calls it a production dictionary);

b. those who speak the target language and wish to know what a lexical item means in their language (Atkins and Rundell call this a decoding dictionary and
Svensén a passive dictionary).

Most bilingual dictionaries are constructed for one or other of these target groups. While the main differences between the dictionary types occur in the microstructure of entries, e.g. part of speech, reading, explanation, examples, it can also affect the choice of lexical items. When the Kodansha’s Furigana English-Japanese Dictionary (Yoshida and Nakamura 1996) was produced in the mid-1990s, effectively by converting a coding dictionary (the Kodansha Howdy English-Japanese Dictionary) into a decoding one, the editors added hundreds of additional entries, e.g. kimono and sushi, which had not been relevant to the original.3

When it comes to the selection of candidates for lexical items some are fairly obvious: nouns, verbs, adjectives, etc. Others are less obvious: affixes, compound nouns, expressions. Since very few dictionaries set out to be complete in their coverage, there is usually a “commonness” test applied to filter out such things as less-common words and ephemera. In monolingual dictionaries there is a general rule that all words/terms that appear in definitions must also appear as headwords. In bilingual dictionaries, e.g. Japanese-English (and by extension English-Japanese), it is appropriate to cast the net fairly widely to cater for language learners.

It is instructive to examine the advice in the handbooks on this vital question. Atkins and Rundell categorize lexical items into “single items” and “multiword expressions”, with several sub-categories:

a. Single Items
   i. simple words:

3Partially discussed in the front-matter of the dictionary, and supplemented by private communication from one of the editors.
Chapter 2: Neologisms: Lexicographic Issues and Terminology

25

• lexical words: nouns, adjectives, etc.

• grammatical words: prepositions, conjunctions, pronouns, etc.

  ii. abbreviations and contractions

  iii. partial words: affixes (bound, productive), combining forms

b. Multiword Expressions

  i. fixed and semi-fixed phrases

  ii. other phrasal idioms

  iii. compounds

  iv. phrasal verbs

  v. support verb constructions

In Atkins and Rundell’s words multiword expressions “pose real problems of identification”, and

  “are a central part of the vocabulary of most languages, and need to be accounted for in the dictionary. They are particularly important for learners’ dictionaries, both monolingual and bilingual, since language learners may not recognize them as a significant units of meaning, cannot usually compose them, and will often have problems understanding them.” (p. 167)

They then analyze the issues in considerable depth, devoting ten pages to giving “a very pragmatic lexicographer’s-eye-view of this difficult area of language”.

Landau is rather less systematic when establishing criteria for the inclusion of multiword expressions, however he does describe a “stress test” to determine if a compound is regularly used as a unit. In this test, the pronunciation of the unit typically has a stress on the first word and little or no gap between them. (This seems to work in English when one compares *guinea pig* with *large pig*, but there is no clear parallel in Japanese.)
Svensén, although a little more systematic than Landau, covers the issue of filtering multiword expressions fairly lightly, and does not add anything useful to the breakdown by Atkins and Rundell. A criterion he proposes for the inclusion of multiword expressions as lexical items is that they should have been *lexicalized*, i.e. “stored in our mental lexicon” as a unit. That is rather obvious, and really the whole point of the exercise is to discover what has been, or should be, stored thus.

### 2.4 The Japanese Perspective

The categorization of lexical items proposed by Atkins and Rundell (summarized above) needs to be assessed as to its relevance to Japanese. Despite their “language-specific” warning, it is apparent that much of their structure is quite relevant to Japanese lexicography. The same single item/multiword expression dichotomy applies, and the issues with multiword expressions are very similar.

The following are some points which emerge from a consideration of the application of their categorization to Japanese.

### 2.5 Single Items

While the handling of single items is very similar to that taken in English, there are issues with deciding whether potential lexical units are single items or compounds. For example:

a. the lack of word boundaries in the orthography, and morphological processes as-
sociated with forming kanji compounds can result in uncertainty as to whether, for example, a 2-kanji sequence is an integral morpheme, and hence a candidate for being treated as a single item, or an expression, or of no relationship at all. An example could be the kanji juxtaposition 皆私 in sentences such as 彼らは皆私たちに大変親切にしてくれた，皆私のせいです and 皆私が悪いのです. While at first glance the kanji pair may appear to be behaving as though it were a unit, in fact the two kanji are independent terms, with 皆 being a floating quantifier effectively modifying a dropped subject. Isolating a kanji pair surrounded by particles is a useful filter for exposing potential single items, but deeper analysis is clearly needed to validate the selection.

b. to a certain extent, the decision will be driven by what the morphological analysis lexicons deem to be units (although this will not apply to real neologisms, which are not in such lexicons). A term such as 真っ青 massao “deep blue, pallid” can be regarded as a compound as it consists of the noun 青 ao “blue” preceded by the prefix 真 ma “pure, genuine” and っ tsu which adds emphasis to the prefix. The major Japanese morphological analyzers and their associated lexicons, however, treat it as a single-morpheme noun, along with other 真 formations such as 真水 mamizu “fresh water”. (This situation with strongly-bound 真 formations shows how fuzzy the classification into single items and compounds can be.)

c. the common use of compound verbs in Japanese. Although these behave as regular verbs, and are always potential lexical items, they are probably best

---

4 karera wa mina watashitachi ni taihen shinsetsu ni shite kureta “All of them were kind to us”.
5 mina watashi no sei desu “Everything is my fault”.
6 mina watashi ga warai no desu “All of us are bad”.
treated as compounds when deciding on their inclusion, as their formation is highly productive and their creation and meaning is often predictable. (Here the issue of inclusion on behalf of learners using a decoding dictionary is particularly relevant.)

d. inflected forms, such as 助けて *tasukute* “Help!”. At first glance this might regarded as a single item (inflected simple verb, translates to a single word), but such forms really need to be treated as expressions for the purposes of evaluation.

Apart from the above caveats, there appear to be no particular issues associated with simple words in Japanese.

In the category of abbreviations and contractions the situation is similar to English. Some specific comments are:

a. alphabetisms are widely used and lexicalized in Japanese. They can be divided into several categories:

i. adoptions from other languages, especially English: BBC, DVD, DNA, etc.

ii. abbreviated forms of the English names of Japanese organizations: JAL, JR, etc.

iii. (less commonly) abbreviated forms of the Japanese names of organizations, e.g. NHK which are the initials of Nippon Hōsō Kyōkai (日本放送協会 - the Japan Broadcasting Corporation).

iv. colloquial use of letters to encode Japanese terms, e.g. OL is sometimes written for オフィス・レディー *ofisu.redī* “office lady”, H for 変態 *hentai* “abnormality, pervert” is often used to mean “indecent, lewd” as in Hビデ
オ ecchibideo “explicit video”, KY is a colloquial term formed from 空気読めない kūkiyomenai “unable to read the situation”.

On occasions the katakana transcription of the letters will be used, e.g. ディー ブイディー dībuidī “DVD”, エヌエッチケー en’ecchikē “NHK”.

b. acronyms are commonly used; usually in a manner similar to English, e.g. JASDAQ (pronounced jazudakku), JIS (pronounced jisu).

c. contractions, such as 高校 kōkō “high school” as a contraction of 高等学校 kōtōgakkō, are very common in Japanese and are a major element in word morphology, to the extent that many (such as 高校) are now classified as morphemes.

Affixation is common in Japanese, with a large number of prefixes and suffixes (often written with a single kanji) in use. Most of these are recorded as entries in dictionaries. Japanese also uses an extensive range of counters, again often written with a single kanji such as 匹 hiki “small animals” and 本 hon “cylindrical objects”. They usually follow numerics (三匹, 二本, etc.) and are already lexicalized. In learner dictionaries composed forms are sometimes recorded when the reading is not regular.

Atkins and Rundell also mention combining forms such as the leafed in flat-leafed. There appears to be no real equivalent of these forms in Japanese.

2.6 Multiword Expressions

We will consider multiword expressions using Atkins and Rundell’s categories:

a. fixed and semi-fixed phrases. Atkins and Rundell comment that these “are important and worth recording during the analysis process of dictionary-writing.”
They subdivide them into:

i. very common collocations without idiomatic meanings.

ii. fixed phrases

iii. similes

iv. catch phrases

v. proverbs

vi. quotations

All of these have common parallels in Japanese, and can be found in dictionaries to varying degrees. Some would divide further, e.g. proverbs into ことわざ kotowaza “proverb, maxim” and 四字熟語 yajijakugo “four-character idiomatic compounds”.

b. other phrasal idioms, which Atkins and Rundell comment “are the most difficult multiword expressions to handle in lexicography”. A common test is whether the “meaning is more than the sum of the parts”. Part of the problem is that there is often no canonical form, and words can be substituted without changing the meaning. There can also be parallel idioms with opposite meanings.

There are many examples of these in Japanese, and they display similar properties to English examples, e.g.

i. parallel idioms - とぐろを巻く togurowomaku “to coil itself” vs とぐろを解く togurowohodoku “to uncoil itself”;

ii. multiple inflected variants - そうこうする内に sōkōsuru’uchini vs そうこうしている内に sōkōshiteiru’uchini both of which mean “meanwhile, in the meantime”,
c. compounds. These are primarily compound nouns and adjectives (verbs will be discussed below). They break broadly into two classes:

i. idiomatic compounds. These are extremely common, are usually the “sum of the parts”, are semantically transparent, and conform to the grammatical structures of the language. Typically they would not be included in a dictionary, however there are cases where it may be appropriate to include them in a bilingual dictionary, e.g.

- where the compound is the most appropriate translation out of a range of possibilities of a common term in a source language.
- where the reading is not derivable from the components, e.g. 二十日 hatsuka which means “twenty days” or “twentieth day of the month”. The reading is quite irregular.

ii. for idiomatic compounds, the challenge is to distinguish them from non-idiomatic ones. Atkins and Rundell list three properties shared by many of them:

- they are fixed; words cannot be added or removed;
- they participate in semantic relationships (e.g. hyponymy, synonymy);
- they have a meaning which is more than the sum of the parts or cannot be deduced from the parts.

They then divide those which warrant lexicalizing into three categories:

- figurative compounds. Just as a lame duck is not a duck which is lame, a バージン・ロード bājinrōdo “virgin road (lit.)” is not a road that is a virgin; it is an aisle in the context of a wedding ceremony. (In fact all
of the many 和製英語 waseieigo “pseudo-loanwords” would fall into this category to some degree.)

- semi-figurative compounds. A blind drunk may be drunk, but is not blind.

- functional compounds, e.g. police dog. It is a dog and it has to do with police, but the term means more than just that.

Svensén adds other grounds:

- the compound refers to a single concept, e.g. the French pomme de terre (potato). A Japanese example might be 含嗽剤 gansōzai “mouth-wash” which is a compound of 含嗽 “gargling” and 剤 “medicine”, but which maps to a single English word.

- the compound is a synonym of a single word. Svensén uses the example of the French bien que/quoique. Japanese examples include the expression 凹状 ōjō (discussed in Section 1.2.3) which means “concavity”, and the compound 愛読者 aidokusha “subscriber, reader, admirer” from 愛読 aidoku “reading with pleasure” and 者 sha “someone of that nature, someone doing that work”.

- terms which can form single-word derivatives, e.g. in French bon enfant and bonenfantisme. (This is really an orthographical quirk in French, and not an issue in Japanese.)

Landau separates regular compound lexical units from “specialized nomenclature”, such as Copernican system and listed building, and expresses doubt that the latter should be included, except that users expect to see
them in a dictionary. (This is a very fine point, and is probably taking an unnecessarily narrow view of the scope of a dictionary.)

iii. phrasal verbs. In English these are typically a verb with one or more particles, and are very similar to some compound verbs in Japanese (Wasow 2002; Dehe et al. 2001). Much of Atkins and Rundell’s discussion is not relevant to Japanese, apart from the general position that a division can be made into those which have literal meanings, and hence are less likely to warrant lexicalizing, and those which have figurative or metaphorical meanings. It acknowledged as a difficult area.

iv. support verb constructions. These relate to constructions of verbs such as take, make, have, give and do, as in “took a walk” and “take a vote”. Japanese equivalents would include 取る toru “to take”, する suru “to do”, なる naru “to become”, かける kakeru “to hang, to hoist, to take time, etc.”, and would often lead to constructions which are idiomatic enough to be lexicalized. (Miyamoto 2000)

2.7 Potential Lexical Items

In this section we will attempt to draw on the foregoing discussion and define the characteristics of potential lexical items in Japanese. A two-stage process is envisaged in which:

a. potential lexical items are identified, typically by some automated identification/extraction process, and delivered to a later stage for evaluation. This stage might include evaluation by human annotators to assess the performance
of the extraction process as to the validity of the proposed lexical item in terms of such things as semantics and grammatical correctness.

b. assessment of the potential lexical item as a candidate for lexicalization. This stage would largely be one of filtering items which were deemed unsuitable, e.g. were literal and non-idiomatic constructions, or were not common enough to warrant inclusion.

An example might be 航宇会社 kōugaisha “aerospace company”, which can be found in WWW pages in sentences such as: パスポートや航宇券を航宇会社の係りさんに見せた. Such a compound would most likely pass the first stage as it has all the characteristics of a valid noun-noun compound, but would not pass the second stage as it is entirely compositional, and not at all common.

It is also envisaged that there would be some initial filtering of words and terms which had already been lexicalized, or were on a stop-list as they had already been seen and rejected.

Using the structure described above (after Atkins and Rundell), and addressing the different types of lexical item in turn we can derive the following Japanese-oriented guidelines:

a. simple words (nouns, adjectives, etc.)

It is expected that nouns will make up the largest proportion of this category. Virtually any single kanji, kanji pair, short hiragana sequence and katakana sequence occurring in the appropriate grammatical context would have to be accepted at a potential lexical item. In some cases, e.g. the 皆私 example men-
tioned above, it may turn out to be a juxtaposition of kanji. Longer katakana sequences may well be compounds, and will need to be assessed as such. Onomatopoeic and mimetic words, which occur in large numbers in Japanese, and are usually written in kana would also be included. Sequences with the inflectional characteristics of simple verbs and adjectives would be accepted at this stage.

It is tempting at this stage to attempt to filter on grounds such as affixation, but without deeper analysis this has risks. For example, if the kanji pair 不安 was being seen for the first time it might be treated as combination of the prefix 不 fu “un-, non-” and the suffix 安 yasu “cheap, rash, careless”, but it is in fact a regular 2-kanji compound with the reading fuan and meaning “anxiety; uneasiness”.

Examples of this type of lexical item include 試錐 shisui “boring, drilling” and 愛い ui “good, nice, splendid”.

b. simple words (prepositions, conjunctions, pronouns, etc.)

These are very unlikely to occur as neologisms, as they are very widely lexicalized. Longer prepositional and conjunctive phrases may occur, but they will be considered as multiword expressions.

Examples of this type of lexical item include 及び oyobi “and, as well as” and 此れ kore “this, this person”.

c. simple words (abbreviations and contractions)

Sequences of capital letters would typically be accepted as alphabetisms and acronyms.
Unlexicalized 2-kanji contractions and abbreviated loanwords would be accepted as potential lexical items, however they would probably be assessed as simple words (above).

Colloquial contractions, such the common contraction ちゃった chatta which is derived from the しまいました/しまった shimaimashita/shimatta past tense forms of the verb 仕舞う shimau “to do completely”, would be accepted as these are relevant to language learners.

An example of this type of lexical item is RAM ramu “random-access memory”.

d. bound morphemes

In Japanese these are usually single kanji which for their particular reading and meaning cannot exist as a freestanding word. An example is 者 sha “someone of that nature; someone doing that work”, which is only found in formations such as 為政者 iseisha “statesman; politician” and 斡旋者 assensha “mediator; intermediary”. 者 can be used as a freestanding noun meaning “person”, but then the reading is mono or mon. Such bound morphemes are often not lexicalized.

e. multiword expressions (phrases)

This is a very difficult issue, as it requires judgement as to whether a multiword sequence is worth deeper analysis, i.e. it is a potential lexical item, or whether it is simply something that an extraction system happens to detect. Consider the following two cases which arose during extraction experiments carried out as part of the work described in Chapter 5.

i. the extraction system identified にはない niwanai in the sentence: 100年
前の日本のことどもは私の直接の記憶にはない。\(^8\) as a potential lexical item.

It is probably safe to reject にはない as it is an extremely common collocation (in fact it would probably be placed on a stop-list).

ii. the extraction system identified どうだろう dōdarō in the sentence: XX都銀、長信銀はどうだろう。\(^9\)

どうだろう is not in any Japanese-English dictionaries as a set phrase, but it occurs in several parallel corpora as a solo expression with translations such as “I ask you” and “It’s hard to say”, as well as commonly occurring in sentence endings with meanings such as “how about”. It probably should be accepted as a potential lexical item at this stage.

An example of this type of lexical item is 本の間 hon no aida “between the pages of a book”.

f. multiword expressions (compounds)

Compounds such as noun-noun, prefix-noun and noun-suffix are extremely common in Japanese. Many are written entirely in kanji, but compounds partly in kana are not uncommon. While many are quite non-idiomatic, assessment of unlexicalized compounds can usually only be done by assessing their usage in context. For example:

i. if the sequence 民間賃貸 minkanchintai “private rental” were suggested as a potential item based on its use in ほかに、「適当な民間賃貸住宅が見

\(^8\) hyakunen mae no nihon no kotodomo wa watashi no chokusetsu no kioku niwanai “As for things in Japan over 100 years ago I have no direct recollection.”

\(^9\) XXtōgin, chōshingin dōdarō “As for the XX metropolitan bank (and the) long-term credit bank, it’s hard to say.”
つからない」など。\(^{10}\) there would be temptation to reject it as it is part of
民間賃貸住宅, which typically be regarded as 民間 + 賃貸住宅 (private
+ rental-accommodation), based on the usual syntactical treatment of an
adjective-adjective-noun sequence in Japanese. However without assessing
whether 民間賃貸 occurs by itself or in other compounds as a compound
adjective, early rejection is risky.

ii. the potential compound ズボン姿 zubon sugata “trouser”-plus-“shape” in
白っぽい綿のような着にズボン姿で、リュックを背負っていた。\(^{11}\) is clearer.
As a possible noun-plus-noun sequence it should be initially accepted as a
potential lexical item, but on closer inspection it will be rejected as in this
case 姿, which is quite polysemous, is not the noun meaning “shape”, it is
a form of suffix meaning “wearing, dressed in”.

The assessment of whether a potential compound lexical item is ultimately ac-
cepted will depend on many factors. Clearly those with idiomatic or figurative
meanings should be accepted, dictionary size permitting. In some cases com-
pounds with non-idiomatic meanings will be accepted. Examples include:

i. very commonly encountered terms. Many Japanese-English dictionaries
have 日本政府 nihonseifu “Japanese Government” as an entry, despite it
being a straightforward non-idiomatic compound.

ii. terms which are most commonly associated with certain equivalent terms

\(^{10}\)hokani, ‘tekitō na minkanchintai jūtaku ga mitsukaranai’ nado “Moreover, ‘suitable private
rental housing is not to be found’, etc.”

\(^{11}\)shiroppoi men no yō na uwagi ni zubon sugata de, ryukku wo shotte ita “[He] was wearing a
whitish cotton-like jacket and trousers, and had a rucksack on [his] back”.

in the target language.
For multiword loanwords much the same assessment approach will apply, with all being potential lexical items, and inclusion/exclusion depending on whether they are non-idiomatic, figurative, etc. Pseudo-loanwords would always be accepted.

Composed numbers would typically be excluded as they are quite compositional, however in a few instances, such as the above-mentioned 二十日 hatsuka “twenty days, twentieth day”, they should be included for the benefit of learners as the reading is irregular.

Examples of this type of lexical item include 学級活動 gakkyūkatsudō “home-room activities” and 活字書体 katsujishotai “typeface”.

g. compound verbs. These verbs, which at their widest include verb-stem+verb, adjective-stem+verb and verb-continuative+verb patterns, would all typically be treated as potential lexical items (see Chapter 8.) Decisions on lexicalization would be made on much the same criteria as with compounds, with idiomatic meanings being readily accepted. A further ground for wide inclusion of compound verbs is that although their formation is often quite productive, the process is often not clear to language learners until they have reached a relatively advanced stage.

Examples of this type of lexical item include 思い出す omoidasu “to recall; to remember; to recollect” and 跳び出す tobidasu “to jump out; to fly out”.

h. support verb constructions.

These will typically be encountered in formed phrases, and need to be assessed in that context. Assessment should be sensitive to whether verbs such as the
above-mentioned 取る, なる or かける are used in such phrases.

An example of this is 相手取る aitedoru “to challenge (esp. in a lawsuit); to take on an opponent”.

It is recognized that while printed dictionaries have limits on the numbers of lexical items that can be recorded, electronic dictionaries are less constrained, and hence the selection process can be more relaxed. The topic is often discussed among lexicographers, and it is a common view that in electronic dictionaries it is still best to avoid the “clutter” of uncontroversial and compositional terms, despite them being relatively harmless.\footnote{For examples of such discussion among lexicographers, see \url{http://www.edrdg.org/jmdictdb/cgi-bin/entr.py?svc=jmdict&q=1595330}, \url{http://www.edrdg.org/jmdictdb/cgi-bin/entr.py?svc=jmdict&q=2747030} and \url{http://www.edrdg.org/jmdictdb/cgi-bin/entr.py?svc=jmdict&q=2492520}.}

### 2.8 Summary

In this chapter we have examined the criteria for assessing the suitability of candidate neologisms, first as to whether they are indeed potential lexical items, and then whether they could or should be included in a dictionary. The classification of proposed dictionary entries and associated criteria for inclusion as discussed in three major texts on lexicography were considered, along with their application to Japanese.

Also addressed was the issue of the appropriate terminology to use in the search for neologisms, with the conclusion that the term “lexical item”, which is well-established in lexicography, is more appropriate than alternatives such as “headword”, “lemma” and “lexeme”, all of which have quite specific application and connotations.
Chapter 3

Prior Work and Literature Review

3.1 Introduction

The work reported in this thesis is broken into a number of distinct project areas, with the projects reported in separate chapters. The discussion of the prior work and literature associated with each of the projects has been placed in the respective chapters in order for them to be read in the appropriate context. This chapter concentrates on prior work and literature in more general and related areas of lexicography, and of term discovery and extraction.

3.2 General Lexicography

The general literature of lexicography has relatively limited treatment of the handling of neologisms. The basic texts, e.g. Landau (2001); van Sterkenburg (2003); Atkins and Rundell (2008); Svensén (2009), several of which are discussed exten-
sively in Chapter 2, barely mention their existence. For example in the recent and well-regarded *Oxford Guide to Practical Lexicography* they only receive a passing reference, with the main comments relating to their frequently transient nature and the prevalence of compounds. The limited discussion of neologisms in the lexicography literature, e.g. Steiner (1976), has until recently been confined to the issues such as the rules that should be applied to the selection and publication of neologisms.

More recently the availability of large electronic corpora, and in particular the WWW, has been accompanied by a number of papers on neologism harvesting (Li 2007; Halskov and Jarvad 2009; Falk *et al.* 2014). These have been confined to European languages, where the actual harvesting is relatively simple when compared with Japanese or Chinese. Again the main issues have been those of the selection of which new words to record, the detection and avoidance of transient words and expressions, etc. O’Donovan and O’Neill (2008) provide a useful insight into how this is handled with *The Chambers Dictionary*, a major commercial English dictionary. A more general approach is described in Kerremans *et al.* (2012), which in particular explores the techniques and issues for retrieving and assessing neologisms from the WWW via a hypothetical *NeoCrawler* toolkit.

Another issue in lexicography is the identification of new senses for existing lexical items. Cook *et al.* (2014) report on the development of a dataset of diachronic sense differences and explore a topic-modelling approach to identifying new senses.
3.3 Japanese NLP - Morphological Analyzers and Parsers

Much of the attention given to unlexicalized Japanese words and expressions in the Japanese NLP literature has been in the context of the development and use of parsers and morphological analysis systems. In this context, the detection of out-of-lexicon words has been usually been characterized as a problem to be overcome, rather than a goal of the analysis, and often as an adjunct to part-of-speech (POS) assignment. Part of the relevant research has also been carried out in the area of speech recognition and analysis.

It is appropriate to give an overview of the development and state of research-oriented morphological analyzers. As these analyzers and their associated morpheme lexicons are an extremely important resource in the current work, they will be described in Chapter 4.

3.4 Alternative Segmentation Approach

A statistical and mostly unsupervised segmentation approach for the long sequences of kanji which often occur in Japanese text was investigated by Kubota, Ando and Lee (2003). This method appeared to have considerable potential to resolve problems which often occur in parsing and chunking such sequences, however the approach does not appear to have been investigated further.
3.5 Identification of Unknown Words

A typical early approach was by Nagata (1994) in which possible word boundaries were identified using forward dynamic programming, and a backward $n$-best algorithm was used to settle on the most likely demarcation. Nagata later extended this (Nagata 1996) using a generalized forward-backward search, but the results were generally disappointing. Another early approach was by Mori and Nagao (1996) in which the POS of neighbouring character strings was used as an indicator of word boundaries.

Nagata returned to the problem (Nagata 1999) with an approach in which sets of word models were developed according to the orthographic characteristics, and conditional probabilities used to assess the most likely boundaries. The approach was moderately successful.

Uchimoto et al. (2001) used a combination of a dictionary and a maximum entropy model, and reported more accurate identification of unknown words than analysis systems were achieving at that time.

In a quite different approach, Asahara and Matsumoto (2004) reported a method which combined a chunker with a morphological analyzer. The approach, which had previously been used with Chinese (Goh et al. 2003), and had been used the previous year for detecting fillers and disfluency (Asahara and Matsumoto 2003a), assigned character position features to individual characters based on the three best morpheme/POS combinations reported by the analyzer (ChaSen) then used an SVM-based chunker (YamCha) to derive possible word boundaries. The method reported a high precision for unknown words.

Character-level analysis was combined with ChaSen’s morpheme-based approach
Chapter 3: Prior Work and Literature Review

in a hybrid model by Nakagawa (2004) in an attempt to improve on unknown word identification. It resulted in a higher recall for unknown words. He returned to it (Nakagawa and Uchimoto 2007) with a slightly different hybrid system focusing on POS tagging, and reported generally similar results.

The approach of using a morphological analyzer in combination with a chunker or dependency analyzer is receiving increased attention. Nakagawa et al. (2001) were early experimenters with this technique, when they used an SVM-based classifier in combination with pre-tagged texts. Utsuro et al. (2007a) carried out extensive experimentation with combinations such as Juman/KNP and ChaSen/CaboCha. These systems work well with carefully annotated texts such as the Kyoto Corpus. Murawaki and Kurohashi, noting that unknown morphemes are one of the main sources of errors in morphological analysis, applied the technique to the extraction of unknown morphemes in uncontrolled texts such as WWW pages, with experiments involving morphological constraints (Murawaki and Kurohashi 2008) and orthographic variation (Murawaki and Kurohashi 2010).

In an approach more targeted at neologisms, Kaji et al. (2009) applied Juman to a large diachronic database of collected WWW pages and tracked the use of neologisms over time. Their study was more concerned with the use of neologisms, e.g. the recent verb ググる guguru “to Google”, than actual extraction or harvesting.

Nakazawa et al. (2005) specifically addressed the issue of loanwords in katakana, with a study aimed both at collecting such loanwords, especially neologisms, and developing techniques for determining their accurate segmentation and likely meanings.

Issues of accurate segmentation and extraction of OOV terms occur also with
Chapter 3: Prior Work and Literature Review

Chinese. Much of this is relevant to Japanese, although there are some major differences, for example a morpheme is usually regarded as a single character (hanzi) (Packard 2000). This is a large and active research area, and certainly too extensive to examine in any depth here. As mentioned above, Goh et al. (2003) had success with segmentation and term detection working at the character level, as did Peng et al. (2004) in a CRF-based approach. Zhang et al. (2011) reported some success using a mixture of statistical and supervised learning techniques, as did Sun et al. (2011) who used latent-dynamic conditional random fields (LDCRF).

Very recent work on the detection of OOVs in which a recurrent neural network (RNN) is trained on the character level and used to predict an OOV (Pinter et al. 2017) shows considerable promise. While it is addressing a different issue from that of neologism extraction, it is certainly an area for future exploration in this context.

3.6 Pre-candidature Work

Prior to the commencement of candidature two small “proof of concept” explorations of aspects of Japanese neologism extraction were carried out and reported.

The first experimented with the construction of abbreviations of 4-kanji compounds (Breen 2004a). Using a stock of some 8,000 compounds, 2-kanji abbreviations were constructed, and (when they were not already lexicalized) tested for occurrences in the WWW using the Google API. Surrounding text was examined to determine the orthographic and syntactic context of the usage, from which a confidence score was calculated. Manual inspection of high-confidence cases indicated a precision of

\footnote{No longer available.}
about 0.5 in detecting valid abbreviations.

The second approach experimented with collection of 500 articles from the *Asahi Shim bun* newspaper for possible unrecorded terms (Breen 2005). A harvester was implemented which collected a range of articles each day from the online edition of the newspaper. Two sets of processes were carried out on the text:

a. strings of *katakana* were extracted and compared with a collection of dictionary entries, with non-matching strings recorded for later analysis. The system harvested a large number of *katakana* words, many of which were transcriptions of proper names, however a number of other unrecorded terms were collected, including:

   a. Japanese flora/fauna terms;

   b. many variants of common loanwords e.g. プロフィル *purofiru* for “profile” instead of the more common プロフィール *purofīru*.

   c. a number of words and expressions worth adding to the lexicon, e.g. ピープルパワー *pīpurupāwa* “people-power” and ゼロメートル *zeromētoru* “sea level”, derived from the English words “zero metre”.

b. the text was processed by the *ChaSen/IPADIC* morphological analyzer and the output examined for the presence pairs of single-*kanji* morphemes. These were checked against the base dictionary and against a file of stop-words comprising common unlexicalized combinations. A large number of *kanji* pairs were collected, of which many turned out to be Chinese and Korean named entities. These were added to the stop-words file. Approximately 10% of *kanji* pairs
detected were worth considering for recording. Examples include:

a. previously unrecorded names e.g. 武示 Takeshi, 晃毅 Kōki, 潔重 Yukishige;

b. newly-arrived terms, e.g. 米紙 beishi “American press/newspapers” and 军歴 gunreki “military service record”;

c. many abbreviations, e.g. 日歯連 nisshiren (from 日本歯科医師連盟 nipponshikaishirenmei “Japan Dentists Federation”;

d. newspaper-style formations such as 中韩 chūkan “Chinese-Korean” and 仏誌 futsushi “French publication”;

e. several apparently new formations such as 入境 nyūkyō “border crossing, border entry” and 公助 kōjo “public assistance”.
Chapter 4

Resources

4.1 Introduction

Carrying out a research project on Japanese neologisms inevitably requires access to a considerable quantity of resources including dictionaries, specialized lexicons, text corpora, software, etc. In this chapter the major resources used in the project are described in general terms, with more specific information being provided in the individual chapters describing the work carried out.

The following sets of resources are described in this chapter:

a. dictionaries and general lexicons;

b. text corpora;

c. n-gram corpora;

d. software, including morphological analyzers and machine-learning systems.
4.2 Dictionaries and General Lexicons

Evaluation of extracted potential lexical items necessitates some form of reference lexicon against which such items can be matched. In order to have a baseline lexicon with as wide as possible coverage of Japanese words, including alternative surface forms and readings, the headwords, readings and entry identification information was extracted from a number of public and commercial dictionaries, and a combined lexicon referred to throughout this thesis as “Comblex” was constructed. The lexicon-building script can be easily extended to add more material if needed.

The dictionaries used in this compilation were:

   This is a major and well-regarded Japanese reference dictionary. 229,000 surface-form/reading combinations were extracted from the 5th edition of 1998.

   Another major Japanese reference dictionary, which is regarded as being more up-to-date than Kōjien. 215,000 combinations were extracted from the 2nd edition of 1995.

c. JMdict Japanese-Multilingual dictionary (Breen 2004b).
   An ongoing open-source online project to compile a comprehensive Japanese dictionary database. It currently has over 178,000 Japanese-English entries covering 240,000 surface-form/reading combinations. Other language glosses, which are imported from similar projects, include German (128,000 glosses), Russian (80,000), Spanish (39,000), Italian (38,000) and Dutch (29,000). A
copy was “frozen” in 2011 for this study and 196,000 combinations extracted.


A companion project to *JMdict* (744,000 combinations from 2011 were extracted).


The largest and most authoritative Japanese-English printed dictionary (128,000 combinations extracted).


A Japanese term/reading/POS/etc. lexicon developed by NTT.¹ (296,000 combinations from the 2002 edition).

g. *Japanese Lexical Database* (JLD) (Halpern 2008)

A Japanese term/reading lexicon from the CJK Dictionary Institute.² (302,000 combinations from the 2009 edition).

h. miscellaneous subject-specific glossaries (e.g. legal, biomedical, engineering, etc.) (329,000 combinations).

Comblex currently has over 1.5 million surface-form/reading combinations. Some illustrative examples of entries are:

- 章程 [しょうてい] / (n)JMD:2426300/DAI:90476/JLD:120126;NC/KOJ:100566/- shōtei “law; rule; ordinance”. The tags indicate it is found in the *Kōjien*,

¹Nippon Telegraph and Telephone Corporation
²copy kindly provided by Mr Jack Halpern
Chapter 4: Resources

*Daijirin*, *JMdict* and *JLD* sources.

- *章典 [あきのり] /(g)NAM/GOI:134532;pos=n/
  - *akinori* the given name “Akinori”, found in *JMnedict* and *Goi Taikei*.

In addition to Comblex a large amount of additional lexicographic reference material was assembled. While this material was not well structured, lacked readings, and had many duplications of terms, it proved very useful for checking possible lexical items. It is referred to in the thesis as the Extended Lexicon.

The significant components of the Extended Lexicon are:

- the sub-headwords from the Kenkyūsha 5th edition. (75,000 items)
- example sentences from the Kenkyūsha 5th edition. (410,000 items)
- the Eijiro dictionary file.³ (1.6 million items from the 2005 edition).
- the Japanese WordNet.⁴ (Isahara *et al.* 2008) (160,000 items from version 1.1 of 2010.)
- article titles (and if available English translations of them) from the Japanese Wikipedia. (140,000 items)
- several scientific, medical, etc. glossaries. (640,000 items)

In addition to the dictionaries mentioned above, several projects used information about individual *kanji*, such as the readings and frequency-of-use. This was derived from the *KANJIDIC* dictionary (Breen 2017b), which is a comprehensive database of information about the most common 13,000 kanji.

⁴http://nlpwww.nict.go.jp/wn-ja
Table 4.1: Summary of Lexical Resources

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Surface-form/Reading Combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kōjien</td>
<td>229,000</td>
</tr>
<tr>
<td>Daijrin</td>
<td>215,000</td>
</tr>
<tr>
<td>JMdict</td>
<td>196,000</td>
</tr>
<tr>
<td>JMNedict</td>
<td>744,000</td>
</tr>
<tr>
<td>Kenkyūsha</td>
<td>128,000</td>
</tr>
<tr>
<td>Goi Taikei</td>
<td>296,000</td>
</tr>
<tr>
<td>JLD</td>
<td>302,000</td>
</tr>
<tr>
<td>Misc.</td>
<td>329,000</td>
</tr>
<tr>
<td>Comblex</td>
<td>1,538,000</td>
</tr>
</tbody>
</table>

4.3 Text Corpora

A number of text corpora have been identified and where necessary collected as part of the project.


b. articles from the Nikkei Shimbun (the main Japanese financial newspaper) 1990-1999, containing about 14 million sentences (both these newspaper collections were already held by the University of Melbourne).
c. the Japanese Wikipedia text. As part of an initially planned diachronic analysis of words, copies of the complete Wikipedia dump were collected at 6-month intervals. Meaningful use of the Wikipedia text as a corpus requires the removal of meta-language elements, which in the case of the Japanese Wikipedia has proved problematic.

d. the Kyoto WWW Corpus. This is a collection of about 500 million Japanese sentences, obtained from a WWW crawl in 2004. A copy was provided to the project by Daisuke Kawahara of the Graduate School of Informatics, Kyoto University. The Corpus is 12Gb of compressed files.

e. the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa 2008; Maekawa et al. 2010; Maekawa et al. 2014). In its full form the BCCWJ consists of samples of modern Japanese texts totalling over 100 million words, covering genres such as general books and magazines, newspapers, business reports, blogs, internet forums, textbooks, and legal documents. The University of Melbourne holds some parts of the full Corpus, which can be searched online.

f. the Tanaka Corpus (Tanaka 2001). This corpus is a collection of 150,000 Japanese-English sentence pairs, originally released in 2001. The corpus has since undergone considerable editing, and each Japanese sentence has been annotated with its constituent lexical items to enable the corpus to be used as a source of example sentences in several dictionary servers and applications.\(^5\)

\(^5\)http://www.edrdg.org/wiki/index.php/Tanaka_Corpus
g. a collection of Japanese Twitter texts. These were extracted from the full Twitter collections from 2014 and 2015, and consist of the contents of Tweets where the language meta-data indicated they were in Japanese. The collection totals 870 million text passages, ranging from a few characters up to short paragraphs.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mainichi Shimbun</td>
<td>18,000,000</td>
</tr>
<tr>
<td>Nikkei Shimbun</td>
<td>14,000,000</td>
</tr>
<tr>
<td>Kyōto WWW Corpus</td>
<td>500,000,000</td>
</tr>
<tr>
<td>BCCWJ (est)</td>
<td>10,000,000</td>
</tr>
<tr>
<td>Twitter (2014–2015)</td>
<td>870,000,000</td>
</tr>
<tr>
<td>Tanaka</td>
<td>150,000</td>
</tr>
</tbody>
</table>

Table 4.2: Summary of Text Resources

4.4 n-gram Corpora

n-gram corpora typically consist of counts of the number of occurrences of particular sequences of words, or morphemes in the case of Japanese, in a body of text. The following n-gram corpora have been used in the current project.

a. the Google Japanese n-gram Corpus.\(^6\) (Kudo and Kazawa 2007) The n-grams in the Corpus were compiled by extracting text from a complete crawl of Japanese

---

\(^6\)https://catalog.ldc.upenn.edu/LDC2009T08
WWW pages for the month of July, 2007, and segmenting that text into morphemes using MeCab with the IPADIC morpheme dictionary (see below). Prior to segmentation the text was broken into “sentences” using punctuation and other special characters as delimiters, and the sentences were rejected if they were shorter than 6 characters, fewer than 5% of the characters were hiragana and fewer than 70% were kanji or kana. Resulting sequences of up to 7-grams are recorded in the Corpus if they were identified in 20 or more text segments. This Corpus, which is a very significant resource in Japanese NLP, was acquired via the University of Melbourne’s subscription to the Linguistic Data Corporation (LDC). The size of the full set is 26Gb of compressed files.\footnote{It is unfortunate that the Corpus has only been distributed via a very restrictive arrangement which prevents such things as making its contents available through online searches. This is in contrast to later corpus releases by Google, which are available for public download and less restricted use.}

Examples from the Corpus are in Table 4.3 and Table 4.4.

b. the Kyoto/Melbourne WWW $n$-gram Corpus. This Corpus was constructed from the Kyoto WWW Corpus (above) in 2014 as part of the work reported in this thesis. It follows the same construction principles and structure used in the Google $n$-gram Corpus; the main differences being that the UniDic morpheme lexicon was used with the MeCab morphological analyzer and hence it contains a finer-grained morpheme structure, and the $n$-grams are counted down to single occurrences instead of the cut-off of 20 occurrences used in the Google $n$-gram Corpus. The size of the full set is 59Gb of compressed files.

c. the Google English $n$-gram Corpus.\footnote{http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13} This contains 1-grams to 5-grams collected...
from the WWW in 2006, along with frequency counts.

As the Japanese $n$-gram corpora consist of multiple large compressed text files, it is necessary to convert them into alternative forms to enable rapid and efficient extraction of frequencies for particular $n$-grams and $n$-gram sequences. While it would have been possible to load them into, say, a relational database, it was decided that it would be quicker and less space-consuming to merge the $n$-grams into a single sorted file and establish a simple index file for rapid access. Table 4.3 and Table 4.4 contain samples of 4-grams and 5-grams from the Google Corpus, and Table 4.5 contains the matching entries in the combined file. The combined files of morpheme sequences and counts from the Google Corpus is 39Gb in size and 1.3 billion entries, and the Kyoto Corpus equivalent in 45Gb and 1.9 billion entries.

<table>
<thead>
<tr>
<th>Morphemes</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>あげ は の ひとり</td>
<td>915</td>
</tr>
<tr>
<td>あげ は の ふく</td>
<td>305</td>
</tr>
<tr>
<td>あげ は の よ</td>
<td>25</td>
</tr>
<tr>
<td>あげ は の よう</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 4.3: Sample of 4-grams from the Google Corpus

This approach to using the $n$-gram corpora proved to be very effective, and much of the use of the corpora took place via a simple API associated with these forms, particularly in the work described in Chapters 7 and 9. A WWW interface is available
Chapter 4: Resources

Table 4.4: Sample of 5-grams from the Google Corpus

<table>
<thead>
<tr>
<th>Morphemes</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>あげ は の ひとり ごと</td>
<td>912</td>
</tr>
<tr>
<td>あげ は の ふく あ</td>
<td>22</td>
</tr>
</tbody>
</table>

for the API.\(^9\) The use of the corpora via the API has also been very useful in the ongoing lexicographic tasks associated with the development of the \textit{JMdict} dictionary, such as determining the frequency of use of potential lexical items in various contexts; an issue discussed in Sections 2.3 and 2.7. In this it provides a similar service to the Japanese edition of Sketch Engine corpus tool (Kilgarriff \textit{et al.} 2014).

4.5 Software

4.5.1 Morphological Analyzers

It is appropriate to give a brief overview of the development and state of research-oriented morphological analyzers, the operation of which is described in Section 5.2. There are four morphological analyzers that are widely used:

a. \textit{Juman}, which was developed by the group headed by Makoto Nagao at Kyoto University in the early 1990s (Matsumoto and Nagao 1994; Kurohashi and Nagao 1998). The first release was in 1992, and the most recent in 2009. \textit{Juman} uses hand-compiled grammar and morpheme dictionaries, and selects the can-

### Table 4.5: Sample of Combined Morphemes from the Google Corpus

<table>
<thead>
<tr>
<th>Morpheme Sequence</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>あげはのひとり</td>
<td>915</td>
</tr>
<tr>
<td>あげはのひとりごと</td>
<td>912</td>
</tr>
<tr>
<td>あげはのふく</td>
<td>305</td>
</tr>
<tr>
<td>あげはのふくあ</td>
<td>22</td>
</tr>
<tr>
<td>あげはのふくあお</td>
<td>22</td>
</tr>
<tr>
<td>あげはのふくあおむし</td>
<td>22</td>
</tr>
<tr>
<td>あげはのよ</td>
<td>25</td>
</tr>
<tr>
<td>あげはのよう</td>
<td>22</td>
</tr>
</tbody>
</table>

Candidate parse which has the smallest number of unknown words, morphemes and independent words.

b. ChaSen, which was developed by Yuji Matsumoto’s group at NAIST (Matsumoto et al. 2003).\(^\text{10}\) It is largely a further development of *Juman*, and uses similar grammar and morpheme dictionaries. Hidden Markov Model (HMM) techniques are used to select candidate parses based on costs associated with each morpheme. The first version was released in 1997 and the most recent version (2.4.2) in 2007.

\(^\text{10}\)Nara Institute of Science and Technology
<table>
<thead>
<tr>
<th>Corpus</th>
<th>n-grams (billions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Japanese</td>
<td>3.7</td>
</tr>
<tr>
<td>Kyoto WWW</td>
<td>5.6</td>
</tr>
<tr>
<td>Google English</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 4.6: Summary of n-gram Resources

c. *MeCab*,\(^\text{11}\) which was also developed at NAIST (Kudo et al. 2004; Kudo 2008) and is essentially a reworking of *ChaSen*, supplementing the use of HMM with CRF-based machine learning techniques (Lafferty et al. 2001). Specifically it uses a bi-gram Markov model for analysis and CRF for discrimination. The first version (0.96) was released in 2006, and the most recent (0.996) in 2013.

d. *Kuromoji*,\(^\text{12}\) which has been developed in the Java language primarily for use on mobile devices.

Each of these analyzers is in regular use. In general *MeCab* is regarded as the most powerful and effective, and it has been used extensively in the work described in this thesis. (Further discussion of this topic can be found in Chapter 5.2.)

### 4.5.2 Morpheme Lexicons

All of the analyzers described above rely on morpheme lexicons to carry out their tasks. Several such lexicons are available:

\(^{11}\)http://taku910.github.io/mecab/

\(^{12}\)https://github.com/tilika/kuromoji
a. the *Juman* dictionary. This was hand-compiled as part of the project and now has approximately 250,000 morphemes. It is available for all four analyzers.

b. *IP ADIC*, which was developed at NAIST from a POS tagset established by the Information-technology Promotion Agency of Japan (IPA) (Asahara and Matsumoto 2003b). It was initially compiled in 1998 for *ChaSen*, and the most recent version (2.7.0) in 2003. It is available for *ChaSen*, *MeCab* and *Kuromoji*.

c. *NAIST-JDIC*, which was developed as a successor to *IPADIC*, mainly to avoid problems with the original IPA tagset licence. It was first released in 2008 and the most recent version (0.6.3) was in 2010. It also is available for *ChaSen*, *MeCab* and *Kuromoji*.

d. *UniDic*, which is a completely new compilation based on a much stricter definition of morphemes and including a much wider coverage of multiple surface forms and orthographical variants (Den et al. 2007). The project to develop UniDic is part of a large project including corpora of written and spoken Japanese (Maekawa 2008; Maekawa et al. 2010; Maekawa et al. 2014) and involves several universities and government agencies. The most recent version (2.1.2) released in 2013 has approximately 750,000 morphemes.\(^{13}\) It is available for *ChaSen*, *MeCab* and *Kuromoji*.

