Lexical Knowledge Acquisition: Towards a Continuous and Flexible Representation of the Lexicon

Pierre Marchal* and Thierry Poibeau**

*ERTIM-INALCO,
Université Sorbonne Paris Cité
1 rue de Lille 75007 Paris, France
pierre.marchal@inalco.fr

**LATTICE (CNRS, École normale supérieure and Université Sorbonne nouvelle),
PSL Research University and Université Sorbonne Paris Cité
1 rue Maurice Arnoux 92120 Montrouge, France
thierry.poibeau@ens.fr

Abstract

The automatic acquisition of lexical knowledge is an important issue for natural language processing. Lots of work has been done since two decades in this domain, but we think there is still room for improvement as we need to develop both efficient and cognitively plausible models. In this paper, we focus on verbs since verbs is the pivot of the sentence and 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. This paper aims at showing that lexical descriptions based on multifactorial and continuous models can be used both by linguists and lexicographers, and provide a cognitively interesting model for lexical semantics. Our results are available online at: `http://marchal.er-tim.fr/ikf/`.

1 Background and Motivations

"You shall know a word by the company it keeps" [Firth, 1957]. This too well known citation from J.R. Firth motivates any lexicographic work today: it is widely accepted that word description cannot be achieved without the analysis of a large number of contexts extracted from real corpora. But this is not enough.

The recent success of deep learning approaches have shown that static representations of the lexicon are no longer appropriate. Continuous models offer a better representation of word meaning, because they encode intuitively valid and cognitively plausible principles: semantic similarity is relative, context-sensitive and depends on multiple-cue integration. However, these models have not been used for representing meaning in dictionaries written for humans.

One may think that these models are complex and convenient for machines, but that they are too abstract for humans. In this paper we defend the opposite idea. If continuous models offer a better representation of the lexicon, we must conceive new lexical databases that are usable by humans and have the same basis as these continuous models. There are arguments to support this view.

For example, it has been demonstrated that semantic categories have fuzzy boundaries and thus the number of word meanings per lexical item is to a large extent arbitrary [Tuggy, 1993]. Although this still fuels lots of discussions among linguists and lexicographers, we claim that a description can be more or less fine-grained while keeping the same accuracy and validity. Moreover, it has been demonstrated that lexical entries in traditional dictionaries overlap and different word meanings can be associated with a same example [Erk and McCarthy, 2009], showing that meaning cannot be sliced in separate and exclusive word senses.

The same problem also arises when it comes to differentiate arguments and adjuncts. As said in [Manning, 2003]: '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". A proper representation thus need to be based on some kind of continuity and should take into consideration the verb, the object, but also the preposition used as well as the wider context.

Some applications already address some of the needs of lexicographers in the era of big data, *i.e.* big corpora in this context. The most well known one is the SketchEngine [Kilgarriff *et al.*, 2014]. This tool has already provided invaluable services to lexicographers and linguists. It gives access to a synthetic view of the different usages of words in context. For example, the SketchEngine may give a direct view of all the subjects or complements of a verb, ranked by frequency or sorted according to various parameters. By exploding the representation, this tool provides an interesting view on the lexicon. However in our opinion it fails short to show the continuous nature of meaning.

Here we propose a system that combines the advantages of existing tools (a wide coverage database offering a synthetic view of a large vocabulary) with those of a dynamic representation. We focus on verbs since these lexical items offer the most complex syntactic and semantic behaviors. We also focus on Japanese that present a complex system of case markers that are generally semantically ambiguous. From this point of view Japanese is a lot more challenging than English (and the system could be easily adapted to English by substituting prepositions to case markers).

