Clustering Hungarian Verbs on the Basis of Complementation Patterns

Kata Gábor
Dept. of Language Technology
Linguistics Institute, HAS
1399 Budapest, P. O. Box 701/518
Hungary
gkata@nytud.hu

Enikő Héja
Dept. of Language Technology
Linguistics Institute, HAS
1399 Budapest, P. O. Box 701/518
Hungary
eheja@nytud.hu

Abstract

Our paper reports an attempt to apply an unsupervised clustering algorithm to a Hungarian treebank in order to obtain semantic verb classes. Starting from the hypothesis that semantic metapredicates underlie verbs' syntactic realization, we investigate how one can obtain semantically motivated verb classes by automatic means. The 150 most frequent Hungarian verbs were clustered on the basis of their complementation patterns, yielding a set of basic classes and hints about the features that determine verbal subcategorization. The resulting classes serve as a basis for the subsequent analysis of their alternation behavior.

1 Introduction

For over a decade, automatic construction of wide-coverage structured lexicons has been in the center of interest in the natural language processing community. On the one hand, structured lexical databases are easier to handle and to expand because they allow making generalizations over classes of words. On the other hand, interest in the automatic acquisition of lexical information from corpora is due to the fact that manual construction of such resources is time-consuming, and the resulting database is difficult to update. Most of the work in the field of acquisition of verbal lexical properties aims at learning subcategorization frames from corpora e.g. (Pereira et al., 1993; Briscoe and Carroll, 1997; Sass, 2006). However, semantic grouping of verbs on the basis of their syntactic distribution or other quantifiable features has also gained attention (Schulte im Walde, 2000; Schulte im Walde and Brew, 2002; Merlo and Stevenson, 2001; Dorr and Jones, 1996). The goal of these investigations is either the validation of verb classes based on (Levin, 1993), or finding algorithms for the categorization of new verbs.

Unlike these projects, we report an attempt to cluster verbs on the basis of their syntactic properties with the further goal of identifying the semantic classes relevant for the description of Hungarian verbs' alternation behavior. The theoretical grounding of our clustering attempts is provided by the so-called Semantic Base Hypothesis (Levin, 1993; Koenig et al., 2003). It is founded on the observation that semantically similar verbs tend to occur in similar syntactic contexts, leading to the assumption that verbal semantics determines argument structure and the surface realization of arguments. While in English semantic argument roles are mapped to configurational positions in the tree structure, Hungarian codes complement structure in its highly rich nominal inflection system. Therefore, we start from the examination of case-marked NPs in the context of verbs.

The experiment discussed in this paper is the first stage of an ongoing project for finding the semantic verb classes which are syntactically relevant in Hungarian. As we do not have presuppositions about which classes have to be used, we chose an unsupervised clustering method described in (Schulte im Walde, 2000). The 150 most frequent Hungarian verbs were categorized according to their comp-

Proceedings of the ACL 2007 Student Research Workshop, pages 91–96,
Prague, June 2007.

lementation structures in a syntactically annotated corpus, the Szeged Treebank (Csendes et al., 2005). We are seeking the answer to two questions:

1. Are the resulting clusters semantically coherent (thus reinforcing the Semantic Base Hypothesis)?
2. If so, what are the alternations responsible for their similar behavior?

The subsequent sections present the input features [2] and the clustering methods [3], followed by the presentation of our results [4]. Problematic issues raised by the evaluation are discussed in [5]. Future work is outlined in [6]. The paper ends with the conclusions [7].

2 Feature Space

As currently available Hungarian parsers (Babarczy et al., 2005; Gábor and Héja, 2005) cannot be used satisfactorily for extracting verbal argument structures from corpora, the first experiment was carried out using a manually annotated Hungarian corpus, the Szeged Treebank. Texts of the corpus come from different topic areas such as business news, daily news, fiction, law, and compositions of students. It currently comprises 1.2 million words with POS tagging and syntactic annotation which extends to top-level sentence constituents but does not differentiate between complements and adjuncts.

