A Continuum-based Model of Lexical Acquisition

Pierre Marchal, Thierry Poibeau

A Continuum-based Model of Lexical Acquisition

Pierre Marchal* and Thierry Poibeau**

* ER-TIM, INaLCO
2 rue de Lille, 75007 Paris, France
** LaTTiCe (CNRS, ENS and University Paris 3)
PSL Research University and Univ. Sorbonne Paris Cité
45 rue d'Ulm, 75005 Paris, France

Abstract The automatic acquisition of verbal constructions is an important issue for natural language processing. In this paper, we have a closer look at two fundamental aspects of the description of the verb: the notion of lexical item and the distinction between arguments and adjuncts. Following up on studies in natural language processing and linguistics, we embrace the double hypothesis *i*) of a continuum between ambiguity and vagueness, and *ii*) of a continuum between arguments and adjuncts. We provide a complete approach to lexical knowledge acquisition of verbal constructions from an untagged news corpus. The approach is evaluated through the analysis of a sample of the 7,000 Japanese verbs automatically described by the system.

1 Introduction

Natural language applications have shown the need for new kinds of lexical resources. Speech transcription or machine translation do not use hand crafted dictionaries as their basic source of knowledge any more, but lexical resources automatically built from the statistical analysis of very large corpora. More precisely, these systems do not usually integrate a component identified as a resource *per se* but make use of very large statistical sources of knowledge (generally called "language models") that incorporate different kinds of linguistic information. Models are generally not readable by humans and are very different from any human readable dictionary.

This does not mean that hand crafted dictionaries are now obsolete, since humans still need practical and usable dictionaries. As a consequence, there seem to be a big divide between these two kinds of lexical resources (those used by computers and those used by humans) although the work of lexicographers relies more and more on the automatic processing of very large corpora. Lexical descriptions produced by lexicographers are now generally established after taking into consideration corpus-based and statistical information.

In this paper we propose a model that takes into consideration very large corpora so as to obtain fine-graine information about lexical items. We focus on verbs since this category of words exhibit different features that make their

description highly challenging. Like most lexical items, verbs can be ambiguous and one lexical item have most of the time different meanings (*i.e.* different word senses). Describing a verb and determining the different relevant word senses is known to be especially difficult and largely depends on the purpose of the lexical resource (for example, is the resource for a language learner or for a language expert?).

The same problem also arises when it comes to differentiate arguments and adjuncts. As said in [1]: "There are some very clear arguments (normally, subjects and objects), and some very clear adjuncts (of time and 'outer' location), but also a lot of stuff in the middle. Things in this middle ground are often classified back and forth as arguments or adjuncts depending on the theoretical needs and convenience of the author."

Following Manning, we support the idea of gradience in grammar and more generally in languages. Except for practical reasons (*e.g.* in the case of a paper dictionary), there is no reason to determine a fix number of word senses per word or to decide out of context what should be an argument or an adjunct. Of course someone elaborating a paper dictionary has to take this kind of decisions for obvious reasons, but it is not the case of modern dictionaries in electronic format. We believe that lexical descriptions can be more or less fine-grained depending on the goal or the application, and different lexical descriptions of a same lexical item can be equally valid (as long as they are linguistically motivated, of course).

In this paper, we describe a system able to dynamically produce different kinds of dictionaries depending on the user's need. The main source of information are large corpora gathered from the Web. The system collects different kinds of information on verbs and on their complements from these corpora and aims at producing meaningful lexical representations based on an accurate statistical analysis of these data. The end user can then explore more or less fine grained descriptions through different variable settings. Among the parameters that the end user can explore are the number of word senses per verb and the information taken into account to calculate the argumenthood of the different complement.

The system we have developed has been applied to a large corpus of Japanese news stories. Japanese offers specific and interesting challenges since arguments are specified by case particles that are most of the time ambiguous. Various other features (order of the constituents, zero anaphora, etc.) make Japanese a highly challenging language for NLP. In the course of this paper we present a complete system with a very large coverage since information is produce for more than 7,000 Japanese verbs with a high accuracy.

The paper is organized as follows. We first describe the state of the art in lexical acquisition. We then describe our approach to the problem, before giving some details on our experiments on Japanese. We conclude with an evaluation and a discussion on our results.