The following example entry from *UniDic* demonstrates the typical contents of morpheme lexicons:

肉づき, 1303, 1303, 8145, 動詞, 一般, *, *, 五段-カ行, 連用形-一般, ニクヅク, 肉付く, 肉づき, ニクズキ, 肉づく, ニクズク, 混, *, *, *, *

\(^{13}\)https://osdn.net/projects/unidic/releases/58338
This entry is for the morpheme 肉づき and it is defining it as the continuative form (連用形 ren’yōkei) of the verb 肉づく nikudzuku “to gain weight, to put on flesh” which is a Type 5 verb (五段 godan) of the ku inflection pattern (カ行 kagyō) and has the canonical form of 肉付く nikudzuku. ニクヅク nikudzuku and ニクズキ nikudzuki are readings in katakana. The “1303,1303,8145” is indicating training weights used with MeCab.

The UniDic lexicon has a number of important features which make it particularly suitable for the current work, when compared with IPADIC and NAIST-JDIC:

a. it has a much larger coverage of recognized morphemes in Japanese;

b. it includes many alternative surface forms of morphemes (as in the 肉づき example above);

c. its coverage is strictly limited to morphemes and does not include the multi-morpheme compounds which can be found in the other lexicons. For example both IPADIC and NAIST-JDIC contain an entry for 日本語 nihongo “Japanese language”, however it is absent from UniDic as it is a compound of two morphemes: 日本 “Japan” and 語 “language”.

4.5.3 Machine-Learning Systems

In addition to the software packages mentioned above, two other NLP-related packages were used in the project:

- the CRF++ Conditional Random Field toolkit.\textsuperscript{14} This is a well-regarded implementation by Taku Kudo of CRF techniques (Lafferty \emph{et al.} 2001; Tseng \emph{et al.}}\textsuperscript{14}https://taku910.github.io/crfpp/
2005; Finkel et al. 2008) for segmenting and labelling sequential data such as text.

- the LIBSVM package (Chang and Lin 2011). LIBSVM is a popular and well-regarded Support Vector Machine (SVM) implementation that has been used successfully in many NLP applications.
Chapter 5

Lexical Item Identification From Morpheme Analysis

5.1 Introduction

This chapter describes work on an exploration into the aggregation of sequences of morphemes into actual or potential lexical items. It addresses the research question of whether it is possible to devise techniques by which such sequences of morphemes in Japanese can be reliably aggregated for that purpose. The design and development of such techniques, the assembly of training and test data, and the evaluation of the results is described below.

The investigation makes use of machine-learning (ML) techniques to identify sequences of morphemes. It is not the object of the research to develop such techniques, or to explore a range of alternative ML approaches; instead a well-established toolkit (CRF++) with a track-record in chunking text has been employed.
5.2 Morphological Analysis of Japanese Text

As Japanese is written without spaces or other word-boundary markers, accurate automated parsing and segmenting is a non-trivial task, requiring a combination of advanced NLP and ML techniques, and an extensive lexicon. Over the past two decades a considerable amount of research and development has been devoted to morphological analyzers for Japanese, and, as with English POS tagging, the field is now generally regarded as mature with the available systems performing at a very high level of token accuracy (Den et al. 2008). Morphological analysis systems, which reduce Japanese text to a stream of morphemes and associated information such as POS tags, are extremely important tools in Japanese NLP, with virtually all NLP analysis of Japanese being carried out on text that has been processed by such systems. Japanese NLP typically uses the sequences of morphemes identified by the morphological analysis systems along with associated POS and other information. (A brief introduction to Japanese morphological analysis systems is in Section 4.5.1.)

The work described in this chapter used the MeCab morphological analyzer (Kudo et al. 2004). MeCab, which uses hidden Markov model (HMM) and conditional random field (CRF) techniques, is probably the most highly-regarded freely-available morphological analyzer at present.

To illustrate briefly the operation of a morphological analysis system, we will process the sentence その教師は講堂に学生を集めた sono kyōshi wa kōdō ni gakusei wo atsumeta “that teacher gathered the students in the hall” through the MeCab analyzer using the UniDic lexicon (Table 5.1). Columnar output has been used with addition of romanization of the readings of the morphemes and translations of the
Table 5.1: MeCab Analyzer Example

<table>
<thead>
<tr>
<th>Morpheme</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>その</td>
<td>sono pre-noun adjectival</td>
</tr>
<tr>
<td>教師</td>
<td>kyōshi noun-common-general</td>
</tr>
<tr>
<td>は</td>
<td>wa particle-binding</td>
</tr>
<tr>
<td>講堂</td>
<td>kōdō noun-common-general</td>
</tr>
<tr>
<td>に</td>
<td>ni particle-case-marking</td>
</tr>
<tr>
<td>学生</td>
<td>gakusei noun-common-general</td>
</tr>
<tr>
<td>を</td>
<td>wo particle-case-marking</td>
</tr>
<tr>
<td>集め</td>
<td>atsume verb-general, Type 1-m row, continuative-general</td>
</tr>
<tr>
<td>た</td>
<td>ta auxiliary verb, aux. v. “ta”, predicative-general</td>
</tr>
</tbody>
</table>

As can be seen, the sentence has been segmented into all the morphemes and the readings given for the morphemes with *kanji*. The conjugated verb 集めた *atsumeta* “gathered” has been split into a base 集め *atsume* and the plain past-tense inflection た *ta*.

In our work we make a clear distinction between *morphemes*, which are the minimal semantic-bearing elements in a language, and *lexical items*, which are combinations of one or more morphemes into lexical units, as would be listed in a dictionary. The distinction between morphemes and lexical items is particularly relevant in Japanese NLP as many of the earlier morphological analyzers and associated lexicons did not make a clear separation. For example, 日本語 *nihongo* “Japanese” is a lexical item made up of two morphemes: 日本 *nihon* “Japan” and 語 *go* “language”, however...
earlier morpheme lexicons such as *IPADIC* and *NAIST-JDIC* treated it as a single morpheme.

Our focus is on lexical items, with specific emphasis on those which could be expected to be included in Japanese–English bilingual dictionaries, in particular the highly productive class of compound nouns (noun-noun, adjective-noun, prefix-noun, etc.), as well as multiword expressions (Baldwin and Bond 2002). See Table 5.2 for examples. It is observed that such lexical items are usually constructed from combinations of simpler morphemes (nouns, adjectives, verb-stems, single-kanji affixes), and that some morphemes are particularly common components of lexical items. The immediate implication of this is that high-frequency expressions that are completely compositional in Japanese, and hence unlikely to occur as lexical items in a monolingual dictionary (e.g. 赤信号 *akashingō* “red light (traffic)”), are often explicit lexical items in bilingual dictionaries to aid learners in understanding the limits of usage of the Japanese expression (e.g. the fact that it has a derived usage meaning “warning sign”).

When investigating the output of morphological analyzers in search of unrecorded lexical items, there are several approaches that can be taken. They include:
a. drawing upon the features of some analyzers to report on out-of-lexicon morphemes. In these features, character strings which appear to be morphemes but do not appear in the morpheme lexicon can be reported as possible morphemes. In fact, these features, which initially seemed promising as a technique for detecting and reporting out-of-lexicon morphemes, have been rendered largely ineffective by the development of large morpheme lexicons which include most single *kanji* likely to occur in text as separate morphemes;

b. examining the morphemes and associated information for features and patterns which may indicate lexical items. For example, in the sentence used in the illustration in Table 5.1 the noun 学生 *gakusei* “student(s)” is flanked by two case-marking particles に *ni* and を *wo*. These, along with the POS of 学生 (noun), are a strong indication that the encapsulated characters are the object of the sentence. In this case 学生 is a single morpheme, but the analysis would apply equally to a multi-morpheme sequence, and could ameliorate the impact of over-segmentation by the analyzer;

c. training a machine learning system with text which has been passed through a morphological analyzer so that morphemes, POS tags, etc. can be used as training features. If lexical items are appropriately labelled in the test data a model can be developed which would enable other passages for text which has also been passed through a morphological analyzer to have possible lexical items identified.

This chapter reports on investigations into the latter two approaches outlined above. These investigations include trials of systems based on the latter two ap-
proaches in identifying lexical items, especially potential but unrecorded lexical items, in text.

5.3 Prior Work

Considerable work has been carried out on the identification of out-of-vocabulary (OOV) terms in Japanese, particularly in the context of POS-tagging and the development of morphological analysis systems (Nagata 1994; Nagata 1996; Mori and Nagao 1996; Nagata 1999; Uchimoto et al. 2001). In general this work has been superseded by modern morphological analysis systems, and as the focus of this early work was more on morpheme identification, is of limited relevance to this project.

The recent compilation of the UniDic lexicon (Den et al. 2007), based on a very strict, fine-grained approach to morpheme separation, has provided greater clarity in the lexical item/morpheme issue, and an almost iron-clad guarantee of morphological analyzers never under-segmenting. Of particular interest is the use of the MeCab morphological analysis system (Kudo et al. 2004) with the UniDic lexicon, as this combination has been shown to produce highly accurate morpheme-level segmentation (Den et al. 2008). The combination of MeCab and UniDic forms the backbone of this research, in providing extremely high-quality morphemes to chunk together into lexical items.

The approach of using a morphological analyzer in combination with a chunker or dependency analyzer has received considerable attention in studies of both OOV term and named-entity recognition (Nakagawa et al. 2001; Asahara and Matsumoto 2004; Utsuro et al. 2007b). The approach has considerable potential in lexical item
extraction. Of particular interest is Asahara’s technique, which used character-based chunking in which text was reduced to single characters, with features of the character-type and the three best POS estimates. A high level of word segmentation accuracy was reported. Ultimately, we apply a similar approach but do not resort to character-level analysis because of the availability of accurate morpheme-level segmentation.

While not immediately relevant to this research, it is worthwhile mentioning the body of work on POS tagging of unknown words in English and other segmenting languages. In languages such as English, word identification is largely trivial, because of the presence of explicit word delimiters. The greater challenge is in the determination of the POS tag of unknown words, where a variety of lexical and context features have been proposed (Brill 1995; Ratnaparkhi 1996; Toutanova et al. 2004). What is, in fact, more relevant to this work is the task of token-level identification of unknown MWEs. There is a significant body of work on type-level extraction of unknown MWEs, under the banner of collocation extraction, but this is generally applied to pre-determined constructions such as adjective-nouns or verb-nouns (Evert 2004; Pecina 2008).

At the time the work described in this chapter was carried out (2010–2011) the only paper we were aware of on the token-level identification of unknown MWEs was Fazly et al. (2009), but even here, only “verb noun idiomatic constructions” (e.g. shoot the breeze) were targeted. Since that time considerable advances have been made on MWE identification and extraction, for example in Schneider et al. (2014) a supervised lexical semantic segmentation approach is described which achieves considerable success (now available in the STREUSLE toolkit). In Qu et al. (2015) a
number of word embedding approaches, i.e. distributed word representations learned from unlabelled data, are explored and evaluated in their application to MWE identification.

Identification of lexical items has some similarities with methods adopted in automatic term recognition (ATR) (Kageura and Umino 1996). ATR has traditionally been focused on two main application areas: indexing documents (including glossary extraction) and information retrieval. The work reported by Nakagawa and Mori on the isolation of compound nouns has some similarity to the rule-based approach described below, but has a much narrower focus (Nakagawa and Mori 2002; Nakagawa and Mori 2003).

5.4 Using Unknown Word Functions in Morphological Analyzers

As mentioned above, it may be possible to draw upon the feature of some morphological analyzers to report on out-of-lexicon morphemes. In this feature, character strings which appear to be morphemes but do not appear in the morpheme lexicon can be reported as possible morphemes.

To give an illustration of how this approach may be employed, the sentence used in Table 5.1 above has been modified as follows:

a. the noun 講堂 was replaced with 全堂. This is a rare term in Japanese formed from the prefix 全 zen “all, whole” and the noun 堂 ど “temple, shrine, hall”.

The term is not in the analyzer’s morpheme lexicon;
Table 5.3: Morphological Analyzer Output for Unknown Morphemes

<table>
<thead>
<tr>
<th>Morpheme</th>
<th>Reading</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>その</td>
<td><em>sono</em></td>
<td>pre-noun adjectival</td>
</tr>
<tr>
<td>教師</td>
<td><em>kyōushi</em></td>
<td>noun-common-general</td>
</tr>
<tr>
<td>は</td>
<td><em>wa</em></td>
<td>particle-binding</td>
</tr>
<tr>
<td>全</td>
<td><em>zen</em></td>
<td>prefix</td>
</tr>
<tr>
<td>堂</td>
<td><em>dō</em></td>
<td>noun-common-general</td>
</tr>
<tr>
<td>に</td>
<td><em>ni</em></td>
<td>particle-case-marking</td>
</tr>
<tr>
<td>鼡黽</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>を</td>
<td><em>wo</em></td>
<td>particle-case-marking</td>
</tr>
<tr>
<td>集め</td>
<td><em>atsume</em></td>
<td>verb-general, Type 1-m row, continuative-general</td>
</tr>
<tr>
<td>た</td>
<td><em>ta</em></td>
<td>auxiliary verb, aux. v. “ta”, predicative-general</td>
</tr>
</tbody>
</table>

b. the noun 学生 was replace by a kanji pair: 鼡黽. This is not a word in Japanese; it is the collocation of two rare kanji: 鼡 sho “rat, mouse” and 黽 bō “green frog”.

Neither 鼡黽 nor the two individual kanji appear in the analyzer’s morpheme lexicon.

The resulting sentence, その教師は全堂に鼡黽を集めた, is thus syntactically sound; two nouns have been replaced by what could be other nouns, however it no longer makes much sense.

When that sentence is passed through MeCab/UniDic with the unknown word option enabled, the result is as shown in Table 5.3.

As can be seen, 全堂 was treated as two morphemes (prefix-plus-noun) and 鼡黽 as an unknown word. This result was not unexpected, as 全 and 堂 are present in
the morpheme lexicon, but neither 鼡 nor 鼸 is present.

Whether these sorts of signals are of significant value when searching text for unrecorded lexical items needs to be considered. The correct identification of 鼡鼴 indicates that for unknown morphemes constructed from kanji which do not function as independent morphemes the approach can be successful. However such formation of new 2-kanji compounds, especially from rare kanji, is not at all common. Moreover the most recent lexicons used by morphological analysis systems contain virtually all the kanji likely to be encountered in text. The handling of 全堂 is more typical, and an indication that the unknown word reporting function by itself is not likely to be adequate, and that some post-processing, such as aggregation of morphemes and checking against a lexicon, is likely to be necessary.

The example above only uses 2-kanji compounds. A more appropriate test is to consider the situation with three and 4-kanji lexical items, which are very common in Japanese. The latter are often noun-noun compounds, although some are prefix-noun-suffix, etc. An example might be 還元不良 formed from 還元 kangen “reduction, resolution” and 不良 fūyō “failure, defect”. The term is used in processes such as steel-making and does not appear in major dictionaries. One industry glossary translates it as “imperfect reduction”. If we test a sentence containing the term, e.g. 高炉の還元不良という問題を調べました kōro no kangenfūyō to iu mondai wo shirabe-mashita “the problem known as imperfect blast-furnace reduction was investigated”, we get the analysis shown in Table 5.4.

As can be seen, 還元不良 was correctly identified as being two morphemes, 還元 and 不良, but the fact that it is an “unknown” compound cannot be deduced by the
Chapter 5: Lexical Item Identification From Morpheme Analysis

<table>
<thead>
<tr>
<th>Morpheme</th>
<th>POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>高炉</td>
<td>kōro noun-common-general</td>
</tr>
<tr>
<td>の</td>
<td>no particle-case-marking</td>
</tr>
<tr>
<td>還元</td>
<td>kangen noun-common-verbal</td>
</tr>
<tr>
<td>不良</td>
<td>furyō noun-common-adjectival</td>
</tr>
<tr>
<td>と</td>
<td>to particle-case-marking</td>
</tr>
<tr>
<td>いう</td>
<td>iu verb-general, Type 5-ua row, continuative-general</td>
</tr>
<tr>
<td>問題</td>
<td>mondai noun-common-general</td>
</tr>
<tr>
<td>を</td>
<td>wo particle-case-marking</td>
</tr>
<tr>
<td>調べ</td>
<td>shirabe verb-general, Type 1-b row, continuative-general</td>
</tr>
<tr>
<td>まし</td>
<td>mashi auxiliary verb, aux. v. “masu”, continuative-general</td>
</tr>
<tr>
<td>た</td>
<td>ta auxiliary verb, aux. v. “ta”, predicative-general</td>
</tr>
</tbody>
</table>

Table 5.4: Example of Morphological Analysis

Much depends on the morpheme lexicons used by the morphological analysis systems. As mentioned above, some of the older lexicons contained both morphemes and some higher-order constructions. One possible approach could be to extend the lexicon with known lexical items, then examine the output for unaggregated noun-noun, noun-suffix, etc. sequences as these may indicate unrecorded lexical items.

A brief experiment was carried out in which the NAIST morpheme lexicon was extended by a large number of lexical items from *JMdict*, then tested with sentences containing those lexical items. The results were not very satisfactory: while the lexical items were quite often reported as single “morphemes”, often they were not. (This can be explained by considering the way in which *MeCab* operates, with its
use of a bi-gram Markov model for analysis and CRF for discrimination. Operation is dependent on weightings assigned to morphemes in the morpheme lexicon which have come from corpus-based training, and words such as 日本語 with overlapping morpheme sets can easily produce different outcomes depending on the contexts in which they appear.) The approach of loading the lexicon with known morphemes was considered to be insufficiently reliable as it would still require considerable post-processing of the analyzer output and there were alternatives (described below) which offered more promise.

Also, while it may be useful for identifying unrecorded morphemes, it is moving away from the reporting of out-of-lexicon items by the analyzer itself.

5.5 Rule-Based Lexical Item Identification

As was demonstrated in Table 5.4 with the sentence 高炉の還元不良という問題を調べました, the morphological analyzer was able to identify the morphemes in the compound noun 還元不良 correctly, but was not by itself able to make any suggestions about 還元不良. With modern Japanese morphological analyzers providing extensive information about the parts-of-speech of the morphemes into which text has been segmented, combined with the nature of Japanese grammar with its extensive use of case-marking particles and affixation, we have an avenue for possibly identifying lexical items within text from the context of neighbouring morphemes of the appropriate part of speech. For example, if we look at the sentence above we see that the morphemes in 還元不良 are both tagged as nouns, and the immediately preceding and following morphemes (の and と) are both tagged as case-marking particles.
In Japanese syntax there is little alternative to 還元不良 being a multi-morpheme noun, and hence a potential lexical item. In this example a simple profile of <case-particle, noun*, case-particle> was applied. It is, of course, possible to develop a wide-ranging set of profiles that could be used to identify potential lexical items in morpheme-segmented text.

To test this hypothesis, an extraction system for potential lexical items was developed based on the analysis of the sequence of part-of-speech tags from a morphological analyzer. The analysis uses sets of POS tags to profile the morphemes which make up the potential lexical item, and the morphemes which encapsulate it and establish its syntactic context. The analyzer used was MeCab with the UniDic morpheme dictionary.

The extraction system is quite simple in concept: a sequence of morphemes is identified as a potential lexical item if:

a. all the morphemes within the potential lexical item have POS tags appropriate to the lexical item type;

b. the immediately preceding and following morphemes match an encapsulation rule for that lexical item type.

The complexity of the extraction system lies in the sets of POS tags making up the lexical item sets and the encapsulation rules. The UniDic lexicon uses a hierarchical system of POS tags, for example noun tags include:

- 名詞, 普通名詞, 一般 meishi,futsūmeishi,ippan noun, common noun, general
- 名詞, 普通名詞, サ変可能 meishi,futsūmeishi,sahenkanō noun, common noun, verbal
Chapter 5: Lexical Item Identification From Morpheme Analysis

- 名詞, 数詞 meishi, sūshi noun, numeral
- 名詞, 固有名詞, 地名, 一般 meishi, koyūmeishi, chimei, ippan noun, proper noun, place name, general

The development of the sets of rules was carried out by processing the Japanese sentences from the Tanaka Corpus, which as described in Section 4.3 have been annotated with their constituent lexical items, through MeCab/UniDic, then aligning the texts and recording the POS tags for both the constituent morphemes of each lexical item, and those of the preceding and following words, particles, etc. This harvesting process collected several thousand different pairs of encapsulating POS tags, which were ranked in descending frequency, and then used to process some segmented texts drawn from newspaper articles. The resulting potential lexical items were examined for suitability, accuracy, etc. and the sets of POS tag rules adjusted in an iterative process in order to establish an effective working model.

In establishing the most useful sets of tags, it was concluded that:

a. for the POS tags of the morphemes making up a lexical item, only the first part of the hierarchy was needed. The POS types identified were:
   i. nouns: 名詞 meishi noun, 接頭辞 settōji prefix, 接尾辞 setsubiji suffix, 形状詞 keijōshi adjectival noun, 代名詞 daimeishi pronoun;
   ii. verbs: 動詞 dōshi verb, 助動詞 jodōshi auxiliary verb;
   iii. adjectives: 形容詞 keiyōshi adjective, 接尾辞 setsubiji suffix, 助動詞 jodōshi auxiliary verb.

b. for the POS tags of the encapsulating morphemes, a slightly finer-grained identification of POS performed best, and the first two levels of the hierarchy are
Table 5.5: Examples of Encapsulating POS Tags for Nouns

<table>
<thead>
<tr>
<th>Left POS</th>
<th>Right POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>助詞, 接続助詞 (particle, conjunctive)</td>
<td>補助記号, 句点 (punctuation, period)</td>
</tr>
<tr>
<td>助詞, 副助詞 (particle, adverbial)</td>
<td>助詞, 格助詞 (particle, case-marking)</td>
</tr>
<tr>
<td>助詞, 副助詞 (particle, adverbial)</td>
<td>補助記号, 読点 (punctuation, comma)</td>
</tr>
</tbody>
</table>

used. Table 5.5 shows some of the POS pairs for nouns.

At present the system handles three lexical item types: nouns, verbs and adjectives. The current rules contain 103 pairs for nouns, 9 pairs for adjectives and 37 pairs for verbs. The large number of POS types associated with nouns is due to them often being adjacent to punctuation, parentheses, etc. The full set of encapsulating rules is in Appendix A.

During testing the extraction system, which we will refer to as the “rule-based chunker”, demonstrated good levels of precision and recall for the three target lexical item types. For example, with the sentence 高炉の還元不良という問題を調べました it correctly identified 還元不良 and 調べました as potential lexical items. One potential weakness lies with long agglutinated noun sequences, which are common in Japanese and which it simply aggregates.

Comparative tests of the rule-based chunker with various machine learning models are described below.
5.6 Application of Machine Learning to Lexical Item Extraction

5.6.1 Basic Model

As has been demonstrated above, there is potential for applying a rule-set derived from the rules of Japanese syntax to the POS-tagged morphemes from a morphological analyzer in order to identify potential lexical items in text. Machine learning is generally accepted as an effective technique for chunking Japanese text (Kudo and Matsumoto 2001; Utsuro et al. 2007b), and has had success in such areas as morpheme and named-entity extraction. It is appropriate that we explore the suitability of ML for lexical item identification. The particular motivations are:

a. to determine the overall efficacy of the technique;
b. to see if it can handle rarer constructions as conjunctions and noun-phrases with relative clauses which are currently outside the scope of the rule-based chunker;
c. to determine whether it can improve the handling of long noun collocations;
d. to explore additional training features, such as dictionary and corpus frequencies associated with sequences of morphemes.

In this section, investigation of the application of machine learning techniques to the extraction of lexical items in Japanese text is described. In broad terms, the goal of the investigation has been to develop and test ML models which are most effective in identifying lexical items, with particular emphasis on lexical items which are not already recorded in dictionaries.

Given the availability of the UniDic morpheme lexicon with its very extensive cov-
verage and strict recording of basic morphemes, it was decided that it was appropriate to work entirely at the morpheme level, and to experiment with training text which had been segmented into morphemes and marked up with features associated with the possible membership of higher-level lexical items.

As with the rule-based chunker, *MeCab/UniDic* was used to carry out morpheme segmentation, and the *CRF++* conditional random field toolkit\(^1\) was used for training and testing different feature models. As discussed below, the single-best MIRA (Margin Infused Relaxed Algorithm) training option was used. Conditional random field techniques have proved very successful in related ML work (Lafferty *et al.* 2001; Tseng *et al.* 2005; Finkel *et al.* 2008).

The general approach taken in the work described below has been to segment text with *MeCab* into a stream of morphemes with POS tags, add other features which may assist with lexical item identification, and to label lexical items in the training data.

### 5.6.2 Labelling

The labelling of sequences of morphemes in the training data enables CRF++, when it encounters a sequence in the test data which matches its model, to indicate the match by applying those labels. The labels are not, of course, features, and play no role in the training itself.

The label scheme used here is IOB2 (Ratnaparkhi 1998), which is very commonly used in NLP to indicate chunks which are not embedded or overlapped. In this

\(^1\)https://taku910.github.io/crfpp/
Table 5.6: Example of Training Features and IOB2 Labels

<table>
<thead>
<tr>
<th>Morphemes</th>
<th>POS Tags</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>春子 haruko</td>
<td>名詞, 固有名詞, 人名, 名,<em>,</em></td>
<td>O</td>
</tr>
<tr>
<td>は wa</td>
<td>助詞, 係助詞,<em>,</em>,<em>,</em></td>
<td>O</td>
</tr>
<tr>
<td>大学 daigaku</td>
<td>名詞, 普通名詞, 一般,<em>,</em>,*</td>
<td>B</td>
</tr>
<tr>
<td>生 sei</td>
<td>接尾辞, 名詞的, 一般,<em>,</em>,*</td>
<td>I</td>
</tr>
<tr>
<td>です desu</td>
<td>助動詞,<em>,</em>, 助動詞-デス, 終止形-一般</td>
<td>O</td>
</tr>
<tr>
<td>。</td>
<td>補助記号, 句点,<em>,</em>,<em>,</em></td>
<td>O</td>
</tr>
</tbody>
</table>

scheme B is used to indicate the first morpheme in a chunk, I to indicate the second and subsequent morphemes in a chunk, and O to indicate all other morphemes. For example in the sentence 春子は大学生です haruko wa daigakusei desu “Haruko is a university student”, 大学生 daigakusei “university student” is a multi-morpheme lexical item (大学 + 生), and thus a training file using that sentence and with the morphemes and the full UniDic POS tags as feature columns would contain labels as shown in Table 5.6.

Labelling is performed by a general tagging/labelling utility used in this investigation. Multi-morpheme lexical items are identified by stepping through the morpheme sequence and applying a longest-match algorithm against a target lexical item dictionary. When two or more adjacent morphemes match a complete dictionary entry the B and I tags are applied to the sequence, otherwise an O tag is applied. The algorithm caters for inflected forms of verbs and adjectives. The lexical item dictionary used is the Comblex compilation described in Section 4.2.
5.6.3 Feature Development

There is a range of features which can be extracted from segmented text, or associated with that text, and which could be used for training and testing text for the presence of lexical items. The most obvious of these are the morphemes themselves and their associated POS tags. In addition, features associated with the locations of known lexical items within the text, and with the characteristics of the morphemes making up known lexical items can be identified and added. The additional features which have been developed and tested in the present study are:

a. lexical frequency features. This is a set of features derived from the frequencies with which the morphemes occur in known lexical items, both within a large lexicon and in a major corpus of Japanese text. It is based on the intuition that if a particular morpheme occurs frequently in one or more recognized multi-morpheme lexical items, then when it is detected in appropriate contexts with other morphemes it is likely to be forming another multi-morpheme lexical item. The derivation of these features is described below.

In developing features based on the occurrence of morphemes in a lexicon and in a corpus it is postulated that the ML system may be assisted by information indicating that for any given morpheme there is an identified likelihood of it being in a lexical item. To implement and test this hypothesis the following steps were taken:

i. each headword in the Comblex lexicon was processed by MeCab/UniDic to reduce it to its constituent morphemes;

ii. the morphemes thus produced were classified as A (the lexical item was a
single morpheme), B (the morpheme was in the initial position in the lexical item), C (the morpheme was in an intermediate position in the lexical item), or D (the morpheme was in the final position in lexical item). After this classification the morphemes and their categories were aggregated. For example, the word/morpheme 賞与 shōyo “bonus, reward, prize” occurs once solo, once in the initial position (賞与金 shōyokin “incentive, bonus”) and four times in the final position: 業績賞与 gyōsekishōyo “performance bonus”; 団体賞与 dantaishōyo “group bonus”; 年未賞与 nenmatsushōyo “year-end bonus” and 役員賞与 yakuinshōyo “directors’ bonus”. Thus for 賞与 the category counts are: A-1, B-1, C-0 and D-4.

iii. the step above was repeated using the frequency counts for the lexical items drawn from the Google Japanese n-gram Corpus. For 賞与 this resulted in counts of 613,309 for a solo lexical item, 5,554 in the first position and 44,202 for the final position.

iv. a set of features was compiled based on the counts themselves and of the proportions of counts falling into the solo/first/intermediate/final categories, as the proportions will approximate to the conditional probability of a category for each morpheme.

From these counts 16 feature elements were compiled for each morpheme, with letter codes indicating if the feature was lexicon (L) or n-gram (N), count (C) or proportion (P), position in lexical items (A: solo, B: initial, C: intermediate, D: final), and a count/proportion value. For the latter the counts have been

---

2As that Corpus consists of n-grams resulting from segmentation using MeCab/IPADIC, and hence often has a different segmentation from that produced by MeCab/UniDic, the n-grams making up lexical items were recombined and the counts extracted.
Figure 5.1: Sample Feature Dictionary Entry

placed in 10 bins using a logarithmic scale and proportions rounded to a single digit.

The features described above were assembled in a morpheme-feature dictionary in the format shown as an example in Figure 5.1 for the morpheme 賞与. The LPB2 feature indicates that 賞与 occurs in the lexicon approximately 20% of the time as the initial part of longer lexical items.

The morpheme-feature dictionary thus produced is used by the tagging utility mentioned above. Table 5.7 contains an example for the training text fragment 下の図は業績賞与について shita no zu wa gyōsekishō ni tsuite “the table below concerning performance bonuses”.

As testing using all of the 16 frequency-based features singly and in combinations would have been time-consuming, it was decided to divide them into four groups each of four features, and to test using combinations of the groups. The groups are:

- Group 1: lexicon counts
- Group 2: n-gram counts
- Group 3: lexicon proportions
- Group 4: n-gram proportions

This division was made intuitively, as it was considered that there was likely to
be a high degree of complementarity within the four groups and more independence between them.

b. encapsulating morphemes. In this feature, a feature column consisting of the morphemes immediately preceding and following a known multi-morpheme lexical item is used. The column is set to O for all other morphemes.

c. encapsulating POS tags. This is similar to (b) above, except that the POS tag is used as the feature rather than the morpheme itself.

<table>
<thead>
<tr>
<th>Morpheme</th>
<th>POS Tags</th>
<th>Frequency Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>下</td>
<td>名詞, 普通名詞, 一般,*,**,</td>
<td>LCA1 [....] NPD2</td>
<td>O</td>
</tr>
<tr>
<td>の</td>
<td>助詞, 格助詞,<em>,</em>,**,</td>
<td>LCA1 [....] NPD2</td>
<td>O</td>
</tr>
<tr>
<td>図</td>
<td>名詞, 普通名詞, 一般,*,**,</td>
<td>LCA1 [....] NPD9</td>
<td>O</td>
</tr>
<tr>
<td>は</td>
<td>助詞, 係助詞,<em>,</em>,**</td>
<td>LCA1 [....] NPD0</td>
<td>O</td>
</tr>
<tr>
<td>業績</td>
<td>名詞, 普通名詞, 一般,*,**,</td>
<td>LCA1 [....] NPD0</td>
<td>B</td>
</tr>
<tr>
<td>賞与</td>
<td>名詞, 普通名詞, 一般,*,**,</td>
<td>LCA1 [....] NPD0</td>
<td>I</td>
</tr>
<tr>
<td>に</td>
<td>助詞, 格助詞,<em>,</em>,**,</td>
<td>LCA1 [....] NPA9</td>
<td>B</td>
</tr>
<tr>
<td>つい</td>
<td>動詞, 一般,<em>,</em>, 五段 [...</td>
<td>LCA1 [....] NPD0</td>
<td>I</td>
</tr>
<tr>
<td>て</td>
<td>助詞, 接続助詞,<em>,</em>,**,</td>
<td>LCA1 [....] NPD0</td>
<td>I</td>
</tr>
</tbody>
</table>

Table 5.7: Text example with POS and frequency features
5.7 Evaluation

5.7.1 Overview

As described above, we have developed two approaches to identifying and extracting potential lexical items in Japanese text which has been segmented to a strict morpheme level: a rule-based approach using POS tags, and an ML approach using a range of features. In evaluating these approaches we decided to conduct three sets of tests:

a. using test and training texts which had been automatically marked up at the lexical item level using the Comblex lexicon. (A greedy morpheme-matching algorithm was used to detect known lexical items in the morpheme stream. The algorithm also tracks inflected verbs and adjectives, which appear as multiple morphemes, back to the base forms.) These tests focused on the identification of a selected set of lexical items which had been omitted from the Comblex version used in the markup;

b. using text which had been hand-annotated to identify and confirm lexical items. As above, the tests also focused on the identification of a selected set of lexical items;

c. using a set of test texts from different genres to evaluate the efficacy of the approaches in detecting potential unrecorded lexical items.
5.7.2 Testing with Automatically Marked-up Texts

The ultimate aim is, of course, to detect potential lexical items in text, however as the system would, if operating successfully, label all lexical items, it is possible to do some testing on the recovery of known lexical items in texts.

To illustrate this, consider the following short passage from the Asahi Shimbun: 同年9月の国交省の回答では、dōnen kugatsu no kokkōshō no kaitō dewa “in the Ministry of Transport’s reply of September the same year” which was used as test data (a different set of text was used for training). This text was segmented by MeCab and processed by our tagger to indicate known lexical items in Comblex. This tagging of known lexical items is for information and comparison purposes only, and is not seen by CRF++. When the text is examined by CRF++ using a model trained on morphemes and POS tags, it adds a column of labels considered appropriate to the model. The results from CRF++ are in Table 5.8.

As can be seen, in this example CRF++ correctly identified 9月 and 国交省 as multi-morpheme lexical items.

Utility software was developed to analyze output from CRF++ and report both on the matching of all known and proposed lexical items and the performance for specific lists of target lexical items. This utility was used for all the detailed testing.

5.7.3 Initial Testing

Initial tests were carried out to test the various combinations of features and options within the CRF++ chunker. These tests focused on the labelling of lexical items by CRF++. At first a 20,000-sentence collection of text from the Mainichi
<table>
<thead>
<tr>
<th>Morphemes</th>
<th>POS Tags</th>
<th>Known Lexical items</th>
<th>Lexical items Proposed by CRF++</th>
</tr>
</thead>
<tbody>
<tr>
<td>同年</td>
<td>名詞, 普通名詞, 副詞可能,*,**</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>9</td>
<td>名詞, 数詞,<em>,</em>,**</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>月</td>
<td>接尾辞, 名詞的, 助数詞,*,**</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>の</td>
<td>助詞, 格助詞,*,**</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>国交</td>
<td>名詞, 普通名詞, 一般,*,**</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>省</td>
<td>接尾辞, 名詞的, 一般,*,**</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>の</td>
<td>助詞, 格助詞,*,**</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>回答</td>
<td>名詞, 普通名詞, サ変可能,*,**</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>で</td>
<td>助詞, 格助詞,*,**</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>は</td>
<td>助詞, 係助詞,*,**</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>,</td>
<td>補助記号, 読点,*,**</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

Table 5.8: Example of CRF++ Labelling

Shimbun newspaper was used for training and a 100,000-sentence collection from the same newspaper was used for testing, with a simple test criterion of determining how many lexical items were correctly labelled by CRF++. At this time the training options of CRF++ were evaluated. As mentioned above (Section 5.6.1), the single-best MIRA training option produced the best results. Tests were also made of various values of the Gaussian hyper-parameter that defines the balance between under- and over-fitting. No conclusive indication was received from these tests at this stage, so further analysis of this was deferred. These relatively crude tests indicated that 89% of lexical items were being correctly labelled using a simple training model consisting of just the morphemes and POS tags as features, and a slightly higher proportion
Table 5.9: Recall of target lexical items for morpheme/POS alone and added features when the lexical item-frequency features were included.

In a more focused test, a set of 50 multi-morpheme lexical items was selected from a financial glossary and two sets of sentences were extracted from the Nikkei Shimbun (financial newspaper): a 5,000-sentence training set which did not contain any of the target lexical items, and a 1,500 test set which contained at least 10 occurrences of each of the target lexical items. In total there were 1,289 occurrences of these lexical items.

The model trained with just lexical item and POS features exactly tagged 519 (48%) of these. With the addition of all the frequency features this rose to 799 (62%), a significant improvement. As a comparison, the rule-based chunker described earlier in this chapter correctly identified 695 (54%) of the lexical items.

As mentioned above, the 16 frequency-based features were aggregated into 4 groups according to whether they were counts or proportions, and whether they were based on lexicon or n-gram counts. Combinations of these feature groups were used in training and testing. In general the combinations containing groups 2 (n-gram counts) and 4 (n-gram proportions) performed best, with the highest result (830) coming from just groups 2 and 4 in combination. Note that these are from a total of 1,289 occurrences in the test data. The results are summarized in the Table 5.9.
Table 5.10: Lexical item recall for different hyper-parameter values

It must be noted that these aggregated results conceal considerable variability across the different models for the 50 target lexical items. The “24” model (n-gram features only), which received the highest total score, was only best or equal-best in 10 of the 50 lexical items. The poor aggregate result for the “13” model (lexicon features only) was due entirely to its poor performance on two frequently-occurring lexical items; for the rest it was comparable with other models.

A similar variability was revealed when the tests were rerun for the “24” model and partially for the “1234” model (all features) for different values of the Gaussian hyper-parameter. The results are in the Table 5.10. (In the cases for C values of 50 and 100 the learning stage terminated without completion at 100,000 iterations.)

These results are difficult to explain. One would intuitively expect that varying the parameter would result in a steady change towards or away from a generally optimal value. Instead a series of quite abrupt changes is observed, with several local maxima/minima. This is probably indicating a high degree of instability in the
interaction of the models and the individual target lexical items.

A further test was also carried out for several of the models without using the MIRA training option. In all cases the level of recall was lower.

5.7.4 Testing with Hand-Annotated Texts

It became apparent during the testing described above that there were several inadequacies with the training and test data. The sentences had been selected relatively simplistically using string matching, and in a number of cases target lexical items were in fact embedded in longer compounds. While it had been possible to extract broad estimates of recall, no method was available to determine the precision of the various models. It was decided that a more robust and verifiable set of sentences was required for more rigorous testing.

A number of texts have recently been made available from the BCCWJ (Balanced Corpus of Contemporary Written Japanese) Project (Maekawa 2008) and it was hoped that as these were reported to be segmented and POS-tagged they could serve as test and training data for the current project. Unfortunately on inspection it was clear that the segmentation and tagging had largely been carried out using MeCab and the IPADIC lexicon, and that no attempt had been made to mark lexical items. This meant that there was no opportunity to use these texts without completely revising the segmenting and tagging.

Accordingly a sub-project was conducted in which approximately 2,000 sentences were annotated by a team of volunteers. The annotation consisted of identifying sequences of morphemes which could reasonably be regarded as lexical items. A full
description of the annotation exercise is included in Appendix B, and the following
is a brief summary.

a. the lexical items added to the *JMdict* dictionary in the preceding months were
extracted. (A copy of that dictionary was “frozen” for testing purposes.) As a
number of these were idiomatic expressions written partly in *hiragana*, it was
decided that it would be best to focus on nouns and compound nouns written
primarily in *kanji*. Applying this restriction, and removing lexical items that
were already in other lexicons in the Comblex file, the total was 624 lexical
items;

b. the Nikkei and Mainichi text collections were scanned for sentences containing
the target lexical items. 287 of the lexical items appeared 10 or more times in
the text collections. Sentences containing the lexical items were selected if they
were between 15 and 100 characters in length, contained a mix of both *kana* and
*kanji* characters, and were terminated by a valid punctuation character. These
restrictions were to ensure that the sentences were not headings, captions, table
entries, etc. which had been a problem with the earlier testing;

c. 10 sentences were chosen arbitrarily for each of 100 target lexical items. A
further 1,000 sentences were chosen which did not contain any of the target
lexical items; these additional sentences were to pad out the text and to ensure
there was an adequate quantity of training data;

d. the 2,000 sentences were presented to the annotators via a WWW system which
enabled the morpheme segmentation and the identification of lexical items in
Comblex to be seen for each sentence (the “frozen” version of Comblex was used
for this, so the target lexical items were not marked). Annotators were asked to mark any additional multi-morpheme lexical items above those already marked. Each sentence was examined by two annotators, at which stage it was removed from consideration. Prior annotations were not visible to a later annotator. Where there was disagreement between the annotators as to which additional lexical items should be marked a third annotator adjudicated.

e. the additional lexical items identified by annotators were extracted. These included both the target lexical items plus quite a large number of other potential lexical items noticed by the annotators.

The outcome of the annotation process was a set of 2,000 sentences with lexical items and potential lexical items fully marked and verified. The following observations can be made about the lexical items under examination:

a. of the 100 target lexical items, 4 turned out to consist of a single morpheme; 現預金 (genyokin: cash equivalent), 致死傷 (chishishō: fatal wound), 長信銀 (chōshingin: long-term credit bank) and 二人暮らし (futarigurashi: two people living together). These were a surprise as morphemes with three adjacent kanji are very rare in UniDic.

b. only 50 of the original 100 target lexical items (now reduced to 96) had 10 or more free-standing instances validated. The remainder had smaller numbers, mainly because they were embedded in longer compounds. For example, 音響機 onkyōki, a variant surface form of 音響器 “sounder” turned out to be always part of the common term 音響機器 onkyōkiki “audio equipment”. Similarly 発展途上 hattentojō “developing” was always part of 発展途上国 hattentojōkoku
Table 5.11: Frequencies of Unknown Lexical items

<table>
<thead>
<tr>
<th>Occurrences</th>
<th>Lexical items</th>
<th>Occurrences</th>
<th>Lexical items</th>
</tr>
</thead>
<tbody>
<tr>
<td>10+</td>
<td>50</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1</td>
<td>624</td>
</tr>
</tbody>
</table>

For testing ML models in these sorts of situations it is usual to use either a cross-validation approach, dividing the sets of target lexical items, or to do a total “hold-out”, which would mean building and testing a model excluding each of the 755 lexical items in turn. As the training times for 2,000 sentences was not excessive, it was decided to use a ten-way cross-validation as this would provide a reasonable balance between the two approaches. The 755 possible lexical items were arbitrarily divided into 10 groups by ranking them according to frequency and allocating 1, 11, 21, etc. to the first group; 2, 12, 22, etc. to the second group; and so on. This maintained an equal spread of the higher-frequency lexical items across the groups, and when sentences were selected from the full 2,000 according to the presence/absence of the
lexical items in a group, the sentence division was always approximately 150:1,850, thus providing a reasonable basis for the cross-validation.

The first step in testing using these sentences and the ten-way cross-validation was to repeat the tests using the lexical features used in previous testing. The results of the test are shown in Table 5.12. With regard to these results:

a. the “morph/POS” model simply uses morphemes and POS tags as features. All the other models use lexical features in addition to morphemes and POS tags.

b. the reported precision is simply based on the accuracy with which the model labelled morpheme sequences as possible lexical items, when compared with the identification in the Comblex lexicon. It is not a true indication of the precision that can be achieved by the model as it is based on all 7,416 tagged lexical items, including those that were in the training data. A more detailed analysis of precision is reported later in this chapter.

c. the recall is divided into token recall, which is based on treating each occurrence of a lexical item as a countable event, and type recall which treats the occurrence of a lexical item as a binary event regardless of how many times the lexical item is detected. In addition a second type recall category has been calculated based on lexical items which are successfully detected 5 or more times. This clearly only applies to the 89 target lexical items which occur that often, and is of interest as in a functioning lexical item harvesting system some frequency threshold would probably be employed.

d. for comparison the results from the rule-based chunker are included.

The following observations can be made about these results:
Table 5.12: Precision and Recall for ML Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Morphi POS</th>
<th>Freq 1</th>
<th>Freq 2</th>
<th>Freq 3</th>
<th>Freq 4</th>
<th>Freq 12</th>
<th>Freq 13</th>
<th>Freq 14</th>
<th>Rule-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>.775</td>
<td>.772</td>
<td>.787</td>
<td>.782</td>
<td>.787</td>
<td>.780</td>
<td>.770</td>
<td>.787</td>
<td></td>
</tr>
<tr>
<td>Recall (Token)</td>
<td>.609</td>
<td>.602</td>
<td>.567</td>
<td>.587</td>
<td>.578</td>
<td>.582</td>
<td>.575</td>
<td>.574</td>
<td></td>
</tr>
<tr>
<td>Recall (Type)</td>
<td>.566</td>
<td>.565</td>
<td>.553</td>
<td>.557</td>
<td>.557</td>
<td>.562</td>
<td>.539</td>
<td>.554</td>
<td></td>
</tr>
<tr>
<td>Recall (Type5+)</td>
<td>.933</td>
<td>.911</td>
<td>.900</td>
<td>.900</td>
<td>.922</td>
<td>.889</td>
<td>.878</td>
<td>.911</td>
<td></td>
</tr>
</tbody>
</table>

a. in a major contrast with the results from the earlier tests, the addition of lexical features has not improved the recall. The relatively simple model using morphemes and POS tags has achieved the best results of all the models. There is no simple explanation for this, however it must be recognized that these latest results are based on more carefully-selected sentences and human-verified markup of lexical items, and thus have to be given greater credence;

b. the straight type recalls are generally lower than the token recalls. This is to be expected, as the impact of the high-scoring potential lexical items (10/10, 9/10, etc.) is reduced to simply 1/1, and hence the long tail of singly-occurring lexical items has more weight;

c. in general the type recalls stayed much the same with the addition of lexical features, whereas the token recalls fell. Inspection of the raw results does not
throw much light on this. It seems the recall of higher-ranking lexical items drops off a little, and the overall recall from the tail stays much the same on average, although there is a lot of variation. Since the tail is of singly-occurring lexical items, not too much can be made of this;

d. the performance of the rule-based chunker is interesting. Its comparatively poor precision is to be expected as it is not designed to detect a range of lexical items such as conjunctions, adverbs, etc. which are being detected by the supervised models. It has, however, achieved higher levels of recall than the supervised models, despite its known tendency to collect long unsegmented compound nouns, etc.