When applying a classification or clustering algorithm to a corpus, a crucial question is which quantifiable features reflect the most precisely the linguistic properties underlying word classes. (Brent, 1993) uses regular patterns. (Schulte im Walde, 2000; Schulte im Walde and Brew, 2002; Briscoe and Carroll, 1997) use subcategorization frame frequencies obtained from parsed corpora, potentially completed by semantic selection information. (Merlo and Stevenson, 2001) approximates diathesis alternations by hand-selected grammatical features. While this method has the advantage of working on POS-tagged, unparsed corpora, it is costly with respect to time and linguistic expertise. To overcome this drawback, (Joanis and Stevenson, 2003) develop a general feature space for supervised verb classification. (Stevenson and Joanis, 2003) investigate the applicability of this general feature space to unsupervised verb clustering tasks. As unsupervised methods are more sensitive to noisy features, the key issue is to filter out the large number of probably irrelevant features. They propose a semi-supervised feature selection method which outperforms both hand-selection of features and usage of the full feature set.

As in our experiment we do not have a pre-defined set of semantic classes, we need to apply unsupervised methods. Neither have we manually defined grammatical cues, not knowing which alternations should be approximated. Hence, similarly to (Schulte im Walde, 2000), we represent verbs by their subcategorization frames.

In accordance with the annotation of the treebank, we included both complements and adjuncts in subcategorization patterns. It is important to note, however, that not only practical considerations lead us to this decision. First, there are no reliable syntactic tests for differentiating complements from adjuncts. This is due to the fact that Hungarian is a highly inflective, non-configurational language, where constituent order does not reveal dependency relations. Indeed, complements and adjuncts of verbs tend to mingle. In parallel, Hungarian presents a very rich nominal inflection system: there are 19 case suffixes, and most of them can correspond to more than one syntactic function, depending on the verb class they occur with. Second, we believe that adjuncts can be at least as revealing of verbal meaning as complements are: many of them are not productive (in the sense that they cannot be added to any verb), they can only appear with predicates the meaning of which is compatible with the semantic role of the adjunct. For these considerations we chose to include both complements and adjuncts in subcategorization patterns.

Subcategorization frames to be extracted from the treebank are composed of case-marked NPs and infinitives that belong to a children node of the verb's maximal projection. As Hungarian is a non-configurational language, this operation simply yields a non-ordered list of the verb's syntactic dependents. There was no upper bound on the number of syntactic dependents to be included in the frame. Frame types were obtained from individual frames by omitting lexical information as well as every piece of morphosyntactic description except for the POS tag and the case suffix. The generalization yielded 839 frame types altogether.[1]

3 Clustering Methods

In accordance with our goal to set up a basis for a semantic classification, we chose to perform the first clustering trial on the 150 most frequent verbs in the Szeged Treebank. The representation of verbs and the clustering process were carried out based on (Schulte im Walde, 2000). The data to be compared were the maximum likelihood estimates of the probability distribution of verbs over the possible frame types:

$$p(t|v) = \frac{f(v,t)}{f(v)} \tag{1}$$

with $f(v)$ being the frequency of the verb, and $f(v,t)$ being the frequency of the verb in the frame. These values have been calculated for each of the 150 verbs and 839 frame types.

Probability distributions were compared using *relative entropy* as a distance measure:

$$D(x\|y) = \sum_{i=1}^{n} x_i \cdot \log \frac{x_i}{y_i} \tag{2}$$

Due to the large number of subcategorization frame types, verbs' representation comprise a lot of zero probability figures. Using relative entropy as a distance measure compels us to apply a smoothing technique to be able to deal with these figures. However, we do not want to lose the information coded in zero frequencies - namely, the presumable incompatibility of the verb with certain semantic roles associated with specific case suffixes. Since we work with the 150 most frequent verbs, we wish to use a method which is apt to reflect that a gap in the case of a high-frequency lemma is more likely to be an impossible event than in the case of a relatively less frequent lemma (where it might as well be accidental). That is why we have chosen the smoothing technique below:

$$\begin{gathered} f_{\mathrm{e}} = \frac{0,001}{f(v)} \quad \text{if} \\ f_{\mathrm{c}}(t,v) = 0 \end{gathered} \tag{3}$$

[1] The order in which syntactic dependents appear in the sentence was not taken into account.

where f_e is the estimated and f_c is the observed frequency.