2 Previous Work

Previous work on the automatic acquisition of lexical data dates back to the early 1990s. The need for precise and comprehensive lexical databases was clearly identified for most NLP tasks (esp. parsing) and automatic acquisition techniques was then seen as a way to solve the resource bottleneck. However, first experiments [2, 3] were limited (the acquisition process was dealing with a few verbs only and a limited number of predefined subcategorization frames). The approach was based on local heuristics and did not take into account the wider context.

The approach was then refined so as to take into account all the most frequent verbs and subcategorization frames possible [4–6]. A last step will consist in letting the system infer the subcategorization frames directly from the corpus, without having to predefined the list of possible frames. This approach is supposed to be less precise but most errors are automatically filtered since rare and unreliable patterns can be discovered by a linguistic and statistical analysis. Most experiments have been on verbs that have the most complicated subcategorization frames, but the approach can also be extended to nouns and adjectives [7].

Most developments so far have been done on English, but more and more experiments are now done for other languages as well. See for example, experiments on French [8], German [9] or Chinese [10], among many others. The quality of the result depends of course on the kind of corpus used for acquisition, and even more on the considered language and on the size of the corpus used. Dictionaries obtained with very large corpora form the Web generally give the best performances. The availability of accurate non lexicalized parser is also a key feature for the quality of the acquisition process.

As for Japanese, different experiments have been done in the past, especially by Kawahara and Kurohashi [11, 12]. Their approach relies on the idea that the closest case component of a given predicate helps disambiguate its meaning, and thus serves as a clue to merge a set of predicate-argument structures into a case frame. Obtained case frames are further merged based on a similarity measure which combines a thesaurus-based similarity measure between lexical heads and a similarity measure between subcategorization patterns. Their resource has been successfully integrated to a dependency parser; however, we found it failed at describing the continuous aspect of lexical meaning (case frames are organized into a flat structure and no indication on the similarity between them is provided) as well as the continuous aspect of argumenthood (except for the closest case components, no indication on the importance of complements is provided).

3 Description of our Approach

Although our approach has been applied and evaluated for Japanese, the theoretical framework to calculate the argumenthood of a complement or the structure

of lexical entries is partially language independent (although actual case or function markers are of course language dependent and have to be specified for each language considered).

3.1 Calculating the Argumenthood of Complements

We suppose a list of verbs along with their complements that have been automatically extracted from a large representative corpus. In our framework, a complement is a phrase directly connected to the verb (or is, in other words, a dependency of the verb), while the verb is the head of the dependents. In what follows we assume that complements are in fact couples made of a head noun and a dependency marker, generally a preposition or a case particle (in the case of Japanese, we will have to deal with case particles but the approach can be generalized to languages marking complement through other means).

Different proposals have been made in the past to model the difference between arguments and adjuncts. For example, [13] and [14] try to validate linguistic criteria with statistical measures. [15] proposes to estimate the probability of a subcategorization frame associated to verb. Lastly, [16] following [17] propose to characterize the link between verbs and complements based on productivity measures.

Building on these previous works, we propose a new measure combining the prominent features describe in the literature. Our measure is derived from the famous TF-IDF weighting scheme used in information retrieval, with the major difference that we are dealing with complements instead of terms, and with verbs instead of documents. We chose this measure for two main reasons:

1. it is a well documented statistical measure, widely used, and which has already proven effective in numerous information retrieval tasks;
2. it implements common rules of thumb for distinguishing between arguments and adjuncts.

The measure applied to a verb and a complement is thus the following:

$$\text{argumenthood}_{v,c} = (1 + \log \text{count}(v, c)) \log \frac{|V|}{|\{v' \in V : \exists (v', c)\}|} \qquad (1)$$

where c is a complement (*i.e.* a tuple made of a lexical head and a case particle); v is a verb; $\text{count}(v, c)$ is the number of cooccurrences of the complement c with the verb v; $|V|$ is the total number of unique verbs; $|\{v' \in V : \exists (v', c)\}|$ is the number of unique verbs cooccurring with this complement.