The next tests were to carry out some feature ablation in which the morpheme and POS features were selectively removed from some of the tests. In this case the tests were carried out on the morpheme/POS, 24 and 1234 models. The results are in Table 5.13.

As can be seen, removing the morphemes as a feature improves the recall somewhat, but reduces the precision significantly. Removing the POS tags as a feature

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall (Token)</th>
<th>Recall (Type)</th>
<th>Recall (Type5+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS (no morph)</td>
<td>.653</td>
<td>.629</td>
<td>.613</td>
<td>.944</td>
</tr>
<tr>
<td>Morpheme (no POS)</td>
<td>.759</td>
<td>.450</td>
<td>.449</td>
<td>.856</td>
</tr>
<tr>
<td>Freq 24 (no morph)</td>
<td>.709</td>
<td>.540</td>
<td>.557</td>
<td>.889</td>
</tr>
<tr>
<td>Freq 24 (no POS)</td>
<td>.767</td>
<td>.464</td>
<td>.462</td>
<td>.822</td>
</tr>
<tr>
<td>Freq 1234 (no morph)</td>
<td>.721</td>
<td>.559</td>
<td>.553</td>
<td>.900</td>
</tr>
<tr>
<td>Freq 1234 (no POS)</td>
<td>.762</td>
<td>.486</td>
<td>.458</td>
<td>.789</td>
</tr>
</tbody>
</table>

Table 5.13: Precision and Recall for Models with Feature Ablation
reduces both precision and recall, although it does not reduce the precision as much as morpheme removal. The best performer on recall was the model with morphemes alone, which achieved levels comparable with the rule-based chunker. The next test was to insert a feature column based on the encapsulating morphemes or POS tags, as discussed earlier. In such a feature either the morpheme itself or the POS tag would be inserted as a feature for the morphemes immediately before and after a known lexical item. There is a risk with such a feature in that it is very strongly associated with the labelling, and the ML model may avoid labelling morpheme sequences in the test data which lacked the feature. This proved to be the case, and in tests where all known lexical items were associated with these features, the recall for untagged lexical items dropped to below 0.2.

A further approach was tested in which the encapsulation features were based on the results from the rule-based chunker. In this case the feature column could be constructed for both the training data and the test data as the chunking is not based on any prior lexical information. The POS tag was used as the feature. The results of these tests are in Table 5.14.

The conclusions that can be drawn from these tests, bearing in mind that they

<table>
<thead>
<tr>
<th>Model</th>
<th>Morph/POS</th>
<th>POS</th>
<th>Freq 24</th>
<th>Freq 1234</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>.758</td>
<td>.641</td>
<td>.770</td>
<td>.765</td>
</tr>
<tr>
<td>Recall (Token)</td>
<td>.616</td>
<td>.628</td>
<td>.586</td>
<td>.589</td>
</tr>
<tr>
<td>Recall (Type)</td>
<td>.583</td>
<td>.605</td>
<td>.559</td>
<td>.558</td>
</tr>
<tr>
<td>Recall (Type5+)</td>
<td>.933</td>
<td>.956</td>
<td>.933</td>
<td>.889</td>
</tr>
</tbody>
</table>

Table 5.14: Precision and Recall for Models with Encapsulation Features
Chapter 5: Lexical Item Identification From Morpheme Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Morph/POS</th>
<th>Morphemes</th>
<th>POS</th>
<th>Feature 24</th>
<th>Rule-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.655</td>
<td>0.598</td>
<td>0.541</td>
<td>0.678</td>
<td>0.633</td>
</tr>
<tr>
<td>Recall</td>
<td>0.649</td>
<td>0.498</td>
<td>0.636</td>
<td>0.654</td>
<td>0.420</td>
</tr>
</tbody>
</table>

Table 5.15: Precision and Recall with Strict Evaluation

are focusing on a target set of potential lexical items, is that in general the rule-based chunker achieves the highest levels of recall, both for tokens and for type. When only the higher frequency lexical items are considered, the ML model using POS and encapsulating information from the rule-based chunker achieves a marginally higher recall.

5.7.5 Precision Issues

As discussed above the precision reported in the tests of the ML models, while probably useful for inter-model comparisons, is not a true precision in that many lexical items in the test data have been labelled in the training data. Also the precision measure was based on all lexical items whereas the recall was only based on a target set. In order to get a more accurate measure of the precision, several tests were rerun with the scoring modified to exclude lexical items which occurred in both the training and test data. The results of these tests are reported in Table 5.15. Note that the measures are based on all lexical items, not on the target set.

The following observations can be made:

a. the suppression of the lexical items in the training data for the ML models reduced the number of lexical items by about half. There was no reduction for
the rule-based chunker;

b. when we are looking at all the lexical items rather than the target set, the recall is actually slightly better. The precision has fallen somewhat, but this is not surprising given that common lexical items have been excluded;

c. in this test the 24 lexical feature model performed the best, which mirrors earlier tests when we were looking at the overall recovery of lexical items;

d. the rule-based chunker performed quite well on precision, but not as well on recall. This is to be expected; it extracts far fewer possible lexical items than the ML models, but the ones it does are of reasonable quality. Also the rule-based chunker will be identifying such things as unlexicalized compound nouns and verbs, which may well be of interest in neologism extraction, but will be treated as a failure when assessing precision based on known lexical items.

5.7.6 Impact of Training Texts

The cross-validation-based testing described above used training data which had been comprehensively marked up at the morpheme level. To test the sensitivity of the process to the degree of markup, a trial was conducted using a training model built using 20,000 lines of newspaper text which had only been automatically tagged. When tested on the sentences containing the target lexical items, the recall was significantly lower than was achieved in the cross-validation-based testing. The conclusion must be that the supervised approach is quite sensitive to the state of the labelling, and that it is advisable to have models built using training data with comprehensive labelling at the lexical item level. This is consistent with findings reported by others, e.g. in
Qu et al. (2015).

5.7.7 Impact of Training Size

As mentioned above, the sentence division in the cross-validation was approximately 150:1,850. In any ML process there is a question of the adequacy of the size of the training data, so it is appropriate to attempt to evaluate whether the results are sensitive to the amount of training data, and if so to what degree.

In order to assess the sensitivity to training data size, additional runs of the cross-validation were carried out stepping the training size downwards from the full set of sentences (about 1,850), and adding extra sentences to the cross-validation set, e.g. drawn from the Tanaka Corpus. The cross-validation runs were then repeated for the plain (POS and morpheme only) and 24 Feature models with:

a. the default 1,850 training sentences from the cross-validation;

b. the training sentences limited to 500, 800, 1,000 and 1,500;

c. the default plus an additional 1,000 sentences from the Mainichi Shimbun (both models) (“M”)

d. the default plus an additional 1,000, 2,000 and 10,000 sentences from the Tanaka Corpus (the latter for plain model only) (“T”)

The results of these tests are in Table 5.16 and Table 5.17.

As can be seen, truncating the number of training sentences initially improved most of the measures, but then it began to decline. The addition of the Tanaka sentences generally made an improvement, especially to the recall, but this was not universal. A further test was made of the plain model plus 10,000 Tanaka sentences,
but there was no improvement.

The conclusion can be drawn from this is that generally stable results are coming from training data in the range 1,500-3,800 sentences, and there is no real advantage in going above that number.
5.8 Testing for Potential Lexical Items in Unseen Texts

As reported in the previous section, two techniques for analyzing the morpheme sequences from a morphological analyzer and proposing possible lexical items: a supervised approach and a rule-based approach were developed and tested. The two techniques achieved broadly comparable results when tested on text containing target “unknown” lexical items. To a certain degree these tests and results are artificial in that they do not determine which, if either, of the techniques is suitable for analyzing other texts, which have not had the advantage of hand-annotation and verification of lexical items.

A further concern with the supervised approach lies in the training of the ML models. In the cross-validation-based testing described in the previous section, the training data had its lexical items quite comprehensively labelled during the annotation process, and achieved a degree of success, however as reported in Section 5.7.6 above the supervised approach is quite sensitive to the state of the labelling.

In order to test the techniques “in the wild”, 100 sentences were extracted from each of the following sources:

a. transcripts of proceedings in the lower house of the Japanese Diet;

b. a collection of crime short-stories;

c. the Japanese Wikipedia article on Australia.

Short pieces of text such as sub-headings, names of speakers, etc. were removed.

For testing the ML models, the 2,000 sentences used in the previous tests were used
Table 5.18: Precision and Recall for Texts from Different Genres

<table>
<thead>
<tr>
<th>Source</th>
<th>Morph/POS</th>
<th>Morph/POS/24</th>
<th>Rule-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Diet</td>
<td>0.703</td>
<td>0.718</td>
<td>0.732</td>
</tr>
<tr>
<td>Crime Stories</td>
<td>0.763</td>
<td>0.773</td>
<td>0.792</td>
</tr>
<tr>
<td>Wikip.</td>
<td>0.748</td>
<td>0.742</td>
<td>0.788</td>
</tr>
</tbody>
</table>

As training data. Two models: the morpheme/POS model and the morpheme/POS/24-feature model were tested along with the rule-based chunker. The initial tests were to assess the identification of known lexical items, and the results are in Table 5.18.

As before, the reported precision is based on all lexical items, including those labelled in the training data. The results indicate the ML models are proposing approximately the same numbers of lexical items as the lexicon-based markup of the test data has indicated, and the precision/recall levels seem quite satisfactory. At first sight the rule-based chunker has produced much worse results, with a much lower number of proposed lexical items and a smaller proportion matching with the known lexical items.

However the reported precision and recall here is not really the issue; after all we know the strengths of the two approaches when it comes to identifying known lexical items. What needs to be considered is how good the approaches are at identifying possible lexical items which are not already known to be lexical items. In the 100 Diet sentences, for example, there are 366 potential lexical items indicated by the morpheme/POS ML model which do not match known lexical items, and for the rule-based chunker there are 229. The question really is how many of those are
suitable potential lexical items.

In order to test this, the lists of reported potential lexical items for each of the three approaches for the 300 sentences were analyzed to determine which:

a. did not match any marked lexical items in the text;

b. did not match a lexical item in Complex;

c. appeared to be worth further investigation.

The reason for the second criterion can be seen in sentence containing: 第三の証人は古美術商・片岡一郎氏で daisan no shōnin wa kobijutsushō kataoka ichirō shi de “the third witness, antiquarian Mr Ichiro Kataoka”. The text had the lexical item 古美術商 kobijutsushō “antiquarian” marked (the morphemes are: 古, 美術 and 商). The ML models both selected 古美術 kobijutsu “antique”, and the rule-based chunker selected 古美術商. Both are lexical items in the Complex lexicon. Although the 古美術 selected by the ML models was reported as a non-match because the markup was on 古美術商, it should be filtered as it is already in the lexicon and counting it introduces a bias against the other system which (correctly) identified 古美術商.

The selection of potential lexical items “worth further investigation” is quite subjective. To give an example of what was done, consider the following sentence from one of the crime stories:

それはともかく、拷問自体の権力悪を別にしても、こういう捜査手法ではどこかで ワイロ容疑がそのまま殺人にすりかえられたかも知れず、“犯人”は笹川氏ではなく、小松氏であった(とされた)かもしれないのだ。3

3 sore wa tomokaku, gōmon jūtsu no kenyoku aku wo accepta ni shitemo, kō iu sōsa shuhō deva doka ka de, wairo yōi ga sonomama satsujin ni surikaerareta kamoshirenai, “han’in” wa sasagawa shi dewanaku, komatsu shi de atta (to sareta) kamoshirenai node. “Anyhow, even if giving the third-degree is a bad thing, from that sort of investigation it seems that somehow after a bribe the murder was switched, and the ‘culprit’ wasn’t Sasagawa, it was Komatsu who probably did it”.

Chapter 5: Lexical Item Identification From Morpheme Analysis 105
Chapter 5: Lexical Item Identification From Morpheme Analysis

<table>
<thead>
<tr>
<th>Models</th>
<th>Source</th>
<th>Morph/POS</th>
<th>Morph/POS/24</th>
<th>Rule-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet</td>
<td>104</td>
<td>74</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>Crime Story</td>
<td>46</td>
<td>28</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>Wikip.</td>
<td>27</td>
<td>23</td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.19: Counts of Potential Lexical items for Different Genres

The potential lexical items marked by the 3 approaches were:

ML-morph/POS: 拷問自体, 権力悪, 捜査手法, された, かも, しれない

ML-24: 拷問自体, 権力悪, 小松氏, された

Rule-Based.: 拷問自体, 権力悪, 捜査手法, ワイロ容疑, まま殺人, 笹川氏, 小松氏, された

In evaluating these the kana fragments (された, etc.), named entities (小松氏, 笹川氏) and other misconstructions (まま殺人) were eliminated leaving the following for closer examination:

ML-morph/POS: 拷問自体, 権力悪, 捜査手法 (3)

ML-24: 拷問自体, 権力悪 (2)

Rule-Based: 拷問自体, 権力悪, 捜査手法, ワイロ容疑 (4)

In general the longer kanji sequences were treated as “worth further investigation” unless they were quite obvious constructions such as 記者会見以降 kishakaiken izen “since the press conference”.

The number of potential lexical items considered worthy of further investigation resulting from this analysis are given in Table 5.19.
The filtering of potential lexical items from the three approaches is, of necessity, quite subjective. In an attempt to place such filtering on a more rigorous basis, a set of guidelines was developed based on the characteristics of lexical items discussed in Section 2. These guidelines can be found in Appendix B.4. As discussed in the guidelines, in the cases of multiword expressions it is extremely difficult to make an accept/reject decision without a degree of analysis of the text itself. For example, the rejection of まま殺人 (above) was largely a result of it being part of the passage そのまま殺人 and そのまま is a common and integral expression. In isolation まま殺人 may well be an idiomatic expression.

Despite the problems associated with initial assessment of potential lexical items, it is apparent that the rule-based chunker, with its propensity to gather up sequences of noun morphemes, is performing at least as effectively as the ML-based approaches in detecting potential lexical items, and possibly more effectively.

5.9 Summary, Discussion and Conclusions

In this chapter we have reported the results of a set of investigations into the ways in which potential lexical items can be identified by examining the morphemes, POS tags and other information provided by Japanese morphological analysis systems.

The ability of the analysis systems themselves to report out-of-lexicon words was investigated, and it was concluded that given the availability of fine-grained and comprehensive morpheme lexicons such as UniDic, and the fact that out-of-lexicon reporting was confined to the morpheme level, this did not provide much opportunity for supporting the detection of unrecorded lexical items.
Two systems for identifying possible lexical items in segmented texts were developed and investigated:

a. a rule-based method using patterns of POS tags applied to the morphemes within lexical items and in the encapsulating syntactical elements;

b. a CRF-based machine-learning approach using a range of features, including some drawn from frequencies derived from a lexicon and a large corpus.

When tested using texts with hand-annotated lexical items with the goal of identifying target “unknown” lexical items, both approaches achieve comparable levels of recall, however when applied to identifying useful potential lexical items in text, the machine-learning approach performed less well. The more relaxed rule-based approach, although it generated fewer candidates, appears to be generally more successful in identifying words and compounds worthy of further investigation.

While the side-by-side comparison reported here needs to be repeated on a larger scale, with the selected potential lexical items actually taken to the stage of deciding whether or not they are worthy of inclusion in a bilingual dictionary, the overall outcome indicates a relatively limited amount of success in training ML models to recognize potential lexical items.

From the point of view of application to Japanese lexicography, the rule-based chunker appears to have considerable potential. It could easily be applied, in combination with the Complex lexicon or its equivalent, to examine large bodies of text and identify potential lexical items for further investigation.
5.10 Postscript: Alternative Chunking Approach

After the work described in this chapter was carried out, it was decided to make a brief exploration into using a dependency analysis system CaboCha, to ascertain whether it could provide a useful alternative approach to chunking morphemes into actual or potential lexical items. Dependency analysis is addressing a different issue from the target of the current investigation, but in isolating segments of text which are in grammatically dependent relationships it may also be highlighting aggregations of morphemes which are of interest to lexicographers.

The exploration consisted of analyzing several passages, including two used earlier in the chapter. The first was: 高炉の還元不良という問題を調べました。 discussed in Section 5.4 (see Table 5.4). The grammatical elements isolated by Cabocha were:

高炉の 還元不良と いう 問題を 調べました

Three of these elements (高炉の, 還元不良と, 問題を) consist of a noun or NP with a trailing case particle. It suggests that it may be possible to acquire useful chunks by detecting and removing trailing particles.

The second sentence was それはともかく、拷問自体の権力悪を別にしても、こういう捜査手法ではどこかでバイオ容疑がそのまま殺人にすりかえられたかも知れず，“犯人” は笹川氏ではなく、小松氏であった（とされた）かもしれないのだ。 used in Section 5.8 above. The identified grammatical elements were:

それは
ともかく、
拷問自体の
権力悪を
別にしても、
こういう
捜査手法では
どこかで
ワイロ容疑が
そのまま
殺人に
すりかえられたかも知れず、
“犯人”は
笹川氏ではなく、
小松氏であった
(と
された)かもしれないのだ。

It is noted that here too trimming trailing case particles would result in the same four unlexicalized NPs being identified as in the earlier discussion (拷問自体, 権力悪, 捜査手法 and ワイロ容疑).

While this does not immediately indicate any advantage over the rule-based chunking approach for NPs, it may be useful for other elements such as the adverbial clause 別にして betsunishite “aside, apart from” which is not usually lexicalized.

The use of this sort of aggregation is clearly an area for future study.
Chapter 6

Japanese Loanword Multi-Word Expressions: Extraction, Segmentation and Translation

6.1 Introduction

This chapter describes an investigation of techniques for the segmentation of Japanese loanword multiword-expressions (MWEs) and the construction of likely English translations. The investigation continues the theme of analysis of morpheme sequences. In the case of loanword MWEs, where the morphemes are transliterated loanwords, the general lack of clear morpheme boundaries is an issue which impacts on the accurate segmentation and eventual translation of the MWEs. The particular

---

research question being addressed is whether a correct morpheme segmentation of such loanword MWEs can be derived using a combination of a loanword morpheme lexicon, a bilingual dictionary and an \( n \)-gram corpus.

The investigation has led to the development and evaluation of a system, which has been called Corpus-based Loanword Segmentation and Translation (CLST), which leverages the availability of large bilingual dictionaries of loanwords and English \( n \)-gram corpora. As described below, it is achieving high levels of accuracy in discriminating between single loanwords and MWEs, and in segmenting MWEs. It is also generating useful translations of MWEs, and has the potential to be a major aide both to lexicography in this area, and to translating.

### 6.2 Orthographical Aspects of Loanwords in Japanese

Loanwords in Japanese are currently written in the katakana syllabic script. This is an orthographical convention that has been applied relatively strictly since the late 1940s, when major script reforms were carried out. Prior to then loanwords were also written using the hiragana syllabary and on occasions kanji, either by applying the pronunciation and meaning of the loanword to a combination of kanji which convey the same meaning (e.g. the compound 煙草, where the constituent characters mean “smoke” and “grass”, is typically pronounced tabako and means “tobacco”, “cigarettes”, etc.), or by using kanji selected because their reading matches the transliteration of the loanword, e.g. “America” was written 亜米利加 as this
sequence would be pronounced *amerika.*

The *katakana* script is not used exclusively for loanwords. Other usages include:

a. transcription of foreign person and place names, and other named entities. Many Japanese companies use names which are transcribed in *katakana*. Chinese (and Korean) place names and person names, although they are usually available in *kanji*, are often written in *katakana* transliterations, unless they are well-known such as 北京 (*Beijing*) or 上海 (*Shanghai*);

b. the scientific names of plants, animals, etc.

c. onomatopoeic words and expressions, although these are often also written in *hiragana*;

d. occasionally for emphasis and in some contexts for slang words, in a similar fashion to the use of italics in English.

The proportion of *katakana* words that were not loanwords was measured by Brill *et al.* (2001) at about 13%, and this result was broadly confirmed by Igarashi (2007). *(The impact and handling of these is discussed briefly at the end of Section 6.7.)*

Loanwords are taken into Japanese by adapting the source language pronunciation to conform to the relatively restricted set of syllabic phonemes used in Japanese. Thus blog becomes *burogu*, and elastic becomes *erasutikku*. When written, the syllables of the loanword are transcribed in the *katakana* script (ブログ, エラスティック), which as mentioned above in modern Japanese is primarily used for this purpose. This use of a specific script means possible loanwords are generally readily identifiable in text and can be extracted without complex morphological analysis.

---

2This process is known as 当て字 *ateji* in Japanese, and is similar to the process used for transcription of foreign words in Chinese.
6.3 Assimilation of Loanwords in Japanese

Unlike some national languages where there is official opposition to the incorporation of foreign words, Japanese has assimilated a large number of such words, to the extent that they constitute a sizeable proportion of the lexicon (Irwin 2011). For example, over 10% of the entries and sub-entries in the major *Kenkyūsha’s New Japanese-English Dictionary* (5th ed.) (Toshiro et al. 2003) and 26% of the entries and sub-entries in the open-source Japanese-Multilingual Dictionary (JMdict) (Breen 2004b) are wholly or partly made up of loanwords. This proportion is higher in glossaries in technical fields, for example in the Japanese-English *Life-Sciences Dictionary* 44% of the entries contain loanwords. In addition there are several published dictionaries consisting solely of such loanwords, e.g. the Gakken *A Dictionary of Katakana Words* (Kabasawa and Satō 2003). Estimates of the number of loanwords and particularly MWEs incorporating loanwords in Japanese range into the hundreds of thousands. While a considerable number of loanwords have been taken from Portuguese, Dutch, French, etc., the overwhelming majority are from English.

While loanwords have been common in Japanese for many years, recent studies have revealed that the proportion of loanwords to other words in newspapers, magazines, etc. has been rising steadily in recent decades (Igarashi 2007), with anecdotal reports that they play a significant role in neologism formation in Japanese.

\[http://lsd.pharm.kyoto-u.ac.jp/\]
6.4 Loanword Multi-Word Expressions

Of particular interest in the present study are multiword loanwords (MWEs). This is because there are now large collections of basic Japanese loanwords along with their translations, and it appears that many new loanwords are formed by adopting or assembling MWEs using known loanwords. As evidence of this, one can cite the numbers of katakana sequences in the Google Japanese n-gram Corpus (Kudo and Kazawa 2007). Of the 2.6 million 1-grams in that Corpus, approximately 1.6 million (62%) are in katakana or other characters used in loanwords. Inspection of those 1-grams indicates that once the words that are in available dictionaries are removed, the majority of the more common members are MWEs which had not been segmented during the generation of the Corpus. Moreover the n-gram Corpus also contains 2.6 million 2-grams (out of 80.5 million, i.e. 3.2%) and 900,000 3-grams (out of 394 million) written in katakana. Even after allowing for the multiple-counting between the 1, 2 and 3-grams, and the imperfections in the segmentation of the katakana sequences, it is clear that the vast numbers of multiword loanwords in use are a fruitful area for investigation with a view to extraction and translation.

6.5 Prior Work

There has not been a large amount of work published on the automatic and semi-automatic segmentation, extraction and translation of Japanese loanwords. Much that has been reported has been in areas such as back-transliteration (Matsuo et al.

---

\[^4\text{In addition to katakana, loanwords use the} \text{- chōon character for indicating lengthened vowels, and on rare occasions the} \text{ and} \text{ syllable repetition characters.}\]
1996; Matsuo et al. 1997; Knight and Graehl 1998; Bilac and Tanaka 2004), or on extraction from parallel bilingual corpora (Brill et al. 2001). More recently work has been carried out exploring combinations of dictionaries and corpora (Nakazawa et al. 2005), although this lead does not seem to have been followed further.

Both Bilac and Tanaka (2004) and Nakazawa et al. (2005) address the issue of segmentation of MWEs. This is discussed in Section 6.5.1 below.

Although working with translation of Japanese noun-noun compounds and employing different techniques, Baldwin and Tanaka (2004) also explore the combination of a bilingual dictionary and large corpus.

### 6.5.1 Segmentation

As mentioned above, many loanwords appear in the form of MWEs, and their correct analysis and handling often requires separation into their composite words. In Japanese there is a convention that loanword MWEs have a “middle-dot” punctuation character (・) inserted between the components, however while this convention is usually followed in dictionaries, it is rarely applied elsewhere. Web search engines typically ignore this character when indexing, and a search for a very common MWE, トマトソース tomato sauce, reveals that it almost always appears as an undifferentiated string. Moreover, the situation is confused by the common use of the “・” character to separate items in lists, in a manner similar to a semi-colon in English. In practical terms, systems dealing with loanword MWEs must be prepared to do their own segmentation.

One approach to segmentation is to utilize a Japanese morphological analysis sys-
tem. These have traditionally been weak in the area of segmentation of loanwords, and tend to default to treating long katakana strings as 1-grams. In testing a list of loanwords and MWEs using the ChaSen system (Matsumoto et al. 2003), Bilac and Tanaka (2004) report a precision and recall of approximately 0.65 on the segmentation, with a tendency to under-segment being the main problem. Nakazawa et al. (2005) report a similar tendency with the Juman morphological analyzer (Kurohashi and Nagao 1998). The problem was most likely due to the relatively poor representation of loanwords in the morpheme lexicons used by these systems. For example the IPADIC lexicon (Asahara and Matsumoto 2003b) used at that time only had about 20,000 words in katakana, and many of those were proper nouns.

In this study, the MeCab morphological analyzer (Kudo et al. 2004) has been used along with the UniDic lexicon (Den et al. 2007), as discussed below.

As they were largely dealing with unlexicalized words, Bilac and Tanaka (2004) used a dynamic programming model trained on a relatively small list of 13,000 katakana words, and reported a high precision in their segmentation. Nakazawa et al. (2005) used a larger lexicon in combination with the Juman analyzer and reported a similarly high precision.

6.5.2 Non-English Words

A number of loanwords are taken from languages other than English. The JMdict dictionary (Breen 2004b) has approximately 44,000 loanwords, of which 4% are marked as coming from languages other than English. Inspection of a sample of the 22,000 entries in the Gakken A Dictionary of Katakana Words (Kabasawa and
Satō 2003) indicates a similar proportion. (In both dictionaries loanwords from languages other than English are marked with their source language.) This relatively small number is known to cause some problems with generating translations through transliterations based on English, but the overall impact is not very significant.

6.5.3 Pseudo-English Constructions

A number of katakana MWEs are constructions of two or more English words forming a term which does not occur in English. An example is バージョンアップ bājon’appu “version up”, meaning upgrading software, etc. These constructions are known in Japanese as 和製英語 wasei eigo “Japanese-made English”. Inspection of the JMdict and Gakken dictionaries indicate they make up approximately 2% of katakana terms, and while a nuisance, are not considered to be a significant problem.

6.5.4 Orthographical Variants

Written Japanese has a relatively high incidence of multiple surface forms of words, and this particularly applies to loanwords. Many result from different interpretations of the pronunciation of the source language term, e.g. the word for “diamond” is both ダイヤモンド daiyamondo and ダイアモンド daiamondo, with the two occurring in approximately equal proportions. (The JMdict dictionary records 10 variants for the word “vibraphone”, and 9 each for “whiskey” and “vodka”.) In some cases two different words have been formed from the one source word, e.g. the English word “truck” was borrowed twice to form トラック torakku meaning “truck, lorry” and トロッコ torokku meaning “trolley, rail car”. Having reasonably complete coverage of
alternative surface forms is important in the present project.

6.5.5 Polysemy in Loanwords

The relatively limited set of phonemes in the Japanese language results in there being many instances where a single loanword actually derives from several source words in other languages. Examples include:

- ホール hōru “hall, hole, whole”
- フリー furī “free, flea, flee”
- フォーク fōku “fork, folk”

This presents a potential problem for translation, especially with systems using simple back-transliteration. A human translator will attempt to find the most appropriate source word from the context, including the other components in the case of an MWE, however it is a well-known cause of mistranslation.

6.6 Extraction of Loanwords from Japanese Text

As mentioned above, approximately 87% of katakana terms are loanwords. This use of a specific script to write loanwords means that much of the task of loanword extraction becomes a relatively simple process of detecting sequences of the appropriate characters (katakana, alphanumerics, special characters, etc.).

It must be recognized that katakana sequences will not always make up complete potential lexemes. Terms such as アーキテクチュラル ākitekuchuraru “architectural” and クロスランゲージ korosurangēji “cross-language”, neither of which are usually included in published dictionaries, typically occur as part of multiword ex-
pressions including *kanji*, e.g. クロスランゲージ情報検索 *korosurangēji jōhōkensaku* “cross-language information retrieval”. This has some implications for lexicon-based processing, and is discussed further in Section 6.9.

### 6.7 Segmentation and MWE Translation

Given the focus on extraction and translation of loanword MWEs the twin tasks of segmentation of the MWEs into their constituent source-language components, and generation of appropriate translations for the MWEs as a whole, need to be addressed. While the back-transliteration approaches in previous studies have been quite successful, and have an important role in handling single-word loanwords, in this study it was decided to experiment with an alternative approach which builds on the large lexicon and *n*-gram corpus resources which are now available. This approach, which has been labelled “CLST” (Corpus-based Loanword Segmentation and Translation), builds upon a direction suggested in Nakazawa *et al.* (2005) in that it uses a large English *n*-gram corpus both to validate alternative segmentations and select candidate translations. To a certain extent it mimics the approach a human translator would use when encountering an unknown loanword.

The three key resources used in CLST are:

a. a dictionary of *katakana* words which has been assembled from:

i. the entries with *katakana* headwords or readings in the *JMdict* dictionary;

ii. the entries with *katakana* headwords in the *Kenkyūsha’s New Japanese-English Dictionary*;

iii. the *katakana* entries in the *Eijiro* dictionary file (see Section 4.2);
iv. the *katakana* entries in a number of technical glossaries covering biomedical
topics, engineering, finance, law, etc.;
v. the named-entities in *katakana* from the *JMnedict* named-entity database.

This dictionary, which contains both base words and MWEs, includes short
English translations which, where appropriate, have been split into identifiable
senses. It contains a total of 285,000 entries.

b. a collection of 210,000 *katakana* words drawn from the headwords of the dictio-
nary above and the *katakana* morphemes in the UniDic lexicon. Known MWEs
have been split into their components where this can be carried out reliably;

c. the Google English *n*-gram Corpus (see Section 4.4), which contains 1-grams to
5-grams collected from the WWW in 2006, along with frequency counts. In the
present project a subset of the Corpus consisting only of case-folded alphabetic
tokens and associated characters such as hyphens and apostrophes has been
used.

The process of segmenting an MWE and deriving a translation is as follows:
a. using the *katakana* words in (b) above, generate all possible segmentations of
the MWE. A recursive algorithm is used for this. Table 6.1 shows the segments
derived for the MWE ソーシャルブックマークサービス *sōsharubukkumakusābisu*
“social bookmark service”.

b. for each possible segmentation of an MWE, assemble one or more possible
glosses as follows:
   i. take each element in the segmented MWE, extract a gloss from the diction-
      ary and assemble a composite potential translation by simply concate-
nating the glosses. Many dictionary entries have several glosses for a term, and the following two approaches were tested: (1) where there are multiple senses, extract only the first gloss from each in order to assemble all possible combinations. (This first gloss heuristic is being applied as lexicographers typically place the most relevant and succinct translation first, and this has been observed to be often the most useful when building composite glosses); and (2) extract and use all glosses from the entry. The relative effectiveness of these two approaches is discussed below. As examples of the overall process, for ソーシャル・ブックマーク・サービス the element サービス sabisu has two senses “service” and “goods or services without charge”, so the possible glosses were “social bookmark service” and “social bookmark goods or services without charge”. For ソーシャル・ブック・マーク・サービス the element マーク has senses of “mark”, “paying attention”,

| ソーシャル・ブックマーク・サービス |
| ソーシャル・ブックマーク・サー・ビス |
| ソーシャル・ブック・マーク・サービス |
| ソーシャル・ブック・マーク・サー・ビス |
| ソー・シャル・ブックマーク・サービス |
| ソー・シャル・ブックマーク・サー・ビス |
| ソー・シャル・ブック・マーク・サービス |
| ソー・シャル・ブック・マーク・サー・ビス |

Table 6.1: Segmentation Example
“markup” and “Mach”, so the potential glosses were “social book mark service”, “social book markup service”, “social book Mach service”, etc. A total of 48 potential translations were assembled for this MWE.

ii. where the senses are tagged as being affixes, also create combinations where the gloss is attached to the preceding or following gloss as appropriate.

iii. if the entire MWE is in the dictionary, extract its gloss as well. The presence of an MWE in the dictionary would normally indicate it would not be considered further as it is clearly not a neologism. It is useful, however, during testing to retain it for comparison with the results of the later processing.

It may seem unusual that a single sense is being sought for an MWE with polysemous elements, however this is a common occurrence in natural languages (Kim and Baldwin 2007). For example in English the words “civil” and “union” are both polysemous, but the NP “civil union” has one common sense. In Japanese, polysemy in loanwords is most commonly due to them being derived from multiple source words. For example ランプ ranpu has three senses reflecting that it results from the borrowing of three distinct English words: “lamp”, “ramp” and “rump”. On the other hand, MWEs containing ランプ, such as ハロゲンランプ harogenranpu “halogen lamp” or オンランプ onranpu “on-ramp” almost invariably are associated with one sense or another.

c. attempt to match the potential translations with the English n-grams, and where a match does exist, extract the frequency data. For the example above, only “social bookmark service”, which resulted from the ソーシャル・ブックマー
ク・サービス segmentation, was matched successfully. In some cases the glosses needed adjustment to match the segmentation in the n-gram Corpus, as it has been compiled following the convention of treating apostrophes, hyphens, etc. as segment breaks (“women’s” appears as the 2-gram “women” + “’s”, and “full-stop” as the 3-gram “full” + “-” + “stop”); d. where multiple matches result, choose the one with the highest frequency as both the most likely segmentation of the MWE and the candidate translation.

The approach described above assumes that the term being analyzed is an MWE, when in fact it may well be a single word. In the case of as-yet unrecorded words one would expect that either no segmentation is accepted or that any possible segmentations have relatively low frequencies associated with the potential translations, and hence can be flagged for closer inspection. As some of the testing described below involves deciding whether a term is or is not an MWE, the system has been enabled to handle single terms as well by checking the unsegmented term against the dictionary and extracting n-gram frequency counts for the glosses. This enables the detection and rejection of possible spurious segmentations. As an example of this, the word ボールト bōruto “vault” occurs in one of the test files described in the following section. A possible segmentation (ボー・ルト) was generated with potential translations of “bow root” and “baud root”. The first of these occurs in the English 2-grams with a frequency of 63, however “vault” itself has a very high frequency in the 1-grams so the segmentation would be rejected.

As pointed out above, a number of katakana words are not loanwords. For the most part these would not be handled by the CLST segmentation/translation process as
they would not be reduced to a set of known segments, and would typically be reported as failures. The transliteration approaches in earlier studies also have problems with these words. Some of the non-loanwords, such as scientific names of plants, animals, etc. or words written in *katakana* for emphasis, can be detected and filtered prior to attempted processing simply by comparing the *katakana* form with the equivalent *hiragana* form found in dictionaries. Some of the occurrences of Chinese and Japanese names in text can be detected at extraction time, as such names are often written in forms such as “… 金鍾泌 (キムジョンピル)…”.

### 6.8 Evaluation

Evaluation of the CLST system was carried out in two stages: testing the segmentation using data used in previous studies to ensure it was discriminating between single loanwords and MWEs, and testing against a collection of MWEs to evaluate the quality of the translations proposed.

#### 6.8.1 Segmentation

The initial tests of CLST were of the segmentation function and the identification of single words/MWEs. Fortunately the study was able to use the same data used by Bilac and Tanaka (2004), which consisted of 150 out-of-lexicon *katakana* terms from the EDR Corpus (EDR 1995) and 78 from the NTCIR-2 test collection (Kando *et al.* 2001). The terms were hand-marked as to whether they were single words or MWEs. Unfortunately we detected some problems with this marking, for example

---

5 Kim Jong-Pil, a former South Korean politician.
Table 6.2: Results from Segmentation Tests

<table>
<thead>
<tr>
<th>Method</th>
<th>Set</th>
<th>Recall</th>
<th>Precision</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLST</td>
<td>EDR</td>
<td>98.67</td>
<td>100.00</td>
<td>99.33</td>
</tr>
<tr>
<td>MeCab</td>
<td>EDR</td>
<td>92.67</td>
<td>97.20</td>
<td>94.88</td>
</tr>
<tr>
<td>CLST</td>
<td>NTCIR-2</td>
<td>94.87</td>
<td>100.00</td>
<td>97.37</td>
</tr>
<tr>
<td>MeCab</td>
<td>NTCIR-2</td>
<td>82.05</td>
<td>92.75</td>
<td>87.07</td>
</tr>
</tbody>
</table>

シェークスピア  shēkusupia “Shakespeare” had been segmented into “shake” + “spear” whereas ホールバーニング  hōrubāningu “hole burning” had been left as a single word. It was considered inappropriate to use this data without amending these terms. As a consequence of this it is not possible to make a direct comparison with the results reported in Bilac and Tanaka (2004). Using the corrected data the two datasets were analyzed and the results are as given in Table 6.2 (the equivalent results reported in Bilac and Tanaka (2004) were: EDR - 95.45, 95.00, 95.23, and NTCIR-2 - 88.62, 86.51, 87.56.) For CLST both the “first gloss” and “all glosses” approaches were tested, with identical results for the preferred segmentation. The results from analyzing the data using MeCab/UniDic are included as well for comparison. As in Bilac and Tanaka (2004), the scores are calculated as follows: $N$ is the number of terms in the set, $c$ is the number of terms correctly segmented or identified as 1-grams, $e$ is the number of terms incorrectly segmented or identified, and $n = c + e$. Recall is calculated as $\frac{c}{N}$, precision as $\frac{c}{n}$, and the F-measure as $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$.

As can be seen, our CLST approach has achieved a high degree of accuracy in identifying 1-grams and segmenting the MWEs. Although it was not part of the
test, it also proposed the correct translations for almost all the MWEs. The less-than-perfect recall is entirely due to the few cases where either no segmentation was proposed, or where the proposed segmentation could not be validated with the English n-grams. There was no difference in the segmentation outcome between the “first gloss” and “all glosses” options.

The performance of MeCab/UniDic is interesting, as it also has achieved a high level of accuracy. This is despite the UniDic lexicon version used (1.3.12 from mid-2009) only having approximately 55,000 katakana words, and the fact that it is operating outside the textual context for which it has been trained. Its main shortcoming was that it tended to over-segment, which is a contrast to the performance of ChaSen/IPADIC reported by Bilac and Tanaka (2004) where under-segmentation was the problem. The tests were repeated with a later release of UniDic (2.1.1 from late 2012) where the number of katakana words had been increased to 94,000, however the results were identical.

6.8.2 Translation

The second set of tests of CLST was directed at developing translations for MWEs. The initial translation tests were carried out on two sets of data, each containing 100 MWEs. The sets of data were obtained as follows:

a. the 100 highest-frequency MWEs were selected from the Google Japanese 2-grams. The list of potential MWEs had to be manually edited as the 2-grams contain a large number of over-segmented words, e.g. アイコン aikon “icon” was split: アイコ + ン, and オークション òkushon “auction” was split オーク +ション;
Chapter 6: Japanese Loanword Multi-Word Expressions: Extraction, Segmentation and Translation

<table>
<thead>
<tr>
<th>MWE</th>
<th>Segmentation</th>
<th>Possible Translation</th>
<th>Frequency</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>ログインヘルプ</td>
<td>ログイン・ヘルプ</td>
<td>login help</td>
<td>541097</td>
<td>1</td>
</tr>
<tr>
<td>ログインヘルプ</td>
<td>ログ・イン・ヘルプ</td>
<td>log in help</td>
<td>169972</td>
<td>-</td>
</tr>
<tr>
<td>キーワードランキング</td>
<td>キーワード・ランキング</td>
<td>keyword ranking</td>
<td>39818</td>
<td>1</td>
</tr>
<tr>
<td>キーワードランキング</td>
<td>キー・ワード・ランキング</td>
<td>key word ranking</td>
<td>74</td>
<td>-</td>
</tr>
<tr>
<td>キャリアアップ</td>
<td>キャリア・アップ</td>
<td>career up</td>
<td>13043</td>
<td>2</td>
</tr>
<tr>
<td>キャリアアップ</td>
<td>キャリア・アップ</td>
<td>carrier up</td>
<td>2552</td>
<td>-</td>
</tr>
<tr>
<td>キャリアアップ</td>
<td>キャリア・アップ</td>
<td>career close up</td>
<td>195</td>
<td>-</td>
</tr>
<tr>
<td>キャリアアップ</td>
<td>キャリア・アップ</td>
<td>career being over</td>
<td>188</td>
<td>-</td>
</tr>
<tr>
<td>キャリアアップ</td>
<td>キャリア・アップ</td>
<td>carrier increasing</td>
<td>54</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6.3: Sample Segmentations and Translations

b. the *katakana* sequences were extracted from a large collection of articles from 1999 in the *Mainichi Shimbun* (see Section 4.3) and the 100 highest-frequency MWEs extracted.

After the data sets were processed by CLST the results were examined to determine if the segmentations had been carried out correctly, and to assess the quality of the proposed translations. The translations were graded into three groups: (1) acceptable as a dictionary gloss; (2) understandable, but in need of improvement; and (3) wrong or inadequate. An example of a translation graded as 2 is マイナスイオン *mainasuion* “minus ion”, where “negative ion” would be better, and one graded as 3 is フリーマーケット *furīmāketto* “free market”, where the correct translation is “flea market”. For the most part the translations receiving a grading of 2 were the same as would have been produced by a back-transliteration system, and in many cases they
Chapter 6: Japanese Loanword Multi-Word Expressions: Extraction, Segmentation and Translation

Table 6.4: Results from Translation Tests

<table>
<thead>
<tr>
<th>Data</th>
<th>Failed Segmentations</th>
<th>Translation Grades</th>
<th>Precision</th>
<th>Recall</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>6</td>
<td>67 26 2</td>
<td>97.87</td>
<td>92.00</td>
<td>94.84</td>
</tr>
<tr>
<td>Mainichi (Set 1)</td>
<td>3</td>
<td>80 16 1</td>
<td>98.97</td>
<td>96.00</td>
<td>97.46</td>
</tr>
<tr>
<td>Mainichi (Set 2)</td>
<td>1</td>
<td>83 16 0</td>
<td>100.00</td>
<td>99.00</td>
<td>99.50</td>
</tr>
</tbody>
</table>

were the *wasei eigo* constructions described above.

Some example segmentations, possible translations and gradings are in Table 6.3.

The assessments of the segmentation and the gradings of the translations are given in Table 6.4. The precision, recall and F measures have been calculated on the basis that a grade of 2 or better for a translation is a satisfactory outcome.

At this stage a detailed comparison was made between the results of the “first gloss” and “all glosses” options for compiling potential translations. The latter generated approximately twice the number of potential translations, which had some impact on both the chosen segmentations and the highest-ranking potential translations.

a. for the 100 Google 2-grams, one (レポートライバル repōtoraibaru “report rival”) was only segmented using the “all glosses” option. A further 8 resulted in different glosses getting the highest n-gram counts, and there was a marginal improvement in quality.

b. for the high-frequency Mainichi set, the “all glosses” option also resulted in different preferred glosses for 8 MWEs. Several of these were improvements,
e.g. “speed skating” instead of “speed skate”, and one was worse (“no success” instead of “no hit”, which is a baseball term).

In general, the “all glosses” option offers some improvement, and demonstrates that the overall approach is robust. Since the potential translations from the “first gloss” option are a subset of those from the “all glosses” option, and any lexicographic tool using this technique would probably offer a ranked list of alternatives, it was decided to use the “all glosses” option throughout.

A brief analysis was conducted on samples of 25 MWEs from each test set to ascertain whether they were already in dictionaries, or the degree to which they were suitable for inclusion in a dictionary. The dictionaries used for this evaluation were the commercial Kenkyūsha Online Dictionary (KOD)\(^6\) which has eighteen Japanese, Japanese-English and English-Japanese dictionaries in its search tool, and the free WWWJDIC online dictionary,\(^7\) which has the JMdic and JMnedict dictionaries, as well as numerous glossaries.

Of the 50 MWEs sampled:

a. 34 (68%) were in dictionaries;

b. 11 (22%) were considered suitable for inclusion in a dictionary. In some cases the generated translation was not considered appropriate without some modification, i.e. it had been categorized as “2”;

c. 3 (6%) were proper names (e.g. hotels, software packages);

d. 2 (4%) were not considered suitable for inclusion in a dictionary as they were simple collocations such as メニューエリア menyūeria “menu area”.

\(^6\)http://kod.kenkyusha.co.jp/service/
\(^7\)http://www.edrdg.org/cgi-bin/wwwjdic/wwwjdic?1C
Chapter 6: Japanese Loanword Multi-Word Expressions: Extraction, Segmentation
and Translation

As the tests described above were carried out on sets of frequently-occurring
MWEs, it was considered appropriate that some further testing be carried out on
less common loanword MWEs. Therefore an additional set of 100 lower-frequency
MWEs which did not occur in the dictionaries mentioned above were extracted from
the Mainichi Shimbun articles and were processed by the CLST system. Of these 100
MWEs:

a. 1 was not successfully segmented;
b. 83 of the derived translations were classified as “1” and 16 as “2”;
c. 8 were proper names.

The suitability of these MWEs for possible inclusion in a bilingual dictionary
was also evaluated. In fact the overwhelming majority of the MWEs were relatively
straightforward collocations, e.g. マラソンランナー marasonrannā “marathon runner”
and ロックコンサート rokkukonsāto “rock concert”, and were deemed to be not really
appropriate as dictionary entries. 5 terms were assessed as being dictionary candi-
dates. Several of these, e.g. ゴールドプラン gōrudopuran “gold plan” and エースストライカー ēsusutoraikā “ace striker” were category “2” translations, and their possi-
ble inclusion in a dictionary would largely be because their meanings are not readily
apparent from the component words, and an expanded gloss would be required.

It was clear from the analysis of the results of the tests described above that to
some extent the high-frequency Google n-gram test data had a bias towards pseudo-
English constructions commonly used in WWW contexts, such as ショッピングトップ
shoppingutoppu “shopping top” (which refers to the first page of an online shopping
site), タウンページ taunpēji “town page” (the name of an online NTT telephone
directory), etc. This possibly inflated the proportion of translations being scored with a “2”.

