Preliminary Experiments in the Extraction of Predicative Structures from a Large Finnish Corpus

Guersande Chaminade
CNRS & ENS & PSL & Université Sorbonne nouvelle & USPC
LATTICE Lab. & INaLCO
1 rue Maurice Arnoux 92120 Montrouge France
gu.chaminade@gmail.com

Thierry Poibeau
CNRS & ENS & PSL & Université Sorbonne nouvelle & USPC
LATTICE Lab.
1 rue Maurice Arnoux 92120 Montrouge France
thierry.poibeau@ens.fr

Abstract

This paper presents different experiments aiming at automatically extracting lexical information from corpora. We first describe a simple experiment on the nature of direct objects in Finnish (partitive vs total objects) so as to check if data automatically extracted from corpora support traditional studies in the field. Which are the verbs subcategorizing partitive objects? Which are the verbs subcategorizing (more frequently) total objects? We provide a tool and a user interface to browse data automatically collected from a large corpus in order to make it possible to answer this question. We then describe some ongoing work aiming at adapting to Finnish a complex lexical acquisition system initially developed for Japanese [1]. In our approach, 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 describe the system developed for the task and conclude with some perspectives for the evaluation of the produced resource.

1

The 3rd International Workshop for Computational Linguistics of Uralic Languages, pages 37–55,
St. Petersburg, Russia, 23–24 January 2017.

1 Introduction

Natural language processing traditionally requires precise and high coverage resources to provide satisfactory results but it is well known that the development of such resources is long and costly. Recently machine learning has made it possible to develop resources at a lower cost. It is now possible to automatically analyse large corpora, typically made of several million words, and develop parsers based on the observation of surface regularities at corpus level, along with some annotated data used for training. Powerful unlexicalized parsers have been developed for several languages, with a surprisingly high accuracy given the fact no lexical information is provided in input [2, 3].

The output of these parsers has subsequently been used as a new source of knowledge for the development of large-scale lexical resources. This area of research, known as lexical acquisition, have permitted the development of large-scale dictionaries for example for English [4], French [5], Japanese [1] and lots of other languages as well. It has also been shown that the approach provides interesting (although not perfect) results: thanks to this approach, it for example possible to discover new subcategorization frames for particular verbs [6, 7] and to monitor in real time the evolution of word usage or the creation of new words in a language [1], etc. Results obtained with automatic methods are of course still far from perfect: they need to be manually checked but they usually provide lots of new results and new sources of evidence for further work. One of the most obvious application is probably the fact that automatic methods make it possible to complete existing resources at a lower cost, so as to obtain a better coverage [6]. Automatic methods also provide statistical information, which is a key element for any computational linguistic system nowadays.

We think these approaches are important for linguistic analysis as well. Too often, natural language processing only focuses on practical applications, which are of course very important for the field. But we think it can also be highly valuable to use automatic tools to provide results based on large-scale analysis to linguists, who could then base their analyses on data that is otherwise hard to process directly.

Therefore, we describe in this paper different experiments concerning verb subcategorization frames in Finnish. Our goal is twofold:

1. provide tools and data for linguists, based on the analysis of large corpora. Linguists are for example interested in the study of differential object marking in Finnish: which verbs subcategorize partitive complements? Which verbs subcategorize total (accusative) objects? Which verbs subcategorize both kinds of complements? Although lots of studies have already explored this question, technology provides today an easy way to obtain quantified data computed over

very large corpora and thus may shed new light on such topics as differential object marking in Finnish (or any other Finno-Ugric language since the method can easily be transferred to another language).

2. use advanced clustering techniques to observe large-scale verb families based on their usage in corpora. Based on the hypothesis put forward by Levin [8], we think that, to a certain extent at least, syntactic behaviour can serve as a footstep to semantics. More precisely, we want to cluster verbs having the same syntactic behaviour and observe if in doing so we get relatively homogeneous semantic classes as a result.

The resource comes along with a user interface making it possible to navigate the data: we think a lexical resource should not be frozen but should be easily adaptable depending on the user need. More specifically we want the end user to be able to navigate and explore the data so as to get more or less fine-grained lexical descriptions in the lexicon.