Two alternative bottom-up clustering algorithms were then applied to the data:

1. First we employed an agglomerative clustering method, starting from 150 singleton clusters. At every iteration we merged the two most similar clusters and re-counted the distance measures. The problem with this approach, as Schulte im Walde notes on her experiment, is that verbs tend to gather in a small number of big classes after a few iterations. To avoid this, we followed her in setting to four the maximum number of elements occuring in a cluster. This method - and the size of the corpus - allowed us to categorize 120 out of 150 verbs into 38 clusters, as going on with the process would have led us to considerably less coherent clusters. However, the results confronted us with the *chaining effect*, i.e. some of the clusters had a relatively big distance between their least similar members.

2. In the second experiment we put a restriction on the distance between each pair of verbs belonging to the same cluster. That is, in order for a new verb to be added to a cluster, its distance from all of the current cluster members had to be smaller than the maximum distance stated based on test runs. In this experiment we could categorize 71 verbs into 23 clusters. The convenience of this method over the first one is its ability to produce popular yet coherent clusters, which is a particularly valuable feature given that our goal at this stage is to establish basic verb classes for Hungarian. However, we are also planning to run a top-down clustering algorithm on the data to get a probably more precise overview of their structure.

4 Results

With both methods we describe in Section 3, a big part of the verbs showed a tendency to gather together in a few but popular clusters, while the rest of them were typically paired with their nearest synonym (e.g.: *zár* (close) with *végez* (finish) or antonym (e.g.: *ül* (*sit*) with *áll* (stand)). Naturally, method 1 (i.e. placing an upper limit on the number of verbs within a cluster) produced more clusters and gave more valuable results on the least frequent verbs. On the other hand, method 2 (i.e. placing an upper limit on the distance between each pair of verbs within the class) is more efficient for identifying basic verb classes with a lot of members. Given our objective to provide a Levin-type classification for Hungarian, we need to examine whether the clusters are semantically coherent, and if so, what kind of semantic properties are shared among class members. The three most popular verb clusters were investigated first, because they contain many of the most frequent verbs and yet are characterized by strong inter-cluster coherence due to the method used. The three clusters absorbed one third of the 71 categorized verbs. The clusters are the following:

C-1 VERBS OF BEING: *marad* (remain), *van* (be), *lesz* (become), *nincs* (not being)

C-2 MODALS: *megpróbál* (try out), *próbál* (try), *szokik* (used to), *szeret* (like), *akar* (want), *elkezd* (start), *fog* (will), *kíván* (wish), *kell* (must)

C-3 MOVEMENT VERBS: *indul* (leave), *jön* (come), *elindul* (depart), *megy* (go), *kimegy* (go out), *elmegy* (go away)

Verb clusters C-1 and C-3 exhibit intuitively strong semantic coherence, whereas C-2 is best defined along syntactic lines as 'modals'. A subclass of C-2 is composed of verbs which express some mental attitude towards undertaking an action, e.g. (*szeret* (like), *akar* (want), *kíván* (wish)), but for the rest of the verbs it is hard to capture shared meaning components.

It can be said in general about the clusters obtained that many of them can be anchored to general semantic metapredicates or one of the arguments' semantic role, e.g.: CHANGE OF STATE VERBS (*erősödik* (get stronger), *gyengül* (intransitive weaken), *emelkedik* (intransitive rise)), verbs with a beneficiary role (*biztosít* (guarantee), *ad* (give), *nyújt* (provide), *készít*(make)), VERBS OF ABILITY (*sikerül* (succeed), *lehet* (be possible), *tud* (be able, can)). Some clusters seem to result from a tighter semantic relation, e.g. VERBS OF APPEARANCE or VERBS OF JUDGEMENT were put together. In other cases the relation is broader as verbs belonging to the class seem to share only aspectual characteristics, e.g. AGENTIVE VERBS OF CONTINUOS ACTIVITIES (*ül* (be sitting), *áll* (be standing), *lakik* (live somewhere), *dolgozik* (work)). At the other end of the scale we find one group of verbs which 'accidentally' share the same syntactic patterns without being semantically related (*foglalkozik* (deal with sg), *találkozik* (meet sy), *rendelkezik* (dispose of sg)).