The first part of the formula, $1 + \log \text{count}(v, c)$, takes into account the cooccurrence frequency of a verb with a given complement (which transposes the idea that arguments are more closely linked to a given verb than a random adjunct). The second part of the formula, $\log \frac{|V|}{|\{v' \in V : \exists (v', c)\}|}$ takes into account the dispersion of a complement, that is, its tendency to appear with different kinds of verbs. In other words, the more a complement is used with different verbs the more likely it is an adjunct.

The proposed measure assigns a value between 0 and 1 to a complement. 0 corresponds to a prototypical adjunct; 1 corresponds to a prototypical argument.

3.2 Enriching verb description using shallow clustering

We introduce a method for merging verbal structures, that is a verb and a set of complements, into minimal predicate-frames using reliable lexical clues. We call this technique *shallow clustering*.

A verbal structure corresponds to a specific sense of a given verb; that is the sense of the verb is given by the complements selected by the verb. Yet a single verbal structure contains a very limited number of complements. So as to obtain a more complete description of the verb sense we propose to merge verbal structures corresponding to same meaning of a given verb.

Our method relies on two principles:

1. Two verbal structures describing the same verb and having at least one common complement might correspond to the same verb meaning;
2. Some complements are more informative than others for a given verb sense.

As for the second principle, the measure of argumenthood, introduced in the previous section, serves as a tool for identifying the complements which contribute the most to the verb meaning. Our method merges verbal structures in an iterative process; beginning with the most informative complements (*i.e.* complements yielding the highest argumenthood value). Algorithm 1 describes our method for merging verbal structures.

3.3 Modeling word senses through hierarchical clustering

We propose to cluster the minimal predicate-frames built during the *shallow clustering* procedure into a dendrogram structure. A dendrogram allows one to define an arbitrary number of classes (using a threshold) and thus fit in with the goal to model a continuum between ambiguity and vagueness. A dendrogram is usually built using a hierarchical clustering algorithm and a distance matrix as the input of the hierarchical clustering algorithm. So as to measure the distance between minimal predicate-frames, we propose to represent minimal predicate-frames as vectors which would serve as the parameters of a similarity function.

We must first define a vector representation for the minimal predicate-frames. Following B. Partee and J. Mitchell, we suppose that "the meaning of a whole is a function of the meaning of the parts and of the way they are syntactically combined" [18] as well as all the information involved in the composition process [19]. The following equation summarizes the proposed model of semantic composition:

$$p = f(\mathbf{u}, \mathbf{v}, R, K) \tag{2}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two lexical components; R is the syntactic information associated with $\mathbf{u}$ and $\mathbf{v}$; K is the information involved in the composition process. Following the principles of distributional semantics [20, 21] lexical heads

Data: A collection $\mathbf{W}$ of verbal structures $(\mathbf{v}, \mathbf{D})$ with $\mathbf{v}$ a verb and $\mathbf{D}$ a collection of verbal complements
Result: A collection $\mathbf{W}'$ of minimal predicate-frames
$W' \longleftarrow [\,]$;
foreach *verb* $\mathbf{v}$ *such as* $\exists(v, D) \in W$ **do**
  `/* Be C the set of complements` c `cooccurring with` v `*/`
  $C \longleftarrow \{c : c \in D \wedge \exists(v, D) \in W\}$;
  `/* Be C' the elements of` C `sorted by decreasing TF-IDF value */`
  $C' \longleftarrow [c : c \in C \wedge$ `argumenthood(`v`,C'[i]) ⩾ argumenthood(`v`,C'[i+1])`$]$;
  foreach *complement* $\mathbf{c}'$ *of* C' **do**
    `/* Be D' a partial classification of` v `*/`
    $D' \longleftarrow [\,]$;
    foreach $D : \exists(v, D) \in W$ **do**
      if $c' \in D$ **then**
        add all the complements in D to D';
        remove (v, D) from W;
      end
    end
    foreach $D : \exists(v, D) \in W$ **do**
      if $\forall c \in D \longrightarrow c \in D'$ **then**
        add all the complements in D to D';
        remove (v, D) from W;
      end
    end
    if $|D'| \geqslant 2$ **then**
      add (v, D') to W';
    end
  end
end

Algorithm 1: Shallow clustering of verbal structures

can be represented in a vector space model [22]. Case markers (or prepositions) can be used as syntactic information. Finally, we propose to utilize our argumenthood measure to initialize the K parameter as it reflects how important is a complement for a given verb.

