Automatic Acquisition of Hungarian Subcategorization Frames

András Serény,[1] Eszter Simon,[2] Anna Babarczy[2]

1. Applied Logic Laboratory, H-1022 Budapest, Hankóczy J. u. 7, sandris@all.hu

2. Budapest University of Technology, Cognitive Science Department, H-1111 Budapest, Stoczek utca 2. babarczy@cogsci.bme.hu, esimon@cogsci.bme.hu,

Abstract: Research on the automatic acquisition of lexical knowledge has recently gained increased attention in computational linguistics. Certain linguistic phenomena (e.g., ambiguity) pose computational problems that can only be solved if the system has access to lexical knowledge in general and to verb subcategorization frames in particular. Automatic subcategorization learning systems have been developed for most European languages. In the present paper a Hungarian adaptation of successful models constructed for other languages is discussed. The key approach adopted for our model is a statistical learning mechanism originally devised by Brent [6] and applied in a number of systems. In addition to, and in parallel with, the development of a system for the automatic acquisition of subcategorization frames, our project in progress has the broader aim of modelling the mechanisms of child language acquisition, specifically the process of learning argument structures (subcategorization frames) from the input available to young children. The outcome of our computational model will be tested against child language corpora: the model is taken to be successful if the development curves characterizing the machine learning algorithms match the characteristic U-shaped acquisition curves observed in child language.

Keywords: computational linguistics, automatic subcategorization frame acquisition, language acquisition, psycholinguistics

1 Introduction

A major part of the literature on natural language processing concerns the machine acquisition of some form of lexical knowledge, that is, the automatic construction of dictionaries. There are two main reasons why this is important. First, manual

development is tedious, time-consuming, cannot account for the constant growth and changes in the lexicon and do not provide information easily obtained by computational means (e.g. frequency data). That is, automatic acquisition of lexical knowledge is invaluable in maintaining a consistent lexicon with wide coverage and in keeping its contents up to date. Second, machine-readable dictionaries are needed by higher-level computational tools; for instance, only by exploiting lexical information is it possible to resolve attachment ambiguities. Ambiguity appears at every stage in language processing and the lexicon plays an important role in its resolution. For example, the following sentence admits two interpretations which correspond to two different syntactic structures: *Salespeople sold the dog biscuits*. The source of ambiguity is that the verb *sell* has (at least) two frames: *sell + indirect object + direct object* and *sell + direct object*. In the first case salespeople sold biscuits to the dog, in the second case salespeople sold the small hard biscuits fed to dogs. However, for the sentence *Salespeople gave the dog biscuits* only the analogue of the first interpretation is possible since the verb *give* has no frame *give + direct object*; in this case, a potential ambiguity is resolved in view of our knowledge of frames.

Early attempts at automatic lexicon building used electronic versions of dictionaries made for non-computational purposes as their main resource (e.g., [7], [25]). This is the automatic procedure most closely paralleling manual processing and, as such, it is burdened with the main shortcomings of non-automatic methods: it is not flexible enough and does not allow automatic expansion, and it is therefore of limited applicability.

A more robust approach is to attempt to retrieve information on verb subcategorization frames from large corpora. Text corpora suitable for machine learning applications are now available in most European languages, including Hungarian (e.g., the Szeged Corpus [10]).

Corpus-based argument structure retrieval is a major research topic mostly applied to English (e.g., [27]) but also several other European languages ([13], [19], [30]). In our paper a Hungarian adaptation of successful models constructed for other languages is discussed. The key approach adopted for our model is a classic statistical learning mechanism originally devised by Brent [6] and later applied in a number of systems. Our model can be seen as complementing a recent Hungarian language system [26], which is aimed at retrieving idiomatic, non-compositional verbal constructions containing specific lemmas.

Subcategorization frames, or argument frames, are defined here as the linguistic information showing the case roles of verbal arguments, such as morphological markings. The task is therefore to decide for each verb given in the initial lexicon whether it can be mapped onto a given subcategorization frame or, more precisely, to assign a probability of a verb occurring with a given subcategorization frame.

To solve the problem of argument frame retrieval, the first step is to define relevant statistical features. Most approaches involve computing the probability of the co-occurrence of a given verb and a given potential argument ([8], [21]). Co-occurrence probabilities can be more or less refined by varying the amount of information the model is sensitive to (e.g., morphology, word order, sentence structure, etc.), which is, of course, contingent on the amount of annotation detail given in the corpus.

In addition to Brent's method we have implemented two more procedures: a likelihood ratio test and a decision technique based on relative frequencies. These techniques are presented in Section 2. The methods were tested on two Hungarian corpora: in Section 3 our evaluation method and the results are described. Last but not least we discuss the conclusions and the psycholinguistic aspects of automatic subcategorization frame acquisition.