The paper is structured as follows. We first give a quick overview of previous work in lexical acquisition. We then provide a brief reminder of Finnish objects and grammatical cases in Finnish. The following section describes the analysis of our corpus (the Finnish section of the Europarl corpus) using the Turku Parser [9] and how relevant information is identified and extracted. The following section describes the Web interface giving access to corpus information about the nature of objects (total vs partitive) depending on the verb considered. The last section describes some ongoing work on the extraction of families of verb constructions using clustering techniques. We describe how the system is derived from a previous implementation for Japanese [1], with of course an adaptation of all the language-dependent modules to Finnish.

2 Previous Work

The first works in automatic lexical acquisition date back to the early 1990s. The need for precise and comprehensive lexical databases was clearly identified as a major need for most NLP tasks (esp. parsing) and automatic acquisition techniques was then seen as a way to solve the resource bottleneck. However, the first experiments [10, 11] were limited (the acquisition process was dealing with a few verbs only and a limited number of predefined subcategorization frames). They were 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 [12, 13, 14]. A last innovation consisted

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 than the previous one, but most errors can be automatically filtered out since they tend to produce patterns with a very low frequency. Most experiments so far have been made on verbs (since verbs are supposed to have the most complex subcategorization frames), but the approach can also be extended to nouns and adjectives without too many problems [4].

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 [5], German [15], Chinese [16], or Japanese [1] 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 unlexicalized parser is also a key feature for the quality of the acquisition process.

To the best of our knowledge, there has not been any large-scale experiment for Finnish yet. However we are lucky to have access to large corpora of Finnish, as well as to relevant parsers. For our experiments on Finnish, we have used the Finnish part of the Finnish-English pair of the Europarl corpus ($6th$ version of the parallel corpus, `http://www.statmt.org/europarl/`), containing more than 29 million words. Europarl is a corpus extracted from the proceedings of the European Parliament between 1997 and 2011. The corpus addresses heterogeneous topics, which makes it possible to acquire a quite varied lexicon, although the style of the corpus is quite regular and formal.

In the near future we plan to use bigger corpora that are now available, especially the Turku Dependency Treebank (TDT) [9] that consists of 181K tokens and 13.5K sentences. Bigger corpora allow one to cover different verb usage and more verbs, since it is necessary to get a minimum number of occurrences in order to provide relevant information. The contrast between partitive and total objects can already be observed taking a threshold of 10 or 15 occurrences of a given verb (but more occurrences will of course give more robust and accurate results). For verb clustering, it seems hard to get relevant results for less than 100 occurrences per verb.

3 Finnish object and grammatical cases

We base our description on the general and widely available Finnish grammar by Fred Karlsson [17].

Finnish is (among many other things) characterized by a linguistic phenomenon

called “differential object marking”. In other words, the object of a given verb may be marked by different cases, depending on the verb, the noun and the overall meaning one wants to convey. The basic opposition is between partitive objects and accusative (or “total”) objects.

Partitive object occurs in three instances:

1. in negative sentences,
2. when the action expressed by the verb is irresultative,
3. when the object expresses an indefinite quantity. [17]

We have to take into account quantifiers since with a quantifier, the case does not depend on the verb but on the quantifier. We also chose to isolate negative verbal structures, since partitive is then mandatory for the object, thus neutralizing the free opposition between partitive and total object.

Along with partitive objects, Finnish also has another kind of direct complement known as accusative (or “total object”). The accusative expresses:

1. a resultative action,
2. a whole or a definite quantity in affirmative sentences.

According to [17], in Finnish “the accusative is not a uniform morphological case form as such, but a collective name given to a certain set of cases when they mark the object of the sentence. These cases are: nominative singular, which of course has no ending (Ø); genitive singular, with the ending -n; the -t accusative ending peculiar to personal pronouns; and the nominative plural in -t. The accusative, i.e. this set of case forms, appears as the case of the object in opposition to the partitive”.

The object takes t-accusative in the following cases : the object is a pronoun, the object is total (in which case, it takes accusative plural, identical to the nominative plural). The object takes n-accusative in all other cases. The object takes nominative when the object is a total object in a passive construction (identical to non-marked accusative), or an object of a verb conjugated in the imperative.