6.9 Summary, Discussion and Possible Improvements

In this chapter an investigation of the automatic segmentation and translation of Japanese MWEs has been described. The investigation has led to the development and evaluation of the CLST (Corpus-based Loanword Segmentation and Translation) system to segment Japanese loanword MWEs and construct likely English translations. For segmentation the CLST system has performed consistently better than previously reported approaches, and it is also producing good quality potential translations.

The response to the initial research question is that an effective loanword MWE segmentation is quite feasible using the approach explored in the work reported in this chapter.

It is worth noting that the apparent success of an approach based on a combination of large corpora and relatively simple heuristics is consistent with the conclusions reached in a number of earlier investigations (Banko and Brill 2001; Lapata and Keller 2004).

Although the CLST system has been demonstrated to perform at a high level, there are at least two aspects which could be improved considerably.

a. at present the loanword dictionary used is simply a concatenation of several
source dictionaries and glossaries, and is based largely on individual loanwords.

The katakana word list used for segmentation is similar. There are, however, a
number of katakana constructs used in loanword MWEs which do not occur in
isolation.

i. one group consists of plurals and possessives derived from English words
and which typically end in ズ (zu). Examples of these include ウイメン
ズ uimenzu “women’s”, アイテムズ aitemuzu “items” and ヘッズ hezzu
“heads”. Although some members of this group do appear in dictionaries,
e.g. メンズ menzu “men’s”, the majority do not. A similar situation exists
with some English “ed” word endings which are transliterated by adding ド
(do) to the base form, e.g. グループドタイヤ gurūbudotaiya “grooved tyre”.
While the base word グループ gurūbu “groove” is often in dictionaries, グ
ループド gurūbudo “grooved” never is.

ii. another smaller group consists of terms which have a particular meaning
when used alone, but may have other meanings when used in MWEs. A
typical example is the word ホース hōsu “hose”, which has been borrowed
from the Dutch hoos. As its pronunciation is similar to the transliteration
of the English word “horse”, it occurs thus in a number of compound
nouns and MWEs, e.g. ホースパワー hōsupawā “horsepower”, ホースラ
ディッシュ hōsuradishu “horseradish” and ホースレース hōsurēsu “horse
race”, but no published Japanese-Japanese or Japanese-English dictionary
records “horse” as a possible meaning of ホース.

Systematic addition of these forms to the lexicon used by CLST would re-
duce the numbers of failed segmentations, and improve the quality of generated translations.

b. in contrast to the cases above, a number of loanword MWEs involving the past tense of English verbs do not transliterate the “ed” ending. Examples include ミックスグリル mikkusuguriru “mixed grill” and ミックスメディア mikkusumedia “mixed media”. As the standard glosses for ミックス are “mix” and “mixture”, for the CLST technique to perform adequately in cases such as these, an additional gloss of “mixed” is required. There are a number of other verbs where this treatment would lead to more accurate generation of glosses.

6.9.1 Online CLST Interface

An online interface to the CLST system was developed and is available at: http://nlp.cis.unimelb.edu.au/jwb/gairaigo.html. The interface displays the leading candidate for segmentations and translations, and also shows the segmentation by MeCab/UniDic. The interface has proved to be useful in confirming the segmentation and translations of potential MWE loanword neologisms.

Example pages from the online interface are in Appendix E.
Chapter 7

Neologism Synthesis

7.1 Introduction

This chapter reports on the initial part of the investigation into the use of synthesis of potential Japanese terms as a method for extraction of neologisms.

The nature of Japanese orthography, with its significant use of a relatively constrained set of kanji characters for such processes as forming nouns and carrying out affixation, opens up the potential to experiment with novel methods for detecting neologisms and other unrecorded terms. The general method, which has been investigated in the work described in this chapter, involves:

a. synthesizing possible neologisms by mimicking typical Japanese morphological processes;

b. using a large corpus of Japanese text to evaluate the synthesized terms, especially to determine if they are being used in the syntactic (and if possible semantic) contexts that would confirm whether or not they are actually used
as terms in Japanese. If they are, they can then be assessed as potential lexical items.

This synthesize-and-test approach was also used with considerable success in the investigation of Japanese compound verbs (JCV) which is reported separately in the thesis (see Chapter 8).

### 7.2 Prior Work

Prior work using term synthesis techniques followed by corpus validation has been limited. Tanaka (2002) carried out some initial work in this area, and Baldwin and Tanaka (2004) used the techniques to generate and validate translations of noun-noun compounds. The present author in 2004 carried out a short study in which possible abbreviations were synthesized from longer compounds and tested against WWW text (Breen 2004a), an approach which is examined further in the work reported below.

### 7.3 Approaches to Neologism Synthesis

In the present study, a number of different approaches to the synthesis of potential neologisms have been investigated. They were:

- **Abbreviation** or **clipping**. This is a very common and productive process in Japanese, wherein the (usually) leading character of each of the components of a composite are taken to form an abbreviated compound (Tsujimura 2006:p 153). Examples of this abound, e.g. 学割 *gakuwari* “student discount” from the
full compound 学生割引 gakuseiwaribiki or 電卓 dentaku “calculator” from 電子卓上計算機 denshitakujoukeisanki (literally “electronic desktop calculator”). In many cases these abbreviations are not lexicalized and native speakers would typically recognize the source and meaning.

b. affixation. The addition of prefixes and suffixes, often written with a single kanji, is a very common morphological process in Japanese (Tsujimura 2006:p 149). Vance (1991) describes 63 single-kanji affixes commonly employed. The process is very productive and the resulting terms are not usually lexicalized unless they have an idiomatic meaning or unusual reading.

c. compounding. As in many languages, the formation of terms by combining two or more words or morphemes is very common. The components can be independent words, as in 秋空 akizora “autumn sky” where both 秋 aki and 空 sora can be used independently, or bound morphemes as in 警告 keikoku “warning” where neither component can be used alone.

Two general groups of synthesized compounds are being investigated here:

i. 2-kanji compounds, as in the examples above;

ii. composites formed by aggregating known 2-kanji compounds, for example combining 警告 (above) with 射撃 shageki “firing, shooting” forms a composite 警告射撃 keikoku shageki meaning “warning shot”. This process is very common and the resulting terms often have a clear sum-of-the-parts meaning, which results in them often not being lexicalized. (警告射 撃 is in several major bilingual dictionaries, but is not in most Japanese
dictionaries.¹ (See Section 4.2.)

Synthesizing these types of potential compounds will generate large numbers of candidates, for example generating 2-kanji compounds from the mostly commonly-used 2,500 kanji would result in over 6 million compounds, and the effectiveness of the approach will depend on the availability of a fast and effective evaluation technique.

It is expected that most synthesized terms will be nouns, and the evaluation techniques have been designed accordingly. In some cases the terms may also be adjectival nouns (形容動詞 keiyōdōshi) which may take the particle な na to act prenominally, or verbal nouns which may take the auxiliary verb する suru, and the analysis will need to take these into account.

### 7.4 Resources

As the experimentation with synthesized terms takes the form of create-and-test, a key requirement is access to appropriate large-scale corpora to test for the presence and usage patterns of the terms. Corpora such as the BCCWJ which contain relatively small quantities of specially selected documents, or collections of newspaper articles (e.g. the Kyoto Corpus), are unlikely to include new or ephemeral terms in quantities that would enable automated detection and evaluation. The largest collection of accessible text is the WWW, and several studies have shown it is broadly representative in areas such as term frequency (Keller and Lapata 2003). It is appropriate, therefore, to concentrate on WWW-based text collection for analyzing synthesized terms.

¹警告射撃 is in the large 大辞泉 Daijisen dictionary but not in 広辞苑 Kōjien or 大辞林 Daijirin.
Large-scale direct searching of the WWW, as was used in the prior work described above (Section 3.6), is no longer an option; no search engine continues to provide useable and free search APIs, and all detect and block repeated searches from the one host.

The main accessible Japanese corpus for this type of testing is the Google \( n \)-gram Corpus, which has been described earlier (see Section 4.4). It is very large, having been based on approximately 20 billion text segments, and is provided in the form of sets of 1-grams to 7-grams with counts of the numbers of occurrences. In order to make effective use of this resource in the present task, the \( n \)-grams were combined back into text strings (which overcame many of the issues with the somewhat erratic nature of the original segmentation) and sorted into a single collection. In addition, \( n \)-grams were discarded if they did not contain kana or kanji, which resulted in a total file of about 36Gb. Hash tables were compiled based on the first three characters of each string, enabling rapid access.

An issue which emerged with the Google \( n \)-gram Corpus was the use of a threshold count of 20 for each \( n \)-gram, with \( n \)-grams occurring less often being discarded. As the evaluation techniques described later in this chapter use features based on counts of \( n \)-grams using candidate terms in various syntactical contexts, and as such \( n \)-grams may occur relatively infrequently, having a truncation of the possible counts was seen as potentially a problem.

As a possible solution to this problem, a second corpus was obtained; the Kyoto University WWW Corpus, which contains about 500 million text segments. (See Section 4.4.) Although this is much smaller than the text from which the Google
Chapter 7: Neologism Synthesis

Corpus was built, it had the potential to be a source of \( n \)-grams with counts that were not subject to an arbitrary cutoff and hence could possibly enable detection and counting of text sequences below the threshold applied in the Google \( n \)-gram Corpus. The \( n \)-gram Corpus constructed from the Kyoto Corpus was used in parallel with the Google Corpus, as described later in this chapter.

A reference lexicon of known Japanese lexical items is also essential, both for filtering candidate synthesized compounds, and for establishing training data for the evaluation processes. In this task the Comblex file described in Chapter 4 has been used.

Also the most recent UniDic morpheme lexicon (756k entries), described in Chapter 4, is a useful resource in this task as it has a very comprehensive coverage of known bound morphemes in modern Japanese.

7.5 Evaluation of Synthesized Compounds

7.5.1 Initial Investigations

At the heart of this task is the development of techniques which will detect if a sequence of characters is being used as a term in written Japanese. The approach which is being evaluated is based on the assumption that if a compound is actually being used, it will occur in textual syntactic contexts which are typical for such terms. For example, if a term is being used as a noun, it could be expected that it is observed in text followed by the \( \text{が} \) \( ga \) subject-marking particle, or the \( \text{は} \) \( wa \) topic-marking particle, and will occur in encapsulations such as \( < \text{が}, \text{を} > \) where \( \text{を} \) \( wo \) marks the
object of a clause or sentence.

To illustrate this consider the number of occurrences of the noun 号令 gōrei “order, command” in the \( n \)-grams, with selected post-positional particles and verb inflections. (Table 7.1) (The Google \( n \)-gram Corpus has been used in these illustrations unless otherwise indicated.)

These \( n \)-gram frequencies confirm that 号令 is quite common, is a noun, can be used prenominally with the genitive particle の, and can occasionally form a verb with the auxiliary する suru “to do” (して shite, しない shinai and します shimasu are common inflections of する).

This contrasts with a synthesized abbreviation from the 2004 study: 国補 kokuho, which was proposed as a possible contraction of 国庫補助 kokkahojo “government subsidy”. In this case the non-zero Google and Kyoto \( n \)-gram frequencies are as shown in Table 7.2.

These frequencies point to a similar conclusion to the one reached in the earlier study; that while not uncommon 国補 is not being used as a noun (analysis of its use in text indicates that it is usually a collocation of 国 koku as a counter for countries

<table>
<thead>
<tr>
<th>Term</th>
<th>( n )-gram count</th>
<th>Term</th>
<th>( n )-gram count</th>
</tr>
</thead>
<tbody>
<tr>
<td>号令</td>
<td>187,886</td>
<td>号令を</td>
<td>40,973</td>
</tr>
<tr>
<td>号令は</td>
<td>5,544</td>
<td>号令する</td>
<td>2,190</td>
</tr>
<tr>
<td>号令が</td>
<td>15,021</td>
<td>号令して</td>
<td>849</td>
</tr>
<tr>
<td>号令な</td>
<td>287</td>
<td>号令しない</td>
<td>28</td>
</tr>
<tr>
<td>号令の</td>
<td>14,462</td>
<td>号令します</td>
<td>58</td>
</tr>
<tr>
<td>号令に</td>
<td>9,222</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1: \( n \)-gram Counts for 号令
Table 7.2: \( n \)-gram Counts for 国補

<table>
<thead>
<tr>
<th>Term</th>
<th>Google</th>
<th>Kyoto</th>
</tr>
</thead>
<tbody>
<tr>
<td>国補</td>
<td>4766</td>
<td>143</td>
</tr>
<tr>
<td>国補は</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>国補の</td>
<td>81</td>
<td>7</td>
</tr>
<tr>
<td>国補に</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 7.2: \( n \)-gram Counts for 国補

and 補 ho as a suffix meaning “probationary”.

These were tests just using post-positions. The next step was to investigate terms encapsulated in sets of common pre/postpended particles. The particles tested were:

pre: は (wa), が (ga), に (ni), の (no), な (na), て (te), や (ya)
post: を (wo), が (ga), に (ni), の (no), な (na), や (ya)

The tests were carried out using all combinations apart from repeated particles (e.g. が ... が), giving 37 combinations. The \( n \)-gram counts for 10 common encapsulations are given in Table 7.3 for three compounds: the common noun 男子 danshi “youth, young man” which occurs 5,682,486 times in the \( n \)-grams, the less common noun: 人魚 ningyo “mermaid, merman” (607,772) and the rather less common 間者 manja “spy” (24,543).

The comparatively rare noun 共編者 kyōhensha “co-editor”, which occurs 478 times in the Corpus, had zero counts for the encapsulations and positive counts for only four of the one-sided combinations (共編者は - 28, 共編者の - 85, 共編者に - 24 and の共編者 - 137). It is not a totally obscure term, and is included in several major Japanese-English dictionaries. The low counts are indicative of the problems inherent in using corpus-based techniques to detect and verify low-frequency terms.
7.5.2 Classification of Synthesized Compounds

The investigation reported above supports our expectation that the patterns of nouns encapsulated by typical particle pairs occur in the \( n \)-grams in relatively high numbers. The question is whether this information can be used to classify synthesized compounds into those which are identifiably being used as nouns and those which are not.

In carrying out such a classification, two broad approaches were considered, based on using features constructed from the counts associated with each of the combinations and encapsulations:

a. a machine-learning approach, in which models were trained on the features of associated with known terms and non-terms;

b. a heuristic approach in which rules were generated based on such things as the number of features.

<table>
<thead>
<tr>
<th>Encapsulation</th>
<th>男子</th>
<th>人魚</th>
<th>間者</th>
</tr>
</thead>
<tbody>
<tr>
<td>はXを</td>
<td>2631</td>
<td>616</td>
<td>100</td>
</tr>
<tr>
<td>はXが</td>
<td>18242</td>
<td>1002</td>
<td>47</td>
</tr>
<tr>
<td>はXに</td>
<td>9117</td>
<td>1150</td>
<td>26</td>
</tr>
<tr>
<td>はXの</td>
<td>24367</td>
<td>4290</td>
<td>54</td>
</tr>
<tr>
<td>はXな</td>
<td>1135</td>
<td>247</td>
<td>0</td>
</tr>
<tr>
<td>はXや</td>
<td>164</td>
<td>103</td>
<td>0</td>
</tr>
<tr>
<td>がXを</td>
<td>5676</td>
<td>285</td>
<td>33</td>
</tr>
<tr>
<td>がXに</td>
<td>4340</td>
<td>1015</td>
<td>35</td>
</tr>
<tr>
<td>がXの</td>
<td>6363</td>
<td>1703</td>
<td>35</td>
</tr>
<tr>
<td>がXな</td>
<td>423</td>
<td>75</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.3: \( n \)-gram Counts for 男子, 人魚 and 間者 with Encapsulations
To evaluate whether a machine-learning approach to classifying synthesized compounds can be used in this case we tested the application of a support vector machine (SVM) to the problem. SVMs have been demonstrated to be a very effective approach to this form of classification problem (Joachims 1999; Steinwart and Christmann 2008). The SVM package used is LIBSVM (Chang and Lin 2011) (See Section 4.5.3). Appendix C.1 describes the structure of feature files used by LIBSVM, and has examples of the features sets for several compounds and the associated combinations and encapsulations and their counts.

With SVM-based classification, and in particular with the LIBSVM package, there are several options and issues which need to be addressed. Several of these required considerable investigation and experimentation to determine the most appropriate options and approach to use. The issues included:

a. **one-class** or **multi-class** classification. An option with LIBSVM is to specify a one-class classification, and to provide training data describing just that class. Initially this seemed a very attractive option use as it simplified some of the decisions relating to choice and classification of training data. On testing this option did not function satisfactorily, being highly sensitive to the size of training data and giving quite erratic results. Accordingly the more usual multi-class classification was chosen.

b. LIBSVM offers two types of multi-class classification: C-support based on a regularization parameter (C) and \( \nu \)-support based on a boundary between training errors and support vectors (\( \nu \)). On testing it was determined that the C-support approach delivered the most reliable results.
c. **choice of kernel.** Often datasets are not well classified by linear classifiers, and SVM implementations typically provide a range of kernel functions which can map the data into a higher dimensional space where linear classification will be unsuitable. Experimenting with the various kernel types available with LIBSVM indicated that the commonly-used radial basis function delivered the best classification. (LIBSVM provides a function to tune various parameters, including the value of gamma for the radial basis kernel, and the value of the regularization parameter (C) for the classification. This function has been used throughout.)

d. **data scaling.** LIBSVM has a recommended option to scale the training and test data dynamically, and as expected, the classification functioned more satisfactorily when this had been done.

e. **counts or proportions.** A question arises as to whether the values of the features should be the actual counts of the various encapsulations, the proportions of each encapsulation for a particular compound, or a combination of the two. In initial testing it was apparent that the classification outcomes only differed slightly between these three, but classification based on the proportions alone functioned marginally better.

### 7.5.3 Initial Testing

In the initial stages of the evaluation of the classification techniques, it was decided to use only the encapsulating particles as features, and to defer consideration of using prepended or postpended particles later. The 37 combinations of prepended
and postpended particles mentioned above were used at first, however as it became clear that some combinations only occurred rarely, this was trimmed back to 30 combinations by the removal of <に,な>, <な,や>, <て,な>, <て,や>, <や,な>, <が,や> and <に,や>. In later testing six postpended particles (は,を,が,に,の,な) were added. The results of these two approaches is discussed below.

For initial training data, a random selection of 2-kanji compounds was chosen from JMdict and feature sets assembled by extracting the counts of the encapsulations from the n-grams. The full selection had 395 compounds, and several subsets were used to test the sensitivity to the size of the training data.

For initial test data, a batch of 841 possible 2-kanji compounds from the 2004 abbreviation experiment was used. This contained a small number of known and accepted 2-kanji compounds. At this stage of the evaluation only the Google Corpus was being used.

A major question arises as to how to classify the training data. Of the 395 known kanji pairs, all are valid nouns, however when they were compared with the n-grams using the chosen encapsulations in order to establish feature sets for each of them, 86 recorded no counts and hence no features. While it would be simple to declare these to be the negative items in the training, this would in effect be stating that compounds with any counts and features at all will be classified, and this would render the machine-learning approach redundant as we may as well simply classify any compound that received a count for any encapsulation.

In order to establish a basis for classifying the training data, an initial analysis of the test data was carried out, experimenting with the compounds which recorded low
numbers of different encapsulations. It appeared that when the number of encapsulations achieving counts was less than 5, the compound was usually not a valid term. Accordingly, the training data was classified so that only sets with 5 or more encapsulations were classified. This resulted in 232 of the 395 compounds in the training data being classified, and is referred to below as the “basic model”. A further “extended model” was built in which the six postpended particle counts were included as features, and compounds were classified if the sum of the counts of these features exceed 1,000. This added 36 compounds, resulting in 268 being classified of the 395.

7.5.4 Initial Testing - Results

The initial test using the basic model classified 71 of the 841 synthesized compounds. On checking these it emerged that 68 were valid words recorded in the reference lexicon (5 of these were proper names). The remaining 3, on investigating their usage in text passages in WWW pages, turned out to be valid, albeit rare terms. (Two were also recorded as morphemes in the UniDic morpheme lexicon.) This indicates a very high precision for the process.

In order to assess the recall, we need to estimate how many of the remaining 770 kanji pairs were valid terms. Of these 713 had no counts of encapsulations in the n-grams (i.e. the counts, if any, were below 20). 57 had non-zero counts, with the median being 3 encapsulations. (One, the common noun 概要 gaiyō “outline, summary” had 30 encapsulations counted, and was probably not classified because 71% of the total count was for a single encapsulation < の, に >.)

A random sample of 20 of the unclassified kanji pairs was checked, and none were
found to be functioning as nouns. A detailed analysis of the 57 compounds with non-
zero encapsulation counts was made which determined that 27 of these (49%) were
valid nouns. There was no discernible bias towards compounds with greater numbers
of encapsulations counted.

If we assume that there remained a small undetected number of valid compounds
in the 713 compounds for which there were no encapsulation counts, we can conclude
that the total number of valid compounds in the test data was around 100 and that
the overall recall of SVM classification was around 70%.

The trial was repeated using the extended model, with the result that 83 com-
pounds were classified of which 80 were in the reference lexicon and the other 3 were
rare but seemingly valid terms. Thus this model achieves a similar high precision and
a higher recall.

Another approach to testing the recall was carried out by processing samples of
known terms. This was done with three different sets of terms:

a. a collection of 71 terms chosen at random from JMdict. (These were taken
from the set of 75 used elsewhere in this study to carry out the analysis of the
morpheme structure of abbreviations described below (Section 7.6.1). 4 were
removed as they were in the training set.)

In this case 49 of the 71 terms were classified by both the basic and extended
models (69%), which confirms the general level of recall reported above.

b. a collection of 145 terms chosen from the more common entries in JMdict (this
dictionary tags about 25,000 of its 170,000 entries as “common words” according
to their frequency of use in various corpora and their inclusion in small-medium
dictionaries). Of these 140 (97%) were classified by the basic model and 145 (100%) by the extended model.

c. a second set of 201 term chosen from JMDict from entries which were not tagged as “common words”. Of these 103 (51%) were classified by the basic model and 115 (57%) by the extended model.

7.5.5 Initial Testing - Discussion

From this stage of the analysis there are two general conclusions that can be reached about the approach being tested:

a. it is achieving a very high precision;

b. while the recall is generally satisfactory at an average of approximately 70%, there is clearly a problem validating terms which are only occurring rarely and hence are not appearing in the encapsulation counts and in the feature sets.

The remarks in (b) above are, of course, intuitively obvious: if terms are not in a corpus, the corpus cannot be used for validating them. In this case the problem is possibly being exacerbated by the exclusion from the Google Corpus of n-gram counts below 20, as it may be masking out potentially useful information about the usage of less common terms.

The perceived limitation of the Google Corpus in having a cutoff in n-gram counts was the reason that a copy of the Kyoto WWW Corpus was obtained and a set of n-grams and associated counts extracted from it. The Kyoto Corpus and the Google Corpus were used in parallel in later stages of the investigation.
Chapter 7: Neologism Synthesis

7.6 MorpHEME-based Abbreviation Construction

7.6.1 Background

The testing in the previous section was on a set of 2-kanji compounds generated by taking known 4-kanji sequences from JMdict and creating potential abbreviations using the first and third kanji. This approach has some validity as most 4-kanji sequences are collocations of 2-kanji compounds, for example the common term for air-conditioning, 空調 kūchō is abbreviated from the full term (空気調節) by taking the first kanji from each of the components 空気 kūki “air, atmosphere” and 調節 chōsetsu “regulation, adjustment”. This pattern, however, is by no means universal. The common short term for “high school”: 高校 kōkō is an abbreviation of 高等学校 kōtōgakkō in which consists of the morphemes 高等 and 学校. The abbreviated form is derived from the first kanji of the first morpheme and the second kanji of the second.

As abbreviation in Japanese often follows a pattern of taking the first part of each component, it is appropriate to consider looking more closely at the internal structure of longer compounds and base the synthesis on the leading element of each morpheme rather than following a fixed pattern.

To analyze this, a set of 75 known abbreviations was extracted from JMdict, where the full term is also indicated (e.g. the abbreviation 携番 keiban “mobile telephone number” is cross-referenced to the full term 携帯番号 keitaibangō), and analyzed as to their relationship with the constituent morphemes.

1. 49 were from 4-kanji compounds which were formed from pairs of 2-kanji mor-
phemes (as in 携番 above);

2. 10 were from longer compounds (five to seven kanji) made up of 3 or more morphemes, but the abbreviations had been based on the first kanji of the first two morphemes;

3. 6 were from longer compounds which clearly had an internal structure, e.g. 原発 genpatsu “nuclear power plant” is from 原子力 発電所, which is formed from 原子力 genshiryoku “nuclear power” plus 発電所 hatsudensho “power plant”, although the strict morpheme breakdown is 原子 + 力 + 発電 + 所;

4. 10 did not follow predictable patterns, e.g. 烏兎 uto “sun and moon” from 金鳥 玉兎 kin’ugyokuto and 共聴 kyōchō “community” from 共同視聴 kyōdōshichō.

The first two groups above, totalling 78%, consistently follow the pattern of taking the first kanji of successive morphemes, indicating this abbreviation approach based on morphemes has a reasonable chance of success.

### 7.6.2 Initial Testing

To test the hypothesis that constructing abbreviations using the underlying morpheme structure would be likely to produce valid terms, a file of terms with 3 or more kanji was extracted from JMDict. The terms were processed by the MeCab/UniDic morphological analyzer to reduce them to their constituent morphemes. For example 亜鉛中毒 aenchūdoku “zinc poisoning” was split into 亜鉛 (zinc) and 中毒 (poisoning).

Where the terms had 2 or 3 morphemes, possible abbreviations were constructed from the leading kanji of each morpheme (in the 3-morpheme cases both 2-kanji and
three-kanji versions were constructed). The synthetic abbreviations were checked against the n-gram Corpus, and extracted where the count was 100 or more, as terms with lower counts were unlikely to pass classification.

The first 300 terms in the file yielded 210 (157 unique) possible abbreviations, which were tested using the multi-class SVM described above. Of the 157, 19 passed the classification.

The 19 turned out to be:
1. 14 existing JMdict entries, practically none of which were abbreviations;
2. 4 proper names;
3. one possible unrecorded term: 壓差 generated from 壓力差 atsuryokusa “pressure differential”. This term was not found in any reference dictionaries but it can be found embedded in longer expressions in medical glossaries, e.g. 肺胞動脈ガス圧差 haihōdōmyakugasuatsusa “alveolar-arterial gas pressure difference”. It is just as likely to be a distinct compound formed from 壓 atsu “pressure” and 差 sa “difference, variation” as an abbreviation of 壓力差.

The detection of one new term from examining about 300 synthesized compounds may be reasonable, but it needs to be compared on a larger scale with alternative approaches. It is of some concern that many of the synthesized terms being classified are not abbreviations, as the point of this particular approach is to mimic the abbreviation process in the hope or expectation that it will be a productive method of discovering new or unrecorded terms. If the positive results are not actually abbreviations, there may be more suitable pathways to discovering them.
7.6.3 Extended testing

As the initial sample of 300 was really too small to reach any useful conclusion, the morpheme-based abbreviation synthesis was run over a larger batch of some 33,000 terms. Of these 7,900 were not already in Comblex, met the criterion of occurring at least 100 times in the \( n \)-grams, and were processed by the SVM classification system described above. The result was the classification of 162 of the compounds as potential new nouns (2.0%).

Hand-checking of 60 of these determined:

- 6 were abbreviations. Interestingly 3 were actually truncations of the original compound, e.g. the term 一步前進 ippozenshin “step forward” has the morphemes 一, 歩, 前進 so the term generated was 一步前 ippomae, which turns out to be a not-uncommon alternative;
- 15 were valid words, but not abbreviations. For example 一話 was generated from 一夕話 issekiwa “short story” but it is a different term meaning “episode (of a series, serial, etc.)”;
- the remaining 41 were unclear. They were not abbreviations, and many might not be valid terms. Many were just on-the-fly compound collocations such as 五年 go’nen “five years”, 三万 sanman “thirty thousand”, 主論 shuron “principal theory”, etc. Constructions using productive affixes such as 的 teki “-like”, 化 ka “-ification”, 用 yō “use, service”, etc. were quite common too.
7.6.4 Discussion

It must be concluded that this revisiting of the generation of abbreviations has not been particularly successful, except in a negative way. To find 6 abbreviations after processing about 33,000 source compounds certainly indicates it is not a highly productive technique. Taking into account all unrecorded terms found, 21 new terms is an improvement, but one must conclude there are probably better ways of searching for new unrecorded lexical items (which was much the same conclusion reached in the 2004 study).

That said, the possible terms coming through the SVM classification, when examined in context in WWW pages, are certainly behaving like real Japanese words or expressions. As a technique for achieving some precision in the analysis, the approach seems to work satisfactorily.

Had the approach for synthesizing abbreviations generated reasonable numbers of possible abbreviations, it was intended that a distributional similarity method based on Jensen–Shannon divergence (Fuglede and Topsoe 2004) be used on the feature vectors of the abbreviations and the matching source terms as a method for confirming that they were appearing in similar textual contexts. With the number of actual abbreviations detected being so low, this approach does not seem very appropriate.
7.7 Affixation

As mentioned above, affixation, often through the use of single-kanji prefixes and suffixes, is a very common and productive morphological process in Japanese. For example the noun 衛学 gengaku “pedantry” can take suffixes such as the personalizing suffix 者 sha to form 衛学者 gengakusha “pedant” or the adjective-forming suffix 的 to form 衛学的 gengakuteki “pedantic”.

There is the potential to synthesize extended compounds using these affixes and test for their presence in Japanese texts. However the productiveness of this morphological process is so high, and in general the meanings of the extended compounds are so predictable, that the resulting terms are usually lightly lexicalized. The reasons for considering the identification and possible lexicalization of such terms include:

a. idiomatic meanings which are not readily apparent from the components;

b. covering targets for English-Japanese dictionary entries;

c. special meanings in certain contexts, e.g. 挟抗 kikkō “rivalry, antagonism” where 挟抗的 kikkōteki is used in medical contexts for “antagonistic” and can be found in formations such as 挾抗薬 kikkōyaku “antagonist drug”.

The challenge is to determine whether this particular investigation could bring up anything useful. Certainly a straight extraction process is virtually guaranteed to identify large numbers of unlexicalized formations, however identifying those worthy of lexicalization, e.g. those with idiomatic or special meanings, is largely outside the scope of the present project, and is considered best left for future work.
7.8 Generalized Compound Creation - 2-kanji Compounds

7.8.1 Introduction

The development of the techniques described earlier allow for the rapid testing of large numbers of synthesized compounds against the $n$-gram Corpus, and are achieving high levels of precision and generally satisfactory levels of recall. While there has been some success with targeted compound synthesis in the form of potential abbreviation formation, there is also the possibility of simply generating kanji combinations using commonly occurring kanji, and using the classification system as a filter to detect combinations which are actually being used as terms.

While the full number of kanji in use in written Japanese is very large, only about 2,500 are in common use, leading to 6.25M possible 2-kanji compounds. While this is a relatively large number, it is a quite manageable processing task to:

a. create synthetic compounds simply by combining kanji;

b. filter out known compounds;

c. classify the remaining compounds using the techniques described above.

7.8.2 Initial Tests

Preliminary testing was carried out using the 2,500 most common kanji from the KANJIDIC dictionary (Breen 2017b) for which relative frequencies are available.
Initially the most common 100 *kanji* (日, 一, 国, etc.) were used to generate all 10,000 2-kanji combinations. Of these approximately 2,700 were in the in the reference lexicon, and of the remainder 97 were classified using the SVM model described above.

Inspection of the 97 classified compounds revealed that many involved numerics (二円 *ni’nen* “two yen, circles, etc.”, 一五 *jūgo* “fifteen”, etc.) and several involved very common compositional suffixes such as 氏 *shi* “surname”, e.g. 東氏 *higashishi* “Mr Higashi”. This resulted in a form of noise that tended to obscure some valid compounds in the set, e.g. 化調 *kachō* which is an unrecorded abbreviation of 化学調味料 *kagakuchōmiryō* “chemical seasoning”.

Sufficient valid results were detected that it was decided to investigate more thoroughly using less common *kanji* and excluding the major single-kanji affixes, which it was hoped would not produce as many obvious compositional compounds.

### 7.8.3 Extended Investigations

For a more thorough investigation of synthesized 2-kanji compounds a number of extensions were made to the experimentation:

a. both the Google and Kyoto *n*-gram corpora were used. This provided the opportunity to evaluate whether the latter, with significantly lower *n*-gram counts which had not been truncated, had any advantage over the former, or whether the much larger corpus would prevail.

b. as well as the 395-item *JMdict*-based training data, training models were also

\[2\text{ It was interesting to note that some not-uncommon typographical errors: 時間 (for 時間 *jikan* “time”) and 問題 (for 問題 *mondai* “problem”) were identified and classified.}\]
built using synthesized compounds, as described below.

c. the possible use of heuristics based on numbers of features was considered.

For the 395-item training data the Google $n$-grams version was unchanged from the earlier investigation. For the Kyoto $n$-grams version the features needed to be regenerated. The heuristics that were used for the classification could have been revisited, but it was decided to leave the classification in place as it would enable a more comparable evaluation.

For the training data based on synthesized compounds, six models were built for each corpus as follows:

a. three sets of training compounds were generated from the list of the 2,500 most common kanji as follows:
   
   - small model: 10,000 compounds using two 100-kanji ranges 800-899 and 1,100-1,199;
   - medium model: 40,000 compounds using two 200-kanji ranges 800-999 and 1,100-1,299;
   - large model: 90,000 compounds using two 300-kanji ranges 800-999 and 1,100-1,399;

b. these compounds were then filtered against the respective $n$-gram corpora, and only accepted if they had a non-zero count (20 or more for Google, 1 or more for Kyoto) and a minimum number of encapsulation features. Two sets were generated; one with at least 1 encapsulation feature, and another with at least 5 encapsulation features.
Chapter 7: Neologism Synthesis

Table 7.4: Training Data: Total Numbers of Compounds and Classifications

c. the sets of compounds were then checked against the reference lexicon to determine if they were known lexical items, and this was used as the basis for in-out classification.

The numbers of compounds in the resulting training files, and the numbers classified are given in Table 7.4.

Two trials were run and evaluated using the training models described above. Each trial was of 40,000 synthesized compounds; the first generated from the 1,600-1,799 and 1,400-1,599 ranges, and the second from the 400-599 and 1,600-1,799 ranges. The same count and feature number criteria were applied as for the training data. As with the training compounds, these were matched against the n-gram corpora and only accepted if they had positive counts and the required number of features.

For the first trial four sets of test data were generated:

a. Google n-grams, 1 or more features. 767 compounds accepted, 295 in the lexicon.

b. Google n-grams, 5 or more features. 202 compounds accepted, 135 in the lexicon.

<table>
<thead>
<tr>
<th>Model</th>
<th>Google</th>
<th>Kyoto</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥ 1 Feature</td>
<td>≥ 5 Features</td>
</tr>
<tr>
<td>Small</td>
<td>536</td>
<td>193</td>
</tr>
<tr>
<td>Medium</td>
<td>1792</td>
<td>818</td>
</tr>
<tr>
<td>Large</td>
<td>3819</td>
<td>1543</td>
</tr>
</tbody>
</table>
c. Kyoto $n$-grams, 1 or more features. 1,341 compounds accepted, 338 in the lexicon.

d. Kyoto $n$-grams, 5 or more features. 319 compounds accepted, 175 in the lexicon.

The greater number of compounds passing the thresholds with the Kyoto $n$-gram Corpus can be explained by the longer tail due to the lower cutoff. The increased numbers of existing lexical items for the Kyoto Corpus and the low feature counts is interesting as it indicates a reasonable number of real terms with measured usage can be found right down in the noise floor.

The four sets of compounds were then classified by the ML models; both the “original” model and the generated models. The results are in Table 7.5 through Table 7.8. Table 7.9 contains some examples of synthesized compounds from the “Kyoto $n$-grams, 5 or more features” batch and their classifications.

The precision and recall figures shown are crude, and are simply based on the assumption that the purpose of the classification was simply to identify known lexical items.

A number of comments can be made about the results of this trial.

a. of some concern is the outcome of three of the tests using the $\geq 5$ feature threshold, which resulted in all the compounds being classified. This raises something of a question-mark over the use of an SVM for classification as it seems it can be confused or overwhelmed in some cases.

b. on one level the trial can be regarded as a test of how well the models identify
established lexical items among the synthesized compounds. In general the "original" hand-classified training data is resulting in reasonably good precision but a comparatively low recall, whereas the generated training data generally shows a lower precision and a higher recall. In fact there is nothing particularly surprising about this as the hand-classified training data came from known lexical items and could be expected to set a higher barrier.

c. it is difficult to see a significant difference between the results from the two corpora;
Table 7.7: Kyoto Corpus (≥ 1 features)

<table>
<thead>
<tr>
<th>Model</th>
<th>In lexicon</th>
<th>Not in lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class</td>
<td>Not Class</td>
</tr>
<tr>
<td>Small</td>
<td>109</td>
<td>229</td>
</tr>
<tr>
<td>Medium</td>
<td>80</td>
<td>258</td>
</tr>
<tr>
<td>Large</td>
<td>81</td>
<td>257</td>
</tr>
<tr>
<td>Original</td>
<td>52</td>
<td>286</td>
</tr>
</tbody>
</table>

Table 7.8: Kyoto Corpus (≥ 5 features)

<table>
<thead>
<tr>
<th>Model</th>
<th>In lexicon</th>
<th>Not in lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class</td>
<td>Not Class</td>
</tr>
<tr>
<td>Small</td>
<td>175</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td>79</td>
<td>96</td>
</tr>
<tr>
<td>Large</td>
<td>93</td>
<td>82</td>
</tr>
<tr>
<td>Original</td>
<td>52</td>
<td>123</td>
</tr>
</tbody>
</table>

d. once the anomalous results are excluded the results for the two cut-off feature counts are similar, despite the “≥ 1” having more candidates initially.

It is inappropriate, however, to focus too much on the effectiveness of the models in recovering known lexical items as it loses sight of the actual goal, which is to identify lexical items which are not in references. To determine the effectiveness and usefulness of the techniques being tested we must concentrate on the Classified/Not-in-Lexicon columns, and examine the compounds therein to determine the degree to which they are valid terms.
Table 7.9: Kyoto Corpus (≥ 5 features) examples

Before proceeding to analyze these terms, it is interesting to consider briefly some other statistics associated with the classified and non-classified compounds, in particular the total n-gram counts for the compounds and the number of features. In Table 7.10 the average number of occurrences of the compounds and the average number of features have been added to the data from Table 7.5.

As can be seen, there is a very significant difference between the number of features associated with the compounds in the classified and non-classified columns. It is a reasonable assumption that it is the number of features that mainly contribute to the classifications.

Repeating this exercise using a ≥ 5 feature threshold and as classified using the Kyoto Corpus (Table 7.11), we see a very similar pattern. (The results from the Small model have been excluded as that model classified all compounds.)

To probe this issue a little further, an analysis was carried out to see what classifi-
Table 7.12 shows the results for various feature-count thresholds using the Google Corpus, \( \geq 1 \) feature data.

As can be seen in Table 7.12, the number of classifications for thresholds of 7 and 8 is comparable with those for some of the SVM models. This strongly suggests that there may be relatively simple heuristic techniques which are as effective as using a machine-learning classification system.

The real question, of course, is whether these techniques are actually identifying unrecorded lexical items. To evaluate that properly, we need to look in detail at the compounds which have been classified but do not occur in the reference lexicon, i.e. those that are in the Classified/Not-in-Lexicon columns.

To investigate this, the compounds from the Google n-grams (\( \geq 1 \) feature) set which were not in the reference lexicon, but which had been classified either by
one of the models or by the $\geq 7$ feature heuristic, were collated giving a total of
38 compounds. These compounds were then investigated in detail using reference
material and analyzing the text in WWW pages in which they occurred. A detailed
analysis of each compound, including the count, classifier(s), reading and meaning (if
any) is in Appendix C.2.1.

The 38 compounds turned out to consist of 11 nouns, 23 named-entities and 4
non-terms (apparently resulting from mis-parses in the $n$-gram generation).

The kanji ranges used happened to include \textit{kanji} such as 灵 (spirit), 魔 (demon),
姫 (princess), 鬼 (ogre), 龍 (dragon), 雷 (lightning), etc. thus virtually guaranteeing
the generation of a number of manga, anime, game, etc. character names. Since
the investigation is not attempting to make value judgements about the compounds,
but needs to assess them as to whether they are valid terms in use which may be
incorporated in some lexicon (even one of manga characters), it has to be concluded
that 34 of the 38 are valid, and hence the overall precision achieved was 89.5%. The
precision achieved by the individual models is shown in Table 7.13.

<table>
<thead>
<tr>
<th>Model</th>
<th>In lexicon</th>
<th>Not in lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class</td>
<td>Not Class</td>
</tr>
<tr>
<td>Medium</td>
<td>79</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>4182/17.43</td>
<td>778/9.43</td>
</tr>
<tr>
<td>Large</td>
<td>93</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>1890/16.90</td>
<td>2797/8.66</td>
</tr>
<tr>
<td>Original</td>
<td>52</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>7431/22.63</td>
<td>152/8.98</td>
</tr>
</tbody>
</table>

Table 7.11: Kyoto Corpus ($\geq 5$ features) with Average Compound and Feature Counts
Table 7.12: Google Corpus (≥ 1 features) with Compound Counts and Average Feature

<table>
<thead>
<tr>
<th>Feature</th>
<th>In lexicon</th>
<th>Not in lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Threshold</td>
<td>Class</td>
</tr>
<tr>
<td>14</td>
<td>29 : 23.83</td>
<td>266 : 4.24</td>
</tr>
<tr>
<td>12</td>
<td>41 : 20.54</td>
<td>254 : 3.85</td>
</tr>
<tr>
<td>10</td>
<td>49 : 18.86</td>
<td>246 : 3.64</td>
</tr>
<tr>
<td>8</td>
<td>69 : 15.83</td>
<td>226 : 3.22</td>
</tr>
<tr>
<td>7</td>
<td>79 : 14.71</td>
<td>216 : 3.04</td>
</tr>
</tbody>
</table>

Measurement of recall will take more investigation. Since there were at least 34 valid compounds on offer, the upper bound on the recall ranges from 70.5% for the heuristics technique down to 32.3% for the “original” SVM model.

Of the three SVM models based on generated training data, the Small model did not classify any compounds which were not also classified by another SVM model. There may be a case for excluding this size of model. While no individual model dominated, it is noted that the SVM models classified 12 compounds with feature numbers below that set as the threshold for the heuristics system (7). Of the 12, 9 were valid terms, and most of these had 5 or 6 features. Two of the invalid classifications had only one feature. This may indicate that there is scope for lowering the threshold for the heuristics and achieving a similar outcome using heuristics alone.

To explore the impact of lowering the threshold for the heuristic classification, the compounds with 5 and 6 features were extracted. There were 38 such compounds of
which 8 had already been classified by the SVM models. A sample of 10 of the 30 was evaluated and revealed there were 3 nouns, 6 named-entities and 1 non-term, which are similar proportions to the previous analysis (the details of these are in Appendix C.2.2). This indicates that it may be possible to operate the heuristic technique successfully at a lower threshold, which would eliminate most of the advantage of the SVM models and give an overall better result.

Lowering the threshold to include feature counts of 3 and 4 classified a further 77 compounds. Again a sample of 10 was evaluated, resulting in 5 named-entities and 5 non-terms (see Appendix C.2.3). This indicates that, as could be expected, the results are noisier the lower the threshold.

The analysis carried out above was done using the Google Corpus with an acceptance threshold of 1 feature, which resulted in a collection of 767 compounds of which 295 were in the reference lexicon. The equivalent set using the Kyoto Corpus resulted in 1,341 compounds of which 338 were in the reference lexicon. Repeating the classification processes resulted in 99 compounds being classified by one or more
Chapter 7: Neologism Synthesis

of the SVM models or having $\geq 7$ features or both. 32 of the 99 were among the 38 compounds identified using the Google Corpus. A sample of 17 of the remaining 67 compounds was analyzed, indicating that 9 were nouns, 7 were named-entities, and one was not a valid term. Overall, there did not seem to be much difference between the two sets. There were more compounds generated using Kyoto $n$-grams, but the quality was much the same.

In order to verify that the results in the investigation reported above were not influenced by the choice of $kanji$ range, part of the analysis was repeated using the ranges 400-599 for the first $kanji$ and 1600-1799 for the second $kanji$. This time the analysis focused on unlexicalized compounds which were classified by one or more of the models or had $\geq 5$ features, and was carried out using both corpora.

For the set analyzed using the Google Corpus, 220 compounds were classified. A sample of 23 were evaluated (see Appendix C.2.4), of which 9 were nouns, 11 were named-entities and 2 were non-terms.

For the Kyoto $n$-grams, 353 compounds were classified. A sample of 35 was evaluated, and the assessment was that 10 were nouns, 19 were named-entities and 6 were not valid terms (see Appendix C.2.5).

As the analysis was done on samples, it is difficult to make a comparison between the sets from the two corpora. Based on the samples we could estimate that the precision is 91% with the Google $n$-grams and 83% with the Kyoto $n$-grams, but it is not really valid to make such claims based on relatively small samples. It was noted that 164 compounds were common to both sets, and as many of the valid sampled compounds in the Kyoto set lay outside the compounds common to both, it is quite
likely that the results using the Kyoto Corpus deliver a higher level of recall.

Intuitively one would expect that candidate compounds with higher counts and/or a greater number of features to be more likely to be valid terms, and hence would display a higher precision. To test this compounds which had higher values of those metrics were sampled and examined. The parameters used were:

- Google $n$-grams: counts $\geq 20,000$ (14 compounds), features $\geq 10$ (30 compounds)
- Kyoto $n$-grams: counts $\geq 500$ (19 compounds), features $\geq 15$ (33 compounds)

These parameters were chosen fairly arbitrarily to result in approximately the same proportion of compounds being selected from each set.

There was a degree of overlap, as one would expect. 6 compounds in the Google $n$-gram set were in both samples and 12 in the Kyoto $n$-gram set were in both samples.