Practically, our system extract verbs along with their complements from a very large corpus. A complement is a lexical head (generally a noun) with a case marker. The system first extracts and stores a comprehensive set of information about verbs and complements. Hierarchical clustering techniques then makes it possible to dynamically group together lexical items with a similar behavior into a dendrogram. Since the representation is dynamic, the interface makes it possible to navigate the data and interactively explore the 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 [Manning, 1993; Brent, 1993] 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 [Briscoe and Carroll, 1997; Korhonen, 2002]. 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 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 [Messiant *et al.*, 2008], German [im Walde and Müller, 2013] or Chinese [Han *et al.*, 2004], 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 [Kawahara and Kurohashi, 2006a; 2006b]. 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, [Merlo and Esteve Ferrer, 2006] and [Abend and Rappoport, 2010] try to validate linguistic criteria with statistical measures. [Manning, 2003] proposes to estimate the probability of a subcategorization frame associated to verb. Lastly, [Fabre and Bourigault, 2008] following [Fabre and Frérot, 2002] 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:

$$\mathrm{Arg}_{v,c} = (1 + \log \mathrm{cnt}(v, c)) \log \frac{|V|}{|\{v' \in V : \exists(v', c)\}|} \quad (1)$$

where c is a complement (*i.e.* a tuple made of a lexical head and a case particle); v is a verb; $\mathrm{cnt}(v, c)$ 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, $1 + \log \mathrm{cnt}(v, c)$, 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.

```
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
    /* Let C be the set of complements $c$
       cooccurring with $v$                              */
    $C \longleftarrow \{c : c \in D \land \exists(v, D) \in W\}$;
    /* Let C' be the elements of $C$ sorted by
       decreasing TF-IDF value                           */
    $C' \longleftarrow [c : c \in C \land$ argumenthood($v$,$C'[i]$) $\geqslant$
    argumenthood($v$,$C'[i+1]$)];
    foreach complement $\mathbf{c}'$ of $C'$ do
        /* Let D' be 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

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" [Partee, 1995] as well as all the information involved in the composition process [Mitchell, 2011]. 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 [Firth, 1957; Harris, 1954] lexical heads can be represented in a vector space model [Salton *et al.*, 1975]. 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 cosine similarity, since it is (as for the TF-IDF weighting scheme) simple, efficient and perfectly suited to our problem.

The equation (3) shows how the cosine similarity can be calculated for two vectors $\mathbf{x}$ and $\mathbf{y}$ (the cosine similarity 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 The acquisition pipeline

4.1 Acquisition and preprocessing of textual data

We gathered a large collection of Japanese text from a selection of RSS feeds. We then filtered these feeds using XPath expressions in order to discard HTML markup and irrelevant content, such as navigation menus. To comply with external NLP tools (*i.e.* a POS tagger and a parser), we then applied specific preprocesses to the raw textual data: fullwidth form conversion, sentence splitting, etc. In the end, our corpus is made of more than 294 million 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] [Kurohashi and Nagao, 1994] and CaboCha[2] [Kudo and Matsumoto, 2002] (we are aware other parsers exist as well like EDA[3] [Flannery *et al.*, 2012]). In this work, we have decided to use CaboCha, for efficiency, among other reasons. Since CaboCha is faster than KNP [Sasano *et al.*, 2013], it seems more convenient to process large textual data. We use the default settings.