2 Experiments

2.1 Binomial Hypothesis Test

Brent was the first to use the following algorithm to extract verb frames from text corpora. Suppose we have a fixed set *F* of frames and a set *V* of verbs and for each pair $(f,v) \in F \times V$ we want to make a decision based on statistical evidence whether the verb *v* takes the frame *f*. First, for each frame $f \in F$ let us define a pattern of words and syntactic categories which indicate the presence of the frame with a high certainty. We call such form patterns *cues* for frame *f*. For example, the obvious cue for the English *transitive* verb frame might be written as *VERB NP* meaning that the verb must be followed by an NP in the sentence. (We shall shortly see examples of cues for Hungarian frames.) Clearly, cues are no infallible indicators of frames, hence we assign a probability of error to each cue: this is the probability that the cue appears in a sentence even thought the frame does not appear in the sentence. The method requires that cues belonging to the same frame should have the same probability of error.

Once the cues have been chosen, we perform hypothesis testing to decide, whether a frame *f* is appropriate for a verb *v*. Our null hypothesis is that the frame is not appropriate for the verb; we reject this null hypothesis if there is sufficient statistical evidence against it. Suppose that the verb *v* occurs *n* times in the corpus and there are *C(v,f)* occurrences together with a cue for frame *f*. Now,

$$p_e = P(C(v,f) \geq m \mid v \text{ does not take } f) = \sum_{r=m}^{n} \binom{n}{r} \varepsilon^r (1-\varepsilon)^{n-r} \qquad (1)$$

is the probability that cues for f occur m or more times together with v, where ε is the error probability for f. If p_e is smaller, than a given threshold (the significance) then we reject our null hypothesis and decide that the verb can take the frame.

There are several principled and less principled ways to choose the error probabilities ε. If we had a hand annotated corpus with true occurrences of frames marked on sentences, we would be able to estimate the error probabilities by the relative frequency of inaccurately occurring cues. With no annotated corpus at our disposal, we either make an educated guess or use more complicated estimation techniques. Korhonen et al. [14] estimated the error probability for a frame f by counting verbs that take f in the ANLT dictionary [3] and applying the formula

$$\varepsilon = \left(1 - \frac{|\text{verbs taking } f|}{|\text{all verbs}|}\right) \frac{|\text{cues for } f|}{|\text{cues}|} \tag{2}$$

We favoured a different approach, based on the ideas of Brent. For some fixed number N let us consider the first N occurrences of each verb that occurs at least N times in a corpus. For $1 \leq i \leq N$, let us count how many distinct verbs occurred with a cue for f exactly i times. An example with the cue *CAS<ACC>* (the morphological annotation code for accusative case) for the Hungarian transitive verb frame shown in Figure 1.

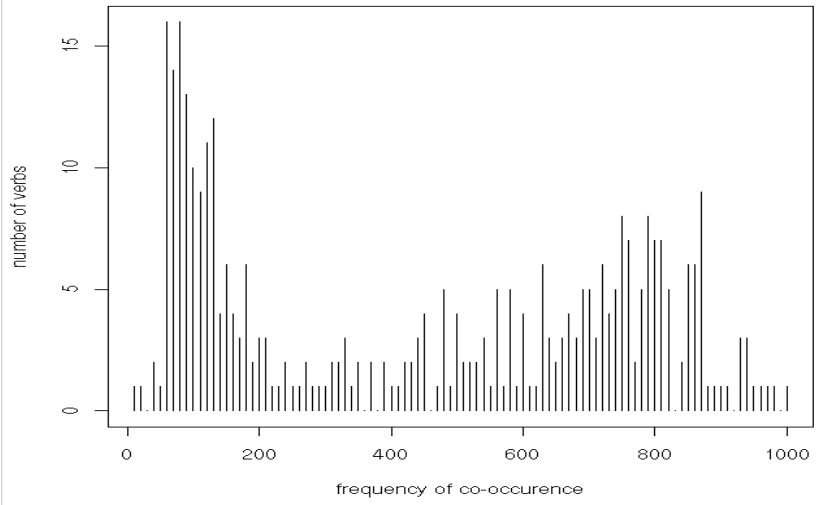

Figure 1
The probability of accusative case

Having a reasonable cue, we may assume that there is a number i_0 such that most intransitive verbs co-occur with the cue i_0 times or fewer, while truly transitive verbs co-occur with the cue i_0 times or more, hence intransitive verbs contribute to

columns on the left hand side of the figure. If our model is correct, the left hand side of the figure should have a roughly (skewed) binomial shape and this leads to an estimate of i_0, which, in turn, yields an estimate of the error probability ε.