The evaluation of the samples is in Table 7.14.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Nouns</th>
<th>Names</th>
<th>Other</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Counts</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>79</td>
</tr>
<tr>
<td>High Features</td>
<td>11</td>
<td>15</td>
<td>4</td>
<td>87</td>
</tr>
<tr>
<td>Kyoto</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Counts</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>74</td>
</tr>
<tr>
<td>High Features</td>
<td>17</td>
<td>13</td>
<td>3</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 7.14: Sample Results for High Counts and Features

Again, it is difficult to draw firm conclusions from relatively small samples, but it appears that having a high feature count may be a better indicator of a candidate compound being an actual term than having a higher $n$-gram count. That said, there is not a lot of difference between the results from these samples and the simple
samples reported above. All are showing quite reasonable levels of precision.

Table 7.15 show some examples of valid unlexicalized 2-kanji compounds detected during the project.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>桃豆 momomame (?)</td>
<td>bean-based sweet rolled in peach powder</td>
</tr>
<tr>
<td>低弦 teigen</td>
<td>low-pitched string instruments (abbr)</td>
</tr>
<tr>
<td>周堤 shūtei</td>
<td>circular bank, e.g. round Jōmon-period grave sites</td>
</tr>
<tr>
<td>整膚 seifu</td>
<td>alternative therapy involving pinching and pulling the skin</td>
</tr>
<tr>
<td>移弦 igen</td>
<td>string-crossing (violin, etc. technique)</td>
</tr>
<tr>
<td>紙筒 kamidzutsu</td>
<td>paper tube (var. of 紙管 shikan)</td>
</tr>
<tr>
<td>製縫 seihō</td>
<td>suffix to business names meaning clothes-making</td>
</tr>
<tr>
<td>賞暦 shōreki</td>
<td>awards, list of years prizes were won</td>
</tr>
<tr>
<td>丸刀 gantō/marutō</td>
<td>gouge with U-shaped blade</td>
</tr>
<tr>
<td>士紳 shinshi</td>
<td>archaic term for ranking official</td>
</tr>
<tr>
<td>帰寮 kiryou</td>
<td>returning to one’s accommodation, etc.</td>
</tr>
<tr>
<td>春苗 shunbyō</td>
<td>spring seedlings (also a girl’s name)</td>
</tr>
<tr>
<td>母珠 moshu</td>
<td>large bead(s) in a Buddhist rosary</td>
</tr>
<tr>
<td>白粒 shirochibu</td>
<td>type of ceramic used in Kitaniware, etc.</td>
</tr>
<tr>
<td>紙函 kamibako</td>
<td>paper box</td>
</tr>
<tr>
<td>親鴨 oyagamo</td>
<td>can mean “parent duck”, but mostly used in 親鴨会, a society of former IBM staff</td>
</tr>
</tbody>
</table>

Table 7.15: Examples of Valid 2-kanji Compounds
7.8.4 Discussion of 2-kanji Synthesis and Evaluation

The approach investigated for synthesizing and evaluating 2-kanji compounds, namely to filter the candidates on n-gram counts and encapsulating/trailing particle syntactic contexts appears to be quite successful in identifying compounds which, if they are not in already in a lexicon, are worth further examination. The proportion of synthesized compounds passing the threshold selection criteria is about 1%.

The use of a machine learning approach in which an SVM is trained on typical compound counts and encapsulation features has produced some interesting results. In summary:

a. in general the precision is reasonably high, and in terms of indicating compounds worth further investigation is probably quite satisfactory;

b. it is noted, however, that much the same result can be achieved by employing relatively simple heuristics based on the n-gram counts of the compounds themselves and the possible features. This casts considerable doubt over whether a machine learning approach is appropriate in this type of classification problem;

c. it appears that in general the number of encapsulation features, (that is, different contexts) associated with a synthesized compound is the best indication that it is likely to be a valid term.
Chapter 7: Neologism Synthesis

7.9 Generalized Compound Creation - 4-kanji Compounds

7.9.1 Introduction

The testing approach described above can be applied to any compounds; not just 2-kanji ones. In this section an investigation into generating 4-kanji compounds from known 2-kanji compounds and their subsequent evaluation is described.

The work described in this section was carried out prior to the completion of the investigation of synthesized 2-kanji compounds described earlier. At the time the Kyoto n-gram Corpus was not available, and the Google Corpus alone was used. Some of the approaches developed in the 2-kanji compound investigation could be applied here too, and the investigation could be revisited on that basis, however it was decided that it would be best left as a future project.

7.9.2 Compound Synthesis

The (then) current JMdict contained approximately 48,500 2-kanji compounds. As there are $2 \times 10^9$ possible 4-kanji combinations of these, there are some logistical and processing challenges associated with testing all of them. (A text file of the combinations would take over 30Gb.)

As the actual number of 4-kanji combinations used in Japanese, even quite rare ones, is in the order of $10^5$, which is a very small fraction of the total number of possible combinations, it is appropriate to combine the compound generation with the creation of the feature sets for the combinations, and to take advantage of the
ability to filter the combinations against the $n$-gram Corpus. In any case in this study we will only be using sample sets of the possible combinations.

The heuristics used in the compound generation are:

a. on generating a potential compound, immediately check its frequency in the $n$-grams, and if it is below a threshold ignore it. (This will speed up the generation/analysis process, as the majority of synthesized compounds have either zero or very small counts in the $n$-grams, and it means there is usually only one check of the $n$-grams file per compound instead of around 40 for the full feature set. The size of the threshold is discussed below.)

b. for the potential compounds that pass step (a) above, only pass them to the classification stage if they have 5 or more encapsulation features. This is based on the observation from the work reported earlier that nothing is being classified with fewer than 5 features, and thus it will reduce the amount of time and processing spent carrying out the classification.

In considering the value of the threshold in (a) above, it was initially set to 1,000, which may be on the high side, but it was observed that nothing was being classified by the SVM model with a total count less than this, probably because the features are few and have low counts. In fact lower values would probably not reduce the classification processing time significantly, as it was observed when running a test using 10,000 synthesized compounds that only a relatively small number had counts at all. For thresholds of 20 (the lowest possible with the Google Corpus), 100 and 1,000 the number of compounds exceeding them were: 20: 31 compounds, 100: 15
Chapter 7: Neologism Synthesis

compounds, 1,000: 5 compounds. In all cases only 4 compounds achieved 5 or more features.

7.9.3 Initial Investigation

A series of tests were carried out using combinations generated from samples of 1,000 2-kanji compounds, i.e. 1,000,000 compounds were synthesized in each test.

a. the first test used compounds 1001-2000 in the source collection (一廓 to 曳網), with the count threshold value set at 1,000.

For the first 70,000 compounds only 24 passed step (a) and of them only 15 met the criterion of 5 or more features. A selection of these are given in Table 7.16.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>Features</th>
<th>Lexicon</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>一軒一軒 ikken’ikken</td>
<td>48601</td>
<td>22</td>
<td>Y</td>
<td>house-to-house</td>
</tr>
<tr>
<td>一言一句 ichigon’ikku</td>
<td>29723</td>
<td>15</td>
<td>Y</td>
<td>word-by-word</td>
</tr>
<tr>
<td>一向一揆 ikkōikki</td>
<td>30691</td>
<td>27</td>
<td>Y</td>
<td>Jodo Shinshu Buddhist uprising</td>
</tr>
<tr>
<td>一個一個 ikkoikko</td>
<td>72951</td>
<td>24</td>
<td>N</td>
<td>one-by-one</td>
</tr>
<tr>
<td>一曲一曲 ikkyokaikkyoku</td>
<td>49027</td>
<td>23</td>
<td>N</td>
<td>piece-by-piece (in music)</td>
</tr>
</tbody>
</table>

Table 7.16: Sample Synthesized Compounds with n-gram and Feature Counts

Overall this batch of 1 million synthesized compounds saw 310 with n-gram counts of over the 1,000 threshold and of these 181 with 5 or more features. The 181 compounds were then tested using the SVM model trained on a modified version of the 395-compound “basic model” training file discussed above. The modifications consisted of reducing the file to contain only items with counts
In Lexicon | Not In Lexicon
---|---
Classified | 16 | 64
Not Classified | 7 | 94

Table 7.17: Classification and Lexicon Checks for Synthesized Compounds

> 100 and ≥ 5 features in order to be consistent with the items being tested. This resulted in a 325-item file with an in/out ratio of 268/57. (Although this training data was assembled using 2-kanji compounds, and on a somewhat different basis, it was considered adequate for these in initial tests.) Of the 181 compound 80 were classified as possible terms and 23 were already in the reference lexicon. These results are summarized in Table 7.17.

b. the next test was essentially a repeat of the one above using a threshold n-gram count of 100. 1,267 of the generated compounds exceeded the threshold, of which 210 (29 more) had 5 or more features. The same 80 as in the previous test were classified by the SVM model, however of the 130 not classified two more were in the reference lexicon:

一言一行 ichigen‘ikkō “every word and act” - count 533, 9 features
一齣一齣 hitokomahitokoma “frame by frame” - count 258, 8 features

As there was a negligible processing time difference between the runs with different threshold values, it was decided to continue to use the lower value, as it provided the opportunity to test unclassified compounds in the reference lexicon.

3Both of these are examples of proverb-like 4-kanji idiomatic compounds, known as 四字熟語 yojijukugo, which are popular in Chinese and Japanese.
Table 7.18: Classification and Lexicon Checks for Synthesized Compounds

<table>
<thead>
<tr>
<th></th>
<th>In Lexicon</th>
<th>Not In Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified</td>
<td>16</td>
<td>64</td>
</tr>
<tr>
<td>Not Classified</td>
<td>9</td>
<td>123</td>
</tr>
</tbody>
</table>

The results are summarized in Table 7.18.

The next task was to determine the precision that had been achieved by examining the 64 out-of-lexicon compounds that had been classified. Some, such as 一曲一曲 and 一個一個 mentioned above, are apparently valid terms. This process involves some quite painstaking and time-consuming analysis of texts, e.g. WWW pages, containing the compounds in typical usage.

With hindsight, working on a sample that contained 357 2-kanji compounds beginning with 一 ichi, hitotsu “one” was somewhat unfortunate. All but 7 of the classified compounds began with 一, and many contain pairs of the 2-kanji compounds, e.g. 一戦一戦 issen’issen “day-to-day, game by game”, etc.”. It appears there are many non-idiomatic expressions of the 一 X 一 X and 一 X 一 Y variety with meanings such as “X occurring one after another” or “one X per Y” which are not necessarily lexicalized. As this distorted the results estimation of the precision in this sample was not completed.

c. a further test was carried out using 1,000 2-kanji compounds in the 2001-3000 range, again using 100 as the threshold for further analysis. This time only 406
Table 7.19: Classification and Lexicon Checks for Synthesized Compounds

<table>
<thead>
<tr>
<th>In Lexicon</th>
<th>Not In Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified</td>
<td>3</td>
</tr>
<tr>
<td>Not Classified</td>
<td>1</td>
</tr>
</tbody>
</table>

exceeded the threshold, 37 had 5 or more features, and of these 10 were classified. Of the 10, 3 were in the reference lexicon, as was one of the unclassified compounds. See Table 7.19.

Examination of the classified out-of-lexicon compounds determined that several are real terms:

英国王室 *eikokuōshitsu* - “British royal family"
液晶演出 *eishōenshitsu* - name of a video/pachinko game
欧州遠征 *ōshūensei* - “European campaign” (esp. with sporting teams)
横浜駅前 *yokohamaekimaе* - a typical address, in this case at the front of Yokohama Station

The one non-classified compound which was in the lexicon was 王子駅前 *ōjieki-mae*, which is also an address.

d. a further test was carried out using compounds in the 3001-4000 range. 78 compounds were passed to the classification stage, of which 16 were classified (Table 7.20). In checking against the reference lexicon, 6 of the classified compounds matched, as did 7 of the unclassified compounds, which is an indication of a fairly low recall.

Many of the out-of-lexicon classified compounds are genuine, e.g. 下限価格
Table 7.20: Classification and Lexicon Checks for Synthesized Compounds

<table>
<thead>
<tr>
<th></th>
<th>In Lexicon</th>
<th>Not In Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classified</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Not Classified</td>
<td>7</td>
<td>55</td>
</tr>
</tbody>
</table>

*kagenkagaku* “price floor”,  化粧下地 *kenshōshitaji* “foundation base”, etc. Most of the in-lexicon but unclassified compounds were rather obscure, e.g. 窪価格 *kasenkakaku* “oligopolistic price” has a low count and barely made the list to be tested.

The results of this initial exploration are quite positive, and it is apparent that the approach is reasonably effective in trawling for terms that are actually in use. It was decided that a more thorough investigation was justified, and in particular, if possible, a more consistent set of training data for the model be developed.

### 7.9.4 Development of Training Data

It will be noted that in the testing described above, the synthesized compounds which met the $n$-gram count and feature-number criteria were divided into about 10% which were in the reference lexicon and 90% which were not. Although some of the 90% turned out to be valid lexical items, a split based on presence/absence in the reference lexicon could a viable method for automatically compiling sets of training data for the classification model. In any case it would be useful to compare it with the “basic model” used above, which had been classified according to a simple heuristic based on numbers of features.
A training set was compiled by generating approximately 200 million 4-kanji compounds. Of these 1,627 passed the filters (> 100 count in Google n-grams, ≥ 5 features). Of these 172 matched with the reference lexicon (7 had n-gram counts below 1,000, with the lowest at 524). As expected, this had the approximate 1:9 classification ratio.

On checking some of the non-matched but high-scoring compounds with other reference material, in particular the Extended Lexicon (see Section 4.2), it was noticed that many were in this compilation. While this material is not in the main reference lexicon, largely because it is poorly structured and lacks readings, it does contain valid Japanese terms and is probably valid for such purposes as classifying training data and assessing the recall of results.

Accordingly the 1,455 non-matched compounds were checked against the Extended Lexicon. For the purposes of the check, the Extended Lexicon was segmented into two parts:

a. the text of the sub-entries and examples from the *Kenkyūsha’s New Japanese-English Dictionary* (Toshiro et al. 2003) (the 128,000 headwords and readings from this are in the reference lexicon, but not the sub-entries, examples, etc.). 114 of the 1,455 compounds were matched. Most were exact matches, with a few matching with strings of more than 4 kanji;

b. a large text collection consisting of the *Eijiro* lexicon, a number of industry and subject-specific glossaries, and the files of term/reading pairs from several word-processing conversion files. 385 of the 1,455 compounds were matched. 86 of the compounds matched in both collections.
Obviously if “known terms” are to be used as a discriminator for setting up the training data, account needs to taken of some or all of these additional matches. Accordingly it was decided to test out three alternative sets of training data against a batch of test data as follows:

a. Model A. Only the reference lexicon matches flagged as “in”, i.e. 172 in and 1455 out.

b. Model B. The reference lexicon matches and the overlapping Kenkyūsha text collection matches classified. 258 in and 1369 out.

c. Model C. All matches classified, i.e. 585 in and 1042 out.

7.9.5 Testing

For test data a batch of 1M compounds from a different range was generated. 128 compounds passed the filters. Of these 128:

i. 15 matched the reference lexicon;

ii. 3 of the remaining 113 matched the Kenkyūsha text;

iii. 27 of the 113 matched the other collection, and overlapped with the 3 in the Kenkyūsha text.

This provided an initial basis against which to assess recall.

The test data was classified using the three training models above, and also for comparison with the earlier basic model. The results are in Table 7.21
Chapter 7: Neologism Synthesis

Table 7.21: Multi-model Classification and Lexicon Checks

<table>
<thead>
<tr>
<th>Model</th>
<th>In Lexicon</th>
<th>Not In Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model A</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Model B</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Model C</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Basic Model</td>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>66</td>
</tr>
</tbody>
</table>

The 5 out-of-lexicon classifications from Models B and C were all confirmed as valid terms, and an inspection of the 20 out-of-lexicon classification from the Basic Model indicated they too were almost all valid.

As can be seen from Table 7.21 the precision stays high as we move from A to C, and recall gradually improves albeit to a not very high level. The interesting result is the heuristic-based Basic Model, which has performed significantly better, particularly on recall.

To probe the application of heuristic-based classification of the training data, the same criteria were applied to the 1,627-item training data file, resulting in an in/out ratio of 362/1265, i.e. similar proportions to the other models which have been used (Model D).

And finally a Model E was built by taking the Model C file (all reference terms classified) and additionally classifying all the terms which met the frequency and
feature heuristics. This file has an in/out ratio of 756/871. The results from these two models are Table 7.22.

### 7.9.6 Discussion of 4-kanji Synthesis and Evaluation

The experimental results reported above are quite sobering. The best efforts at using live data to train the SVM have ended up with models which at best deliver about half the performance of models based on quite simplistic heuristics. It is only by tweaking the models using basic heuristics that results comparable with those of models using the heuristics alone to mark the training data are achieved.

Inspection of the training data as tagged according to the presence of the compounds in the various dictionary and glossary resources, reveals a possible reason for the relatively poor performance of these models. The features associated with the tagged items vary wildly. While many are on high-frequency widely-used compounds, a significant number are on terms with low usage. For example 調停努力 chōteidoryoku “mediation effort” was tagged because it is in a glossary, but it only has a few features and a low count. There are many like this; some from the glossaries, but also

<table>
<thead>
<tr>
<th>Model</th>
<th>In Lexicon</th>
<th>Not In Lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model D</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Not Classified</td>
<td>26</td>
<td>74</td>
</tr>
<tr>
<td>Model E</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td>Not Classified</td>
<td>22</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 7.22: Multi-model Classification and Lexicon Checks
quite a lot from the Japanese-Japanese dictionaries. One can envisage the SVM having a lot of trouble establishing a sound boundary based on this type of model. The models using the heuristics alone are being much more consistent in stating what is in and out, and that seems to help the SVM identify many compounds with relatively high counts and features.

Indeed simple application of the classifying heuristics directly to the test data results in 41 compounds being selected. Most of these are the same as those classified by SVM using the heuristic-based models.

The hypothesis being tested in developing and testing these models based on known words is the expectation/assumption/hope that they have a discernible syntactical environment in which they are used, and this can be detected by the nature and distribution of the features, the model(s) trained to it, and then be used for identifying terms with a similar syntactical environment. If training models constructed using heuristics are achieving the best results, and the application of those heuristics directly to the test data is achieving comparable outcomes, it does raise the question why we are bothering to train a model at all.

While on the subject of heuristics, it is worthwhile taking stock of what this investigation is about. Potential compounds are being synthesized in the expectation that some can be identified which are actually being used as terms in Japanese, but are not yet in dictionaries. For example in the training data generation described above, 100 million compounds were generated, of which only about 15,000 (0.015%) exceeded the (arbitrary) threshold of $n$-gram counts $> 100$. Of those only about 1,500 passed a second criterion of having 5 or more features, i.e. appear in appropriate syntactic
contexts.

When investigated in detail, e.g. examining text passages where they occur, virtually all have some degree of validity. This is not that surprising when one considers the process of forming these sorts of extended compounds in Japanese - one has considerable freedom and flexibility. It is not unreasonable to conclude that almost of the 4-kanji compounds passing the (heuristic-driven) filters are actual compounds being used.

Quite possibly this part of the study needs to be viewed from the position of discovering 4-kanji compounds that are in common enough use to be considered for lexicalization. That consideration needs to take into account issues which are really outside the SVM-based classification which has been tested here, as it involves such questions as:

a. whether the compound has an idiomatic meaning (e.g. 援助交際 enjōkōsai, where the component compounds mean “assistance” and “company”, but the full compound is a common euphemism for teenage prostitution.)

b. whether it is a useful inclusion reasons such as for English-to-Japanese reverse searches.

Certainly the process has exposed some very common compounds which are not in reference materials, e.g.:

- 商品詳細 shōhinshōsai - n-grams: 5,318,613 - “product details, etc.”
- 商品状態 shōhinjōtai - n-grams: 215,476 - “product/stock condition/status”
- 初期衝動 shokishōdō - n-grams: 19,385 - “initial impetus”
Overall the investigation has demonstrated a valid process for determining unlexicalized common 4-kanji compounds which can be examined further. It has also demonstrated that attempting to classify them into useful or not-useful by machine learning techniques is not particularly productive when compared with the application of relatively simple heuristics.

7.10 Summary, Discussion and Conclusions

In this chapter a major investigation into the possibility of identifying and extracting unlexicalized Japanese kanji compounds has been described. The work consisted of mimicking typical morphological processes in Japanese to produce large sets of possible compounds, then using NLP techniques with large corpora to determine whether the synthesized compounds were in fact being used.

Two type of compounds were considered:

a. 2-kanji compounds formed by such techniques as affixation of common single-kanji prefixes and suffixes to single-kanji terms and abbreviation of longer multimorpheme compounds, as well as consideration of a “brute force” approach of simply generating all combinations of possible compounds;

b. 4-kanji compounds formed by the collocation of known 2-kanji compounds.

In the case of affixations the initial investigation revealed that while many unlexicalized cases were apparent, it was also clear that the process is highly productive and the resulting formations are generally quite non-idiomatic and hence not candidates for inclusion in a typical dictionary. Investigation of possible techniques for determining cases with idiomatic meanings has been left for future work.
The investigation of artificial abbreviations revealed that the process only achieved a very low yield of valid terms, and in fact many were not actually abbreviations of the source compounds. It was concluded that this approach was not worth pursuing further.

The investigation of the bulk generation of 2-kanji compounds and their subsequent evaluation indicated that the approach appears to be quite successful in identifying compounds which are worth further examination.

The evaluation techniques applied to synthesized compounds, which involved basic \( n \)-gram counts and features based on encapsulating particles, explored the possible use of SVM models to classify the candidates. While the outcome of this was generally satisfactory it was concluded that it had little or no advantage over applying some relatively simple heuristics to the same problem.

**7.10.1 Future Work**

As mentioned above the techniques investigated here have been quite successful in identifying compounds which are worth further examination, as shown in Table 7.15. Several compounds identified here have been incorporated into the JMdict dictionary. The investigation has of necessity only considered a small proportion of the possible synthesized compounds. There remains the lexicographic task of exploring the rest of the compounds which pass the evaluation criteria established here with a view to adding valid unrecorded terms to the lexicon.

In addition, the development of encapsulation patterns of the はXXの, がXXX Xを, etc. variety (see Table 7.3), which have been used here as evaluation features,
opens the possibility that they could be deployed as filters on bodies of text to detect unlexicalized compounds which occur in sufficient quantities to warrant further investigation. In fact $n$-gram corpora are ideally suited for the initial part of such an investigation.
Chapter 8

Generation and Extraction of Compound Verbs

8.1 Introduction

This chapter describes work conducted to generate and extract Japanese compound verbs (JCVs) from corpora and corpus-based resources. The work lies within the synthesis approach to Japanese neologism extraction although the techniques developed and evaluated apply equally to well-recognized and lexicalized JCVs as well unrecorded and new verbs.¹

Compound verbs in Japanese (described in some detail below) have attracted considerable attention in Japanese linguistics as they are a highly productive and flexible element of the language (Shibatani 1990; Baldwin and Bond 2002; Tsujimura

Apart from some manually-prepared verb lists they have received relatively little attention in corpus linguistics.

The reasons for collecting and studying Japanese compound verbs include:

a. the development of reliable methods for extraction of the verbs from corpora;

b. investigation of the distribution of the verbs and their constituents;

c. investigation of the coverage of the verbs in the major lexicons.

In particular, it is hoped that by isolating JCVs which are in use, but are not currently recorded or lexicalized, and eventually by developing and verifying Japanese meanings and English translational equivalents for these verbs, the lexicon of JCVs can be expanded.

As part of this work, two methods for extracting compound verbs have been developed and applied over a major Japanese corpus. The result has been the identification of a large number of potential JCVs, which we show to have a high level of precision, relative to a sample of JCVs across varying frequency bands. We also investigate the distribution of the frequency of the verbs and their components.

### 8.2 Overview of Japanese Compound Verbs

The compound verb in Japanese (複合動詞 *fukugōdōshi*, hereafter JCV) is a concatenation of two or more verbs which function as a single multiword verb. There are several classes of JCV, however in this work we concentrate of the largest and most common class in which the first verb is in the continuative form, which is also
known as the *masu*-stem because it forms the base for the polite spoken -*masu* group of inflections (Uchiyama et al. 2005; Kubota 1992). In common with most studies of JCVs, we concentrate on verbs where both components are native Japanese verbs, not loanwords or Sino-Japanese words. This exclusion is because these latter verbs are much less common and have a different morphology.

As a JCV consists of two adjacent verb components, we will refer to these components as the V1 and V2. A typical JCV is 行き過ぎる *ikisugiru* “to go too far”, where the V1 is the continuative form of 行く *iku* “to go” and the V2 is 過ぎる *sugiru* “to be excessive; to be too much”. 過ぎる is a particularly productive V2. A less productive V2 is 汚す *yogosu* “to make dirty”, found in the JCV 食べ汚す *tabeyogosu* “to eat messily”.

JCVs play a role in Japanese which is analogous to several different structures in other languages. English equivalents include verb-plus-verb-phrase-complements (e.g. *to start to eat* and *to start swimming*) and verb particle constructions (e.g. *to kick up (ball, fuss)*, *to pull down*).

The JCV is a highly productive form, with some particular V1s and V2s being strongly represented, however there is no real restriction on a verb being used within a JCV, subject to issues such as aspect and valency (Kubota 1992), and the result being meaningful. Some popular references list many hundreds of JCVs (Tagashira and Hoff 1986) and major dictionaries typically include several thousand as entries, however it is generally recognized that many more JCVs are in use than are lexicalized. The incompleteness of the lexicalization of JCVs arises not only from their productivity,
but from the fact that their meaning is often obvious to a Japanese speaker, and hence dictionary editors usually concentrate on JCVs which are polysemous, or have idiosyncratic meanings. Extension of the coverage of recorded and translated JCVs would be of assistance in areas such as language learning, and in lexicons used by morphological analysis and machine translation systems.

An example of a polysemous JCV is 引き抜く hikinuku, from 引く hiku “to draw; to pull”, and 抜く nuku “to extract”. It means both “to uproot” and “to pull out”, and less obviously “to head-hunt” and “to lure away”.

8.3 Prior Work

Japanese compound verbs have been studied intensively, e.g. Tagashira and Hoff (1986); Kubota (1992); Nomura and Ishii (1987), however they have had relatively little attention in natural language processing. In recent studies Hashimoto and Bond (2005) developed a computational grammar for these verbs, and tested the approach against a small corpus. Uchiyama et al. (2005) studied techniques for disambiguating polysemous compound verbs, evaluating the techniques against a newspaper corpus.

8.4 General Approach and Resources

8.4.1 Approaches

As discussed in an earlier chapter, a fundamental problem with searching Japanese corpora for unrecorded words is that Japanese text does not usually have spaces or any other marking between words. Thus the identification of words in text necessi-
tates the use of a morphological analysis process to separate the words, and all such processes currently rely on extensive lexicons. The absence of a word in the lexicon usually results in the analysis software defaulting to producing a sequence of untagged morphemes until it can resynchronize.

An approach that has had some previous success is to synthesize possible words by mimicking Japanese morphological processes, and then testing, e.g. using a WWW search engine, to determine whether the word is in use (Breen 2004a). A variant of this approach has been applied in the work described in this chapter. A second approach, in which the Google \textit{n}-gram Corpus (Kudo and Kazawa 2007) was scanned using a filter designed to detect the character patterns consistent with JCVs, was developed and tested. These approaches are described in detail below. Figure 8.1 shows a diagram of the two approaches.
8.4.2 Resources Used

The investigation uses several lexical resources to assist with the identification and extraction of JCVs. The *JMdict* Japanese-English dictionary database (Breen 2004b) and the associated *KANJIDIC* database (Breen 2017b) were used to establish sets of possible V1 and V2 components, and Comblex was used to check JCVs. (See Section 4.2.)

A key resource has been the Google Japanese *n*-gram Corpus (Kudo and Kazawa 2007). (See Section 4.4.) It had been noticed that the *MeCab/IPADIC* analysis system used in the creation of the Corpus usually split JCVs into up to three morphemes depending on the inflection of the verb. Therefore only the 1-gram, 2-gram and 3-gram sections of the Corpus were used in this study, and then only those *n*-grams which began with a *kanji* character.

A similar Google *n*-gram corpus has been used successfully in the extraction of verb-particle constructions in English (Kummerfeld and Curran 2008).

8.4.3 Synthesis of Compound Verbs

In this approach, a set of JCVs was synthesized as follows:

a. The *JMdict* dictionary was examined and JCVs identified. Including alternative surface forms, some 2,900 JCVs in which *kanji* were used in both the V1 and V2

---

3 This work was carried out before the Kyoto WWW Corpus was obtained and an *n*-gram corpus compiled from it.

4 Japanese has considerable flexibility as to whether words are written in *kanji*, *kana* or a mixture. Also, alternative *kanji* are often used. For example, the JCV 詰め合わせる *tsume-awaseru* “to pack an assortment of goods, etc.” can also be written: 詰め合せる, 詰合わせる, 詰め合わせる, 詰めあわせる, つめあわせる, and 並び変える *narubikaeru* “to put things in order” can also be written 並び替える, 並替える, 並替える, etc.
positions were extracted. These were divided into the V1 and V2 components, yielding approximately 700 unique V1s and 600 V2s.

b. Using the V1 and V2 components, 420,000 synthetic JCVs were created using all combinations of the V1s and V2s. For each verb two forms were generated: the form in which the V2 used kanji as the root of the verb, and the form in which the V2 was entirely in the hiragana script. Both these forms are freely used in Japanese, for example 抱き付く dakitsuku “to cling to; to embrace” can equally well be written 抱きつく, and in fact the latter is more commonly used. For each of these, as well as the plain non-past tense (which is considered to be the reference form of Japanese verbs and is used for dictionary headwords) two inflections were generated: the continuative te-form and the plain past tense. These three are the most commonly used inflections in written Japanese, and it was considered appropriate to focus on them in order to detect whether words were in use.

Each synthetic JCV was initially checked against Comblex, resulting in a total of 6,094 matches.

Each synthetic JCV was then checked against the Google n-gram Corpus. As there were three inflections of two written forms of 420,000 JCVs, a total of 2,520,000 words were tested. The sections of the n-gram files which began with a kanji were preprocessed to recombine each 2-gram and 3-gram into a single character string, then sorted, resulting in a file of 270M unigrams in a file of 5.8GB. This facilitated processing in a single pass against the sorted verb file, thus enabling a rapid comparison and collation of results.
Initially, approximately 26,000 of the synthesized JCVs were matched in one or more of their inflections. On inspection, the JCV form in which the V2 was in hiragana did not contribute significantly to the matches, and as this form has an increased chance of homophones which cannot be resolved without textual context, it was removed from the analysis. Also removed were a number of JCVs which were effectively alternative conjugations (passive, potential, etc.). This reduced the matched JCVs to 22,692.

Of the 6,094 JCVs which were found in Comblex, 4,779 matched n-grams in the Corpus, i.e. 1,315 which were found in Comblex were not in the Corpus. On inspection it was noted that many of the 1,315 were archaic and literary words.

The distribution of the counts of occurrences in the Corpus is sharply asymptotic, with a small number having very high counts and declining to a long tail with over 15,000 having counts below 500.

8.4.4 Direct n-gram Search

In addition to the synthesis approach, an alternative approach was devised in which the n-gram Corpus was scanned for character strings which conformed to the structural pattern of JCVs. From the examination of known JCVs, it can be determined that the common structural pattern is:

- a V1 consisting of one or two kanji followed by one to three hiragana;
- a V2 consisting of one or two kanji followed by one to four hiragana.

As there are many other valid text fragments which also conform to this pattern, e.g. noun/particle/verb, noun/particle/adjective, etc. filters were applied as follows:
Chapter 8: Generation and Extraction of Compound Verbs

a. the V1 component was limited to the *masu*-stems of known or potential verbs.

To do this, a list of verb *masu*-stems was created and merged from:

i. all the verbs in the *JMdict* dictionary;

ii. all the V1 components used in the synthesis methods;

iii. all the kanji in the *KANJIDIC* database which had the potential to form a verb (this information is detailed in the database).

A total of 6,023 actual or potential verb stems were identified and used to filter the potential JCVs.

b. the inflecting part of the V2 was limited to the *hiragana* strings associated with valid verbs in the plain non-past, plain past and *te*-form inflections. A list of 208 such inflections was compiled and used as a filter.

A scan of the *n*-gram Corpus for unigrams which conformed to the structural model and passed the filters yielded just on 135,000 potential JCVs. As these included many inflected forms, a considerable amount of post-processing was carried out to reduce them to a consolidated set of reference (plain non-past) forms. Some of the processes involved included:

a. matching the inflected and reference forms and combining the counts;

b. detecting and removing additional inflections such as the potential and passive forms, which share some of the inflection patterns of the reference form;

c. detecting and removing adjectives. In Japanese, adjectives inflect in a manner similar to verbs, and a number had been collected in the scan.
Figure 8.2: Analysis of the token counts and dictionary matches of JCVs, ranking in decreasing order of token frequency

From this a reduced list of 64,776 potential JCVs was produced. When tested against Complex 6,203 matched, an increase of 1,424 over the synthesis method. It is clear that this approach has an improved recall, i.e. the number of actual JCVs identified as a proportion to the number in existence, relative to the published lexicons, but possibly at the price of a reduced precision, i.e. the proportion of actual JCVs among the potential JCVs. As with the synthesized JCVs, the distribution of the counts is asymptotic with a long tail.
8.5 Analysis of the Potential Compound Verbs

A detailed comparison of the potential JCVs compiled in the two approaches revealed that all the synthesized JCVs which had matched unigrams in the Corpus has also been collected in the search, and moreover the \( n \)-gram counts were almost always identical. This meant that a combined set could safely be used for further analysis, with tagging as to whether a JCV had been detected by both methods, or by the search alone.

Figure 8.2 shows the distribution of the \( n \)-gram counts and also the proportion of the potential JCVs which were found in a lexicon.\(^5\)

A key issue is the extent to which these methods have revealed actual JCVs as opposed to character sequences which simply share symbolic characteristics with JCVs. To examine this aspect fully would require an evaluation of each JCV candidate in context, e.g. as it is used in WWW texts, to determine its status. In order to estimate the effectiveness of the JCV extraction approaches, samples of 50 potential JCVs were selected at random from each of three bands based on token frequency

a. High: JCVs with over 5,000 counts in the \( n \)-gram Corpus (3,795 JCVs)

b. Medium: JCVs with 1,000 to 4,999 counts (4,886 JCVs)

c. Low: JCVs with 20 to 999 counts (56,095 JCVs)

The sample JCV candidates were classified as to whether they were in a lexicon or not, and if not, whether they were actually verbs. For the latter analysis, each

\(^5\)For the purposes of depicting this, JCVs were examined in batches of 500, and the percentage which matched were plotted.
potential JCV was manually checked against WWW pages via a search engine to verify whether it was being used as a verb. (At some later stage it may be possible to employ deeper linguistic analysis to carry out this process automatically.)

The summary of this classification is in Table 8.1. For JCV candidates not in the lexicon the counts are broken down into verb and non-verb candidates. The figures in parentheses are numbers of JCVs in each category resulting from the search approach alone.

The JCV candidates which were classified as “other”, i.e. not verbs, fell into several categories. The most common were inflected adjectives which had not been detected in filtering, adverbs such as 再び futatabi “again; once more”, V1s such as 見て (continuative form of 見る miru “to see; to look; etc.” which probably should have been filtered out (see below)), and apparent typographical or grammatical errors associated with other verbs. Some were other constructs such as noun/verb without the usual intervening particle.

Of considerable interest is the relative performance of the JCVs identified by the synthesis approach; in each of the three bands almost all of these JCVs were valid.
verbs.

In terms of the precision of the different approaches, the full set of potential JCVs achieved precisions of 0.68, 0.76 and 0.56 respectively across the three selected bands, however within this the synthesis approach alone, i.e. after subtracting the parenthesized figures in Table 8.1, achieved precisions of 0.91, 1.00 and 0.93.

The comparative recall is difficult to measure as there is no gold standard for the number of JCVs in use or able to be used. Certainly the direct search approach achieved a greater recall but at the price of a lower precision.

As reported above, 6,203 of the potential JCVs matched entries in Comblex. It was noticed that while the high-ranking JCVs tended to match all the dictionaries, lower-ranking JCVs tended to match more sparsely, with often only one or two dictionaries matching. The specific dictionary matches were extracted for each of the high, medium and low bands. These are shown in Table 8.2. The figures in parentheses are the proportions of the dictionary matches against the total dictionary matches for the band.

8.6 Productivity Measures of V1 and V2 Components

The productivity of the JCV is well known, as is the frequency with which some V1 and V2 components appear. It is useful, having established a reasonably large collection of JCVs, to use this to analyze the frequency of usage of the components.

For the purpose of this analysis, the V1s and V2s were extracted, counted and
Table 8.2: Occurrence of potential JCVs in the different dictionaries across the three frequency bands, in terms of the raw type count and proportion of overall types (in parentheses)

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>High</th>
<th>Med</th>
<th>Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>JMdict</td>
<td>1,788 (0.47)</td>
<td>453 (0.09)</td>
<td>671 (0.01)</td>
<td>2,912</td>
</tr>
<tr>
<td>Kōjien</td>
<td>1,420 (0.37)</td>
<td>401 (0.08)</td>
<td>976 (0.01)</td>
<td>2,797</td>
</tr>
<tr>
<td>Daijirin</td>
<td>1,626 (0.43)</td>
<td>491 (0.10)</td>
<td>970 (0.01)</td>
<td>3,087</td>
</tr>
<tr>
<td>GoiTaikei</td>
<td>1,375 (0.36)</td>
<td>377 (0.08)</td>
<td>661 (0.01)</td>
<td>2,413</td>
</tr>
<tr>
<td>JLD</td>
<td>2,172 (0.57)</td>
<td>932 (0.19)</td>
<td>2,023 (0.03)</td>
<td>5,126</td>
</tr>
</tbody>
</table>

Table 8.3: V1 and V2 frequencies for the two proposed methods and in the lexicon ranked from:

- the full collection of possible JCVs collected from the n-gram Corpus;
- the synthesized JCVs which had matches in the n-gram Corpus;
- the JCVs extracted from the combined lexicon;
- the highest-ranked 10,000 JCVs in the full collection (to see if there is a bias in component use according to the how common the JCV is).

In addition, V2 rankings collected by Kubota (1992) from her own corpus analysis
and from an earlier published collection (Nomura and Ishii 1987) were added for comparison.

The number of V1s and V2s which were extracted are shown in Table 8.3. As can be seen, over 80% of the V2s in the combined file are only found in the relatively low-frequency JCVs.

While there was some correlation of the frequency rankings of the V1s and V2s, there were also some notable differences. This can be seen in Tables 8.4 and 8.5, which show the 20 most common components as they were found in the combined JCV list, with their comparative rankings in the other lists.

With regard to the V1s in Table 8.4, it will be noted that there are some which do not appear in all lists, or have very different rankings. These almost all relate to differing interpretations at to what comprises a JCV. For example:

a. the V1s 限り kagiri “restricted”, 余り amari “remain” and 及び oyobi “and, as well as”, while deriving from verbs, are almost invariably use as conjunctions or adverbials in modern Japanese, and hence would not normally be part of a JCV. They should be added to the filter rules.

b. 再び futatabi “again” is more commonly regarded as an adverb, and should also probably be excluded.

c. the two “te-form” V1s (見て mite and 出て dete) could be classed as either part of JCVs, or as the common (V1 te-form, V2) sequence which has the sense of simultaneous occurrence or activity. They are often excluded from JCV classifications.
Table 8.4: V1 rankings for the two proposed methods, the lexicon and the most frequent 10,000 JCVs

<table>
<thead>
<tr>
<th>V1</th>
<th>Combined n-gram</th>
<th>Synth. n-gram</th>
<th>Lexicon</th>
<th>First 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>通り</td>
<td>1</td>
<td>2</td>
<td>190</td>
<td>1</td>
</tr>
<tr>
<td>再び</td>
<td>2</td>
<td>1</td>
<td>642</td>
<td>3</td>
</tr>
<tr>
<td>思い</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>見て</td>
<td>4</td>
<td>4</td>
<td>520</td>
<td>4</td>
</tr>
<tr>
<td>限り</td>
<td>5</td>
<td>–</td>
<td>857</td>
<td>9</td>
</tr>
<tr>
<td>同じ</td>
<td>6</td>
<td>27</td>
<td>–</td>
<td>51</td>
</tr>
<tr>
<td>入り</td>
<td>7</td>
<td>13</td>
<td>52</td>
<td>29</td>
</tr>
<tr>
<td>余り</td>
<td>8</td>
<td>–</td>
<td>726</td>
<td>18</td>
</tr>
<tr>
<td>使い</td>
<td>9</td>
<td>11</td>
<td>45</td>
<td>13</td>
</tr>
<tr>
<td>出て</td>
<td>10</td>
<td>7</td>
<td>823</td>
<td>37</td>
</tr>
<tr>
<td>取り</td>
<td>11</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>入れ</td>
<td>12</td>
<td>19</td>
<td>49</td>
<td>24</td>
</tr>
<tr>
<td>行き</td>
<td>13</td>
<td>16</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>買い</td>
<td>14</td>
<td>24</td>
<td>24</td>
<td>11</td>
</tr>
<tr>
<td>持ち</td>
<td>15</td>
<td>22</td>
<td>25</td>
<td>19</td>
</tr>
<tr>
<td>作り</td>
<td>16</td>
<td>15</td>
<td>68</td>
<td>30</td>
</tr>
<tr>
<td>考え</td>
<td>17</td>
<td>17</td>
<td>113</td>
<td>25</td>
</tr>
<tr>
<td>言い</td>
<td>18</td>
<td>9</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>及び</td>
<td>19</td>
<td>–</td>
<td>1242</td>
<td>132</td>
</tr>
<tr>
<td>感じ</td>
<td>20</td>
<td>30</td>
<td>428</td>
<td>41</td>
</tr>
</tbody>
</table>

*Chapter 8: Generation and Extraction of Compound Verbs*


<table>
<thead>
<tr>
<th>V2</th>
<th>Combined n-gram</th>
<th>Synth. n-gram</th>
<th>Lexicon</th>
<th>First 10,000</th>
<th>Kubota</th>
</tr>
</thead>
<tbody>
<tr>
<td>始める</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>続ける</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>過ぎる</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>出す</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>合う</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>得る</td>
<td>6</td>
<td>13</td>
<td>152</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>行く</td>
<td>7</td>
<td>10</td>
<td>59</td>
<td>12</td>
<td>–</td>
</tr>
<tr>
<td>初める</td>
<td>8</td>
<td>8</td>
<td>106</td>
<td>43</td>
<td>–</td>
</tr>
<tr>
<td>終わる</td>
<td>9</td>
<td>7</td>
<td>74</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>切る</td>
<td>10</td>
<td>11</td>
<td>9</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>難く</td>
<td>11</td>
<td>–</td>
<td>–</td>
<td>28</td>
<td>–</td>
</tr>
<tr>
<td>込む</td>
<td>12</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>多く</td>
<td>13</td>
<td>–</td>
<td>–</td>
<td>56</td>
<td>–</td>
</tr>
<tr>
<td>直す</td>
<td>14</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>見る</td>
<td>15</td>
<td>14</td>
<td>109</td>
<td>27</td>
<td>–</td>
</tr>
<tr>
<td>忘れる</td>
<td>16</td>
<td>15</td>
<td>18</td>
<td>16</td>
<td>39</td>
</tr>
<tr>
<td>上げる</td>
<td>17</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>歩く</td>
<td>18</td>
<td>19</td>
<td>50</td>
<td>25</td>
<td>34</td>
</tr>
<tr>
<td>来る</td>
<td>19</td>
<td>22</td>
<td>291</td>
<td>31</td>
<td>–</td>
</tr>
<tr>
<td>頑張る</td>
<td>20</td>
<td>–</td>
<td>–</td>
<td>68</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 8.5: V2 rankings for the two proposed methods, the lexicon and the most frequent 10,000 JCVs
It will be noted that few of the high-ranking V1s in the lexicon appear in Table 8.4. On inspection it proved that they mostly lie in the 20–40 range in the \( n \)-gram lists. Given that dictionary compilers concentrate on JCVs which are polysemous or have idiosyncratic meaning, a lack of frequency alignment with corpus-based lists is to be expected.

While there is generally good agreement between the rankings of the lists of potential V2s in Table 8.5, some attract similar comments:

a. two of the potential V2s (難く gataku and 多く ōku) are clearly derived from adjectives, and should be added to the filter rules.

b. V2s which rank lower in the lexicon list (得る eru “to attain”, 初める hajimeru “to start”, etc.) are usually parts of semantically regular JCVs, and hence are less likely to be in a dictionary.

c. the appearance of 頑張る ganbaru “to persist; to insist on; to stand firm” is of interest. It is not usually regarded as a JCV component, yet from its occurrence in the \( n \)-gram lists, it is clearly being used as such.

The foregoing comments are confined to the V1 and V2 components appearing among the 20 highest ranking counts in the combined list; similar comments can be made about a number of lower-ranking components. There is scope for considerable analysis of the ranking lists of V1 and V2 components.
8.7 Summary, Conclusions and Future Work

In this chapter an investigation into the synthesis of Japanese compound verbs has been described. The primary synthesis technique was to combine confirmed candidate verb components and test the resulting construction against a large \( n \)-gram corpus. This resulted in the identification of a set of 22,692 potential verbs with a high level of precision. The second technique explored was to scan the \( n \)-gram corpus using a template based on the orthographic pattern of typical JCVs. This resulted in a list of 64,776 potential verbs, which included all members of the list produced by the first technique. The scan technique had a lower level of precision, but a somewhat greater recall. 6,203 of the potential verbs identified by the two techniques matched entries in a reference lexicon.

The work described here has demonstrated that large numbers of JCVs are in regular use and can be detected through the application of NLP techniques to corpora.