CaboCha is based on a tagger called MeCab[4] [Kudo *et al.*, 2004] that requires a dictionary of surface forms for tagging. Among the different possible dictionaries, we chose IPAdic [Asahara and Matsumoto, 2003], which 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 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

[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/

り (*yori*), まで (*made*), and と (*to*) [Nihongo Kizyutu Bunpô Kenkyûkai, 2009]. 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 [Martin, 1975] or [Nihongo Kizyutu Bunpô Kenkyûkai, 2009], 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 million verbal structures, 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.

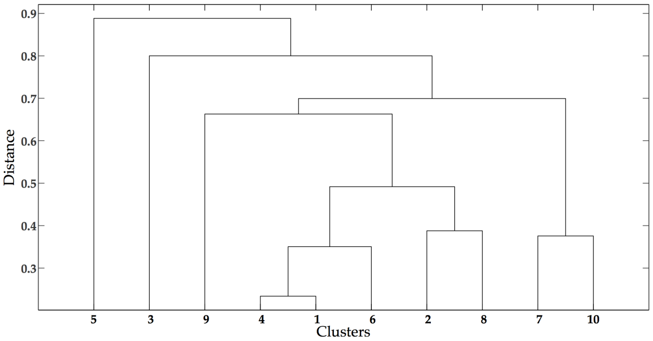


Figure 1: Dendrogram obtained after the hierarchical clustering of the ten first minimal predicate-frames of the verb 積む (*tumu*).

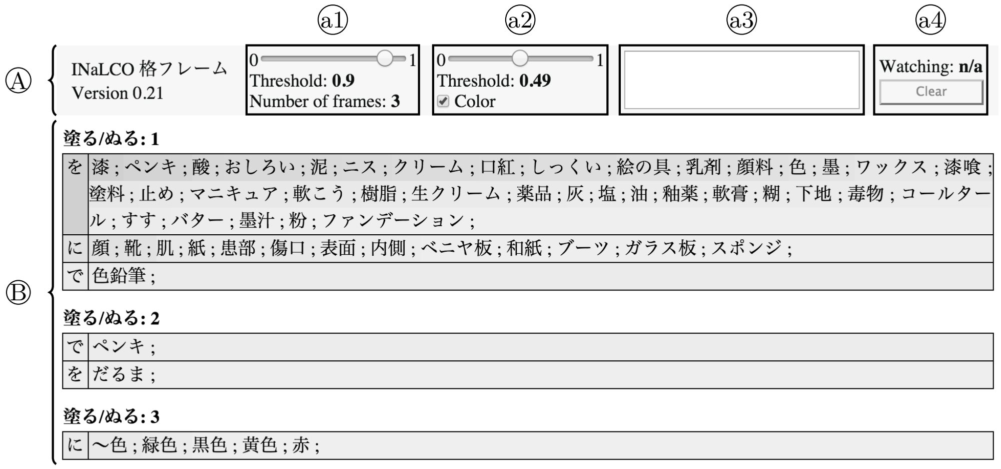
Each verb is thus described through a variable number of word senses, each word sense being itself defined by the different arguments attached to the verb. It is possible to explore the resource by navigating the hierarchy of word senses, *i.e.* by examining more or less fine-grained description. The 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 A visual interface to navigate the data

Lexical resource are traditionally evaluated through a comparison with a reference resource [Briscoe and Carroll, 1997; Korhonen, 2002]. Although this approach is intuitive, in our opinion 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 comparison with a manually built resource: IPAL [Information-technology Promotion Agency (IPA), 1987]. The results show similar results as for other languages *e.g.* [Messiant *et al.*, 2008]: 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 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.

Our challenge is thus twofold: we want to *i)* produce a resource that reflects the subtleties of continuous models but avoids the complexity of a multifactorial analysis and *ii)* offer a simple interface that allows a lexicographer or a linguist to navigate easily the data collection. The goal is



INaLCO 格フレーム
Version 0.21

Threshold: **0.9**
Number of frames: **3**

Threshold: **0.49**
☑ Color

Watching: **n/a**

塗る/ぬる: 1

を	漆 ; ペンキ ; 酸 ; おしろい ; 泥 ; ニス ; クリーム ; 口紅 ; しっくい ; 絵の具 ; 乳剤 ; 顔料 ; 色 ; 墨 ; ワックス ; 漆喰 ; 塗料 ; 止め ; マニキュア ; 軟こう ; 樹脂 ; 生クリーム ; 薬品 ; 灰 ; 塩 ; 油 ; 釉薬 ; 軟膏 ; 糊 ; 下地 ; 毒物 ; コールタール ; すす ; バター ; 墨汁 ; 粉 ; ファンデーション ;
に	顔 ; 靴 ; 肌 ; 紙 ; 患部 ; 傷口 ; 表面 ; 内側 ; ベニヤ板 ; 和紙 ; ブーツ ; ガラス板 ; スポンジ ;
で	色鉛筆 ;

塗る/ぬる: 2

で	ペンキ ;
を	だるま ;

塗る/ぬる: 3

に	〜色 ; 緑色 ; 黒色 ; 黄色 ; 赤 ;

Figure 2: Screen capture of our visualization tool – Ⓐ control panel: ⓐ1 slider for partitioning of sub-entries; ⓐ2 slider for selection of complements; ⓐ3 notification zone; ⓐ4 sub-entry identifier. – Ⓑ sub-entry panel.

of course to make the user discover interesting facts: new constructions, new idioms, and above all semantically related linguistic sequences made of words that would otherwise (*i.e.* in isolation) not be related.

After many attempts, we managed to propose a simple interface where the multifactorial analysis is abstracted as a double continuum: a continuum between ambiguity and vagueness [Tuggy, 1993], and a second continuum between arguments and adjuncts [Manning, 2003]. This double continuum is metaphorized just through two simple sliders.