Due to the variable word order characteristic of Hungarian, we cannot rely on exploiting particular linear configurations alone when we create cues. On the other hand, Hungarian is an agglutinative language with rich case marking, and morphological case markers and postpositions lend themselves to being used as building blocks for cues. So the cues we use are regular expressions over the alphabet of the KR-code used for morphological annotation in our corpora [16], which we match against strings of morphological description. For example, the cue for the Hungarian *ditransitive* verb frame (verbs taking a complement in the accusative and a complement in the dative in either order, e.g. *ad vkinek vmit (give someone*-dat *something*-acc*)*) has the following pattern: (CAS<ACC>.*CAS<DAT>) | (CAS<DAT>.*CAS<ACC>). In this paper only simple clauses are considered, but we try to take into account all the case inflections as well as many postpositions.

2.2 Likelihood ratio test

The likelihood ratio test is a widely used, general parametric statistical test. We apply it in the following way. Let us fix a frame *f* and a verb *v*. Let I_f denote the following random variable: $I_f = 1$, if a cue for *f* occurs in a sentence and $I_f = 0$ otherwise; similarly, $I_v = 1$, if *v* appears in a sentence and $I_v = 0$. Essentially, we would like to determine whether the random variables I_f and I_v are independent; if so, then we infer *v* does not take *f*, if not, then we infer *v* takes *f*. It is easily seen that I_f and I_v are independent if and only if the conditional distributions $I_f \mid I_v$ and $I_f \mid (1 - I_v)$ coincide. We shall use the likelihood test to make the decision. Let k_1, n_1, k_2, n_2 denote the number of occurrences of *v* and a cue for *f* together, the number of verbs in the corpus, the number of occurrences of a cue for *f* with any other verb than *v* and the number of verb occurrences other than *v*, respectively. Then the logarithm of the likelihood ratio is calculated as

$$\lambda = l(\frac{k_1 + k_2}{n1 + n_2}, k_1, n_1) + l(\frac{k_1 + k_2}{n1 + n_2}, k_2, n_2) - l(\frac{k1}{n_1}, k_1, n_1) - l(\frac{k_2}{n_2}, k_2, n_2) \qquad (3)$$

where $l(q,n,k) = k \log q + (n-k) \log (1-q)$. As it is well known, λ tends to the chi-squared distribution, so we can compare λ to critical values of X^2 given a significance level.

2.3 Relative frequencies

This straightforward method was suggested by Korhonen et al. [14] as a baseline and it does away with the notion of significance completely. For each verb, count the occurrences of cues and choose those frames whose relative frequency of co-

occurrence with the verb exceeds a threshold; this threshold is determined empirically.

3 Evaluation

3.1 Corpora

Our methods were tested on two Hungarian text corpora: the Szeged Corpus and the Hungarian Webcorpus. The Szeged Corpus is a Hungarian treebank, containing approximately 82 thousand sentences along with full morphological and syntactical annotation. We retained only the information concerning morphology and postpositions.

The other corpus we used is the Hungarian Webcorpus [11] [15], which, with over 1.48 billion words unfiltered (589 million words fully filtered), is by far the largest Hungarian language corpus. Only a section of the corpus was used here containing 832 thousand sentences. As this corpus is not annotated, we needed a part-of-speech tagger to extract the morphological information. We used the hunpos [12], which is a Hidden Markov Model-based open source part-of-speech tagger, with the Hungarian language resources of morphdb.hu [28].

3.2 Methodology

To measure the accuracy of a machine learning algorithm, its output has to be compared to a gold standard test dataset. The standard method to quantify the similarity between the gold standard and the output is the CoNLL F-measure [9], [24]. In the present work the gold standard is a verb list: the 1000 most frequent verbs from the Szeged Corpus and their subcategorization frames as specified by a linguist expert. The evaluation method used the following procedure: a subcategorization frame was taken to be correctly assigned to a verb if the given frame was specified for this verb in the gold standard list. Based on this, precision and recall values can be calculated for the experiments. Performance of learning algorithm-based natural language processing modules is traditionally measured in precision, which is the ratio of the correct answers to the produced answers, and recall, which is the ratio of the correct answers to the total expected answers. The F-measure is, as usual, the harmonic mean of these two values.

3.3 Results

By comparing the results of our measurements on the two corpora we see that even though the Webcorpus is noisy and automatic morphological parsing is a source of further errors, the sheer size of the Webcorpus outweighs these disadvantages: we obtain better results here than on the Szeged Corpus (see Table 1 below).