A substantial collection of potential JCVs which are not recorded in commonly-used dictionaries has been identified for further study and lexicographic analysis. The collection, which includes corpus-based frequency-of-use and the V1 and V2 for each JCV has been made available for other Japanese NLP and lexicographic projects.\(^6\)

Future lexicographic work based on the project could include the development and testing of the meanings of unrecorded JCVs.\(^7\) The approach developed in earlier work (Uchiyama et al. 2005), which employs rule-based and statistical methods based on extensive classification of the V1 and V2 components, could be a starting point in

---

\(^6\)https://aclweb.org/aclwiki/Resources_for_Japanese#Monolingual

\(^7\)An exploratory investigation into this was carried out with promising initial results, however it was decided to leave the full investigation to future work.
mapping compound verbs onto semantic classes in Japanese, and learning translation rules for each semantic class. Also the use of a large English n-gram corpus, as explored in Chapter 6, could be considered.
Chapter 9

Neologism Identification through Language Contexts

9.1 Introduction

This chapter reports on an investigation as to whether it is possible to identify and extract neologisms from text, based on the language patterns in which they occur.\footnote{The material in this chapter formed the basis of the following paper: Breen, James, Timothy Baldwin and Francis Bond. 2018. The Company They Keep: Extracting Japanese Neologisms Using Language Patterns. In Proceedings of the Global Wordnet Conference 2018. Singapore. (Accepted - to appear.)}

The genesis of the project is the observation that one often encounters in Japanese text terms which the writer thinks needs some amplification, either because they are new or are not common. This is signalled by following the term with phrases such as というのは (to iu no wa, literally “as for that which is said <term>”) and とは (to wa, literally “as for <term>”), sometimes combined with the reading in parentheses, and then followed by an explanation. The phenomenon is well known to Japanese
Chapter 9: Neologism Identification through Language Contexts

translators, who often will do a WWW search for “<term> というのは”,”<term> とは” etc. when encountering an unfamiliar term in order to identify cases where the term is being described, discussed or otherwise highlighted.²

The research question being investigated here is whether it is possible to use these sorts of linguistic patterns in a systematic way as a means for harvesting possible lexical items from large bodies of text, which presumably could then be filtered against a reference lexicon to expose possible neologisms. Possible approaches for classifying text include using string matching for the linguistic patterns and training a machine learning model with the patterns.

The investigation broadly breaks into two components:

a. the identification of the sorts of language patterns used to describe, discuss, highlight, etc. terms;

b. the extraction and evaluation of terms so targeted by those language patterns.

9.2 Prior Work

Research into the use of text patterns to detect terms of interest has taken place in several language-processing contexts. A major field is keyphrase extraction, i.e. the extraction of sets of phrases that are related to the main topics of given documents.

In general the purpose of keyphrase extraction is to identify documents which satisfy some particular criterion, e.g. subject matter, from large sets of documents, or to identify the general topic of a document. The process is used for many tasks such as

²Such WWW searches need to employ the full phrase without it being segmented by the search engine; this is typically achieved by encapsulating the expression in quotation marks.
text summarization, document clustering, etc. Hasan and Ng (2014) have produced a wide-ranging survey of the various techniques used in keyphrase extraction and their relative effectiveness, and Kim et al. (2013) evaluate the performance of a variety of supervised and unsupervised approaches deployed in the SemEval-2010 workshop on the topic.

Another field is term extraction, which is a major part of the broader field of terminology, usually in technical contexts (Kageura 2000). In this area Takeuchi et al. (2009) adapted the French ACABIT system, which detects morpho-syntactic sequences, to isolate terms in Japanese for later analysis. Le et al. (2013) used patterns of phrases to identify particular Japanese legal documents of interest. Mathieu (2013) successfully adapted a keyphrase extractor for use with Japanese, although its use was restricted to kanji sequences.

A relationship between a text pattern and a term of interest is a form of collocation, i.e. lying between idiomatic expressions and free word combinations. In their survey of collocations in language processing, McKeown and Radev (2000) explore the role of the extraction of collocations in lexicography, although the focus is on the identification of general terms rather than those which are highlighted as being of interest.

Another case is the use of lexico-syntactic “Hearst patterns” to detect the semantic relationship between nouns, which has proved useful in areas such as hypernymy (Hearst 1992; Snow et al. 2005). (Similar patterns were explored earlier by Calzolari (1984).) The application of this approach to lexical item extraction in general, particularly in Japanese, would be an interesting study.
Prior published research into the use of Japanese text patterns which target general terms of interest appears to be quite limited. Sato and Kaide (2010) employed a technique for extracting English-Japanese name pairs by scanning texts for nearby occurrences of “Mr”, “Mrs”, etc. and the Japanese equivalents, e.g. さん san.

### 9.3 Text Corpora

Testing for the presence and usage of selected text patterns requires the availability of reasonably large quantities of Japanese text, preferably from a variety of sources.

One possibility is the Google Japanese n-gram Corpus itself. というのは (described above) is a 4-gram (と + いう + の + は) and という意味 to iu imi “said term’s meaning” and の意味は no imi wa “as for the meaning of” are 3-grams (と + いう + 意味, の + 意味 + は) in that Corpus. (See Section 4.4 for details.) It would be possible to examine, for example, the 7-grams section of the Corpus and examine the text which precedes those patterns. While this would deliver possibly useful frequencies for targeted terms, it would provide virtually no context for the usage of the term and hence would be unlikely to be of any real use.

While there are number of Japanese corpora available for use in NLP work, most are actually quite small. For example the pattern という造語 to iu zōgo “said coinage” (discussed below), which occurs 10,042 times in the Google n-gram Corpus, occurs 398 times in the Kyoto WWW Text Corpus, 340 times in the Twitter text collection, 42 times in the collection of Mainichi Shimbun articles, 31 times in the collection of Nikkei Shimbun articles, and 9 times in the BCCWJ. (See Section 4.3 for details of these corpora.)
In fact the number of available large-scale collections of Japanese text is quite limited. After investigation of the options, it was decided to concentrate the testing and evaluation in this project to two corpora:

a. the Kyoto WWW Corpus (Section 4.3). At around 500 million sentences this is a reasonable body of text. The main problem is that it was compiled in 2004 and is getting dated, and hence what may have been neologisms at the time of its collation may well be recorded and accepted now, or have totally faded from use.

b. the Twitter data collection which as described in Section 4.3 is a collection of 870 million Japanese Twitter text passages from 2014 and 2015. This data provides the opportunity to see how the techniques under investigation perform with contemporary and at times slangy text, both of which have the potential to contain neologisms.

9.4 Initial Exploration

9.4.1 Pattern Frequencies

An early issue to be explored was whether the text patterns typically associated with the discussion of particular terms occur in sufficient quantities to make them useful search keys. A simple method for testing this is to examine their frequencies in an \( n \)-gram corpus, such as the Google Japanese \( n \)-gram Corpus. Patterns such as と言うのは to iu no wa “as for the said” and という言葉 to iu kotoba “said term” are 3-grams and 4-grams respectively, and a tool is available to query the Corpus for
Chapter 9: Neologism Identification through Language Contexts

Table 9.1: Google n-gram Corpus Frequencies of Text Patterns

<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>とは to wa “as for”</td>
<td>169,756,339</td>
</tr>
<tr>
<td>というのは/と言うのは to iu no wa “as for the said”</td>
<td>19,134,679/1,207,555</td>
</tr>
<tr>
<td>という言葉/ということば to iu kotoba “said term”</td>
<td>5,360,613/167,095</td>
</tr>
<tr>
<td>という意味/といういみ to iu imi “said term’s meaning”</td>
<td>4,544,800/10,364</td>
</tr>
<tr>
<td>という意味は to iu imi wa “as for the said term’s meaning”</td>
<td>51,726</td>
</tr>
<tr>
<td>の意味は/のいみは no imi wa “as for the meaning of”</td>
<td>1,979,108/1,169</td>
</tr>
</tbody>
</table>

the frequency of occurrence of particular n-grams. The frequencies for a selection of typical patterns is in Table 9.1. (Text strings such as というのは and と言うのは are paired because they essentially the same, differing only in their surface forms in written Japanese.)

The high-scoring とは is really a common form of topic marker without any particular association with new or unusual terms, and almost certainly would produce very noisy results if used as a search pattern. On the other hand というのは, という言葉, という意味 and の意味は are typically associated with particular terms and are probably worth further investigation.

9.4.2 Testing Contexts of Known New Terms

As a preliminary to developing a set of linguistic patterns for testing, an investigation of the sorts of contexts in which known new terms are being used to see if any useful additional patterns could be identified. As an initial exploration 5 terms
were chosen from recent additions to the JMdict database which had been noted as popular new words/expressions. The 5 terms were:

a. マタハラ *matahara*. This is an abbreviation of マタニティハラスメント *mataniti harasumento*, which is from the English “maternity harassment” and means workplace discrimination against pregnant women;

b. こじらせ女子 *kojirase joshi*; a colloquial expression meaning “girl who has low self-esteem”;

c. ナマポ *namapo*, which derives from the abbreviation 生保 *seiho*. As 生保 can be an abbreviation of both 生命保険 *seimei hoken* “life insurance” and 生活保護 *seikatsu hogo* “public assistance; welfare”, some people pronounce it in the latter context as なまほ *namaho* (the kanji 生 can be pronounced both sei and nama). From this ナマポ has emerged as rather derogatory slang for a welfare recipient;

d. 美魔女 *bimajo*; a colloquial expression, meaning “middle-aged woman who looks very young for her age”.

e. 隠れメタボ *kakure metabo*; an abbreviation of a longer term meaning “normal weight obesity” which emerged in the press in Japan in early 2016.³

Sample texts using these terms were identified by carrying out a Google WWW search, then stepping through the results selecting every 10th search result. If the resultant text did not contain the term in a sentence, the following result was examined

³https://kotobank.jp/word/%E9%A0%EA%E3%82%8C%E3%82%BF%E3%83%9C-460942
until one was found. 10 sentences were collected for each term; the full 50 sentences can be found in Appendix D.1.

While this is clearly a small number of samples, a few points emerge:

a. there were relatively few of the という/とは/etc. sorts of patterns used. Only four occurred a total of seven times in the 50 sentences. These are shown in Table 9.2.

<table>
<thead>
<tr>
<th>Term</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>という言葉</td>
<td>皆さん是「こじらせ女子」という言葉を... “Everyone is [...] the expression ‘kojirase joshi’”</td>
</tr>
<tr>
<td>という言葉</td>
<td>実年齢よりずっと若く見える女性を指す“美魔女”という言葉もすっかり... “Indicating women who look younger than their age, the term ‘bimajo’ is also completely [...]”</td>
</tr>
<tr>
<td>とは</td>
<td>「マタハラ」とはマタニティハラスメントの略で “As for ‘matahara’, an abbreviation of maternity harassment, [...]”</td>
</tr>
<tr>
<td>とは</td>
<td>こじらせ女子とは、女性としての自分に... “As for ‘kojirase joshi’, women themselves [...]”</td>
</tr>
<tr>
<td>とは</td>
<td>隠れメタボとは、日本のメタボリックシンドロームの... “As for ‘kakure matabo’, the metabolic syndrome in Japan [...]”</td>
</tr>
<tr>
<td>というと</td>
<td>美魔女というと、最近は「年甲斐もなくイタい」... “The term ‘bimajo’ has recently [become ...] disgracefully painful.”</td>
</tr>
<tr>
<td>といって</td>
<td>最近では、隠れメタボといふて細い人でも知らない... “Recently, when the term ‘kakure matabo’ is said, thin people also don’t know [...]”</td>
</tr>
</tbody>
</table>

Table 9.2: WWW Text Samples using Selected Patterns

b. quite a number of the terms being tested occurred encapsulated by some form of parentheses, either “...” (5 occurrences), 「...」 (10 occurrences) or『...』 (1
occurrence).\textsuperscript{4} As can be seen in Table 9.2 both occurrences of the pattern という言葉 follow the term with such marking. This indicates that it is quite possible that the combination of these sorts of parentheses plus a suitable text pattern is a stronger signal that a term is of interest than either on their own.

9.4.3 Explicit Neologism Labelling

The next element of the search for suitable text patterns associated with neologisms was to carry out some testing to see whether passages could be found with explicit text stating that it contains a neologism. There are a number of terms in Japanese which can mean neologism; these are shown in Table 9.3, along with their relative frequencies from the Google \(n\)-grams.

As the first three account for almost all the usage, these were investigated for their use in combination with the という and と言う ("as said") patterns. The Google \(n\)-gram frequencies for the combinations are shown in Table 9.4.

As the frequencies for という造語 and いう新語 looked promising, a sample of 10 sentences for each were identified via a Google WWW search. These sentences are included at Appendix D.2.

These sample sentences indicate the approach seems to have considerable promise. Quite a few relatively new terms, such as ブロマンス buromansu "bromance", were in the samples. It is also interesting to note that all the terms referenced by the patterns were encapsulated in some form of parentheses.

\textsuperscript{4}Japanese orthography uses a variety of symbols for text encapsulation, with the 「」 pair commonly used where inverted commas are used in English. Other symbols used for this include: ○, ‖, ◄,  Declare,  Declare,  Declare,  Declare and  Declare.
Table 9.3: Google \( n \)-gram Frequencies for Words Meaning “Neologism”

<table>
<thead>
<tr>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
</table>
| 造語  
zōgo “neologism, coinage” | 232,837 |
| 新語  
shingo “neologism, new word” | 152,785 |
| 現代用語  
gendai yōgo “neologism, recent word” | 62,705 |
| 新造語  
shinzōgo “neologism, new coinage” | 3,978 |
| 言語新作  
gengo shinsaku “neologism (esp. medical)” | 220 |
| 造語症  
zōgoshō “neologism (esp. medical disorders)” | <20 |
| ネオロジズム  
neorojizumu “neologism” | <20 |
| ネオレジズム  
neorejizumu “neologism” | <20 |

9.4.4 Parenthesized \( kana \)

It has been observed that explanations of terms in Japanese are often accompanied by the reading of the term. For example the Wikipedia article on 家制度 (the former legal framework for households and families) begins 家制度(いえせいど)とは..., where いえせいど ieseido gives the reading of the term in hiragana.

It is appropriate to consider whether parenthesized readings are present in association with the sorts of language patterns under consideration, and if so whether they are in sufficient quantities to include them in the text analysis.

To check this, a scan was made of the Kyoto Corpus to extract all sentences containing the patterns described above (という言葉, という造語, etc.). Approximately 2.4 million sentences were extracted, and these were analyzed to determine if they contained parenthesized strings of \( kana \). Only 116 text lines contained “(\( kana \))”
patterns, and on analysis these were classified as follows:

- 1 reading associated with the target of the pattern;
- 15 readings of individual kanji and kanji compounds. These were not associated with the target of the pattern (it is common to append text with the readings of less-known terms);
- 5 readings of names (this is commonly done with Chinese and Korean names);
- 95 asides, amplifications, comments, etc., e.g. らしい rashii and でしょう deshō. Both these mean “it seems”, “perhaps”, etc.

Finding only one passage containing “term (reading)” pattern in a reasonably large quantity of text indicates that while it may represent some sort of gold standard for establishing a reading, it may not be common enough to make it worth a lot of attention.
9.5 Further Investigation

9.5.1 Expansion of Linguistic Patterns

As discussed above, a number of text patterns had been identified with initially positive test results. They were:

- **xx という言葉/ということば** \(to\ iu\ kotoba\)
- **xx と言うのは/というのは** \(to\ iu\ no\ wa\)
- **xx といういみ/という意味** \(to\ iu\ imi\)
- **xx というと** \(to\ iu\ to\)
- **xx という造語** \(to\ iu\ zōgo\)
- **xx という新語** \(to\ iu\ shingo\)
- **xx という流行語** \(to\ iu\ ryūkōgo\)

In all of these the term to which they refer precedes the text as “xx”.

Discussions were held with several native speakers of Japanese in order to identify possible patterns which may be used with new terms. From these discussions a number of additional patterns were identified. Some also typically followed the term in question, as follows.

- **xx という言葉を聞き** \(to\ iu\ kotoba\ wo\ kiki\) “Hearing the said word xx”
- **xx という言葉を耳に** \(to\ iu\ kotoba\ wo\ mimi\ ni\) “Hearing the said word xx”
- **xx という言葉が話題に** \(to\ iu\ kotoba\ ga\ wadai\ ni\) “The said word is the topic (or subject)”
- **xx という言葉がはやって** \(to\ iu\ kotoba\ ga\ hayatte\) “The said word is in vogue/fashion”
- **xx という言葉が流行って** \(to\ iu\ kotoba\ ga\ hayatte\) “The said word is in vogue/fashion”
- **xx という言葉が流行して** \(to\ iu\ kotoba\ ga\ ryūkō\ shite\) “The said word is in vogue/fashion”
- **xx という不思議な** \(to\ iu\ fushigi\ na\) “The said xx is strange/curious”
- **xx という謎の** \(to\ iu\ nazo\ no\) “The said xx is enigmatic/mysterious”
- **xx という新しい** \(to\ iu\ atarashii\) “The said xx is new”
As can be seen, some of these extend the という言葉 pattern by adding modifiers such as 聞き \text{\textit{kiki}}/耳に \text{\textit{mimi ni}} “heard”, 話題 \text{\textit{wadai}} “topic”, はやって/流行って/流行 \text{\textit{hayatte/ryūkōshite}} “vogue, fashion”, 不思議な \text{\textit{fushigi na}} “curious, strange”, 謎の \text{\textit{naza no}} “enigmatic, mysterious”, and 新しい \text{\textit{atarashii}} “new”.

In addition a set of phrases which would precede a target word were identified as follows:

このごろ/この頃よく聞く \textit{xx kono goro yoku kiku} “The often heard recently \textit{xx}”
近頃よく聞く \textit{xx chikagoro yoku kiku} “The often heard recently \textit{xx}”
最近よく聞く \textit{xx saikin yoku kiku} “The often heard recently \textit{xx}”
このごろ/この頃(よく)耳にする \textit{xx kono goro (yoku) mimi ni suru} “The often heard recently \textit{xx}”
近頃(よく)耳にする \textit{xx chikagoro (yoku) mimi ni suru} “The often heard recently \textit{xx}”
最近(よく)耳にする \textit{xx saikin (yoku) mimi ni suru} “The often heard recently \textit{xx}”
このごろ/この頃(よく)話題になる \textit{xx kono goro (yoku) wadai ni naru} “The recently topicalized \textit{xx}”
近頃(よく)話題になる \textit{xx chikagoro (yoku) wadai ni naru} “The recently often-raised topic \textit{xx}”
最近(よく)話題になる \textit{xx saikin (yoku) wadai ni naru} “The recently often-raised topic \textit{xx}”
近頃はやりの \textit{xx chikagoro hayari no} “The recently popular \textit{xx}”
最近はやりの \textit{xx saikin hayari no} “The recently popular \textit{xx}”
このごろ/この頃はやりの \textit{xx kono goro hayari no} “The recently popular \textit{xx}”
このごろ/この頃流行(り)の \textit{xx kono goro hayari no} “The recently popular \textit{xx}”
近頃流行(り)の \textit{xx chikagoro hayari no} “The recently popular \textit{xx}”
最近流行(り)の \textit{xx saikin hayari no} “The recently popular \textit{xx}”

In these the alternative prefixes (このごろ/この頃 \textit{kono goro}, 近頃 \textit{chikagoro}, and 最近 \textit{saikin}) all mean “recently”. The full expressions have meanings such as “the
recently-(often)-heard ...” and “the recent fad ...”, etc.

This resulted in an overall set of 37 text patterns, some of which have alternative surface forms, e.g. このごろ and この頃.

### 9.5.2 Comparative Evaluation of the Language Patterns

The text patterns were tested against the Kyoto WWW Corpus. (As a comparison the counts of occurrences in text in a collection of one week’s Twitter text from January 2015 were also extracted.) The counts of matching lines of text are shown in Table 9.5 and Table 9.6.

For each pattern a sample of 20 sentences was examined in detail. The sampling method was to examine every N-th sentence, where N was approximately no-of-lines/20. Where the sentence was too fragmentary to evaluate, a common problem with that Corpus, the following one was used instead. In the examination each sentence was classified into one of three groups:

a. Group 1: the sentence did not directly discuss any identifiable word or term;

b. Group 2: the sentence focused on a word or term which is already established in one or more lexicons;

c. Group 3: the sentence focused on a word or term which is not in an accessible lexicon, and which warrants further investigation.

For each of Groups 2 and 3 the classifications were separated according to whether or not the word or term was in some form of parentheses.
The classification process is inevitably somewhat subjective, especially in establishing the difference between members of Group 1 on one hand, and members of Groups 2 and 3 on the other. It was also very time-consuming with about 750 sentences to be examined.

As initial inspection of the samples indicated that some of the patterns were more productive with the addition of 言葉 (kotoba “word”), which probably provided more focus, these expanded forms were added to the evaluation (the romanization and glosses of these patterns is given above in Section 9.5.1). The expansions were:

- という新しい言葉 (from という新しい)
- 最近よく聞く言葉 (from 最近よく聞く)
- という不思議な言葉 (from という不思議な)
- という謎の言葉 (from という謎の)

The classification counts for each pattern are in Table 9.5 and Table 9.6. Patterns which only differed in surface forms, e.g. この頃/このごろ and 流行り/流行/はやり, were grouped in the evaluation. (See Section 9.5.1 above for the romanization and glosses of these patterns.)

9.5.3 Discussion

From the results recorded in Table 9.5 and Table 9.6, and from the analysis of the sampled text passages, a number of observations can be made:

a. it is clear that some of the text patterns are quite effective in identifying text passages which focus on words or terms of interest.

b. In some cases the precision appears to be quite high: in three of the sets of samples (という造語, という新語, という新しい言葉) all of the passages had such
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Kyoto count</th>
<th>Twitter count</th>
<th>Grp 1</th>
<th>Grp 2</th>
<th>Grp 2 (P)</th>
<th>Grp 3</th>
<th>Grp 3 (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>と言うのは</td>
<td>1218088</td>
<td>7122</td>
<td>17</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ということの</td>
<td>216917</td>
<td>2052</td>
<td>11</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>という意味</td>
<td>637177</td>
<td>4720</td>
<td>15</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ということ</td>
<td>398</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>という新語</td>
<td>187</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>の意味悩み</td>
<td>36154</td>
<td>721</td>
<td>13</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>という言葉文</td>
<td>266646</td>
<td>5426</td>
<td>7</td>
<td>5</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>という言葉を聞き</td>
<td>636</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>という言葉を耳にする</td>
<td>1327</td>
<td>8</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>という言葉が話題に</td>
<td>49</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>という言葉がはやってという言葉が流行って</td>
<td>238</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>という言葉が流行して</td>
<td>74</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>という不思議な</td>
<td>5170</td>
<td>84</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>という不思議な言葉</td>
<td>43</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>いう言葉の</td>
<td>3003</td>
<td>397</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>いう言葉の言葉</td>
<td>205</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>いう新しい</td>
<td>16992</td>
<td>178</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>いう新しい言葉</td>
<td>127</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>いう流行語</td>
<td>176</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ごろごろよく聞くこの頃よく聞く</td>
<td>17</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9.5: Classification Counts for Text Patterns (initial part)
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Kyoto count</th>
<th>Twitter count</th>
<th>Grp 1</th>
<th>Grp 2</th>
<th>Grp 2 (P)</th>
<th>Grp 3</th>
<th>Grp 3 (P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>近頃よく聞く</td>
<td>20</td>
<td>0</td>
<td>11</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>最近よく聞く</td>
<td>725</td>
<td>45</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>最近よく聞く言葉</td>
<td>80</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>このごろよく耳にするこの頃よく耳にする</td>
<td>12</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>近頃よく耳にする</td>
<td>27</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>最近よく耳にする</td>
<td>790</td>
<td>20</td>
<td>4</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>このごろ耳にするこの頃耳にする</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>近頃耳にする</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>最近耳にする</td>
<td>158</td>
<td>2</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>このごろよく話題になるこの頃よく話題になる</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>近頃よく話題になる</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>最近よく話題になる</td>
<td>50</td>
<td>1</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>このごろ話題になるこの頃話題になる</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>近頃話題になる</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>最近話題になる</td>
<td>669</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>このごろはやりのこの頃はやりのこのごろ流行りのこの頃流行りのこのごろ流行のこの頃流行の</td>
<td>193</td>
<td>14</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>近頃はやりの近頃流行りの近頃流行の</td>
<td>304</td>
<td>0</td>
<td>14</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>最近はやりの最近流行りの最近流行の</td>
<td>6629</td>
<td>268</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 9.6: Classification Counts for Text Patterns (continued)
a focus, and in another five (という言葉を聞き, という言葉を耳に, という言葉が話題に, という言葉がはやって, という流行語) 85% or more had that focus. (See Section 9.5.1 above for the romanization and glosses of these patterns.)

c. in attempting to distinguish between the Group 2 and 3 classifications, it was obvious that the Kyoto Corpus is now over 12 years old, and what may well have been quite new and unrecorded in 2004 is likely to have been added to some lexicon since then. Ideally a more modern text collection will be desirable.

d. in many ways it is not so important if the words/terms being identified are already lexicalized as any harvesting system should be checking them off against a reference lexicon anyway.

e. it is noted that approximately half of the sampled passages (349) were classified into Groups 2 and 3, and these were about evenly split between those where the target term was in parentheses (177) and those where it was not (172).

f. overall the numbers of sentences extracted with the selected patterns only made up a very small proportion of the sentences in the Corpus. The Kyoto Corpus contains approximately 500 million sentences, however the sentence patterns mentioned in (b) above, which appear to have high precision, only extracted 2,600 sentences. On a more relaxed measure, e.g. the patterns which led to 10 or more in Groups 2 and 3, the numbers extracted came to about 280,000 (about 0.06%), and it was observed that most of these were from one pattern (という言葉).

As an additional brief test, a sample of terms which had emerged as Group 3
in the evaluation (化身話 keshingo “avatar language”, 光の速度に近い hikari no sokudo ni chikai “near light-speed”, 等生化 tōseika “normalization (in treatment of the disabled)”, ガイノイド gainoido “gynoid”) was selected, and all of their occurrences in the Corpus were examined to see if there were any other text patterns which may have been associated with them. There were a few cases of というと, but apart from that no other useful patterns were apparent.

### 9.6 Locating Target Terms

An automated extraction process needs to be able to identify the particular terms that are the focus or target of the text patterns being investigated here. As discussed above, in about half the cases these are in some form of parentheses, which should simplify their extraction. Examination of some of the samples indicates that, as one would expect, the target term generally precedes a という.... pattern, and in the other patterns generally follows the pattern. There are exceptions, for example in the passage: もう一つ「PDP」ということを最近よく聞くのですが... mōhitotsu PDP to iu koto wo saikin yoku kiku no desu ga, which means “also the term ‘PDP’ is a recently heard one, however ..” contains two patterns of interest but only the first (ということ) is actually targeting the term.

For the non-parenthesized terms things are more complex. For the という.... patterns the term is usually immediately preceding the pattern, and from inspection it appears that collecting all the contiguous nouns, verbs and adjectives between the last preceding particle and the pattern will usually work. For example in 私もニュースで鳥インフルエンザという言葉を耳にするたびに... watashi mo nyūsu de tori
Chapter 9: Neologism Identification through Language Contexts

infuruenza to iu kotoba wo mimi ni suru tabi ni, which means “whenever I hear the expression ‘bird flu’ in the news ...” the expression 鳥インフルエンザ tori infuruenza would (correctly) be extracted. In some cases this may lead to over-extraction, e.g. in 番組の中でおもわずファンドルという造語を使ってしまいました bangumi no naka de omowazu fandoru to iu zōgo wo tsukatte shimaimashita, which means “in the middle of the program the neologism ‘fundle’ was unintentionally used”, the word pair おもわずファンドル omowazu fandoru “unintentionally fundle” would be extracted, whereas it should be just ファンドル (fandoru “fundle”) alone. Similarly in 最近デー・トレーダーという言葉を聞きますが... saikin dē torēdā to iu kotoba wo kikimasu ga, which means “recently the term ‘day trader’ is heard” the extracted phrase 最近デー・トレーダー saikin dē torēdā “recently day trader” should be just デー・トレーダー (dē torēdā “day trader”).

For the patterns where the target term would usually follow the pattern, a similar extraction process is possible, although in cases such as お前もこの頃はやりの物質論者だ omae mo kono goro hayari no busshitsu ronsha da), which means “you too are a faddish materialism advocate”, it would need to stop at the だ/です/である da/desu/dearu copula. In fact it would probably be best to stop the extraction at any verb; a passage such as 近頃はやりの体感安全度って、chikagoro hayari no taikan anzen dotte, which means “the recent fad of identity safety” should probably just yield 体感安全度.
9.7 Precision and Recall Issues

The establishment of precision and recall metrics in this area poses an interesting challenge. In terms of precision the testing reported above indicates that some patterns, e.g. という造語/という新語, are likely to result in fairly high levels, however if they result in a relatively small number of lexical items being collected it is of limited use in lexicon building. Casting a wider net and being prepared to sift results is probably a better course.

In terms of measuring recall the typical approach would be to identify how many terms-of-interest there are in a corpus, and test how often they are identified by the extraction method. For example the phrase 光の速度に近い hikari no sokudo ni chikai “near light speed” appears 20 times in the Kyoto Corpus and was identified once (by the という造語 pattern). Obviously far more cases need to be included in the metrics, but it is an indication that the techniques under investigation may have fairly low levels of recall for harvested terms. On the other hand, it is quite unrealistic to expect that a high proportion of the usages of a term such as 光の速度に近い would be associated with patterns such as those under investigation, and one cannot simply equate extractions as a proportion of the total occurrences as a measure of recall.

Obtaining a realistic measure of recall would ideally involve something like identifying a substantial set of out-of-lexicon terms which are in the Corpus being used and which could reasonably be expected to be candidates (new formations, in-words, etc.) and seeing how many were extracted. This would be a very difficult task with any corpus, and well-nigh impossible with one over 12 years old, such as the Kyoto
WWW Corpus.

9.8 Detailed Investigation

The initial exploration described above involved very basic *grep*-based sentence selection and manual identification and extraction of possible target terms. In order to embark on a more comprehensive evaluation, it was necessary to extend the techniques in a number of areas. These included:

a. refining the set of patterns to maximize the possibility of identifying useful target terms;

b. establishing an efficient system for identifying whether a text passage contained one of the chosen patterns;

c. automatically extracting target terms, whether they are in parentheses or not.

The corpora used for the extended testing were the 2004 Kyoto WWW Corpus of about 500 million sentences, and the collection of 870 million Japanese text passages extracted from 2014 and 2015 Twitter data as described earlier.

9.8.1 Pattern Selection

From the original set of 37 patterns, a set of 18 was chosen for further experimentation. The selection process was to choose those patterns which had resulted in the higher proportion of Group 2/3 being detected in the sampling. (See Table 9.5 and Table 9.6.)

The chosen patterns are below:
With regard to the selection of these patterns:

a. excluded from the original set were three of the more commonly occurring patterns: と言うのは/というのは, というと and といういみ/という意味. Although between them they accounted for about 80% of the of the sentence selections, they performed comparatively poorly in being associated with possibly useful terms.

b. of the chosen patterns ということば/という言葉 accounted for over 90% of the remaining extracted lines, and 最近はやりの/最近流行の/最近流行りの accounted for a further 7%. Thus the overwhelming majority of remaining extractions come from two patterns. They are among the middle-ranking performers according to the sampling, and certainly cannot be ignored. While there are other patterns which performed considerably better in the sampling
in terms of precision, the number of actual extractions associated with them is much lower.

### 9.8.2 Text Scanning and Target Term Extraction

With over a billion lines of text to examine for the presence of the language patterns a reasonably fast searching technique is desirable. The possibility of training a machine learning model was raised in Section 9.1, however since we are dealing with a constrained set of patterns a direct pattern-matching approach is clearly more appropriate. Also the nature of the patterns lends itself to a fast character-by-character search using a search tree as depicted in Figure 9.1. Initially each character in a line of text only has to be compared with こ, と, 近 and 最 to determine whether more of the tree is to be searched. Similarly at each level of the tree only a few characters typically need to be tested.

As the ということば/という言葉 pattern will also encompass the という言葉がはやって/という言葉が流行って and という言葉が流行して cases there is no need to include them in the tree.

The 500 million lines in the Kyoto Corpus had 280,574 matches with these patterns, and the 870 million tweets had 130,310 matches. The hit rate for these patterns in Twitter is thus only about 30% that of the WWW text, which is probably indicative of both the brevity of many tweets, and possibly a very different text style for longer tweets.

From the extracted lines of text, it was necessary to isolate the target terms associated with the patterns. The approach taken was:
a. divide the patterns into those where the target usually precedes the pattern (these always begin with という), and those where the target usually follows (the rest).

b. detect and extract text which occurred in some form of parentheses before or after the pattern. Japanese typically uses a greater number of parenthesis characters than European languages, with the most common, 「 and 」, playing the role of inverted commas in English. Examination of the extracted text indicated that in fact a variety of parenthesizing characters were being used
and accordingly the following parenthesis pairs were used:\(^5\)

「」, 『』, 〈〉, 《》, <>, ０, □, 「」, “”, []

Originally it was considered possible that all the parenthesized terms in a passage may be validly treated as targets, however inspection showed that terms not immediately associated with the patterns were rarely the focus of the pattern. Accordingly the extraction was restricted to parenthesized terms beginning 3 or fewer characters before or after the pattern. This margin was to allow for the occasional punctuation characters and words such as など nado “et cetera”. Also it was clear that there were occasionally quite long strings of parenthesized text, typically quotations, which were not going to be considered valid lexical items, so the extraction was restricted to strings of up to 10 characters.

c. where there are no parenthesized target strings associated with the text patterns, it is necessary to attempt to extract target terms from the text preceding or following the patterns. Inspection of a number of passages indicated that most likely candidates were made up of combinations such as noun-noun, prefix-noun, noun-suffix, adverb-noun, adjective-noun, etc., and that a reasonable heuristic would be to collect morphemes until one which typically lies on the boundary of an expression, such as a particle or a verb, was encountered. This inevitably has some limitations, for example the phrase meaning “mother nature” in Japanese is 母なる自然 hahanaru shizen (literally “mother-becoming-nature”), and would not be extracted because it contains a verb acting prenom-

\(^5\)Despite appearing identical in some fonts, 「」 and 「」 are actually different sets of Unicode codepoints. Both are used in Japanese text.
inally. Fortunately it was apparent from inspection that these particular forms of phrases were not very common.

To implement this approach, the text following or preceding the pattern was passed through the MeCab morphological analyzer and adjacent morphemes which met a limited set of part-of-speech (POS) attributes were aggregated, with any other morpheme type triggering termination. The selected MeCab/UniDic POSs are:

- 名詞 meishi (noun)
- 形容詞 keijōshi (adjectival noun)
- 形容詞 keiyōshi (adjective)
- 副詞 fukushi (adverb)
- 接頭詞/接頭辞 settōshi/settōji (prefix)
- 接尾辞/接尾詞 setsubiji/setsubishi (suffix)
- 感動詞 kandōshi (interjection)

Punctuation symbols (補助記号 hojokigō ) occurring between the pattern and the first target morpheme were ignored, however once accumulation had begun a further punctuation symbol forced its termination.

Some examples of text passages and extracted target terms for both the parenthesized and non-parenthesized cases are shown in Table 9.7 and Table 9.8.

The numbers of target terms extracted from the text collections is shown in Table 9.9.

Some general observations that can be made about these extractions are:

a. the extractions comprise a very small proportion of the text in the two collections. The passages extracted from the WWW Corpus represent only 0.056% of the text and the ones from the Twitter collection only 0.015%.
Table 9.7: Example Text Passages and Extracted Terms: Parenthesized

<table>
<thead>
<tr>
<th>Target Term</th>
<th>Text</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>絵の上手さ</td>
<td>同じ「絵の上手さ」という言葉で議論されていたが、それらは質が異なるものだ。</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>“The same ‘painting skill’ was discussed, but they are of differing qualities”</td>
<td></td>
</tr>
<tr>
<td>ナレッジ経営</td>
<td>それって最近はやりの「ナレッジ経営」のお手本みたいな話ですね。</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>“That story seems to be a model of the popular ‘knowledge management’.”</td>
<td></td>
</tr>
<tr>
<td>蟹の肉</td>
<td>奥様の「蟹の肉」という言葉に猛然と噛み付く3歳児。</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>“The 3-year-old boy who bit fiercely into the lady’s ‘crab meat’.”</td>
<td></td>
</tr>
<tr>
<td>酵素ダイエット</td>
<td>みんな最近流とりの「酵素ダイエット」はチャレンジしてみた？</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>“Everyone tried the challenge of the recently fashionable ‘enzyme diet’.”</td>
<td></td>
</tr>
<tr>
<td>高温注意情報</td>
<td>みなさんは「高温注意情報」という言葉をご存知でしょうか。</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>“Does everyone know the term ‘Warning - High Temperature’?”</td>
<td></td>
</tr>
</tbody>
</table>

9.9 Evaluation of Extracted Target Terms

The next stage in the evaluation was to examine the extracted target terms in order to assess whether they are suitable to be classified as lexical items, and to determine whether there are any other characteristics, e.g. frequency of occurrence, which may signal whether terms are worthy of consideration.
Chapter 9: Neologism Identification through Language Contexts

<table>
<thead>
<tr>
<th>Target Term</th>
<th>Text</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>皆親切</td>
<td>30分や1時間前後する程度の労働は、皆親切という言葉を持ち出すまでもなく,</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>“As for 30 minutes to an hour’s labour, it doesn’t bring about everyone’s kindness.”</td>
<td></td>
</tr>
<tr>
<td>ドラム式洗濯機</td>
<td>最近流行のドラム式洗濯機は本当に使いやすいの？</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>“Is the recently popular drum-type washing machine really easy to use?”</td>
<td></td>
</tr>
<tr>
<td>絵本ブーム</td>
<td>社会現象として絵本ブームという言葉が使われるほどに...</td>
<td>WWW</td>
</tr>
<tr>
<td></td>
<td>“To the extent that the picture book boom is used as a social phenomenon.&quot;</td>
<td></td>
</tr>
<tr>
<td>迎撃オフ</td>
<td>この界隈には迎撃オフという言葉が足ってだな...</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>“There is a term intercept off in this neighborhood...”</td>
<td></td>
</tr>
<tr>
<td>劇中舞踏</td>
<td>ただ最近流行の劇中舞踏とかは知らないので踊れませんが...</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td>“I cannot dance because I don’t know these recently popular mid-drama dances”</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.8: Example Text Passages and Extracted Terms: Not Parenthesized

The process that has been followed for each of the corpora is:

a. run the target term extraction as described above;

b. filter the extracted terms against the reference lexicon, as the aim of the investigation is determine whether the method is extracting new or unrecorded terms;

c. sort the extractions and aggregate them according to how often they occur. This is to enable evaluation of the hypothesis that more frequently-occurring terms are more likely to be potential lexical items. (A major issue with respect to the Twitter data is handling of retweets\(^6\) as these have the potential to greatly amplify the numbers of occurrences. This is discussed below.)

The extracted terms were then categorized according to several criteria:

a. the usefulness of the term as a lexical item. This involved examining the term

\(^6\)Reposted or forwarded Twitter messages.
both in the context of the text passage(s) in which it was detected, in other
text passages such as those discovered from WWW searches, and in reference
material such as glossaries which were not part of the reference lexicon. From
this categorization codes were assigned to the terms as follows:

i. A: in the reference dictionary in different surface form, e.g. partially or
   fully in kana instead of kanji;

ii. B: an inflected or variant form of existing entry;

iii. C: definitely of interest as it has the potential to be a valid lexical item;

<table>
<thead>
<tr>
<th>Source</th>
<th>Total lines</th>
<th>Extractions</th>
<th>None extracted</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW Corpus (all patterns)</td>
<td>280574</td>
<td>124371</td>
<td>45841</td>
</tr>
<tr>
<td>Twitter (all patterns)</td>
<td>130310</td>
<td>37083</td>
<td>21232</td>
</tr>
<tr>
<td>WWW Corpus (という言葉)</td>
<td>270553</td>
<td>122727</td>
<td>44715</td>
</tr>
<tr>
<td>Twitter (という言葉)</td>
<td>119871</td>
<td>36074</td>
<td>19543</td>
</tr>
<tr>
<td>WWW Corpus (最近流行りの)</td>
<td>6711</td>
<td>573</td>
<td>485</td>
</tr>
<tr>
<td>Twitter (最近流行りの)</td>
<td>7635</td>
<td>314</td>
<td>791</td>
</tr>
<tr>
<td>WWW Corpus (the rest)</td>
<td>3310</td>
<td>1071</td>
<td>641</td>
</tr>
<tr>
<td>Twitter (the rest)</td>
<td>2805</td>
<td>696</td>
<td>898</td>
</tr>
</tbody>
</table>

Table 9.9: Target Term Extraction Counts
Table 9.10: Categorizations of Extracted Text: WWW Corpus

<table>
<thead>
<tr>
<th>Categ.</th>
<th>Top 50</th>
<th>5 Times (20)</th>
<th>Once (50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>18</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>8</td>
<td>2</td>
<td>46</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

iv. D: other, e.g. a phrase not of particular interest;

v. E: corrupted text

b. whether the occurrences of the terms were parenthesized or not;

c. which pattern(s) generated the extraction. (This was done for the “C” terms.)

9.9.1 WWW Corpus

Of the 234,733 terms extracted from this Corpus, 68,644 were not in the reference lexicon. Of these 52,277 were terms that occurred only once, and the remainder occurred multiple times (the maximum was 55 times).

A detailed analysis of 120 terms was carried out as follows: the most common 50 terms (13–55 occurrences), a sample of 20 terms which occurred 5 times each, and a sample of 50 terms which occurred once each. The categorization of the terms is shown in Table 9.10.

Some examples of the extractions that were assigned to the various categories are:
Category A

- がんばれ ganbare: kana form of 頑張れ “go for it!”
- キチガイ kichigai: katakana form of 気違い “madness, etc.”
- カワイイ kawaii: katakana form of 可愛い “cute; adorable, etc.”
- ムカつく mukatsuku: variant of むかつく “to feel irritated”
- ガイジン gaijin: katakana form of 外人 “foreigner”
- エンパワメント enpawamento: variant form of エンパワーメント “empowerment”

Category B

- 愛している aishiteiru: from the verb 愛する and meaning “to be in love”
- 好きだ sukida: from 好き “love, liking” plus the copula
- がんばって ganbatte: variant of 頑張れ (above)
- 感動した kandōshita past tense of 感動する “to be moved”
- 自分らしく jibunrashiku adverbial form of 自分らしい “characteristic”

Category C

- ゲーム性 gēmusei “quality of a video game; game rating”
- 情報モラル jōhō moraru “information ethics”
- 生きる力 ikiruchikara “zest for living”
- 子育て支援 kosodate shien “child-rearing assistance”
- 共創 kyōsō “growing together; joint development”
- 病診連携 byōshin renkei “coordination of local clinics and hospitals”
- 限定品 genteihin “limited-release products”

Category D

- シンプルイズベスト shinpuru izu besuto (“Simple Is Best”: pop song name)
The relatively high proportion of “C” terms in the multiply-occurring sets (36–50%) is interesting. It might seem intuitively obvious that more commonly used or discussed terms would be more likely to be potential lexical items, but it could well not have been the case. Some more sampling of the 2, 3 and 4 batches may be appropriate, but it seems clear that multiple occurrences of a term, at least among the terms extracted by the processing being tested here, is a signal of its likelihood to be of interest.

A selection of the “C” (of interest) terms in the above samples were examined in greater depth in order to get some indication of the efficacy of the technique, and also to see if there were any other characteristics that may be used as signals as to the likelihood that terms are of interest. The selection consisted of 21 terms: the first 10 from the top 50 batch and all the others. The investigation was to determine:

a. the number of occurrences, both from the pattern extractions, and in the total Corpus;

b. the equivalent numbers in the Twitter corpus;

c. the numbers of extractions that occurred in and out of parentheses;
d. the extraction patterns which had identified the terms.

The results are in Table 9.11.

Of the approximately 300 term extractions in the table above, all except one resulted from the ということば/という言葉 pattern. A single non-parenthesized extraction of 情報モラル jōhō moraru “information ethics” was made using the いう造語 pattern.

Some observations that can be made as a result of this analysis are:

a. as stated above, it seems clear that multiple occurrences of a term, at least among the terms extracted by the processing being tested here, is a signal of its likelihood to be of interest.

b. the continued dominance of the ということば/という言葉 pattern, which contributed over 99% of the extractions.

c. about 75% of the “good” extractions came from parenthesized terms. In fact this is not very surprising as about 70% of all extractions came from parenthesized terms, so there is no particular signal associated with it.

d. the side-by-side comparison with the frequencies and extractions from the Twitter corpus seems to be indicating that even when terms are occurring in equivalent numbers in both corpora, e.g. ゲーム性, 生きる力, 子育て支援, インクルージョン, etc., the number of extractions from the Twitter corpus are significantly lower.

e. as can be seen, the proportion of extractions to total occurrences is generally
### Table 9.11: Extractions and Total Counts for Target Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>WWW extr’ns</th>
<th>WWW total</th>
<th>Twitter extr’ns</th>
<th>Twitter total</th>
<th>Non-Paren</th>
<th>Paren</th>
</tr>
</thead>
<tbody>
<tr>
<td>ゲーム性</td>
<td>38</td>
<td>12596</td>
<td>3</td>
<td>10643</td>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td>情報モラル</td>
<td>31</td>
<td>2115</td>
<td>0</td>
<td>569</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>生きる力</td>
<td>29</td>
<td>13381</td>
<td>2</td>
<td>6061</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>還浄</td>
<td>28</td>
<td>166</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>子育て支援</td>
<td>18</td>
<td>26505</td>
<td>0</td>
<td>6498</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>企業は人なり</td>
<td>18</td>
<td>350</td>
<td>0</td>
<td>37</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>インクルージョン</td>
<td>18</td>
<td>1990</td>
<td>0</td>
<td>1018</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>儀礼的無関心</td>
<td>17</td>
<td>302</td>
<td>0</td>
<td>22</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>共創</td>
<td>17</td>
<td>1254</td>
<td>1</td>
<td>960</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>薬食同源</td>
<td>15</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>ラーフラ</td>
<td>5</td>
<td>102</td>
<td>0</td>
<td>462</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>三方よし</td>
<td>5</td>
<td>56</td>
<td>0</td>
<td>170</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>よくやった</td>
<td>5</td>
<td>7841</td>
<td>0</td>
<td>45756</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>あいよかけよ</td>
<td>5</td>
<td>235</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>総合型地域スポーツクラブ</td>
<td>5</td>
<td>2112</td>
<td>0</td>
<td>76</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>ベイオフ解禁</td>
<td>5</td>
<td>2584</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>医療不信</td>
<td>5</td>
<td>501</td>
<td>0</td>
<td>70</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>フラクチュオマティクス</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>バラのつぼみ</td>
<td>5</td>
<td>215</td>
<td>0</td>
<td>283</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>口腔ケア</td>
<td>5</td>
<td>2351</td>
<td>0</td>
<td>641</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>実弾演習</td>
<td>1</td>
<td>473</td>
<td>0</td>
<td>199</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>保守王国</td>
<td>1</td>
<td>258</td>
<td>0</td>
<td>221</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>キャラクタービジネス</td>
<td>1</td>
<td>470</td>
<td>0</td>
<td>192</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
quite low, but as discussed in Section 9.7 above one must be cautious in attempting to interpret this as a low recall.