5 Evaluation and Discussion

As (Schulte im Walde, 2007) notes, there is no widely accepted practice of evaluating semantic verb classes. She divides the methods into two major classes. The first type of methods assess whether the resulting clusters are coherent enough, i. e. elements belonging to the same cluster are closer to each other than to elements outside the class, according to an independent similarity/distance measure. However, relying on such a method would not help us evaluating the semantic coherence of our classes. The second type of methods use gold standards. Widely accepted gold standards in this field are Levin's verb classes or verbal WordNets. As we do not dispose of a Hungarian equivalent of Levin's classification – that is exactly why we experiment with automatic clustering – we cannot use it directly.

We also run across difficulties when considering Hungarian verbal WordNet (Kuti et al., 2005) as the standard for evaluation. Mapping verb clusters to the net would require to state semantic relatedness in terms of WordNet-type hierarchy relations. However, if we try to capture the distance between verbal meanings by the number of intermediary nodes in the WordNet, we face the problem that the semantic distance between mother-children nodes is not uniform.

As our work is about obtaining a Levin-type verb classification, it could be an obvious choice to evaluate semantic classes by collecting alternations specific to the given class. Hungarian language hardly lends itself to this method because of its peculiar syntactic features. The large number of subcategorization frames and the optionality of most complements and adjuncts yield too much possible alterna-

	acc	ins	abl	ela
indul	-	ins/com	source	source
jön	-	ins/com	source	source
elindul	-	ins/com	source	source
megy	-	ins/com	source	source
kimegy	-	ins/com	source	source
elmegy	-	ins/com	source	source

Table 1: The semantic roles of cases beside C-3 verb cluster

	acc	ins	abl	ela
marad	-	com	cause	material
van	-	com	cause	material
lesz	-	com	cause	material
nincs	-	com	cause	material

Table 2: The semantic roles of cases beside C-1 verb cluster

tions. Hence, we decided to narrow down the scope of investigation. We start from verb clusters and the meaning components their members share. Then we attempt to discover which semantic roles can be licenced by these meaning components. If verbs in the same cluster agree both in being compatible with the same semantic roles and in the syntactic encoding of these roles, we consider that they form a correct cluster.

To put it somewhat more formally, we represent verb classes by matrices with a) nominal case suffixes in columns and b) individual verb lemmata in rows. The first step of the evaluation process is to fill in the cells with the semantic roles the given suffix can code in the context of the verb. We consider the clusters correct, if the corresponding matrices meet two requirements:

1. They have to be specific to the cluster.
2. Cells in the same column have to contain the same semantic role.

Tables 1. and 2. illustrate coherent and distinctive case matrices[2].

According to Table 1. ablative case, just as elative, codes a physical source in the environment of movement verbs. Both cases having the same semantic role, the decision between them is determined by the semantics of the corresponding NP. These cases code an other semantic role – cause – in the case of verbs of existence (Table 2).

It is important to note that we do not dispose of a preliminary list of semantic roles. To avoid arbitrary or vague role specifications, we need more than one persons to fill in the cells, based on example sentences.

6 Future Work

There are two major directions regarding our future work. With respect to the automatic clustering process, we have the intention of widening the scope of the grammatical features to be compared by enriching subcategorization frames by other morphological properties. We are also planning to test top-down clustering methods such as the one described in (Pereira et al., 1993). On the long run, it will be inevitable to make experiments on larger corpora. The obvious choice is the 180 million words Hungarian National Corpus (Váradi, 2002). It is a POS-tagged corpus but does not contain any syntactic annotation; hence its use would require at least some partial parsing such as NP analysis to be employable for our purposes. The other future direction concerns evaluation and linguistic analysis of verb clusters. We define well-founded verb classes on the basis of semantic role matrices. These semantic roles can be filled in a sentence by case-marked NPs. Therefore, evaluation of automatically obtained clusters presupposes the definition of such matrices, which is our major linguistic task in the future. When we have the supposed matrices at our disposal, we can start evaluating the clusters via example sentences which illustrate case suffix alternations or roles characteristic to specific classes.