Each verbal construction is transformed into a vector. The distance between two vectors will represent the dissimilarity between two occurrence of a same verb. Among the very large number of metrics available to calculate the distance between two vectors, we chose the cosinus, since it is (as for the TF-IDF weighting scheme) simple, efficient and perfectly suited to our problem.

The equation (3) shows how the cosinus can be calculated for two vectors $\mathbf{x}$ et $\mathbf{y}$ (the cosinus varies between 0 for orthogonal vextors to 1 for identical vectors)

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3}$$

Hierarchical clustering is an iterative process which clusters the two most similar elements of a set into a single element and repeats until there is only one element left. Yet different clustering strategies are possible (*e.g.* single linkage, complete linkage, average linkage). So as to select the best strategy (that is the one which would preserve the most the information from the distance matrix) we propose to apply the cophenetic correlation coefficient.

$$c = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(\mathbf{D}_{i,j} - \bar{d})(\mathbf{C}_{i,j} - \bar{c})}{\sqrt{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(\mathbf{D}_{i,j} - \bar{d})^2 \sum_{i=1}^{n}\sum_{j=i+1}^{n}(\mathbf{C}_{i,j} - \bar{c})^2}} \tag{4}$$

Where $\mathbf{D}$ is the initial distance matrix and $\mathbf{C}$ is the cophenetic matrix that is the inter-cluster distances in the dendrogram. The clustering strategy that maximizes the cophenetic correlation coefficient should be selected.

4 Experiment

4.1 Acquisition and preprocessing of textual data

A large collection of Japanese text is gathered from a selection of news websites using RSS feeds and a set of XPath expressions so as to discard HTML markup and unrelevant content (*e.g.* navigation menu). To comply with external NLP tools (*i.e.* a POS tagger and a parser), specific preprocesses are then applied to the raw textual data: fullwidth form conversion, sentence splitting, etc. In the end, our corpus is made of more than 294 millions characters.

4.2 Verbal structure extraction

The next step is to apply a parser to the corpus in order to get a syntactic analysis of the data. The parser must be unlexicalized since our goal is to calculate the argumenthood of the different complement (an unlexicalized parser attaches all the complement to the verb without making any different between arguments and adjuncts). The two most well-known parsers for Japanese are KNP[1] [23] and CaboCha[2] [24] (we are aware other parsers exist as well like EDA[3] [25]). In this work, we have decided to use CaboCha, for efficiency, among other reasons. Since CaboCha is faster than KNP [26], it seems more convenient to process large textual data. We use the default settings.

CaboCha is based on a tagger called MeCab[4] [27] that requires a dictionary of surface forms for tagging. Among the different possible dictionaries, we chose IPAdic [28], that is the recommended dictionary for MeCab.

The next step consists in extracting verbs, along with their complements and case particles. The process is mainly based on the part-of-speech tags from

[1] `http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP`
[2] `http://taku910.github.io/cabocha/`
[3] `http://plata.ar.media.kyoto-u.ac.jp/tool/EDA/home_en.html`
[4] `http://taku910.github.io/mecab/`

MeCab and on the syntactic links identified by CaboCha. The identification of verbs is not straightforward since some ambiguities or language specificities have to be avoided but we will not detail this part here. As for the particles, nine simple case markers can be identified: が (*ga*), を (*wo*), に (*ni*), へ (*he*), で (*de*), から (*kara*), より (*yori*), まで (*made*), et と (*to*) [29]. However, a large number of complex case markers have been described: the list is not fixed and lots of variation exist among grammars and linguists. In our case we are partly dependent on the list of case markers defined in IPAdic. However, following previous descriptions like [30] or [29], we consider some particles as simple surface variants, like に対して (*ni tai site*), にたいして (*ni tai site*), に対し (*ni tai si*), に対しまして (*ni tai simasite*), and にたいしまして (*ni tai simasite*), that correspond to に対して ni tai site. Last but not least, we consider まで (*made*) as a case particle (and contrary to the choice made by IPAdic). In the end, we have a list of 30 (simple and complex) case particles. Lastly, lexical heads of complement are extracted. When the head can be identified as a named entity, it is replaced by a generic tag; numerical expressions are also replaced by a more generic tags <NUM>.