### 9.9.2 Significance of Multiple Occurrences

As can be seen in Table 9.11 the three single extractions had reasonably high counts of occurrences in the Corpus (258–473). That raises the question of whether the number of Corpus occurrences is linked or correlated to the usefulness of extracted terms. To test this a sample of 10 of the singly-occurring “C” terms was checked to determine the number of occurrences in the Corpus. 6 of these occurred fewer than 10 times and the others occurred 39, 52, 62 and 1,561 times respectively. Also checked were the Corpus counts of the 8 “D” terms in the “top 50” set. While they varied, they were noticeably lower than the “C” counts. This seems to indicate support for a (quite reasonable) hypothesis that high overall occurrence counts are related to the usefulness of extracted terms.

As a further test of this hypothesis, a set of 2,000 of the singly-extracted terms was chosen and their overall counts in the Corpus established. About 160 of these (8%) each occurred 400 or more times. Examination of a sample of 20 of these more commonly occurring terms resulted in the following category counts: B: 1, C: 14, D: 6.

This is a very different outcome to that shown by the randomly selected singly-extracted terms. From this it seems likely that a high extraction count and/or a high overall Corpus count are good indicators that an extracted term has a chance of being a term of interest.
The overall Corpus count of a term may not be a particularly useful metric as it would be difficult to obtain in a general harvesting process. They are only available with the Kyoto WWW Corpus because an $n$-gram corpus and associated utility software are available (Section 4.4). On the other hand, a useful corpus count could well be taken from a different comprehensive corpus such as the Google $n$-gram Corpus.

9.9.3 Twitter Data

A similar analysis was carried out on the text of 2014/15 Twitter data. Some additional analysis was carried out on two aspects of this data:

a. where the text passages were identified as retweets these were aggregated and a separate investigation made of the term to see if occurrence within a retweet was any different to other target terms in terms of usefulness.

b. since the Twitter text was associated with specific dates, an analysis was made to determine if identified terms were clustered and if so whether this was associated with greater usefulness.

9.9.4 Retweets

The fact that Twitter text contains retweets raises a number of issues in terms of the analysis of the text. On the one hand the retweeting can seriously distort any analysis which attempts to use frequency information with regard to such things as extracted terms (Lu et al. 2014). On the other hand the fact that a passage is being relayed by Twitter users may in itself be useful in the analysis of the passage. Neubig
and Duh (2013) have shown that retweets tend to have more information and a more consistent writing style. The impact of retweets in the present project is discussed below.

The actual identification of retweets has proved to be a significant problem. If a user simply passes on a message unchanged this is supposed to be flagged in the metadata accompanying the text. However as discussed below it can be seen that often the users make minor amendments, add or remove text, etc. before sending the message as though it were new. In some cases the (former) practice of adding “RT” to such messages has been applied. As Boyd et al. (2010) note, the conventions for retweeting are “hugely inconsistent”.

As it was considered important to identify complete or partial retweets as far as was possible, both with and without metadata or tagged indications, text passages were aggregated according to the actual Japanese text they contained. This meant removing leading and trailing non-Japanese characters as these often contained material that differed between versions of the repeated message. This was partially successful, but in many situations a significant degree of manual intervention was needed.

To illustrate this, consider the following two examples.

a. the passage これ使って俺の嫁になれ (full text below) was found to occur in a tweet which had been repeated 223 times. The tweet:

友人から「これ使って俺の嫁になれ」という言葉と共にこれを渡された俺はこれから的人生どう生きていけばいいんでしようか

*yujin karā ‘kore tsukatte orenoyome ni nare’ to iu kotoba to tomoni kore wo watasareta ore wa korekara no jinsei dō iikite ikeba ii ndeshō ka*
“I was handed this by a friend along with the expression ‘use this to be my wife’; how I am going to live my life?

occurred 217 times and the same tweet prepended by the heading 【話題の画像】 (wadai no gazō “topic image”) occurred a further 6 times.

b. the phrase これかわいい！ (see below) was found in 32 different aggregations of over 300 tweets. The most common, which occurred 191 times, was:

「これかわいい！」という言葉には、「これをかわいいと思う私ってかわいい？」という意味がある【YOSHIKI】
‘kore kawaii!’ to iu kotoba ni wa, ‘kore wo kawaii to omou watashitte kawaii?’ to iu imi ga aru
“As for the term ‘this cute’, it has a meaning ‘Am I cute in thinking this is cute?’”

Further replications of that text with the addition of names, hashtags and small amounts of text commentary occurred 53, 23, 18 and 14 times.

### 9.9.5 Analysis of Retweets

The terms extracted from retweets were aggregated and ranked according to the numbers of times the tweet was repeated in order to see if greater repetition was associated with the usefulness of the extracted term. 16 terms were extracted from tweets with over 100 repetitions and were examined. Approximately 400 terms were extracted from tweets with between 10 and 99 repetitions and a sample of 25 of these was examined. 91 tweets were repeated 5 times each and a sample of 20 of these was examined. The classification of these terms is in Table 9.12.

Among those classified as C (potential lexical item) were:

Over 100 repeats:
Chapter 9: Neologism Identification through Language Contexts

<table>
<thead>
<tr>
<th>Categ.</th>
<th>Over 100 (16)</th>
<th>10–99 Times (25)</th>
<th>5 Times (20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>14</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9.12: Categorization of Twitter Extractions: Retweets

- ラウリル酸 *raurirusan* “lauryl acid”, a component of cosmetics. This was also identified in a different set of tweets.
- 激おこぷんぷん丸 *gekiokopunpunmaru* female slang for “extreme rage”.

10–99 repeats:
- マスハラ *masuhara* abbreviation of “mathematical harassment”;
- 家事ハラ *kajihara* housework harassment;
- コントラプンクトゥス *kontorapunkutusu* contrapunctus (the tweets concerned Bach’s “Art of the Fugue”, where this term is used);
- 水無月マグダラマリア *minatsuki magudara maria* probably a reference to Mary Magdalene.

5 repeats:
- アベハラ *abehara* abbreviation of “Abe harassment” (reference to the Japanese Prime Minister);
- 足フェチ *ashi fetchi* “foot-fetish”.

These results need to be compared with those obtained from repetitions not arising from retweets (discussed below). It does not appear that the occurrence of extracted terms in retweets is a strong indication of usefulness.
Chapter 9: Neologism Identification through Language Contexts

9.9.6 Multiple Occurrence: Not Retweets

As was carried out with extractions from the WWW Corpus, a detailed analysis of 120 terms was carried out as follows: the most common 50 terms (10–98 occurrences), a sample of 20 of the 91 terms which occurred 5 times each, and a sample of 20 terms which occurred once each. The categorization of the terms is shown in Table 9.13.

Several terms such as アベハラ and 家事ハラ also occur among the retweets.

Among those classified as C (potential lexical item) were:

Top 50
- 自称芸術家 jishō geijutsuka “self-described artist”;
- 放射脳 hōshanō “obsession with the effects of radiation”;
- 女性向け joseimuke “aimed at women”;
- 異物混入 ibutsukon’nyū “contamination (esp. foreign bodies in food)”;
- クリぼっち kuribotchi “spending Christmas alone”

5 occurrences
- 逆レイプ gyakureipu “reverse rape”;
- アホノミクス ahonimikusu “Ahonomics” (idioteconomics: play on “Abenomics”);
- パイスラ paisura woman with a diagonal shoulder strap between her breasts

<table>
<thead>
<tr>
<th>Categ.</th>
<th>Top 50</th>
<th>5 Times (20)</th>
<th>Once (20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>21</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>20</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9.13: Categorization of Twitter Extractions: Multiple Occurrences
(from オッパイ oppai “breasts” and スラッシュ surasshu “slash”;
• 睡眠学習 suimingakushū “hypnopedia”;
• 反知性主義 hanshisei shugi “anti-intellectualism”

Single occurrences
• ミルクガラス miruku gurasu “milk glass”, i.e. opaque glass
• ラジオネーム rajionēmu pseudonym used on talk-back radio (transliteration of “radio name”)
• アラサーメンズ arasāmenzu fashions for men over 30 (アラサー arasā is a contraction from “around thirty” and usually refers to women)

As with the investigation of candidate term extractions from the WWW Corpus discussed above, it does appear that the number of times a term is extracted is correlated with the likelihood it is of interest.

As with the WWW Corpus terms, a sample of singly-occurring terms was checked against an n-gram corpus, in this case the Google n-gram Corpus. A selection of 2,000 singly-extracted candidate terms was matched against the Corpus and a sample of 20 of the higher-ranking terms were evaluated. The results were 7 terms ranked as A or B, 5 as C and 8 as D. While this is only a small sample, it does seem to indicate that a high count in an n-gram Corpus indicates a greater likelihood that a term is of interest.

9.9.7 Classification of Names

In contrast to the terms identified in the WWW Corpus, a significant proportion of the terms extracted from Twitter text were names, e.g. anime characters, Pokemon
characters, singers, etc. In hindsight there probably should have been a category for them. In the classifications reported above they have been treated as “D” (not of interest), but the fact they are being collected is an indication of the efficacy of the approach.

9.9.8 Issue of Parenthesized Terms

As previously described, the method for extracting possible terms involves either collecting a string of text in parentheses associated with the pattern, or collecting a string of morphemes with restricted POSs associated with the pattern. It is worth examining the relative outcomes of these two approaches.

Of the approximately 27,000 strings extracted from the Twitter text, 12,650 were parenthesized and 14,348 were not parenthesized. It is appropriate to compare the outcomes with terms extracted by the two approaches to determine if there was a qualitative difference. Accordingly a sample of 20 from each was examined to see if there were noticeable differences in the classifications. The non-parentheses approach resulted in 10 each in the C and D categories. The parentheses approach resulted in 5 in category B, 6 in category C and 9 in category D.

While these are small samples, it seems there is no clear domination of one approach over the other. The presence of “B” terms in the parenthesized sample and their absence from the non-parenthesized sample is interesting, and actually to be expected. Most of the “B” terms in the parenthesized sample, such as この世に kono yo ni “in this world” and へそが茶を沸かす heso ga cha wo wakasu “what
a joke!\textsuperscript{7}, which both contain particles (に, が, を), could not have been extracted by the non-parenthesized approach as it did not collect target terms with embedded particles.

**9.9.9 Burstiness**

As the Twitter texts have dates in their metadata it was possible to examine whether multiple occurrences were in bursts, and whether this might be associated with greater or lesser relevance.

A sample of ten non-retweet multiply-occurring extractions ranging from 16 to 48 occurrences was examined. Of the 10, 3 were clustered into a relatively short period, e.g. a few days, and the other 7 were spread over the whole period of the data. The 3 clustered terms were:

- 自称芸術家 jishō geijutsuka occurred in an outraged discussion over someone being called a “self-proclaimed artist”;
- バブみ babumi nickname of an anime illustrator;
- 夢を諦めない yume wo akiramenai title of 1987 pop song.

The 7 non-clustered terms were:

- キチガイ kichigai slangy and very common form of 気違い (“enthusiast, freak”);
- 放射脳 hōshanō “radiation obsession”;
- 女性向け joseimuke “towards women”, in a discussion of light fiction, etc. aimed at women;
- カープ女子 kâpu joshi female fans of the Hiroshima Carp baseball team.

\textsuperscript{7}Literally “one’s navel boils water for tea”.

Chapter 9: Neologism Identification through Language Contexts

- ヤバイ yabai variant of やばい/ヤバイ ("dangerous, amazing, cool").
- また涼し matasuzushi part of the kotowaza (proverb) 心頭滅却すれば火もまた
  涼し shintō mekkyaku sureba hi mo mata suzushi “suppress your ‘self’, even a
  fire is cool”.
- 毒親 dokuoya “toxic parents”.

From this is does not appear that clustered multiple occurrences of candidate
terms have any particular advantages. The clustering may indicate a degree of top-
icality of a term, as in the case of 自称芸術家 (see Table 9.14). It may, however,
lead to focus on an ephemeral term, when a greater spread of usage over time may
indicate more general usage.

9.9.10 Terms In and Out of Patterns

As discussed previously in Section 9.7 an issue in the analysis of the extraction
techniques being explored in this project is that of determining how good are they at
finding expressions in the text being used (which may or may not be related to the
recall). To probe this issue the 10 candidate terms examined above were tested to see
how often they occurred in the text, both in and out of the extraction patterns. By
way of comparison they were also checked against the WWW Corpus. The results
are given in Table 9.14.

As can be seen, in 8 of the 10 terms over half of the occurrences were identified, and
in three cases over 95% were identified. The two exceptions were the slangy katakana
versions of common adjectives (キチガイ, ヤバイ). This is a marked contrast with
the results from the WWW Corpus (Table 9.11), where the proportions of extracted
terms in almost all cases was much lower. Although the temptation to regard this as a good level of recall should be resisted, one must conclude that in the context of terms appearing in multiple tweets it seems to be quite effective.

The cross-corpus comparison is interesting, with only 4 of the 10 terms being found in the WWW Corpus. In many ways this is to be expected, as the two corpora are over 10 years apart in generation, so topical issues will be quite different. Also the nature of recurring terms can be expected to differ. One expects that things like song names and *anime* artists will be found more in Twitter text.

<table>
<thead>
<tr>
<th>Term</th>
<th>In-pattern (Twitter)</th>
<th>Total (Twitter)</th>
<th>In-pattern (WWW)</th>
<th>Total (WWW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>キチガイ</td>
<td>48</td>
<td>109</td>
<td>39</td>
<td>62</td>
</tr>
<tr>
<td>自称芸術家</td>
<td>25</td>
<td>27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>放射脳</td>
<td>24</td>
<td>40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>女性向け</td>
<td>24</td>
<td>35</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>バブみ</td>
<td>19</td>
<td>27</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>カープ女子</td>
<td>19</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ヤバイ</td>
<td>18</td>
<td>85</td>
<td>9</td>
<td>32</td>
</tr>
<tr>
<td>夢を諦めない</td>
<td>18</td>
<td>19</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>また涼し</td>
<td>18</td>
<td>19</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>毒親</td>
<td>16</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9.14: In-pattern and Total Term Counts
9.10 Summary, Discussion and Conclusions

9.10.1 Summary

This chapter has described an investigation into techniques for extracting potential neologisms based on the language contexts in which they occur. The development and evaluation of the techniques has made extensive use of two large text collections: the Kyoto WWW Corpus and a collection of Twitter data.

The initial part of the investigation was to establish detectable text patterns typically associated with words and terms under discussion. A set of 37 such patterns was investigated and after examination of sample extractions, 18 of the patterns were used in a complete analysis of the corpora. Two these patterns: ということば/という言葉 and 最近はやりの/最近流行の/最近流行りの accounted for 97% of the total extractions.

Evaluation of the extracted terms established that terms which occurred multiple times in the corpora were frequently of considerable interest as potential neologisms, and were associated with a reasonably high precision.

9.10.2 Discussion and Conclusions

From the investigations described above, a number of conclusions can be drawn and observations made about the techniques being investigated. Among them are:

a. it is clear that the technique is quite effective in highlighting terms suitable for further investigation, as it is identifying candidates that are often very worthy of detailed examination and subsequent lexicalization.
b. it is interesting and not a little frustrating that after all the early work in identifying useful text patterns for identifying possible terms, the outcome has been so totally dominated by two patterns, to the extent that the others may as well be ignored. Several of the other text patterns have demonstrably better precision, but their recall of useful terms is so low as to make them of little use in a practical harvesting exercise. (That is no reason, of course, to exclude them as they add little overhead to process and at the margin can improve the outcome.)

c. the technique can clearly be enhanced by association with an $n$-gram corpus with frequency counts. A term, particularly one which has not been extracted often, is much more likely to be a useful candidate if it has a high $n$-gram count.

d. at present we have no real indication of the recall of the techniques being investigated. Objective analysis of recall would be a major task and best left for future work.

9.10.3 Future Work and Lexicographic Application

In terms of future work on the term identification techniques investigated and reported in this chapter, there are several areas which could warrant further investigation:

a. the identification and testing of additional language patterns which may identify neologisms and other terms of interest. While the net has been cast fairly widely in the present study, there is always the possibility that the approach may be
enhanced by using additional terms.

b. the techniques for identifying and extracting non-parenthesized target terms could probably be improved. The approach taken in this study is relatively basic and based solely on a range of POS tags associated with the morphemes in the target zone. A deeper analysis would probably lead to more complete terms being identified. For example at present the extraction stops when a verb is encountered, however if it were to accept verbs acting prenominally, as the example of 母なる自然 hahanaru shizen “mother nature” quoted earlier, the overall quality of the extraction may improve.

In terms of the application of this term identification technique to lexicographic activities, it would be relatively simple to use it as a tool for scanning large quantities of text and in association with a reference lexicon and an n-gram corpus produce lists of potential lexical items for future analysis. As discussed the search technique is quite fast as it does not require pre-processing such as use of a morphological analyzer. It could also be attached to Internet text sources, such as Twitter, RSS feeds, etc. for more real-time monitoring of contemporary text.

---

8Really Simple Syndication: a standardized system for the distribution of content from online publishers to Internet users.
Chapter 10

Summary, Conclusions and Future Work

10.1 Summary

This thesis presents the outcomes of a research program into the extraction by natural language processing techniques of neologisms from Japanese corpora. The research aim was, in particular, to establish techniques which can be developed and exploited to assist significantly in neologism extraction for compiling monolingual and bilingual dictionaries. The topic straddles three areas; computational linguistics and natural-language processing; lexicography; and aspects of Japanese linguistics.

The particular challenge of the research lies in the nature of Japanese orthography, in particular the lack of word separation in text. This leads to the extensive use of morphological analysis systems to segment text, and relies on lexicons of known morphemes. Thus the focus is on the identification of new or unrecorded morphemes,
and new or unrecorded but relevant morpheme collocations, typically expressions, compounds and phrases.

Three broad approaches have been investigated:

a. the analysis of sequences of morphemes in Japanese text to detect unrecorded potential morphemes or larger structures such as compounds and phrases which may be candidates for lexicalization. The approaches taken include using machine-learning based on training texts marked up with known and potential lexical items along with lexical and corpus-frequency features, and rule-based chunking based on the parts-of-speech of the morphemes.

Morpheme identification and analysis was also carried out on loanword MWEs as a separate sub-project.

b. a synthesis approach which largely mimics Japanese morphological processes.

The particular processes investigated include:

i. morpheme aggregation, e.g. compound verbs, compound nouns;
ii. affixation of known prefixes and suffixes to existing terms;
iii. abbreviation, using common methods of ellipsis in Japanese;
iv. generation of artificial potential morphemes and terms using typical orthographical elements such as kanji;

with Japanese n-gram corpora being used extensively to evaluate the synthesized terms.

c. analysis of Japanese texts for the types of patterns associated with new and unusual words, and extracting and evaluating the “target” terms. The work carried out included the identification of appropriate language patterns, and
the analysis of large WWW and Twitter-derived Japanese corpora.

In general each of the approaches described above results in sets of potential lexical items which are not currently in the reference lexicon, accompanied by additional information such as their frequency of occurrence and where possible samples of text where they are used. These sets could be passed to lexicographers for assessment as to whether they are indeed worthy of inclusion in some lexicon.

The results of the investigations and the efficacy of the various approaches are discussed in the following section.

10.2 Conclusions

As the goal of the work is to extract terms which can potentially be added to monolingual and bilingual dictionaries, the ultimate test is whether the extracted terms are actually useful in this role. The evaluation of such terms can be quite subjective, and is typically based on the types of criteria discussed in Chapter 2, particularly in Section 2.7. Such detailed assessment is beyond the scope of this project, which is of necessity confined to identifying potential lexical items according to broad criteria.

In terms of the effectiveness of each of the three general approaches investigated, the following conclusions can be drawn:

10.2.1 Morpheme Analysis Approach

The following conclusions can be made about the identification of potential lexical items using morpheme analysis:
a. using a machine-learning system to identify lexical items in text was reason-
able success, with generally satisfactory levels of precision and recall being
achieved. Use of additional features involving combinations of lexical and corpus
frequency data resulted in small improvements over models based on morpheme
and POS features alone. These results were based on the identification of known
lexical items in text.
b. a rule-based chunker, which was restricted to identifying nouns, verbs and ad-
jectives, achieved lower levels of precision and recall. This was not unexpected
as the restricted range would inevitably result in a lower recall.
c. when tested with unseen texts, and using an assessment based solely on the
identification of possibly useful unlexicalized terms, the rule-based chunker was
generally more successful than the machine-learning model(s) in selecting such
terms. This outcome was a surprise, but on reflection it is probably to be
expected that a more relaxed approach would be more effective than an ML
system trained on generally more focussed items.

The development of a corpus-based morpheme segmentation system for loanword
MWEs was quite successful, achieving higher levels of precision and recall than pre-
viously reported approaches (e.g. the F-measure in one test improved from 87.56 to
97.37. See Section 6.8.1. It also generated a high level of acceptable English transla-
tions of the MWEs.
10.2.2 Synthesis Approach

In general terms the synthesize-and-test approach, with testing initially using \( n \)-gram corpora to establish frequency of use and associated criteria, followed by more detailed investigation of the potential term, was quite successful. The conclusions from the individual investigations are:

a. the approach of generating artificial 2-\( kanji \) compounds then evaluating those with significant \( n \)-gram corpora occurrences, both for the compound itself and in association with adjoining particles achieved quite satisfactory levels of recall. Candidates were classified both by training an ML model and by heuristics, with similar outcomes in terms of detecting known valid terms and potential unrecorded terms. While the actual proportion of synthesized terms proceeding into evaluation is quite small at about 1\%, the technique appears useful as a means for detecting unrecorded terms which are in use.

b. a similar generation approach for 4-\( kanji \) compounds had a broadly similar outcome, although as expected the yield of useful compounds was lower. In this case the ML models did not perform as well as using heuristics to select compounds. The conclusion was that commonly occurring compounds were quite likely to be valid, and the challenge lay in determining if they were worthy of lexicalization, e.g. because of idiomatic meanings.

c. the investigation of the synthesis of abbreviated terms based on multi-morpheme compound nouns revealed that the approach had very little to offer in terms of identifying unlexicalized abbreviations. This was something of a surprise, as initial indications had been that it may have been a useful technique.
d. similarly the investigation of affixation indicated that there was little prospect of it being a successful technique for neologism generation. Clearly the process is so productive and the resulting lexical items are generally of such predictable meanings that it is not worth pursuing the approach for any distance.

The sub-project dealing specifically with the synthesis of compound verbs was clearly very successful in identifying a large collection of such verbs and validating their usage.

### 10.2.3 Language Pattern Approach

The investigation into the analysis of Japanese text for language patterns associated with new terms achieved a degree of success in that it was able to detect the presence of significant numbers of such terms. Not surprisingly it was demonstrated that multiple occurrences of such terms, either in the texts themselves or in $n$-gram corpora, was a strong signal of the term’s validity. It was clear that the technique only yields small numbers of terms from large quantities of text and that the recall is probably very low, however it may be suitable for use in conjunction with high-volume text flows such as Twitter feeds, blogs and RSS feeds.

### 10.2.4 Closing Observations

An observation that can be made based on the investigations reported in this thesis is that, in general, classification and extraction techniques using established machine-learning systems have not performed significantly better than the application of tailored heuristics, and in some cases have not performed as well. Investigation of
the reasons for this happening are outside the scope of the current work, but it may well be a fruitful topic for future work.

Another observation is that a study of this variety depends on the ready availability of Japanese corpora; both text and \( n \)-gram. It has turned out that access to such corpora, especially to large quantities of text, is quite difficult. Anecdotal evidence is that this is in part due to the relatively restrictive nature of Japanese copyright law. Research requiring quantities of Japanese text has little choice but to use publicly available sources such as the WWW, Twitter, etc.

\textbf{10.3 Future Work}

The future work leading from this thesis breaks into two groups: additional related research, and application of the approaches and techniques which have been investigated to practical lexicography.

\textbf{10.3.1 Additional Research}

A number of areas where the research conducted as part of this project could be extended are readily apparent, and have been discussed in the concluding sections of the relevant chapters. They include:

a. further investigation of the application of machine learning techniques to the extraction of Japanese neologisms. Aspects worth of further attentions are:

i. investigation of alternative techniques for chunking morphemes with the goal of better identification of potential lexical items. More recently reported work such as Schneider \textit{et al.} (2014) and in particular Qu \textit{et al.}
(2015) indicate some possible directions as they develop methodologies for carrying out exhaustive MWE identification which may overcome the weaknesses in ML-based lexical item identification discussed in this thesis;

ii. similar investigation of alternative techniques for classification of synthesized terms. Approaches worth exploring are reported in Schneider et al. (2016) as the SemEval task involved jointly identifying and super-sense tagging MWEs, with some synergy between the two;

iii. investigation of the potential for the application of deep learning to neologism extraction. The recent work on the training of an RNN at the character level to predict OOVs is worth pursuing in this respect (Pinter et al. 2017).

b. further investigation of automatic translation of extracted lexical items. Among these are:

i. generation of translations of Japanese compound verbs, possibly following the approach described in Section 8.7;

ii. further work on the translation of loanword MWEs with a view to improving the success rate of the approach reported here. An expansion into handling named-entities is worth considering, perhaps following some of the approaches indicated in Duan et al. (2016);

iii. investigation of the corpus-based translation validation approach in the context of possible translations of noun-noun compounds.

c. further investigation of the use of language patterns to identify neologisms, in particular:
i. investigation to attempt to identify additional suitable patterns;

ii. investigation into improving the extraction of non-parenthesized target terms;

iii. further and more detailed investigation of text passages containing neologisms to ascertain if there are contexts suitable for training a machine-learning system to assist in the identification of neologisms.

d. investigation of diachronic usage of potential lexical items, both in terms of identifying emerging terms and identifying the changes in the meanings of recorded terms (Cook et al. 2014; Frermann and Lapata 2016). This investigation was initially planned to be a component of the current project had time permitted. The investigation would need to have access to bodies of text associated with particular publishing dates. One possible source of such text is the Japanese Wikipedia, which is available in archival form over a number of years.

10.3.2 Application to Lexicography

As discussed in Section 1.4, a particular focus of the work described in this thesis has been the candidate’s involvement in the establishment and development of the JMdict Japanese-Multilingual dictionary, as it provided the impetus for the research program, and has given the work on neologism extraction a strong link with practical lexicography.

In several areas the techniques which have been identified and investigated, and the tools that have been developed could readily be applied to the task of generating sets of possible lexical items for consideration by lexicographers and assisting lexicog-
raphers with their assessment. Among the tasks that could be undertaken, and are
now part of the forward plans for the JMdict project, are:

a. application of the morpheme analysis techniques described in Chapter 5, in
   particular the rule-based chunker, to larger quantities of contemporary text.
   The particular challenge here will be to obtain access to such quantities of
   text as many of the primary sources, such as newspapers, invariably require
   subscriptions in order to access their archives;

b. carrying out a full generation, filtering and classification of synthesized 2- and
   4-kanji compounds as described in Chapter 7 to produce sets of potential lexical
   items;

c. implementing a trial of the language pattern system described in Chapter 9 on a
   large contemporary text feed, such as Twitter. The trial would typically include
   accumulation and diachronic analysis of target words and cross-checking with
   an n-gram corpus.

The potential lexical items thus identified would then be evaluated, perhaps in
process involving an initial crowd-sourced investigation followed by final assessment
by lexicographers.
Bibliography


——, and STUART McGRAW. 2013. JMdictDB - online development and maintenance system for the Japanese-Multilingual dictionary. In ASIALEX 2013 - Lexicography and Dictionaries in the Information Age.


NAKAZAWA, TOSHIKI, DAI SUKE KAWAHARA, and SADAO KUROHASHI. 2005. Automatic acquisition of basic katakana lexicon from a given corpus. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*.


QU, LIZHEN, GABRIELA FERRARO, LIYUAN ZHOU, WEIWEI HOU, NATHAN SCHNEIDER, and TIMOTHY BALDWIN. 2015. Big data small data, in domain out-of-domain, known word unknown word: The impact of word representations on sequence labelling tasks. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*.

BIBLIOGRAPHY


——, DIRK HOVY, ANDERS JOHANSEN, and MARINE CARPUAT. 2016. Semeval-2016 task 10: Detecting minimal semantic units and their meanings (dimsum). In SemEval@ NAACL-HLT.


Appendix A

Rule-Based Chunker - Encapsulations

A.1 Introduction

This appendix details the encapsulation rules which define the operation of the rule-based chunker described in Section 5.5. For each of the types of lexical items which are the target for the chunker: nouns, verbs and adjectives, the following rules are defined:

a. the POS types acceptable for the components of the chunk;

b. the POS types acceptable to precede and follow the chunk.
A.2 Unidic POS Tags

The Unidic morpheme lexicon has a set of POS tags which are associated with each morpheme. The tags are hierarchical, for example the word for school (学校 gakkō) has the tags 名詞, 普通名詞, 一般 - “noun, common, general” and the surname Suzuki (鈴木) has the tags 名詞, 固有名詞, 人名, 姓 - “noun, proper, person, family”.

The POS tags used by the chunker, along with their translations are in Table A.1 and Table A.2.1

A.3 Nouns

For noun/NP chunks, the accepted morphemes are those with the POS tags: 名詞 (noun), 接頭辞 (prefix), 接尾辞 (suffix), 形状詞 (adjectival noun) and 代名詞 (pronoun).

The morphemes used to identify noun chunks are in Table A.3 through Table A.7.

A.4 Verbs

For verb chunks, the accepted morphemes are those with the POS tags: 動詞 (verb) and 助動詞 (auxiliary verb).

The morphemes used to identify verb chunks are in Table A.8 and Table A.9.

---

1A more complete list with translations is currently at https://gist.github.com/fasiha/59fd33c2075b10fdd0df.
A.5 Adjectives

For adjective chunks, the accepted morphemes are those with the POS tags: 形容詞 (adjective), 接尾辞 (suffix) and 助動詞 (auxiliary verb).

The morphemes used to identify adjective chunks are in Table A.10.
# Appendix A: Rule-Based Chunker - Encapsulations

<table>
<thead>
<tr>
<th>POS</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>空白</td>
<td>white-space</td>
</tr>
<tr>
<td>形容詞</td>
<td>adjective</td>
</tr>
<tr>
<td>形容詞, 一般</td>
<td>adjective, general</td>
</tr>
<tr>
<td>形容詞, 非自立可能</td>
<td>adjective, bound</td>
</tr>
<tr>
<td>助詞</td>
<td>particle</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>particle, case-marking</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>particle, binding</td>
</tr>
<tr>
<td>助詞, 終助詞</td>
<td>particle, phrase-final</td>
</tr>
<tr>
<td>助詞, 準助詞</td>
<td>particle, nominal</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>particle, conjunctive</td>
</tr>
<tr>
<td>助詞, 副助詞</td>
<td>particle, adverbial</td>
</tr>
<tr>
<td>助動詞</td>
<td>auxiliary verb</td>
</tr>
<tr>
<td>接頭辞</td>
<td>prefix</td>
</tr>
<tr>
<td>接尾辞</td>
<td>suffix</td>
</tr>
<tr>
<td>動詞</td>
<td>verb</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>verb, general</td>
</tr>
<tr>
<td>動詞, 非自立可能</td>
<td>verb, bound</td>
</tr>
<tr>
<td>副詞</td>
<td>adverb</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>supplementary symbol, general</td>
</tr>
<tr>
<td>補助記号, 括弧開</td>
<td>supplementary symbol, bracket-open</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>supplementary symbol, bracket-close</td>
</tr>
<tr>
<td>補助記号, 仮点</td>
<td>supplementary symbol, period</td>
</tr>
<tr>
<td>補助記号, 漢字</td>
<td>supplementary symbol, punctuation</td>
</tr>
</tbody>
</table>

Table A.1: POS Tags Used By Chunker
<table>
<thead>
<tr>
<th>POS</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>名詞</td>
<td>noun</td>
</tr>
<tr>
<td>名詞, 固有名詞</td>
<td>noun, proper,</td>
</tr>
<tr>
<td>名詞, 固有名詞, 一般</td>
<td>noun, proper, general</td>
</tr>
<tr>
<td>名詞, 固有名詞, 人名, 姓</td>
<td>noun, proper, person, family</td>
</tr>
<tr>
<td>名詞, 固有名詞, 人名, 名</td>
<td>noun, proper, person, first-person</td>
</tr>
<tr>
<td>名詞, 固有名詞, 地名, 国</td>
<td>noun, proper, place, country</td>
</tr>
<tr>
<td>名詞, 数詞</td>
<td>noun, numeral</td>
</tr>
<tr>
<td>名詞, 普通名詞</td>
<td>noun, common</td>
</tr>
<tr>
<td>名詞, 普通名詞, サ変可能</td>
<td>noun, common, suru-verbal</td>
</tr>
<tr>
<td>名詞, 普通名詞, 一般</td>
<td>noun, common, general</td>
</tr>
<tr>
<td>名詞, 普通名詞, 副詞可能</td>
<td>noun, common, adverbial</td>
</tr>
<tr>
<td>形状詞</td>
<td>adjectival noun</td>
</tr>
<tr>
<td>代名詞</td>
<td>pronoun</td>
</tr>
<tr>
<td>連体詞</td>
<td>adnominal</td>
</tr>
<tr>
<td>EOS</td>
<td>end-of-string</td>
</tr>
</tbody>
</table>

Table A.2: POS Tags Used By Chunker (continued)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>空白</td>
<td>助詞</td>
</tr>
<tr>
<td>空白</td>
<td>空白</td>
</tr>
<tr>
<td>空白</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>空白</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>空白</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>空白</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>空白</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>空白</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>補助記号, 括弧開</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>補助記号, 括弧開</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>助動詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>助動詞</td>
<td>助動詞</td>
</tr>
<tr>
<td>助動詞</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>助動詞</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助動詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 副助詞</td>
</tr>
</tbody>
</table>

Table A.3: Encapsulating Morpheme Pairs (Nouns)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 括弧開</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 一般</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>助動詞</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>補助記号, 括弧開</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>助詞, 副助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 副助詞</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>助詞, 副助詞</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>助動詞</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>助動詞</td>
</tr>
</tbody>
</table>

Table A.4: Encapsulating Morpheme Pairs (Nouns) (continued)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>補助記号, 一般</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>補助記号, 一般</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>空白</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>助動詞</td>
</tr>
<tr>
<td>EOS</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>EOS</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>EOS</td>
<td>助詞, 接続助詞</td>
</tr>
<tr>
<td>EOS</td>
<td>助動詞</td>
</tr>
<tr>
<td>EOS</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>EOS</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>EOS</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>形容詞, 一般</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>形容詞, 非自立可能</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>補助記号, 括弧閉</td>
<td>補助記号, 読点</td>
</tr>
</tbody>
</table>

Table A.5: Encapsulating Morpheme Pairs (Nouns) (continued)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>補助記号, 括弧開</td>
<td>補助記号, 一般</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助詞, 接続助詞</td>
</tr>
<tr>
<td>形容詞, 一般</td>
<td>助詞, 接続助詞</td>
</tr>
<tr>
<td>形容詞, 一般</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>補助記号, 括弧開</td>
<td>助詞, 接続助詞</td>
</tr>
<tr>
<td>補助記号, 括弧開</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>補助記号, 括弧開</td>
<td>助動詞</td>
</tr>
<tr>
<td>副詞</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>副詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>副詞</td>
<td>助動詞</td>
</tr>
<tr>
<td>副詞</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>連休詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>連休詞</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>動詞, 非自立可能</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>動詞, 非自立可能</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>動詞, 非自立可能</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>動詞, 非自立可能</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>助詞, 係助詞</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>助詞, 副助詞</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>補助記号, 一般</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助動詞</td>
</tr>
</tbody>
</table>

Table A.6: Encapsulating Morpheme Pairs (Nouns) (continued)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>助詞, 係助詞</td>
<td>動詞, 非自立可能</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>補助記号, 括弧開</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>助動詞</td>
</tr>
<tr>
<td>動詞, 一般</td>
<td>動詞, 非自立可能</td>
</tr>
</tbody>
</table>

Table A.7: Encapsulating Morpheme Pairs (Nouns) (continued)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 固有名詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>形容詞, 一般</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>EOS</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>名詞, 数詞</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>助詞, 接続助詞</td>
<td>名詞, 普通名詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 括弧閉</td>
</tr>
<tr>
<td>名詞, 普通名詞, 一般</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>名詞, 普通名詞, サ変可能</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>名詞, 普通名詞, 一般</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>名詞, 普通名詞, サ変可能</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 読点</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 一般</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>接頭辞</td>
</tr>
<tr>
<td>EOS</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>EOS</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>EOS</td>
<td>名詞, 普通名詞, サ変可能</td>
</tr>
</tbody>
</table>

Table A.8: Encapsulating Morpheme Pairs (Verbs)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>名詞, 固有名詞, 人名, 姓</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 固有名詞, 人名, 姓</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 準体助詞</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 普通名詞, 副詞可能</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 固有名詞, 一般</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 固有名詞, 地名, 国</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>助詞, 係助詞</td>
</tr>
</tbody>
</table>

Table A.9: Encapsulating Morpheme Pairs (Verbs) (continued)
<table>
<thead>
<tr>
<th>Preceding Morpheme</th>
<th>Following Morpheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>補助記号, 読点</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>名詞, 普通名詞, 一般</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>助詞, 終助詞</td>
</tr>
<tr>
<td>助詞, 係助詞</td>
<td>補助記号, 句点</td>
</tr>
<tr>
<td>補助記号, 読点</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>補助記号, 一般</td>
<td>助詞, 格助詞</td>
</tr>
<tr>
<td>助詞, 格助詞</td>
<td>補助記号, 括弧閉</td>
</tr>
</tbody>
</table>

Table A.10: Encapsulating Morpheme Pairs (Adjectives)
Appendix B

Lexical Item Tagging System

B.1 Introduction

This appendix contains material supporting the task of marking potential lexical items in text, as described in Chapter 5.7.4. The material consists of:

a. the main introductory WWW page to the online mark-up system, which includes an illustration of the process. The page has some colours and shading which will not be evident in the printed version. The original page can be seen at: http://nihongo.monash.edu/cgi-bin/annotate/instructions

b. an FAQ page for the people carrying out the task;

c. an extended set of guidelines for the mark-up process.

Note that in the instructions the term “lexeme” has been used. As discussed in Chapter 2, it was later decided to use “lexical item” for this purpose.
B.2 Introductory WWW Page

Japanese Sentence Annotation: Tagging Lexemes

Getting Started

1. Register your name (link above).
2. Read the following instructions and the FAQ.
3. Go and start looking at sentences ("Sentences to Annotate" link above).

Background

This annotation task has the aim of preparing test data for a Machine Learning experiment in the identification of new lexemes, i.e. words, noun compounds, expressions, etc. in Japanese text which are not (yet) recorded in dictionaries/lexicons, but which could reasonably be expected to be recorded. We need a set chosen-by-humans lexemes with which to test the system.

A set of sentences have been extracted from Japanese newspapers (mostly the Mainichi Shimbun) which may contain unrecorded lexemes. What we want to do is tag such possible lexemes so that the ML system can be tested appropriately.

The Annotation Task

The tagging or annotation process is being carried out via a set of very simple WWW scripts. There is a basic form which shows the sentence and any prior annotations, and lets you propose further annotations of possible lexemes, comment on annotations made by others or simply indicate that you think there are no more lexemes to tag.

Before beginning to add annotations, please register your name or nickname (link at the top of this page). Although this is optional, it will help if I want to follow up any questions about your annotations.

The parts of the form are:

- the sentence itself. For convience the sentence has been processed by WWWJDIC's text glosser and the meanings of words/phrases can be seen by hovering the mouse pointer over the sentence.
- the sentence with the morphemes, i.e. the basic word elements, indicated in alternating colours. There is more about morhemes in the FAQ (link above). The splitting into morphemes has been done using MeCab and the UniDic lexicon, which results in a fine-grained identification of morphemes.
  
  Example: 職場や家庭での喜怒哀楽を、600字ほどにまとめて投稿してください。
  Here 喜怒哀楽, ほどに, まとめて, etc. are identified as existing multi-morpheme lexemes.
Appendix B: Lexical Item Tagging System

294

Japanese Sentence Annotation: Tagging Lexemes http://nihongo.monash.edu/cgi-bin/annotate/instructions

- annotations made so far on this sentence (if any).
- a list of morphemes with checkboxes to enable them to be selected as part of a possible lexeme. To propose additional lexemes, or to make a comment:
  - click on the checkboxes under the first and last morphemes of the sequence you think should make up a lexeme. Typically 2, 3 or 4 morphemes would be selected.
  - add a comment as to why you think it's a useful/valid lexeme (e.g. appears in Eijiro/ALC, has xxx Google hits, etc.)
  - click on Submit.

The process is repeated for each lexeme. Note that we are only interested in lexemes that made up from two or more morphemes.

- If you make a mistake, you can remove your own annotations by clicking on the Delete button. (If you think someone else's annotation is wrong, enter a comment about it.)
- If you see no further lexemes to mark, e.g. a previous annotator may have marked the only ones, you can record your agreement by clicking on the "Mark" button.
- Once you finish with a sentence, click on Finish. This unlocks the sentence and makes it available for others to review. If you navigate away from the annotation page without clicking on Finish or Mark, the sentence will remain locked and inaccessible for 30 minutes. After that time you can return to it.

A Worked Example

Let's say we have the following passage in a sentence:
Sentence: 数年前から花火の紹介記事を新聞や雑誌に書くようになった。
Morphemes: 数年, 前, から, 花火, の, 紹介, 記事, を, 新聞, や, 雑誌, に, 書く, ようになった。
Lexemes: 数年前から花火の紹介記事を新聞や雑誌に書くようになった。

This passage has several lexemes: 花火, 新聞, 雑誌, etc. which are single morphemes so they don't interest us. 数年 is a multi-morpheme lexeme, but it's already marked. A strong possibility is 紹介記事 (introductory article), so we decide to mark it. To do this we go to the section labelled: "Select Lexeme and/or Add Comments" and click the checkboxes under the 紹介 and 記事 morphemes, add a comment such as "gets 5 million Google hits" and click on "Submit".

Select Lexeme and/or Add Comments

数年 から か は 火 の に 介 記 紹 事 部 に 書 に く よ う に な っ た 。

Comments:

gets 5 million Google hits

After that if there are no more lexemes to mark, click on the "Mark" button.

Choosing Sentences

A list of sentences can be reached via the "Sentences to Annotate" link. You will...
return to this list when you finish a sentence. The sentences are in blocks of 100. Simply click on the sentence number to start working on it. Feel free to choose either sentences with no annotations, or ones with prior annotations.

I will be monitoring the sentence collection and removing sentences when the annotation appears to be complete.

Choosing Lexemes

Deciding whether a group of morphemes should be regarded as a lexeme is a value judgement. Some of the reasons for concluding that it should be are:

a. It already appears in one or more dictionaries;
b. It gets a reasonable number of hits in a WWW search engine. (If you are Googling for it, be sure to put it in quotation marks, as otherwise Google may split it back to morphemes and search for them independently.)
c. Its meaning is not immediately apparent from the aggregated meanings of the morphemes themselves.
d. Its meaning is often associated with a single English word, or a very common phrase in English.

Thank you very much for your assistance. Feel free to email me if you have any questions or comments (jimbreen@gmail.com)

Jim Breen
July 2011
B.3 FAQ Page

Japanese Sentence Annotation: Tagging Lexemes

Frequently Asked Questions

[Q] Can you explain a bit about morphemes vs lexemes.
[A] OK:

- morphemes are the smallest meaningful components of a word that have semantic meaning. They really can’t be subdivided further without having just letters, characters, etc. (In Japanese you could argue that kanji are morphemes themselves, but it is more useful to stop short of that.) A word like 高校生 is regarded as having two morphemes: the nouns 高校 and 生.
- lexemes are the words or groups of words you put in a dictionary. They consist of one or more morphemes. 高校生 is a typical lexeme.

[Q] In the sentence “今、私はわが人生最良の日々を過ごしています。” I see that "人生最良" is not coloured as a multi-morpheme lexeme, yet it is in JMdict/EDICT. Why is that?
[A] The version of JMdict used for that flagging/colouring has been frozen as at late 2010. Entries added since then are being used as targets in the experiment. 人生最良 is a relatively recent addition. That’s one of the ones that we’d like to be tagged.

[Q] What about lexemes that are already coloured?
[A] Don’t mark them again.

[Q] Can I make a lexeme out of other marked lexemes?
[A] Certainly. 有機塩素化合物 is a lexeme made from 有機塩素 and 化合物, both of which are multi-morpheme lexemes.

[Q] What about numbers such as 600? It might be useful for a dictionary to have them because of the pronunciation.
[A] Don’t mark them. Numbers like that aren’t in the scope of this experiment.

[Q] I see that 数年 is being called a multi-morpheme lexeme and 雑誌 is not. Surely they are both lexemes?
[A] Yes, they are both lexemes, but they are constructed a little differently. The MeCab morphological analyzer is treating 雑誌 as a single morpheme, i.e. it’s a regular jukugo constructed from two kanji. It’s treating 数年 as a pair of morphemes (prefix plus noun). The difference between these two can be quite subtle. The MeCab/UniDic combination is quite strict in its reduction to morphemes (see below.)

[Q] What are “MeCab” and “UniDic”?
[A] MeCab is an open-source Japanese morphological analysis package. (See its homepage.) It segments Japanese text into morphemes (Wikipedia: a morpheme is the smallest conceptual meaningful component of a word, or other linguistic unit, that has semantic meaning) and attaches part-of-speech tags to each morpheme. Written by Taku Kudo, MeCab uses the relatively new Conditional Random Field (CRF) model. It is the third in a series of research-oriented morphological analysis...
packages developed in Japan, the others being Juman and ChaSen.