7 Conclusions

The experiment of clustering the 150 most frequent Hungarian verbs is the first step towards finding the semantic verb classes underlying verbs' syntactic distribution. As we did not have presuppositions

[2] *Com* is for comitative – approximately encoding the meaning 'together with' , *ins* is for the instrument of the described event, *source* denotes a starting point in the space, *cause* refers to entity which evoked the eventuality described by the verb.

about the relevant classes, neither any gold standard for automatic evaluation, the results have to serve as input for a detailed linguistic analysis to find out at what extent they are usable for the syntactic description of Hungarian. However, as demonstrated in Section 4, the verb clusters we got show surprisingly transparent semantic coherence. These results, obtained from a corpus which is by several orders of magnitude smaller than what is usual for such purposes, is a reinforcement of the usability of the Semantic Base Hypothesis for language analysis. Our further work will emphasize both the refinement of the clustering methods and the linguistic interpretation of the resulting classes.

References

Anna Babarczy, Bálint Gábor, Gábor Hamp, András Kárpáti, András Rung and István Szakadát. 2005. Hunpars: mondattani elemző alkalmazás [Hunpars: A rule-based sentence parser for Hungarian]. *Proceedings of the 3th Hungarian Conference of Computational Linguistics (MSZNY05)*, pages 20-28, Szeged, Hungary.

Michael R. Brent. 1993. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262, MIT Press, Cambridge, MA, USA.

Ted Briscoe and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, pages 356–363, Washington, DC, USA.

Dóra Csendes, János Csirik, Tibor Gyimóthy and András Kocsor. 2005. The Szeged Treebank. *LNCS series Vol. 3658*, 123-131.

Bonnie J. Dorr and Doug Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. *Proceedings of the 14th International Conference on Computational Linguistics (COLING-96)*, pages 322–327, Kopenhagen, Denmark.

Kata Gábor and Enikő Héja. 2005. Vonzatok és szabad határozók szabályalapú kezelése [A Rule-based Analysis of Complements and Adjuncts]. *Proceedings of the 3th Hungarian Conference of Computational Linguistics (MSZNY05)*, pages 245-256, Szeged, Hungary.

Eric Joanis and Suzanne Stevenson. 2003. A general feature space for automatic verb classification. *Proceedings of the 10th Conference of the EACL (EACL 2003)*, pages 163–170, Budapest, Hungary.

Jean-Pierre Koenig, Gail Mauner and Breton Bienvenue. 2003. Arguments for Adjuncts. *Cognition*, 89, 67-103.

Judit Kuti, Péter Vajda and Károly Varasdi. 2005. Javaslat a magyar igei WordNet kialakítására [Proposal for Developing the Hungarian WordNet of Verbs]. *Proceedings of the 3th Hungarian Conference of Computational Linguistics (MSZNY05)*, pages 79–87, Szeged, Hungary.

Beth Levin. 1993. *English Verb Classes And Alternations: A Preliminary Investigation*. Chicago University Press.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics, 27(3)*, pages 373-408.

Fernando C. N. Pereira, Naftali Tishby and Lillan Lee. 1993. Distributional Clustering of English Words. *31st Annual Meeting of the ACL*, pages 183-190, Columbus, Ohio, USA.

Bálint Sass. 2006. Igei vonzatkeretek az MNSZ tagmondataiban [Exploring Verb Frames in the Hungarian National Corpus]. *Proceedings of the 4th Hungarian Conference of Computational Linguistics (MSZNY06)*, pages 15–22, Szeged, Hungary.

Sabine Schulte im Walde. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 747–753, Saarbrücken, Germany.

Sabine Schulte im Walde and Chris Brew. 2002. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223-230, Philadelphia, PA.

Sabine Schulte im Walde. to appear. The Induction of Verb Frames and Verb Classes from Corpora. *Corpus Linguistics. An International Handbook.*, Anke Lüdeling and Merja Kytö (eds). Mouton de Gruyter, Berlin.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised Verb Class Discovery Using Noisy Features. *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL-03)*, pages 71-78, Edmonton, Canada.

Tamás Váradi. 2002. The Hungarian National Corpus. *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 385–389, Las Palmas, Spain.