Karlsson [17] formalizes this through the three following rules:

1. The -t accusative always marks the object
   (a) in the plural
   (b) in personal pronouns.

2. A singular accusative object
   (a) usually takes -n
   (b) takes no ending with verbs in first and second person imperative, passive verbs, and some verbs of obligation

3. Numerals (except yksi ‘one’) have no accusative ending.

Karlsson [17] also says that “when determining the particular case of the object one must first check whether any of the conditions for the partitive hold; if so, the object must be in the partitive. The partitive is thus a ‘stronger’ object case than the accusative. Only after this, if none of the partitive object conditions are fulfilled, can one proceed to determine which of the accusative endings is the correct one”.

The case of the object is therefore accusative only if (a) the sentence is affirmative, and also (b) the action of the verb is resultative, or (c) the object is a whole or a definite quantity. With respect to (c), the accusative may be compared to the nominative when the nominative marks the subject

4 Verbal structure extraction

In this section, we describe how the corpus is analysed with the Turku Parser and how relevant information is automatically extracted for further analysis.

4.1 The parser

The first step consists in analysing the corpus so as to be able to identify the main relations between words. We chose to use the Turku parser as described in [18]. This is probably he most accurate parser for Finnish currently available.

For each sentence, the parser produces a tree where the main verb is the root. According to Haverinen et al. [9] “the annotation scheme of the treebank is a Finnish-specific version of the well-known Stanford Dependency (SD) scheme, originally developed by de Marneffe and Manning [19, 20]. The SD scheme represents the syntax of a sentence as a graph where the nodes represent the words of the sentence, and the edges represent directed dependencies between them. One of the two words connected by a dependency is the head or governor while the other is the dependent. Each dependency is labelled with a dependency type, which describes the syntactic function of the dependent word”.

Results reported in the [9] are “97.3% POS and 94.8% PM” (correct part-of-speech tags / correct part of speech tags + morphological information) and [18] gives results around .86 UAS and LAS (nodes that have the correct incoming arc / node that have the correct incoming arc with the correct label, i.e. the right syntactic tag).

The output has to be slightly modified so as to fit with our problem. The following mappings are defined : N-accusative matches genitive, T-Accusative matches accusative, non-marked accusative matches nominative (and partitive is partitive). Hence in this context, a partitive object takes the partitive case, and a total object

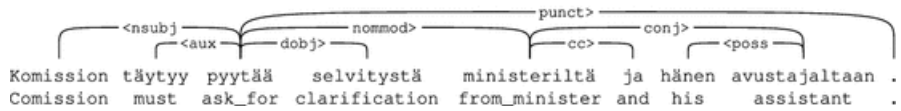

Figure 1: An example taken from [9] using the Stanford dependency scheme. The sentence can be translated as "*The commission must ask for clarification from the minister and his assistant.*"

takes either the genitive, or the accusative (or the nominative, although we mostly limited the process to the active voice).

4.2 Verb structure extraction from the CONLL format

The goal here is to generate a list of predicative structures, each containing three elements: the verb, its complements and the case used for each complement. The program takes as input a CONLL format file, which is the output by the Turku parser. The CONLL format is very convenient for this kind of extraction, as shown in the example below (the two empty slots were originally used to contain gold-standard data) :

```
1 kauden    kausi     NOUN _ Case=Gen|Nb=Sing 2 nmod:poss _ 1.1
2 avaaminen avaaminen NOUN _ Case=Nom|Nb=Sing 0 root      _ 1.1
3 on ...
```

Each line corresponds to one word (aka one lexical entry), and contains the following list of information :

1. an index referring to the word,
2. the inflected form (i. e. the form as it appears in the original text),
3. the corresponding lemma,
4. the POS tag (Universal Dependencies),
5. (empty)
6. morphological features,
7. the index of the word's governor / head (for root elements, this value is 0),
8. the dependency relation ,
9. (empty)
10. an index referring to the sentence where the word appears.

Since each line is a tabulation-separated list, it is quite simple to split it into an easy to handle data structure on which grammatical filters are applied. The CONLL09 format is also handy when it comes to transcribe the filters into sets of selection constraints. To give a very simple example, we state that for a word to be a verb (and thus be selected), the fourth item in the list must be the string 'VERB' (other restrictions are applied, as discussed in what follows).