Finally we filter out verbal structures exhibiting suspicious patterns (*e.g.* two complements marked as direct objects of the verb). In the end we obtain more than 5.5 millions of verbal structure, corresponding to a bit more than 10,000 verbs.

4.3 Measuring the degree of argumenthood of complements

We apply our measure of argumenthood of complements to those obtained during the process of extraction of verbal structures. Here complements are couples made of a lexical head and a case marker. We could assess the suitability of our approach by comparing, for a given verb, complements with the highest degree of argumenthood with complements with the lowest degree of argumenthood. As for the verb 積む (*tumu*, to load, to pill up), the complements with the highest degree of argumenthood all disambiguate the meaning of the verb: 研鑽を[積む] (*kensan wo [tumu]*, to study hard), 修業を[積む] (*syuugyou wo [tumu]*, to train), 経験を[積む] (*keiken wo [tumu]*, to gain experience), etc. On the other hand, none of the complements with the lowest degree of argumenthood help disambiguating the meaning of the verb: ～氏が[積む] (*si ga [tumu]*, Mr. ...+ nominative), <NUM>人で[積む] (*<NUM>-nin de [tumu]*, <NUM> people + manner), etc.

4.4 Shallow clustering of the verbal structures

We apply our shallow clustering method to the collection of verbal structures. After filtering of the most unfrequent minimal predicate-frames, we obtain a collection of almost 386,000 minimal predicate-frames, associated with 7,116 unique lemmas.

4.5 Hierarchical clustering

Minimal verbal classes must then be merged gradually through hierarchical clustering, as shown in section 3.3. Using the cophenetic correlation coefficient we found out that the average linkage was the best clustering strategy. Hierarchical clustering output can be represented as a dendogram, as shown on figure 1.

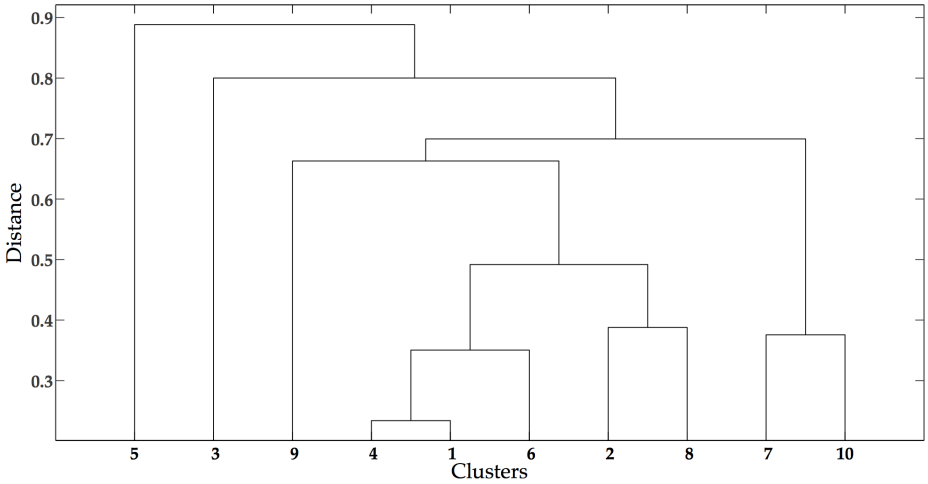

Figure 1: Dendrogram obtained after the hierarchical clustering of the ten first minimal predicate-frames of the verb 積む (*tumu*).

Each verb is thus described through a variable number of word senses, each word sense being itself defined by the different arguments attached to the verb. It is possible to explore the resource by navigating the hierarchy of word senses, *i.e.* by examining more or less fine-grained description. The interface making it possible to explore the data as well as some comments for the evaluation of the resource are presented in the following section.