Japanese morphological analysis packages depend on extensive morpheme lexicons for their operation. Several of these have been developed over the years, including the Juman lexicon, IPADIC and NAIST-JDIC. The most recent and in many ways the most comprehensive is UniDic, which is being developed by a team from several universities and the National Institute for Japanese Language and Linguistics (formerly the National Language Research Institute), familiarly known as "Kokken". In the compilation of UniDic a strict approach has been taken to the identification of morphemes. For example, earlier lexicons usually included 日本語 as a single morpheme, whereas in UniDic, and hence in segmentation by MeCab or ChaSen, 日本 and 語 are the only relevant morphemes.

Q: Some of the flagged lexemes seem odd. For example "梅の" - is that really a lexeme?
A: For various aspects of this work we are using a monster dictionary which combines Kojien, Daijirin, JMDict, Enamdict, the Kenkusha JE (5th ed.), etc. It has been used to flag the lexemes and also to prepare training data for the machine learning. One of the dictionaries is the GoiTaikei from NTT. It is the source of the "梅の". Hopefully there are not too many like that.
B.4 Guidelines for Assessing Potential Lexical Items

B.4.1 Introduction

Several systems for automatically identifying potential lexical items in Japanese text have been developed and are being tested. The overall process is envisaged as having two stages:

a. an automated stage for the identification/extraction of potential items in text. These would typically be filtered to remove items which have already been lexicalized, or which have been placed on a stop-list (e.g. because they have already been considered and rejected).

b. a manual stage for the detailed examination of the potential items to determine their suitability for lexicalization. This can be seen as having two components:

i. checking that the potential lexical items conform to the syntactic structures typically associated with lexical items, e.g. simple words (nouns, verbs, adjectives, etc.), expressions, multiword compounds, etc. This check would consist of examining the items in the context of the text passages where they occur.

ii. checking that the items which pass the step above are in fact suitable for lexicalization, e.g. that phrases or compound nouns have idiomatic meanings and are not purely compositional, and that the item occurs commonly enough to warrant recording.

In operation a system for harvesting lexical items would probably combine the
two components of (b) above, but in testing it is important to separate them as the decisions made in (b) (i) determine the precision of the automated process at (a). These guidelines are intended to assist annotators carrying out the evaluation as per (b) (i) to assess whether potential lexical items are suitable to passed to the next stage of evaluation.

B.4.2 Detailed Guidelines

Nouns, including Compound Nouns

Single kanji, sequences of two or more kanji, hiragana sequences and katakana sequences occurring in an appropriate syntactic context should be accepted, as should sequences consisting of a mixture of kanji and kana.

The contexts indicating acceptance include:

a. at the start of sentences, followed by a particle such as は, が, で, に, etc.;

b. encapsulated by case particles or punctuation characters;

c. preceded by an adjective (-い, -な, -たる) or a verb acting prenominally.

d. followed by する or inflections such as して, しない, etc.

Example: in the sentence なお、右案の趣旨説明は、谷垣文部大臣が行います. the strings 右案, 趣旨説明 and 谷垣文部大臣 were proposed. All should be accepted (the decision that 谷垣文部大臣 is 谷垣 + 文部大臣 cannot be made at this stage.)

Example: in the partial sentence こういうことをおやりになるということは.. the string おやり was proposed. This should also be accepted at this stage.
Appendix B: Lexical Item Tagging System

Example: in the sentence 白っぽい綿のような上着にズボン姿で、リュックを背負っていた。
the string ツボン姿 should be accepted.

Prepositions, Conjunctions, Pronouns, etc.

These are not likely to occur, except as longer expressions (covered below.)

Abbreviations and Contractions

Sequences of capital letters will be accepted at this stage as potential as alphabetisms and acronyms.

Contractions of compound nouns, e.g. a two-kanji contraction of a four-kanji compound, will be treated as though they were regular 熟語, as per A above.

Verbs

Regular verbs consisting of one (sometimes two) kanji plus an inflectional part will rarely appear as potential lexical items as almost are already lexicalized. Compound verbs (複合動詞) are like to appear more often as they can be formed easily and are only partially lexicalized. They can be identified by their constructional pattern of an initial part consisting of the stem or て-form of a verb or adjective, followed by another verb, and by their context, e.g. clause/sentence end, preceded by を, が, etc. All should be accepted in this stage.

Multiword Expressions (Phrases)

Noun compounds have been covered in A above, however there are some phrases which also can effectively be regarded as noun compounds; they are typically formations of noun の noun (e.g. 学問の基礎) or noun な noun (e.g. 法外な安値), will occur in the same syntactic contexts as regular noun compounds, and can be assessed in the same way.
Example: The ML model suggested 物のタ from the sentence: しかしこれは物のタトエでもあるが... This can be safely rejected as the syntactical context in which it occurs indicates that it is a fragment, probably resulting from the model being confused.

More general potential phrases are very difficult to assess without considering the underlying morphemes and proceeding into analysis of the semantics. To illustrate this, consider the following partial sentences.

a. ...と誘導されたとのべている。

Here the ML model suggested とのべて. At first glance this would be rejected - it is simply と + のべて (述べる), but can we be certain

i. that とのべて is not an idiomatic expression we haven’t come across before?

After all there are recorded expressions such as とあって, と比べて, etc.

ii. that とのべて is not a noun written in hiragana? (e.g. ときめき)

Both of these are most unlikely when the whole sentence is read, but it really takes a fairly complete analysis, not just of the syntax, to reach that conclusion.

b. 沼や湖などはその最たるものである。

Here the ML model suggested ものである. In fact the parse will be 最たるもの + である, but that depends largely on us knowing that 最たるもの (最たる物) is an expression. Can we confidently rule out ものである being an idiomatic expression just on the syntax?

c. ほんとうに危険ののはこの“新しい現象”にほかならない。
The ML system suggested なのは. While this sequence is quite common, it can be seen that it’s associated with the preceding 形容動詞, and thus it is 危険なの + は and not a free-standing expression.

d. “この沼のあたりに死体を捨てたんだろう、どこか水のあるところだろう…”と誘導された、とのべている。

The ML system suggested 水のある and ところだろう (not とのべって this time.) Can these be rejected without going into the meaning of the passage? There are plenty of ... のある expressions recorded: 角のある, 愛敬のある, 趣のある, etc., so why not 水のある?

e. 何が危ういかっていうと、佐藤さんがいつもいっているように、“死体を水のなかに捨てた”っていったと、これはもうほとんど....

The ML system suggested っていいると twice. At first glance it’s a fragment, but it’s very common in speech (and dialogue). It probably should be investigated.

f. 警察が頼んで叩かしているみたいな感じが、....

The ML system suggests みたいな. It is not a fragment but the very common みたい + な construction. Can it be rejected initially? It also suggested 叩かして, in fact many て-forms get suggested. A number of て-forms are recorded when they are very common or have idiomatic meanings, but perhaps as a matter of policy they should be filtered at this stage as the SNR will be very high if they all have to be evaluated in detail (叩かせて is simply from 叩かす, a causative of 叩く).
Appendix C

Neologism Synthesis - Data and Examples

C.1 Feature Data

C.1.1 Feature File Structure

LIBSVM uses the same format for both training and test data files. The files consist of lines of text in the following format:

\(<label> <index1>:<value1> <index2>:<value2> ...\)

- \(<label>\) will be either \(-1\) or \(+1\) in training files depending on whether the features are associated with a classified item or not. It is ignored in test data files.
- \(<index1>, <index2>, etc.\) are integer feature numbers for features with a non-zero value.
Appendix C: Neologism Synthesis - Data and Examples

- \(<\text{value1}>, \text{<value2>}, \text{etc.}\) are the values for the individual features.

C.1.2 Feature Examples

The following is an extract of one of the training feature files. The first index field is the count for the compound in the \(n\)-gram corpus. The following fields are the non-zero \(n\)-gram counts for the compound in combination with the encapsulating characters. The compound itself is appended as a comment for documentation purposes.

\[
\begin{align*}
& 35:6003 36:5157 37:51 \\
& 1 1:15 11:1 17:1 34:1 35:2 36:3 \\
& 31:8 32:44 33:93 34:25 35:75 36:126 \\
& -1 & 1:49 10:1 16:1 21:1 22:1 32:22 33:1 34:8 35:3 36:10 \\
& -1 & 1:41 5:1 11:1 29:1 32:4 33:11 34:1 35:3 36:4 37:1 \\
& 1 & 1:8 7:1 21:1 32:2 33:2 36:2 \\
& -1 & 1:51 32:9 33:18 34:1 35:2 36:3 \\
& 31:8 32:44 33:93 34:25 35:75 36:126 \\
& & 31:8 32:44 33:93 34:25 35:75 36:126
\end{align*}
\]

C.2 Evaluation of Synthesized Terms

The following tables contain the synthesized compounds which were classified as part of the investigation reported in Chapter 7. The tables contain the following information about the compounds:

- the compound
the compound’s count in the n-gram corpus

- the number of features associated with the compound

- the classifier set(s) which resulted in the compound being classified. For ML classification the models are indicated: (O)riginal, (S)mall, (M)edium, (L)arge; and for feature-count classifications, the feature-count threshold which was met: F7, F8, etc.

- the reading of the compound

- the translation of the compound, plus comments

C.2.1 Group 1

Table C.1 and Table C.1 contain summary information on the initial sample of 38 classified unlexicalized compounds discussed in Section 7.8.3.
### Appendix C: Neologism Synthesis - Data and Examples

<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>No. feat.</th>
<th>Classifiers</th>
<th>Reading</th>
<th>Valid</th>
<th>Meaning/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>伏雷</td>
<td>2775</td>
<td>7</td>
<td>F7</td>
<td>ふせいがずち</td>
<td>Y</td>
<td>One of the Thunder Gods (games, etc.)</td>
</tr>
<tr>
<td>因眼</td>
<td>2815</td>
<td>8</td>
<td>F8</td>
<td>きょうめ</td>
<td>Y</td>
<td>evil eye (games, etc.) Eijiro</td>
</tr>
<tr>
<td>刀姫</td>
<td>16521</td>
<td>5</td>
<td>S,M,L</td>
<td>かたなひめ</td>
<td>Y</td>
<td>game character</td>
</tr>
<tr>
<td>刀魔</td>
<td>6525</td>
<td>16</td>
<td>F14,S,M,O</td>
<td>とうま</td>
<td>Y</td>
<td>manga character</td>
</tr>
<tr>
<td>刀径</td>
<td>1348</td>
<td>7</td>
<td>F7</td>
<td>はけい</td>
<td>Y</td>
<td>blade diameter (machine tools)</td>
</tr>
<tr>
<td>刀紋</td>
<td>7403</td>
<td>14</td>
<td>F14,O</td>
<td>はもん</td>
<td>Y</td>
<td>blade pattern, esp of tempering</td>
</tr>
<tr>
<td>唐巣</td>
<td>7137</td>
<td>11</td>
<td>F10,L</td>
<td>からす</td>
<td>Y</td>
<td>name of a priest in anime series</td>
</tr>
<tr>
<td>塔也</td>
<td>3112</td>
<td>9</td>
<td>F8</td>
<td>とうや</td>
<td>Y</td>
<td>boy’s name</td>
</tr>
<tr>
<td>忍信</td>
<td>18691</td>
<td>14</td>
<td>F14,S,M,L,O</td>
<td>にんたて</td>
<td>Y</td>
<td>Final Fantasy character</td>
</tr>
<tr>
<td>晶霊</td>
<td>9866</td>
<td>7</td>
<td>F7,M,L</td>
<td>しょれい</td>
<td>Y</td>
<td>name in a game (onsen, etc.)</td>
</tr>
<tr>
<td>朱鬼</td>
<td>4954</td>
<td>8</td>
<td>F8</td>
<td>しゅき</td>
<td>Y</td>
<td>game character</td>
</tr>
<tr>
<td>枯滝</td>
<td>1407</td>
<td>7</td>
<td>F7</td>
<td>かれたき</td>
<td>Y</td>
<td>form of 枯れ滝 (dry waterfall)</td>
</tr>
<tr>
<td>桃肌</td>
<td>5826</td>
<td>6</td>
<td>M</td>
<td>ももはだ</td>
<td>Y</td>
<td>brand of cosmetics</td>
</tr>
<tr>
<td>滝糸</td>
<td>2476</td>
<td>8</td>
<td>F8</td>
<td>うずいと</td>
<td>Y</td>
<td>vortex filament</td>
</tr>
<tr>
<td>溝鏡</td>
<td>2432</td>
<td>13</td>
<td>F12,S,O</td>
<td>??</td>
<td>N</td>
<td>chance collocation?</td>
</tr>
<tr>
<td>潤肌</td>
<td>9193</td>
<td>4</td>
<td>M,L</td>
<td>うるはだ</td>
<td>Y</td>
<td>brand of cosmetics</td>
</tr>
<tr>
<td>狩魔</td>
<td>6773</td>
<td>6</td>
<td>M,L</td>
<td>かるま</td>
<td>Y</td>
<td>(i.e. karma) - game name</td>
</tr>
<tr>
<td>猫姫</td>
<td>43701</td>
<td>6</td>
<td>S,M,L,O</td>
<td>ねこひめ</td>
<td>Y</td>
<td>桜の猫姫 is a manga series</td>
</tr>
<tr>
<td>猿鬼</td>
<td>3247</td>
<td>8</td>
<td>F8</td>
<td>さらおに</td>
<td>Y</td>
<td>monkey demon (manga)</td>
</tr>
<tr>
<td>獣魔</td>
<td>3251</td>
<td>8</td>
<td>F8</td>
<td>じゅうま</td>
<td>Y</td>
<td>animal demon (manga)</td>
</tr>
<tr>
<td>磁粉</td>
<td>3879</td>
<td>9</td>
<td>F8,S,M,L</td>
<td>じふん</td>
<td>Y</td>
<td>magnetic particle</td>
</tr>
<tr>
<td>磁肌</td>
<td>2848</td>
<td>9</td>
<td>F8,M,L</td>
<td>じはだ</td>
<td>Y</td>
<td>surface of porcelain</td>
</tr>
<tr>
<td>緃紫</td>
<td>12080</td>
<td>8</td>
<td>F8,S,M,L</td>
<td>あやむらさき</td>
<td>Y</td>
<td>brand of sake</td>
</tr>
<tr>
<td>緋霊</td>
<td>24675</td>
<td>6</td>
<td>O</td>
<td>ばくれい</td>
<td>N</td>
<td>fragment of 緋霊</td>
</tr>
<tr>
<td>臭菌</td>
<td>10576</td>
<td>6</td>
<td>M,L</td>
<td>しゅうきん</td>
<td>Y</td>
<td>smelly bacteria (mostly in adverts)</td>
</tr>
</tbody>
</table>

Table C.1: Details of the 38 Group 1 Synthesized kanji Compounds
<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>No. feat.</th>
<th>Classifiers</th>
<th>Reading</th>
<th>Valid</th>
<th>Meaning/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>虎姫</td>
<td>941</td>
<td>2</td>
<td>O</td>
<td>とらねえ</td>
<td>Y</td>
<td>manga character</td>
</tr>
<tr>
<td>虎眼</td>
<td>16900</td>
<td>8</td>
<td>F8,S,M,L</td>
<td>とらねえ</td>
<td>Y</td>
<td>manga character</td>
</tr>
<tr>
<td>蛇姫</td>
<td>3687</td>
<td>5</td>
<td>M,L</td>
<td>へびひめ</td>
<td>Y</td>
<td>manga character</td>
</tr>
<tr>
<td>蛇矛</td>
<td>3051</td>
<td>11</td>
<td>F10,M</td>
<td>Y</td>
<td></td>
<td>だぼうじやぼう halberd with snake-like blade (Chinese)</td>
</tr>
<tr>
<td>蝶眼</td>
<td>36546</td>
<td>21</td>
<td>F14,O</td>
<td>じゃし</td>
<td>Y</td>
<td>(another) evil eye</td>
</tr>
<tr>
<td>蝶霊</td>
<td>10237</td>
<td>24</td>
<td>F14,S,M,L,O</td>
<td>じゃれい</td>
<td>Y</td>
<td>evil spirit (mostly in game, manga, etc. contexts)</td>
</tr>
<tr>
<td>郭嘉</td>
<td>22095</td>
<td>23</td>
<td>F14,S,M,L,O</td>
<td>かくか</td>
<td>Y</td>
<td>Guo Jia (Chinese historical character, now in manga, etc.)</td>
</tr>
<tr>
<td>酔鯨</td>
<td>13896</td>
<td>10</td>
<td>F10,S,M,L,O</td>
<td>すいげい</td>
<td>Y</td>
<td>brand of sake</td>
</tr>
<tr>
<td>鎮壇</td>
<td>7961</td>
<td>1</td>
<td>M,L</td>
<td>ちんだん</td>
<td></td>
<td>fragment of 鎮壇具 (mis-parse?)</td>
</tr>
<tr>
<td>陶謨</td>
<td>5695</td>
<td>17</td>
<td>F14,S,M,L,O</td>
<td>とうけん</td>
<td>Y</td>
<td>Tao Qian (Chinese historical character)</td>
</tr>
<tr>
<td>食魔</td>
<td>52287</td>
<td>1</td>
<td>S,M,L,O</td>
<td>きまつ</td>
<td>N</td>
<td>fragment of 帽飢魔 (Seikima-II heavy metal band)</td>
</tr>
<tr>
<td>龍脈</td>
<td>13243</td>
<td>19</td>
<td>F14,S,M,L,O</td>
<td>りゅうみゃく</td>
<td>Y</td>
<td>route taken by spirit flowing underground (fantasy, etc.)</td>
</tr>
<tr>
<td>龍軒</td>
<td>11747</td>
<td>5</td>
<td>M,L</td>
<td>りゅうけん</td>
<td>Y</td>
<td>Chinese restaurant name (chain?)</td>
</tr>
</tbody>
</table>

Table C.2: Details of the 38 Group 1 Synthesized kanji Compounds (continued)
C.2.2 Group 2

Table C.3 contains summary information on the sample of 10 classified unlexicalized compounds discussed in Section 7.8.3. (These compounds resulted from using a heuristic classification with a lowered cutoff.)

<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>No. feat.</th>
<th>Classifiers</th>
<th>Reading</th>
<th>Valid</th>
<th>Meaning/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>凶魔</td>
<td>1381</td>
<td>5</td>
<td>F5</td>
<td>きょうま</td>
<td>Y</td>
<td>evil spirit (manga)</td>
</tr>
<tr>
<td>刀鬼</td>
<td>676</td>
<td>5</td>
<td>F5</td>
<td>とうき</td>
<td>Y</td>
<td>manga character</td>
</tr>
<tr>
<td>呉漢</td>
<td>772</td>
<td>5</td>
<td>F5</td>
<td>ごかん</td>
<td>Y</td>
<td>Chinese commander (1st century CE)</td>
</tr>
<tr>
<td>呉班</td>
<td>966</td>
<td>5</td>
<td>F5</td>
<td>ごはん</td>
<td>Y</td>
<td>Wu Ban Chinese commander (2nd century CE)</td>
</tr>
<tr>
<td>寮棟</td>
<td>713</td>
<td>5</td>
<td>F5</td>
<td>りょうとう</td>
<td>Y</td>
<td>dormitory building; dormitory hall</td>
</tr>
<tr>
<td>弦琴</td>
<td>5302</td>
<td>5</td>
<td>F5</td>
<td>げんきん</td>
<td>N</td>
<td>seems a mis-parse/truncation of 一弦琴, etc.</td>
</tr>
<tr>
<td>掌把</td>
<td>623</td>
<td>6</td>
<td>F6</td>
<td>しょうは</td>
<td>Y</td>
<td>type of fighting grasp in some video games</td>
</tr>
<tr>
<td>朱霊</td>
<td>2746</td>
<td>5</td>
<td>F5</td>
<td>しゅれい</td>
<td>Y</td>
<td>Zhu Ling Chinese general during the Three Kingdoms period</td>
</tr>
<tr>
<td>桐也</td>
<td>1085</td>
<td>5</td>
<td>F5</td>
<td>きりや</td>
<td>Y</td>
<td>given name of 衛藤桐也, character in 金色のコルダ (La Corda d’Oro) RPG</td>
</tr>
<tr>
<td>殲魔</td>
<td>1851</td>
<td>5</td>
<td>F5</td>
<td>おうま なぐりま</td>
<td>Y</td>
<td>Guild in various manga and games. Also 殲り魔</td>
</tr>
</tbody>
</table>

Table C.3: Details of the 10 Group 2 Synthesized kanji Compounds

C.2.3 Group 3

Table C.4 contains summary information on an additional sample of 10 classified unlexicalized compounds discussed in Section 7.8.3 using an even lower cutoff.
### Appendix C: Neologism Synthesis - Data and Examples

#### C.2.4 Group 4

Table C.5 contains summary information on a sample of 23 classified unlexicalized compounds which were generated using an alternative kanji range and evaluated using the Google Corpus.

#### C.2.5 Group 5

Table C.6 contains summary information on a sample of 35 classified unlexicalized compounds which were generated using an alternative kanji range and evaluated using the Kyoto Corpus.
### Table C.5: Details of the 23 Group 4 Synthesized kanji Compounds

<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>No. feat</th>
<th>Classifiers</th>
<th>Reading</th>
<th>Valid</th>
<th>Meaning/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>丸刀</td>
<td>2524</td>
<td>7</td>
<td>F05</td>
<td>がんとう/まるとう</td>
<td>Y</td>
<td>gouge with U-shaped blade</td>
</tr>
<tr>
<td>丸粒</td>
<td>6708</td>
<td>10</td>
<td>F10,L</td>
<td>まるつぶ</td>
<td>Y</td>
<td>prefix meaning &quot;whole grain&quot;</td>
</tr>
<tr>
<td>丸貝</td>
<td>5913</td>
<td>13</td>
<td>F12,M</td>
<td>がんばい</td>
<td>Y</td>
<td>round shell - seems to be used as bait.</td>
</tr>
<tr>
<td>他猫</td>
<td>4007</td>
<td>11</td>
<td>F10,L,M</td>
<td>ほかねこ</td>
<td>N</td>
<td>just seems a juxtaposition</td>
</tr>
<tr>
<td>佐弦</td>
<td>12733</td>
<td>21</td>
<td>F14,L,S</td>
<td>でいげん</td>
<td>Y</td>
<td>low-pitched string instruments (abbr)</td>
</tr>
<tr>
<td>光刃</td>
<td>5415</td>
<td>14</td>
<td>F14,M,S</td>
<td>こうじん</td>
<td>Y</td>
<td>&quot;light sword&quot; used in manga, games, etc.</td>
</tr>
<tr>
<td>医龍</td>
<td>46813</td>
<td>18</td>
<td>F14,L,M,S</td>
<td>いりゆう</td>
<td>Y</td>
<td>name of a manga (Medical Dragon!)</td>
</tr>
<tr>
<td>古鎮</td>
<td>10239</td>
<td>11</td>
<td>F10,L</td>
<td>???</td>
<td>Y</td>
<td>in Chinese means &quot;old town&quot;. In many travel sites</td>
</tr>
<tr>
<td>古龍</td>
<td>48843</td>
<td>24</td>
<td>F14,L,M,S</td>
<td>グロン</td>
<td>Y</td>
<td>Chinese novelist Gu Long</td>
</tr>
<tr>
<td>周堤</td>
<td>980</td>
<td>5</td>
<td>F05</td>
<td>しゅうてい</td>
<td>Y</td>
<td>circular bank, e.g. round Jomon-period grave sites</td>
</tr>
<tr>
<td>夜刀</td>
<td>23510</td>
<td>8</td>
<td>F08,L,M,S</td>
<td>やと</td>
<td>Y</td>
<td>manga name</td>
</tr>
<tr>
<td>夜寮</td>
<td>1345</td>
<td>5</td>
<td>F05</td>
<td>???</td>
<td>N</td>
<td>fragment of a CD name</td>
</tr>
<tr>
<td>太伯</td>
<td>2037</td>
<td>6</td>
<td>F05,M</td>
<td>たいはく</td>
<td>Y</td>
<td>eldest son of King Tai of Zhou</td>
</tr>
<tr>
<td>字溝</td>
<td>21258</td>
<td>6</td>
<td>F05</td>
<td>じこう</td>
<td>Y</td>
<td>from X 字溝, i.e. X-shaped gutter</td>
</tr>
<tr>
<td>師魂</td>
<td>9359</td>
<td>8</td>
<td>F08,L,M</td>
<td>しこん</td>
<td>Y</td>
<td>var. of sake: “蔵の師魂&quot;</td>
</tr>
<tr>
<td>悪劣</td>
<td>607</td>
<td>3</td>
<td>M</td>
<td>あくれつ</td>
<td>Y</td>
<td>seems a reversal of 劣悪 (inferiority) as in Chinese</td>
</tr>
<tr>
<td>整膚</td>
<td>11637</td>
<td>7</td>
<td>F05,L,M,S</td>
<td>せいふ</td>
<td>Y</td>
<td>therapy involving pinching the skin and pulling it</td>
</tr>
<tr>
<td>松潤</td>
<td>174200</td>
<td>36</td>
<td>F14,L,S</td>
<td>まつじゅん</td>
<td>Y</td>
<td>nickname of singer 松本潤.</td>
</tr>
<tr>
<td>極辛</td>
<td>384</td>
<td>6</td>
<td>F05</td>
<td>こあ</td>
<td>Y</td>
<td>from a game: 黒の核晶</td>
</tr>
<tr>
<td>落辛</td>
<td>9189</td>
<td>12</td>
<td>F12,L,M,S</td>
<td>ごくから</td>
<td>Y</td>
<td>flavour style (極辛口 with wine, sake, etc.)</td>
</tr>
<tr>
<td>橋龍</td>
<td>10253</td>
<td>14</td>
<td>F14,L,M,S</td>
<td>はしりゅう</td>
<td>Y</td>
<td>name of ramen shop</td>
</tr>
<tr>
<td>母猫</td>
<td>74534</td>
<td>32</td>
<td>F14,S</td>
<td>ははねこ</td>
<td>Y</td>
<td>mother cat</td>
</tr>
<tr>
<td>母猿</td>
<td>2646</td>
<td>9</td>
<td>F08</td>
<td>ぼえん/はざる</td>
<td>Y</td>
<td>mother monkey</td>
</tr>
</tbody>
</table>

Appendix C: Neologism Synthesis - Data and Examples
<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>No. feat.</th>
<th>Classifiers</th>
<th>Reading</th>
<th>Valid</th>
<th>Meaning/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>丸刀</td>
<td>167</td>
<td>10</td>
<td>F10</td>
<td>がんとう/まるとう</td>
<td>Y</td>
<td>gouge with U-shaped blade</td>
</tr>
<tr>
<td>丸貝</td>
<td>730</td>
<td>24</td>
<td>F14,L,M</td>
<td>がんばい</td>
<td>Y</td>
<td>round shell - seems to be used as bait.</td>
</tr>
<tr>
<td>健衞</td>
<td>32</td>
<td>4</td>
<td>S</td>
<td>たちゅう</td>
<td>N</td>
<td>seems a random construction</td>
</tr>
<tr>
<td>低弦</td>
<td>889</td>
<td>30</td>
<td>F14,L,M</td>
<td>ていげん</td>
<td>Y</td>
<td>low-pitched string instruments (abbr)</td>
</tr>
<tr>
<td>健苗</td>
<td>392</td>
<td>18</td>
<td>F14,L,M,S</td>
<td>けんびょう</td>
<td>Y</td>
<td>good seedlings, healthy seedlings</td>
</tr>
<tr>
<td>光刃</td>
<td>239</td>
<td>18</td>
<td>F14,L,M,S</td>
<td>こうじん</td>
<td>Y</td>
<td>light sword (in manga, games, etc.)</td>
</tr>
<tr>
<td>光狩</td>
<td>1109</td>
<td>27</td>
<td>F14,L,M</td>
<td>ひかり/みかり</td>
<td>Y</td>
<td>recent female name (poss from RPG).</td>
</tr>
<tr>
<td>単郭</td>
<td>77</td>
<td>4</td>
<td>S</td>
<td>たんかく</td>
<td>Y</td>
<td>simple fortification</td>
</tr>
<tr>
<td>単勇</td>
<td>189</td>
<td>16</td>
<td>F14</td>
<td>ふだたみ</td>
<td>Y</td>
<td>old tatami</td>
</tr>
<tr>
<td>高奥</td>
<td>846</td>
<td>3</td>
<td>L,M,S</td>
<td>ふるくさ</td>
<td>N</td>
<td>fragment from 古臭い</td>
</tr>
<tr>
<td>高龍</td>
<td>262</td>
<td>15</td>
<td>F14,L,M,S</td>
<td>こりゆう</td>
<td>Y</td>
<td>Gu Long (Chinese novelist), plus many elderly dragons in manga, etc.</td>
</tr>
<tr>
<td>士紳</td>
<td>6</td>
<td>5</td>
<td>F05</td>
<td>しんし</td>
<td>Y</td>
<td>archaic term for ranking official</td>
</tr>
<tr>
<td>夜刀</td>
<td>620</td>
<td>17</td>
<td>F14,L,M,S</td>
<td>やと</td>
<td>Y</td>
<td>manga game</td>
</tr>
<tr>
<td>狛犬</td>
<td>45</td>
<td>13</td>
<td>F12,M,S</td>
<td>でんじゅう</td>
<td>Y</td>
<td>character in a game.</td>
</tr>
<tr>
<td>字姓</td>
<td>58</td>
<td>5</td>
<td>F05</td>
<td>じせい</td>
<td>N</td>
<td>fragment from 二字姓, 一字名, etc.</td>
</tr>
<tr>
<td>字溝</td>
<td>981</td>
<td>6</td>
<td>F05,L</td>
<td>じこう</td>
<td>Y</td>
<td>letter-shaped gutter, etc.</td>
</tr>
<tr>
<td>帰寮</td>
<td>130</td>
<td>8</td>
<td>F08,M,S</td>
<td>きりょう</td>
<td>Y</td>
<td>returning to accommodation, etc.</td>
</tr>
<tr>
<td>張賓</td>
<td>16</td>
<td>7</td>
<td>F05</td>
<td>ちょうひん</td>
<td>Y</td>
<td>Zhang Bin - Chinese general 4C CE.</td>
</tr>
<tr>
<td>影媛</td>
<td>109</td>
<td>18</td>
<td>F14,L,M</td>
<td>かげひめ</td>
<td>Y</td>
<td>shadow princess, in old stories.</td>
</tr>
<tr>
<td>影龍</td>
<td>22</td>
<td>8</td>
<td>F08</td>
<td>えいりゅう</td>
<td>Y</td>
<td>shadowy dragon in manga, etc.</td>
</tr>
<tr>
<td>捕弦</td>
<td>49</td>
<td>7</td>
<td>F05</td>
<td>せいづえ</td>
<td>Y</td>
<td>found in 捕弦定理 (alternate segment theorem), and appears in guitar texts as &quot;adjacent chords&quot;</td>
</tr>
<tr>
<td>摘粒</td>
<td>229</td>
<td>16</td>
<td>F14,L,M,S</td>
<td>てきりゅう</td>
<td>Y</td>
<td>berry thinning, esp grapes.</td>
</tr>
<tr>
<td>春苗</td>
<td>28</td>
<td>6</td>
<td>F05</td>
<td>しゅんびょう</td>
<td>Y</td>
<td>both girl’s name and spring seedlings</td>
</tr>
<tr>
<td>松潤</td>
<td>749</td>
<td>21</td>
<td>F14,L,M</td>
<td>まつじゅん</td>
<td>Y</td>
<td>nickname of singer 松本潤.</td>
</tr>
<tr>
<td>极辛</td>
<td>304</td>
<td>17</td>
<td>F14,L,M,S</td>
<td>ごくから</td>
<td>Y</td>
<td>flavour style (極辛口 with wine, sake, etc.)</td>
</tr>
<tr>
<td>横毅</td>
<td>810</td>
<td>5</td>
<td>F05,L,M,S</td>
<td>??</td>
<td>N</td>
<td>random juxtaposition</td>
</tr>
<tr>
<td>横芽</td>
<td>8</td>
<td>5</td>
<td>F05</td>
<td>??</td>
<td>N</td>
<td>lateral bud in plants</td>
</tr>
</tbody>
</table>

Table C.6: Details of the 35 Group 5 Synthesized *kanji* Compounds (initial part)
<table>
<thead>
<tr>
<th>Compound</th>
<th>Count</th>
<th>No. feat.</th>
<th>Classifiers</th>
<th>Reading</th>
<th>Valid</th>
<th>Meaning/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>橋龍</td>
<td>611</td>
<td>22</td>
<td>F14,L,M</td>
<td>はしりゅう</td>
<td>Y</td>
<td>restaurant name</td>
</tr>
<tr>
<td>歌魂</td>
<td>16</td>
<td>7</td>
<td>F05</td>
<td>うたたま</td>
<td>Y</td>
<td>title of a 2008 comedy film</td>
</tr>
<tr>
<td>母猫</td>
<td>1730</td>
<td>29</td>
<td>F14,L,M,S</td>
<td>ははねこ</td>
<td>Y</td>
<td>mother cat</td>
</tr>
<tr>
<td>母猿</td>
<td>113</td>
<td>15</td>
<td>F14,L,M,S</td>
<td>ぼえん/ははざる</td>
<td>Y</td>
<td>mother monkey</td>
</tr>
<tr>
<td>母珠</td>
<td>49</td>
<td>7</td>
<td>F05</td>
<td>もしゅ</td>
<td>Y</td>
<td>large bead(s) in a Buddhist rosary</td>
</tr>
<tr>
<td>母貝</td>
<td>666</td>
<td>20</td>
<td>F14,L,M,S</td>
<td>ぼかい</td>
<td>Y</td>
<td>pearl oyster</td>
</tr>
<tr>
<td>火灰</td>
<td>5</td>
<td>6</td>
<td>F05</td>
<td>かかい</td>
<td>N</td>
<td>random collocation.</td>
</tr>
<tr>
<td>火龍</td>
<td>479</td>
<td>19</td>
<td>F14,L,M</td>
<td>かりゆう</td>
<td>Y</td>
<td>fire dragon/salamander/etc. in manga etc.</td>
</tr>
</tbody>
</table>

Table C.7: Details of the 35 Group 5 Synthesized *kanji* Compounds (continuation)
Appendix D

Sample WWW Sentences - Language Contexts

D.1 Sample Sentences Containing Neologisms

This Section contains 10 sentences extracted from the WWW for each of five indicated neologisms. See Section 9.4.2

a. マタハラ

先日、私がアップしたマタハラのブログに関して、「自分はマタハラ加害者側かと思うのですか....

女性の活躍が叫ばれる一方で、深刻な問題となっている“マタハラ”、「マタハラ」とはマタニティハラスメントの略で、働く女性が妊娠・出産を機に嫌がらせを受けたり、雇い止めや自主退職に追い込まれたりすることを指します。

法律事務所でマタハラ被害の相談を受けている大渕愛子弁護士と.....
マタハラに負けない母は強し！
このようなことがあったらあなたはマタハラ（マタニティハラスメント）に合っている可能性があります。
セクハラ」と「マタハラ」の実態調査 労働政策研究・研修機構は１日、昨年実施した「マタハラ・セクハラに関する実態調査」の結果を発表した。
「マタハラをする人はこう考えているという典型的なサンプル」との声も。
日本ではマタハラが社会問題として大きく取り上げられています。
具体的なマタハラの被害のことについて詳しく知りたい方はママテラスへ！
既婚女性の約４割が「マタハラ／イクハラを受けた経験がある」と回答。

b. こじらせ女子

皆さんは「こじらせ女子」という言葉を聞いたことがあるでしょうか？
そんな“ゆるフワ”的な女子像にどうしても近づけない“こじらせ女子”という生態。
耳にしたことがある方もきっと多い「こじらせ女子」ですが。
「隠れこじらせ女子」を見抜く10のチェックリスト。次のうち、いくつ当てはまるかチェックしてみてください。
こじらせ女子の特徴はいろいろとありますが、..
わかりやすく言うと内面的にめんどくさい女性、いわゆる『こじらせ女子』は、こじらせ女子が増殖傾..
こじらせ女子とは、女性としての自分に自信が持てない女性のこと。
恋愛に対してこじれた考え方をしてしまう「こじらせ女子」。
何．千鶴はこじらせ女子っぽいとマキちゃんが言いました。

目下、取り組んでいる役どころは、32歳バツイチの“こじらせ女子”である。

c. ナマポ

生活保護受給者の40代女性が、友人から「アナタはナマポでラクできるかなりいいね。」と言われたと。

ナマポ（なまぽ）は、生活保護を受けながら生活している人たちを指す言葉です。

それくらいのスペックじゃないと・とてもじゃないがナマポの生活費だけで生活できない

ナマポって言葉、ご存知ですか？インターネットスラングのひとつで、生活保護を受給する人を表現する言葉です。

俺は榎本のナマポ奴隷の実態について...

俺は元ナマポだったけど吸ってた

電話の主は、生活保護（ナマポ）の受給者だ

ギリシアへの支援をナマポと考えれば、年金制度が支援国より優遇されていることが許されるはずはありません。

あとごみ処理件ちゃんと貼ってと言ってたら、ナマポは有料ゴミは無料なんだってさ。

兵庫県のある市では、「ナマポ」つまり、生活保護受給者がパチンコ店で遊んでいると、通報されてしまうらしい。

d. 美魔女

実年齢よりずっと若く見える女性を指す“美魔女”という言葉もすっかり定着しましたね。
最近の巷では、実年齢より10〜20歳も若く見えて40代以上の「美魔女」と呼ばれる女性たちが増えている。

本物の美魔女は、筆者のような単なる普通の三十路熟女と違って、...

Amazon公式サイトで美魔女ビューティ20歳若返る魔法のメソッドを購入すると、美魔女というと、最近は「年甲斐もなくイタい」「必死すぎてコワい」と揶揄する人がいたり、....

いつまでも若々しい中年女性の呼称「美魔女」がすっかり定着した今日。

テレビや雑誌で話題になっている美魔女は、ホントに美しくて私も綺麗になりたい！と思う同年代の方も多いですよね。

2013年より放映されている池田聡の美魔女レッスンを、ウェブ動画でご覧いただけます。

香港には年齢不詳の美魔女が多いのです。

20代の私たちも素敵な美魔女になれるように今からケア頑張りましょう。

e. 隠れメタボ

肥満ではないのに高血圧や高血糖などの異常を超数持つ「隠れメタボ」の患者は全国で914万人。

隠れメタボとは、日本のメタボリックシンドロームの判定基準を満たしてはいないが、...

メタボ体型でなくても糖尿病を発症する可能性がある隠れ肥満や隠れメタボを解消させるには、難しいダイエットをする必要はないと言われています。
最近では、隠れメタボといって細い人でも知らないうちにメタボリックシンドロームになっている可能性があるので、若い女性もメタボには注意が必要になります。

いわゆる隠れ内臓肥満、隠れメタボがそれです。

痩せててもお腹だけポッコリ「隠れメタボ」に要注意！ぱっと見は普通の体型で、特に太っている感じもない。

そして、「隠れメタボ」の危険性は「普通のメタボ」の比ではありません！慢性疲労の原因から隠れメタボまで…検診歎番は語る。

日ごろから太っていると自覚している人や、隠れメタボの人なども、定期的に病院での検査を受けないと良いでしょう。

白黒長毛の隠れメタボ猫。

D.2 Sample Sentences Containing という造語 and という新語

This Section contains 10 sentences extracted from the WWW for each of という造語 and という新語. See Section 9.4.3.

a. という造語

- 手打ちに関して、昨今ではマスコミや芸人らによって、「大阪締め」という造語が定着されつつある。(Wikipedia)
- 米国 NBC テレビがハルマゲドンならぬ「アベゲドン」という造語を使用し警笛を鳴らした...
- これを象徴する言葉として「フィギュア萌え族」という造語を作り...
読みにくいのではと思い、皆様に早く覚えて頂けるよう『Pireus』という造語にいたしました。

「FinTech」という造語に注目があります。

「Aquair」(アクエア：【Aqua+Air】)という造語を冠した経営理念に「エデュテインメント」という造語を使っているのが...

私は、このことわざを更に発展させ、「百見は一験にしかず」という造語をつくり、持論を展開しています。

...「アンドロイドみたいなんだけどちょっと違う」という意味の「○○ ロイド」とか「○○ ノイド」という造語を目にする事があります。

さらにそこから「ダブダブ」という造語も作っています。

b. いう新語

「社会生命」という新語に何をイメージできますか？

それから「嘲笑い」「嘲笑」「大笑い」「噛（笑）」「爆笑」などという新語もあります。

「終わりハラ」という新語が話題になっています。

イギリスのオックスフォード辞典も、毎年「今年の英単語（Word of the year）」という新語・流行語大賞を発表しています。

iPhoneに続いて今年登場したサムソン製の「ギャラクシーS」から「ギャルス族」という新語が誕生しました。

島？信長さんが「ボーイフレンド（仮）」という新語爆誕させる！www可愛すぎてやばいwwwwwwの名無しさん

2011年に「メリアム・ウェブスターズカレッジエイト英英辞典」に掲載された「プロマンス」という新語が、この数年の「シャーロックホームズ」人気によって多くの人に認識されはじめました。
「爆買い」という新語も生まれました。

また、イケメンはイケメンでも、「逝け面」という新語がある。

さて、この『アボカド女子』という新語に対して世間の.....
Appendix E

Corpus-based Loanword Segmentation and Translation (CLST) Online Interface

E.1 Introduction

This appendix includes sample pages from the online interface to the Corpus-based Loanword Segmentation and Translation (CLST) system described in Chapter 6.
E.2 Entry Page

Gairaigo Segmentation and Translation

This server attempts to segment Japanese loanwords (gairaigo) into their component parts and to generate basic translations for them. Two methods are used: the CLST (Corpus-based Loanword Segmentation and Translation) method, and the MeCab Japanese morphological analysis package combined with the Unidic lexicon. Both are described below. (Try a couple: クーリングパイプ, アルキレンオキシドポリマー)

Enter Gairaigo:

Generate Reset

CLST Method
This method is described in a recent paper. In summary, it:

a. uses a large list of known gairaigo to find all possible ways of segmenting the input term;
b. looks up each possible segment in a large gairaigo-English dictionary and generates all the combinations of possible meanings of the term;
c. checks each possible translation against the Google English N-gram database to see if that sequence of words is used in WWW pages and with what frequency. (As this database is about 50Gb, checking a large number of combinations may take a few seconds.)

MeCab/Unidic
MeCab is a major open-source Japanese morphological analysis and part-of-speech tagging package. It uses advanced AI techniques and a large morpheme lexicon to find the boundaries between morphemes in Japanese text. In the past Japanese morphological analysis systems did not work very effectively with gairaigo. The Unidic morpheme lexicon in its latest version (2.1.1) has been trained with a large number of gairaigo (about 95,000), and has simple English translations attached to about half of them.

Warning
The server code is still a bit buggy (well, all code is...). Occasionally it will hang or abort. I'm chasing a couple of over-run issues, so if you find it misbehaves, email me the offending gairaigo.

Jim Breen
March 2013
### E.3 Output Examples

**Gairaigo Segmentation and Translation**

Input Gairaigo: ハイレベルデータリンク

**CLST Method**

Testing 120 possible segmentations/translations.

<table>
<thead>
<tr>
<th>Segment(s)</th>
<th>Translation(s)</th>
<th>N-gram Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ハイレベル・データ・リンク</td>
<td>high - level data link</td>
<td>14630</td>
</tr>
<tr>
<td>ハイ・レベル・データ・リンク</td>
<td>high level data link</td>
<td>2365</td>
</tr>
<tr>
<td>ハイレベル・データ・リンク</td>
<td>high level data link</td>
<td>2365</td>
</tr>
</tbody>
</table>

**MeCab/Unidic**

<table>
<thead>
<tr>
<th>Segment(s)</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ハイ・レベル・データ・リンク</td>
<td>high level data link</td>
</tr>
</tbody>
</table>
### Gairaigo Segmentation and Translation

Input Gairaigo: ハウスクリーニング

**CLST Method**

Testing 13 possible segmentations/translations.

<table>
<thead>
<tr>
<th>Segment(s)</th>
<th>Translation(s)</th>
<th>N-gram Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ハウスクリーニング(*)</td>
<td>housecleaning</td>
<td>129912</td>
</tr>
<tr>
<td>ハウス・クリーニング</td>
<td>house cleaning</td>
<td>442977</td>
</tr>
<tr>
<td>ハウ・スクリーニング</td>
<td>how screening</td>
<td>1313</td>
</tr>
<tr>
<td>ハウス・クリーニング</td>
<td>house laundry service</td>
<td>1024</td>
</tr>
<tr>
<td>ハウス・クリーニング</td>
<td>house dry cleaning</td>
<td>68</td>
</tr>
</tbody>
</table>

(*) A dictionary entry exists for the full gairaigo.

**MeCab/Unidic**

<table>
<thead>
<tr>
<th>Segment(s)</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ハウス・クリーニング</td>
<td>house cleaning</td>
</tr>
</tbody>
</table>

Gairaigo Segmentation and Translation

Input Gairaigo: ハウスクリーニング

CLST Method

Testing 13 possible segmentations/translations.

<table>
<thead>
<tr>
<th>Segment(s)</th>
<th>Translation(s)</th>
<th>N-gram Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ハウスクリーニング(*)</td>
<td>housecleaning</td>
<td>129912</td>
</tr>
<tr>
<td>ハウス・クリーニング</td>
<td>house cleaning</td>
<td>442977</td>
</tr>
<tr>
<td>ハウ・スクリーニング</td>
<td>how screening</td>
<td>1313</td>
</tr>
<tr>
<td>ハウス・クリーニング</td>
<td>house laundry service</td>
<td>1024</td>
</tr>
<tr>
<td>ハウス・クリーニング</td>
<td>house dry cleaning</td>
<td>68</td>
</tr>
</tbody>
</table>

(*) A dictionary entry exists for the full gairaigo.

MeCab/Unidic

<table>
<thead>
<tr>
<th>Segment(s)</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ハウス・クリーニング</td>
<td>house cleaning</td>
</tr>
</tbody>
</table>
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Breen, James

Title:
Extraction of neologisms from Japanese corpora

Date:
2017

Persistent Link:
http://hdl.handle.net/11343/211675

File Description:
Extraction of neologisms from Japanese corpora

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.