4.3 Retrieving predicative structures from the CONLL09 format

We extract verbal structures that correspond to different selection criteria expressed though constraints. Verbal structure extraction happens only if both the expectations on the verb and on the complement are met. To achieve this, we use the following selection of constraints on the verb:

- the POS tag is 'VERB',
- the clause must be an independent clause or a subordinate clause (i.e. not an infinitive or participle clause),
- as for mood restriction, it is possible to choose between indicative, conditional and imperative; negative forms of the verb can also be selected but these forms are processed separately;
- finally, the voice of the verb can be active or passive.

These characteristics are then 'transcribed' in the CONLL09 format:

- the POS tag is the fourth slot of the data structure concerning the verb, so this constraint becomes data[3]='VERB'.
- since the verb is the root element of its clause, the nature of the clause is described using the dependency relationship of the verb. Thus the different possibilities are given as a list. Technically, things look like: extract data in ('root', 'ccomp', 'xcomp'), where 'root' corresponds to an independent clause, 'ccomp' to a subordinate clause, and 'xcomp' to a subordinate clause without its own subject.
- depending on which constraints one wants to use, the query is:
  - for the indicative and conditional modes: 'Mood=Ind','Mood=Cnd',

- for the imperative mood: 'Mood=Imp',
- if one only wants to work on negative / non-negative forms, a 'Connegative=Yes' feature is available,
- the voice of the verb can be either active 'Voice=Act', or passive 'Voice=Pass'.

5 Observing differential object marking in Finnish based on corpus data

In this experiment we are only interested in objects depending directly on the verb: the goal here is to classify verbs depending on the case they use for their direct object. Negative sentences (i.e. sentences with a negation) are excluded since the negation automatically entails the use of the partitive for the object.

5.1 Description of the Approach

The classification of verbs is done according to three categories:

1. verbs subcategorizing exclusively the partitive case,
2. verbs subcategorizing exclusively the accusative / genitive case,
3. verbs subcategorizing both cases.

The idea is to find under (1) and (2) verbs that impose specific constraints to their object, thus neutralizing the possible opposition between partitive and total objects, while verbs in the third category (3) are supposed to be really subject to differential object marking. Practically, we categorize a verb as a partitive object verb, if at least 85% of the direct objects appearing with this verb are marked with the partitive. We categorize a verb as a 'total object' verb, if at least 65% of the objects are marked with the genitive / accusative. Other verbs are categorized as mixed object verbs. These thresholds have been fixed empirically and can be modified. The threshold is not the same for the different categories because more verbs have partitive objects due to semantic reasons that are not directly connected to aspectual oppositions (for example when the object denotes an imprecise quantity that constrains the use of the partitive).

The case of (1) and (2) should be monitored thoroughly since (1) for example also contains verbs for which £total objects' can be found but in small quantities.

Technically, the verb complement must be either direct or indirect object ('dobj', iobj'). All the results are stored in a file that is accessible through a simple Web interface.

The following results are obtained using the Europarl corpus described above: 439 total object verbs, 1 269 partitive object verbs, and 683 mixed object verbs. These results are obtained without taking into account verb frequency. In order to obtain more reliable results we did some experiments with different thresholds.

- With a threshold of 20 instances, we get 61 total object verbs, 316 partitive object verbs, and 209 mixed object verbs.
- With a threshold of 50 instances, we get 49 total object verbs, 205 partitive object verbs, and 142 mixed object verbs.
- With a threshold of 100 instances, we get 34 total object verbs, 148 partitive object verbs, and 95 mixed object verbs.

These results show the great number of verbs appearing only a few times in the corpus: from 2436 without threshold, the total number of verbs falls to only 586 with a threshold of 20, 396 with a threshold of 50, and 277 with a threshold of 100. This is a direct consequence of the type of text used in this experiment: there are many different topics discussed during the European Parliament sessions, and only relatively frequent verbs remain when we increase the frequency threshold.

Table 1: Number of verbs depending on the frequency threshold.

Threshold	Verbs	Total Object		Partitive Object		Mixed	
20	586	61	10.4%	316	53.9%	209	35.7%
50	396	49	12,4%	205	51.7%	142	35,9%
100	277	34	12.3%	148	53.4%	95	34,3%

However, note that the results are rather stable once the less frequent verbs are removed. Partitive object verbs always represent more than half of the total, and thus outnumber the two other categories; mixed objects verbs always outnumber total object verbs (these two observations are also true if no threshold is set, as one can see from the figures given previously).