5 Results and Discussion

Lexical resource are traditionally evaluated through a comparison with a reference resource [4, 6]. Although this approach is intuitive, it is not satisfactory since different lexical descriptions can be valid for a same lexical item, as it has been shown previously. We have nevertheless done a quick comparison with a manually built resource: IPAL [31]. The results show similar results as for other languages *e.g.* [8]: our system is able to discriminate relevant word senses, but the description is not fully similar to the one obtained with IPAL. Some differences are caused by errors (parsing errors, undetected ambiguities, etc.) but most

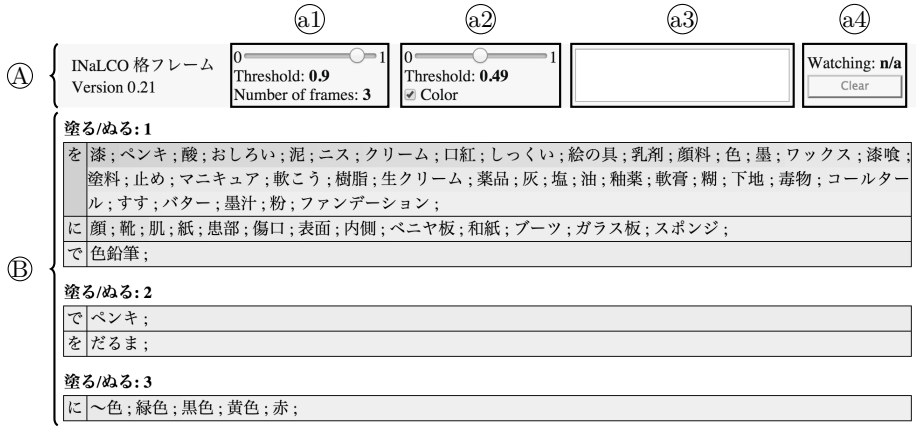

Figure 2: Screen capture of our graphical interface – Ⓐ control panel: (a1) slider for partitioning of sub-entries; (a2) slider for selection of complements; (a3) notification zone; (a4) sub-entry identifier. – Ⓑ sub-entry panel.

differences reveal in fact new or interesting word senses that are not described as such in IPAL.

However, the major novelty of our approach is the description of lexical item through a double continuum. In order to make the resource usable by humans, it is necessary to develop a visual interface allowing the end user to navigate the data and explore them in more details. In doing so, it is possible to have a more fine grained comparison with IPAL, which is not only based on a static arbitrary output of the system.

Figure 2 shows the proposed graphical user interface to access our resource. The control panel is the same for all the sub-entries in the resource. It allows the end-user to navigate the data thanks to:

- the slider for the sub-entry partitioning threshold (a1) ;
- the slider for the complement selection threshold (a2) ;
- the notification zone (a3) ;
- the sub-entry identifier (a4).

Beyond the comparison with IPAL, a thorough evaluation of the resource has been done. According to lexicographers the linguistic description is globally accurate, relevant and motivated. The idea of a dynamic description is well received although it increases the complexity of the proposed description.

Some interesting features have also been noted from a linguistic point of view, like the fact that the first divide, for most verbs, is being concrete and abstract use (and not between more notional word senses). Case particle are more discriminative than head nouns for the definition of word senses at the construction level, which is not too surprising. The most fine grained classes often correspond to very specific word senses that are not always registered in

more general resources. A wide variety of idioms and frozen expressions can also be found at this level, making it possible to semi-automatically enriching existing resources.

6 Conclusion

We have shown in this study that it is now possible to develop new kinds of lexical resources based on continuous models and on the automatic analysis of very large corpora. Our system allows one to produce resources that can be finely tuned depending on the task and the expected precision of the foreseen result. Thanks to a relevant user interface, our resource is to the best of our knowledge the first one implementing the idea of a continuum-based representation usable by a professional lexicographer (contrary to most approaches producing a machine-coded language model that is unreadable by humans). The degree of argumenthood and the number of entries per verb can be tuned very easily through the interface and first experiments have proven that valuable, linguistically-motivated distinctions can be observed this way. The full integration of different versions of our resource in natural language processing tools (*e.g.* syntactic parsers) remains to be done. In the near future, different strategies will be explored to determine the best description for a given task, following previous experiments coupling lexical acquisition with practical tasks.

Acknowledgement

Pierre Marchal's research has been partially supported by a national "contrat doctoral" from the ministry of research.