Figure 2 shows a screen capture of our visualization tool. Slider ⓐ1 represents the continuum between ambiguity and vagueness. It sets a threshold on the dendrogram of the subentries; subentries which distance is less than the threshold are merged so as to make a single subentry. When the threshold is set to 0, each minimal predicate-frame corresponds to a distinct subentry; when set to 1 all minimal predicate-frames are merged into a single subentry. Slider ⓐ2 represents the continuum between arguments and adjuncts. It sets a threshold that selects complements that exhibit an argumenthood value greater than the threshold. Also, a color is associated to each lexical head so as to indicate its degree of argumenthood: a light color indicates a value close to 0 (an adjunct); a dark color indicates a value close to 1 (an argument). When the threshold is set to 0, all complements are displayed; when set to 1, only the complement with the highest degree of argumenthood is visible.

A lexicographer can make use of the two sliders to dynamically increase or decrease the number of subentries and complements. As the number of subentries can be important, we implemented various functionalities to help the end-user track the changes in the subentries. This is visible through the notification panel ⓐ3 that displays information about subentries that have merged or split, and an autofocus that makes it possible to freeze the subentry panel on a particular subentry (see ⓐ4).

First experiments with lexicographers have shown that the exploration of the lexicon makes it possible to find new verb usages. The interface is intuitive enough to allow them to gradually unveil the meanings of verbs, starting with discriminative syntactic patterns (*e.g.* transitive versus intransitive) or broad semantic classes of the complements (*e.g.* literal versus figurative), to finally discover – as constraints on the partitioning of subentries and on the selection of complements are released – more fine-grained and domain-dependant meanings of the verbs. Using this exploration method, one can also observe linguistic phenomena at the syntax/semantics interface – such as diathesis alternations, as shown in Figure 2 with the locative alternation of the verb *nuru* (to smear) – or verify prior assumptions that have been formulated in a different framework, especially the status of certain complements (*i.e.* arguments versus adjuncts).

The resource is publicly available through a Web interface at: `http://marchal.er-tim.fr/ikf/`.

6 Conclusion

In this paper we have proposed a novel approach to lexical acquisition, where meaning representation is represented as a continuum. The lexicographer car navigate the data and obtain more or less fine grained description, depending on his task and on his need. The approach has been evaluated on Japanese but is now being transferred to Finnish, a challenging language since Finnish is both agglutinative and highly inflectional (from 14 to 17 different cases can be distinguished, depending on the grammar taken into consideration). Finally we are also considering a practical evaluation through the integration of this resource into specific natural language applications.

7 Acknowledgements

Pierre Marchal has been partially funded by a PhD scholarship granted by INaLCO. Thierry Poibeau is partially funded by the PSL funded LAKME project, in the framework of the IDEX ANR-10-IDEX-0001-02 PSL*.

References

[Abend and Rappoport, 2010] Omri Abend and Ari Rappoport. Fully unsupervised core-adjunct argument classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 226–236, 2010.

[Asahara and Matsumoto, 2003] Masayuki Asahara and Yuji Matsumoto. Ipadic version 2.7.0 users manual, 2003.

[Brent, 1993] Michael R. Brent. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:203–222, 1993.

[Briscoe and Carroll, 1997] Ted Briscoe and John Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC., 1997.

[Erk and McCarthy, 2009] Katrin Erk and Diana McCarthy. Graded word sense assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 440–449, 2009.