5.2 Results and Evaluation

The result of the classification is presented in a Web interface that shows the three categories (verbs mainly subcategorizing total objects, verbs mainly subcategorizing partitive objects, and verbs subcategorizing both kinds of objects).

The different categories are represented as columns, each one with a different colour: orange for total object verbs, blue for partitive object verbs and green for mixed object verbs. Moreover each verb entry is associated with different information: the list of the most frequent nouns subcategorized by the verb, a few typical examples, etc. It is also possible to look for a specific verb in the database. Figure 2 shows an overview of the graphical interface with some examples displayed for *ymmärtää*.

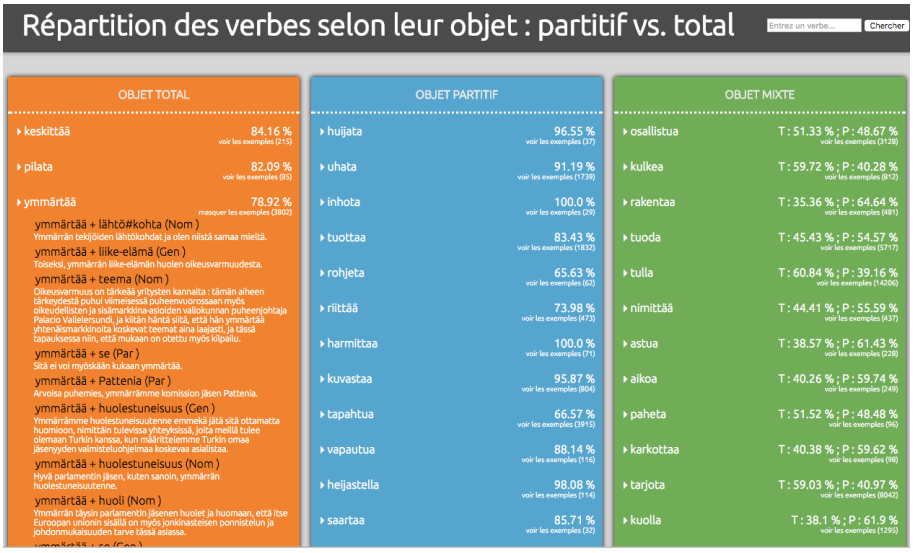

Figure 2: Web interface showing the different verbs sorted in three categories

The interface also provides information on other possible subcategorization frames for each verb. This is not the primary goal of this interface but it makes it possible for lexicographers to have a quick overview of the diversity of possible constructions for each verb.

Our first experiments with end-users prove that this tool is useful for linguists and lexicographers to check the behaviour of verbs in different contexts. It is also

useful for professors and even more students learning Finnish, since the interface provides a large number of examples. Most of the time the verb frames observed in the Europarl corpus support the traditional descriptions found in dictionarires (e.g. "*harmittaa*" subcategorizes the partitive, "*Minua harmittaa kuitenkin se, että...*", "*Tämä harmittaa minua.*", etc.) but language learners may be surprised by certain facts, like the proportion of total objects (i.e. nominative / genitive / accusative objects) for a verb like *ymmärtää*. For several language learners, *ymmärtää* is irresultative and should thus require the use of the partitive case, which is far from being true (e.g. "*Ymmärrämme tämän välttämättömän keskustelun taustat.*", "*Kansalainen ymmärtää nyt EU:n.*").

Our system just extracts information from the output of the Turku parser and thus should not make errors in itself. We noted a few errors in the output of the Turku NLP system2. When these errors were a problem for our analysis, we added some constraints to the extraction algorithm (for example, we noted a frequent error with lexical forms ending with '*-mme*', especially between the verbal and the possessive suffix. This was filtered out by the extraction algorithm).

6 Towards large-scale lexical acquisition for Finnish

Our goal is now to extend the work to the whole argument structure of the verb, following the line of research described in the state-of-the-art section. The input is the same very large corpus of raw text that is parsed by the unlexicalized Turku parser for Finnish (it is necessary to use an unlexicalized parser since we want to learn subcategorization frames and we do not want the process to be biased by pre-defined resources). By observing regularities at surface level, it is possible to infer subcategorization frames, i.e. infer the most probable constructions, separate arguments and adjuncts (also called modifiers) and have statistical information about the possible complements of a given verb.