References

1. Manning, C.D. In: Probabilistic syntax. R. Bod, J. Hay, S. Jannedy (2003) 289–341
2. Manning, C.D.: Automatic acquisition of a large subcategorization dictionary from corpora. In: Proceedings of the Meeting of the Association for Computational Linguistics. (1993) 235–242
3. Brent, M.R.: From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics **19** (1993) 203–222
4. Briscoe, T., Carroll, J.: Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC. (1997) 356–363
5. Korhonen, A.: Subcategorization acquisition. PhD thesis, University of Cambridge (2002)
6. Korhonen, A., Briscoe, T.: Extended lexical-semantic classification of english verbs. In Moldovan, D., Girju, R., eds.: Proceedings of the HLT-NAACL 2004: Workshop on Computational Lexical Semantics, Boston, Massachusetts, USA, Association for Computational Linguistics (May 2 - May 7 2004) 38–45

7. Preiss, J., Briscoe, T., Korhonen, A.: A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In: Proceedings of the Meeting of the Association for Computational Linguistics, Prague (2007) 912–918
8. Messiant, C., Poibeau, T., Korhonen, A.: Lexschem: a large subcategorization lexicon for french verbs. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco. (2008)
9. im Walde, S.S., Müller, S.: Using web corpora for the automatic acquisition of lexical-semantic knowledge. JLCL **28**(2) (2013) 85–105
10. Han, X., Zhao, T., Qi, H., Yu, H.: Subcategorization acquisition and evaluation for chinese verbs. In: Proceedings of the 20th International Conference on Computational Linguistics. COLING '04, Stroudsburg, PA, USA, Association for Computational Linguistics (2004)
11. Kawahara, D., Kurohashi, S.: Case frame compilation from the web using high-performance computing. In: Proceedings of the 5th International Conference on Language Resources and Evaluation. (2006) 1344–1347
12. Kawahara, D., Kurohashi, S.: A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. (2006) 176–183
13. Merlo, P., Esteve Ferrer, E.: The notion of argument in prepositional phrase attachment. Computational Linguistics **32**(3) (2006) 341–377
14. Abend, O., Rappoport, A.: Fully unsupervised core-adjunct argument classification. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. (2010) 226–236
15. Manning, C.D. In: Probalistic Syntax. The MIT Press, Cambridge, MA (2003) 289–341
16. Fabre, C., Bourigault, D.: Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. Journal of French Language Studies **18**(1) (2008) 87–102
17. Fabre, C., Frérot, C.: Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. In: Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2002). (2002) 215–224
18. Partee, B.H. In: Lexical Semantics and Compositionality. The MIT Press, Cambridge, MA (1995) 311–360
19. Mitchell, J.: Composition in Distributional Models of Semantics. PhD thesis, University of Edinburgh (2011)
20. Firth, J.R. In: A Synopsis of Linguistic Theory 1930–1955. Basil Blackwell, Oxford (1957) 1–32
21. Harris, Z.S.: Distributional structure. Word **10** (1954) 146–162
22. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
23. Kurohashi, S., Nagao, M.: Kn parser : Japanese dependency/case structure analyzer. In: Proceedings of the Workshop on Sharable Natural Language Resources. (1994) 48–55
24. Kudo, T., Matsumoto, Y.: Japanese dependency analysis using cascaded chunking. In: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). (2002) 63–69

25. Flannery, D., Miyao, Y., Neubig, G., Mori, S.: A pointwise approach to training dependency parsers from partially annotated corpora. Journal of Natural Language Processing **19**(3) (2012) 167–191
26. Sasano, R., Kawahara, D., Kurohashi, S., Okumura, M.: koubun/zyutugo-kou-kouzou kaiseki sisutemu knp no nagare to tokutyou (2013)
27. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis. In: Proceedings of EMNLP 2004. (2004) 230–237
28. Asahara, M., Matsumoto, Y.: Ipadic version 2.7.0 users manual (2003)
29. Nihongo Kizyutu Bunpô Kenkyûkai: gendai nihongo bunpou 2: dai-3-bu kaku to koubun; dai-4-bu voisu (2009)
30. Martin, S.E.: A reference grammar of Japanese. Yale University Press, New Haven and London (1975)
31. Information-technology Promotion Agency (IPA): Ipa lexicon of the japanese language for computers, basic japanese verbs (1987)