[Fabre and Bourigault, 2008] Cécile Fabre and Didier Bourigault. Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(1):87–102, 2008.

[Fabre and Frérot, 2002] Cécile Fabre and Cécile Frérot. Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus. In *Actes de la $9^{\text{ème}}$ conférence sur le Traitement Automatique des Langues Naturelles (TALN 2002)*, pages 215–224, 2002.

[Firth, 1957] J.R. Firth. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32, 1957.

[Flannery *et al.*, 2012] Daniel Flannery, Yusuke Miyao, Graham Neubig, and Shinsuke Mori. A pointwise approach to training dependency parsers from partially annotated corpora. *Journal of Natural Language Processing*, 19(3):167–191, 2012.

[Han *et al.*, 2004] Xiwu Han, Tiejun Zhao, Haoliang Qi, and Hao Yu. Subcategorization acquisition and evaluation for chinese verbs. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[Harris, 1954] Zellig S. Harris. Distributional structure. *Word*, 10:146–162, 1954.

[im Walde and Müller, 2013] Sabine Schulte im Walde and Stefan Müller. Using web corpora for the automatic acquisition of lexical-semantic knowledge. *JLCL*, 28(2):85–105, 2013.

[Information-technology Promotion Agency (IPA), 1987] Information-technology Promotion Agency (IPA). Ipa lexicon of the japanese language for computers, basic japanese verbs, 1987.

[Kawahara and Kurohashi, 2006a] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1344–1347, 2006.

[Kawahara and Kurohashi, 2006b] Daisuke Kawahara and Sadao Kurohashi. 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*, pages 176–183, 2006.

[Kilgarriff *et al.*, 2014] Adam Kilgarriff, Vít Baisa, Jan Busta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. The sketch engine: Ten years on. *Lexicography*, 1(1):7–36, Jul 2014.

[Korhonen, 2002] Anna Korhonen. *Subcategorization acquisition*. PhD thesis, University of Cambridge, 2002.

[Kudo and Matsumoto, 2002] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 63–69, 2002.

[Kudo *et al.*, 2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of EMNLP 2004*, pages 230–237, 2004.

[Kurohashi and Nagao, 1994] Sadao Kurohashi and Makoto Nagao. Kn parser : Japanese dependency/case structure analyzer. In *Proceedings of the Workshop on Sharable Natural Language Resources*, pages 48–55, 1994.

[Manning, 1993] Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the Meeting of the Association for Computational Linguistics*, pages 235–242, 1993.

[Manning, 2003] Christopher D. Manning. Probabilistic syntax. In S. Jannedy R. Bod, J. Hay, editor, *Probabilistic Linguistics*, pages 289–341. MIT Press, 2003.

[Martin, 1975] Samuel Elmo Martin. *A reference grammar of Japanese*. Yale University Press, New Haven and London, 1975.

[Merlo and Esteve Ferrer, 2006] Paola Merlo and Eva Esteve Ferrer. The notion of argument in prepositional phrase attachment. *Computational Linguistics*, 32(3):341–377, 2006.

[Messiant *et al.*, 2008] Cédric Messiant, Thierry Poibeau, and Anna Korhonen. 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.

[Mitchell, 2011] Jeffrey Mitchell. *Composition in Distributional Models of Semantics*. PhD thesis, University of Edinburgh, 2011.

[Nihongo Kizyutu Bunpô Kenkyûkai, 2009] Nihongo Kizyutu Bunpô Kenkyûkai. gendai nihongo bunpou 2: dai-3-bu kaku to koubun; dai-4-bu voisu, 2009.

[Partee, 1995] Barbara H. Partee. *Lexical Semantics and Compositionality*, pages 311–360. The MIT Press, Cambridge, MA, 1995.

[Salton *et al.*, 1975] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[Sasano *et al.*, 2013] Ryohei Sasano, Daisuke Kawahara, Sadao Kurohashi, and Manabu Okumura. koubun/zyutugo-kou-kouzou kaiseki sisutemu knp no nagare to tokutyou, 2013.

[Tuggy, 1993] David Tuggy. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290, 1993.