We re-use the pipeline developed initially by Pierre Marchal for Japanese [21, 1, 22]. Since the original linguistic pipeline was built to study predicative structures extracted from Japanese, all language-dependent modules had to be changed and adapted to Finnish. This was mostly true for the initial linguistic pipeline: we used the Turku parser for this, as explained above. All the clustering modules (that are language-independent) have been used as they were in the original implementation.

As for grammatical cases, we considered the following cases: partitive, genitive, accusative (traditional object cases), locative cases (inessive, elative, illative, adessive, ablative and allative), and translative (we do not necessarily exclude other cases, although they appear less frequently except the nominative in passive structures).

We slightly modified the original interface to add examples that illustrate the different produced clusters. We think this is necessary because the end user can vary, via two sliders, two thresholds corresponding to two parameters. The first corresponds to the minimum distance between the minimal classes obtained with the first classification step. The second corresponds to the argumentality score of the complement based on a tf-idf measure (see below).

6.1 Description of our Approach

The starting point is 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 case marker.

6.1.1 Calculating the Argumenthood of Complements

Building on previous works [23, 24, 25, 26], Marchal [21, 1] proposes a new measure combining the prominent features describe in the literature. The measure is derived from the famous tf.idf used in information retrieval, with the major difference that we are dealing with complements instead of terms (or keywords), and with verbs instead of documents.

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

6.1.2 Minimal clustering at the verb entry level

Marchal [21, 1] introduces a method for merging verbal structures (i.e. a verb and a set of complements) into minimal predicate-frames structures using reliable lexical clues. He calls this technique *shallow clustering*. The technique is based on two principles: i) two verbal structures describing the same verb and having at least one common complement might correspond to the same verb meaning and ii) some complements are more informative than others for a given verb sense. The merging algorithm is presented at length in [21].

6.1.3 Modelling word senses through hierarchical clustering

Marchal [21, 1] proposes 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 operating at the input of the hierarchical clustering algorithm.

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" [27]. Similarly we propose to calculate the meaning of the verb structure (i.e. the construction) in function of the meaning of its parts. Following the principles of distributional semantics [28, 29] lexical heads can be represented in a vector space model [30]. Case markers (or prepositions) can be used as syntactic information. Finally, we propose to use 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 represents the dissimilarity between two occurrences 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 score) simple, efficient and perfectly suited to our problem.

Hierarchical clustering is an iterative process that clusters the two most similar elements of a set into a single element and repeats the operation 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 (the one that will preserve most of the information included in the distance matrix) we propose to apply the cophenetic correlation coefficient. The details of this technique are presented in [21, 1].

6.2 A visual interface to navigate the data

The major novelty of our approach is the description of predicative vocabulary of a language (here verbs in Finnish) 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.

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

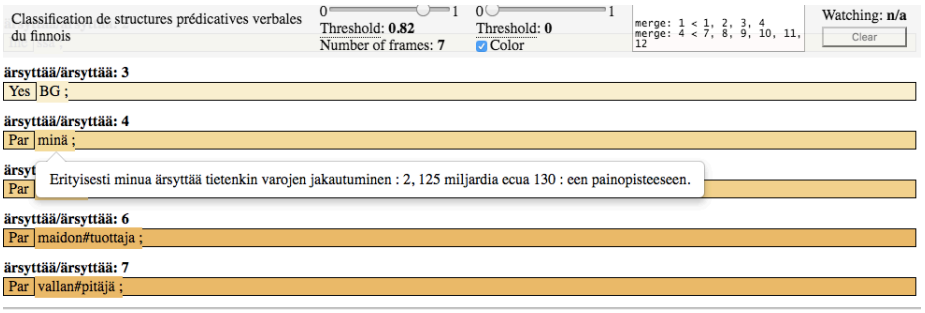

Figure 3: Web interface of the lexical acquisition system

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 [31], and a second continuum between arguments and adjuncts [25]. This double continuum is represented by two simple sliders.

Figure 3 shows a screen capture of our visualization tool. One can see two sliders on top of the interface: the first slider represents the continuum between ambiguity and vagueness (practically, this slider corresponds to a threshold in the dendrogram of the subentries; subentries with a distance inferior to the threshold are merged into a single subentry: when the threshold is set to 0, each minimal predicate-frame corresponds to a distinct subentry; when the threshold is set to 1 all minimal predicate-frames are merged into a single subentry). The second slider represents the continuum between arguments and adjuncts. It sets a threshold that selects complements that exhibit an argumenthood value greater than the threshold.

7 Conclusion

In this paper we have described some ongoing work on lexical acquisition for Finnish. The first application makes it possible to observe the partitive vs total object partition in this language. The second is a broader application aiming at acquiring a large database of verbal predicative structures in Finnish. The direct acquisition from a large corpus means that it is possible to get information on the use of verbs in context and also to collect statistics related to verb use. Statistics are especially important for nowadays applications based most of the time on a statistical approach.

The next stage consists in evaluating the results and checking their quality. This will in turn give us new perspectives to enhance the quality of the system, so as to

take into account more linguistic features that are important for the task. We also plan to use a larger corpus soon (the Turku Dependency Treebank (TDT) [9]) so as to get more results for more verbs. Lastly, we plan to validate our data against a gold standard for Finnish so that lexical acquisition systems for this language can be compared.

Acknowledgments

This work has been mainly developed in the framework of the LAKME project. LAKME is funded by a grant from Paris Sciences et Lettres within the framework of the IDEX (Initiatives d'Excellence) PSL reference ANR-10- IDEX-0001- 02. The authors are also partially supported by a RGNF-CNRS (grant between the LATTICE-CNRS Laboratory and the Russian State University for the Humanities in Moscow).

References

[1] Pierre Marchal and Thierry Poibeau. A Continuum-based Model of Lexical Acquisition. In *CICLing Conference on Intelligent Text Processing and Computational Linguistics*, Konya, Turkey, 2016.

[2] Eugene Charniak. Immediate-head parsing for language models. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01)*, pages 124–131, Toulouse, France, 2001.

[3] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL '03)*, pages 423–430, Sapporo, Japan, 2003.

[4] Judita Preiss, Ted Briscoe, and Anna Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Meeting of the Association for Computational Linguistics (ACL'07)*, pages 912–918, Prague, Czech Rep., 2007.

[5] 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'08)*, Marrakech, Morocco, 2008.

[6] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. Extending verbnet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May 2006.

[7] Cédric Messiant, Kata Gábor, and Thierry Poibeau. Lexical acquisition from corpora: the case of subcategorization frames in french. *Traitement Automatique des Langues*, 51(1):65–96, 2010.

[8] Beth Levin. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London, 1993.

[9] Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531, 2014.

[10] Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Meeting of the Association for Computational Linguistics (ACL'93)*, pages 235–242, Columbus, Ohio, USA, 1993.

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

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

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

[14] Anna Korhonen and Ted Briscoe. Extended lexical-semantic classification of english verbs. In Dan Moldovan and Roxana Girju, editors, *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 38–45, Boston, Massachusetts, USA, 2004.

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

[16] 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)*, Geneva, Switzerland, 2004.

[17] Fred Karlsson. *Finnish: An Essential Grammar*. 2nd ed, Routledge Essential Grammars, London, 2008.

[18] Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richard Farkas, Filip Ginter, and Jan Hajic. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1(4):415–428, 2013.

[19] Marie Catherine De Marneffe and Chris Manning. Stanford typed dependencies manual. *Technical report, Stanford University*, 2008.

[20] Marie Catherine De Marneffe and Chris Manning. Stanford typed dependencies representationl. *Technical report, Stanford University*, 2008.

[21] Pierre Marchal. *Acquisition de schémas prédicatifs verbaux en japonais*. PhD Thesis, INaLCO, 2015.

[22] Pierre Marchal and Thierry Poibeau. Lexical Knowledge Acquisition: Towards a Continuous and Flexible Representation of the Lexicon. In *Workshop IJCAI-Cognitum*, New York, 2016.

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

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

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

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

[27] Barbara H. Partee. Lexical semantics and compositionality. In Lila R. Gleitman and Mark Liberman, editors, *An invitation to cognitive science. Volume 1: Language*, pages 311–360, Cambridge, MA, 1995. The MIT Press.

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

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

[30] Grard Salton, Chung-Shu Yang, and Anita Wong. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

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