The Brent method was tested using a number of different values of error probability. The results reveal that precision improves but recall declines with an increase in the value of error probability. The F measure, of course, balances these values but it remains the case that lower values of error probability lead to better performance. An error probability of 0.1 gave the best results. Performance could not be improved by estimating error probabilities for individual cues.

The likelihood ratio test gave slightly poorer results than the Brent method but the learning curve suggests that performance could be improved by using more training data (i.e., a larger corpus). Surprisingly, the best result was achieved with the method where the decision was made on the basis relative frequency, similarly to the findings presented in [14].

Brent [6] took a very cautious approach to extract subcategorization frames from untagged corpora: he tried to extract just five frames. Manning [18] extended the method by using morphological information, and also extended the number of subcategorization frames to 19. First we tested the model on all of our 11 subcategorization frames. If we work with the same method and parameters, but we use only the 3 most frequent frames (*transitive, dative, ditransitive*), F-measure increases. It is a typical consequence of Zipf's Law of word distribution: a few words occur very often, more words occur somewhat often, and many words occur infrequently. This is the phenomenon called the problem of data sparseness in natural language processing: there is no corpus large enough to find all words at least once in it.. The situation is similar in the case of the number of verbs in the gold standard list. If we take into account only the 200 most frequent verbs in the evaluation, the performance of the system increases.

Corpus	***Method***	***Frames***	***Number of verbs***	***Precision***	***Recall***	***F-measure***
Szeged	Brent estimated	3 frames	1000	63%	50%	56%
Wc	Brent estimated	3 frames	1000	70%	67%	68%
Wc	Freq baseline	3 frames	1000	90%	67%	76%
Wc	Brent estimated	3 frames	1000	70%	67%	68%
Wc	Likelihood test	3 frames	1000	25%	79%	39%
Wc	Brent 0.1	allframes	200	44%	86%	58%
Wc	Brent 0.1	3 frames	200	64%	94%	76%
Wc	Brent 0.2	allframes	100	60%	71%	64%
Wc	Brent 0.2	allframes	1000	57%	57%	57%

Table 1
Precision and recall values with various settings

4 The psycholinguistic aspects of statistical acquisition of subcategorization frames

One of the key issues in international research on child language acquisition is the development of a mental grammar in the mind. Experimental evidence shows that children do not acquire their first language without errors and that these errors are not necessarily arbitrary but may clearly follow rules or patterns. This suggests that at different stages of development the child entertains different, sometimes erroneous hypothesis grammars. Such a hypothesis grammar may be the result of overgeneralization, which surfaces as the occurrence of linguistic constructions which are not part of the adult language.

Experimental and observational data [4], [5] reveal that children frequently overgeneralize verb argument structure alternation patterns. Argument structure alternation is the phenomenon when the argument roles a verb subcategorizes for may be realized by two or more syntactic frames, for instance:

1. a) Mary smeared paint on the wall.
   b) Mary smeared the wall with paint.

Child language data suggest that following a period of correct usage, children between the ages of about 3 and 8 tend to assume that verbs with similar meanings share an argument frame, specifically, they generalize alternation patterns to verbs that do not allow frame alternation in the adult grammar. These errors appear to give a U-shaped learning curve, where correct usage precedes overgeneralization. The Hungarian examples below are from a corpus of child language [1], [2], [23] (the children's ages are given in brackets: years;months).

2. a) *Nekem is kérek egy halat. (Zoli 2;2)
      I-dat too ask-for a fish.
   b) *Kérek mászni ide. (Éva 2;10)
      I-want to-climb here.

These are typical argument frame overgeneralizations: the verb *kér* (ask-for) requires a nominative subject and an accusative object but in the child's language it appears in the argument frame of the verb *kell* (need) (Cf. *nekem is kell egy* hal (I-dat too need a fish) and in the argument frame of the verbs *akar* and *szeretne* (want and would like) (Cf. *szeretnék mászni* (I-would-like to-climb)).

The question to ask is what kind of learning mechanism allows children to correct overgeneralization errors of this kind. One possible solution builds on the concept of pre-emption or blocking [22], [29]. The hypothesis is that children assume that a given meaning can only be encoded by a single sound string. Thus, if the child's mental lexicon lists the verb *kér* with a dative argument frame, when the child observes the nominative-accusative frame in the input, this will pre-empt the old entry provided that the same meaning is assigned to the two forms. If the child faces a conflict where a single meaning appears to be encoded by two

different strings, there are two ways to solve the problem: either modify the meaning assigned to one of the strings or reject one of the strings as erroneous. To arrive at the right decision, the child needs to (a) be satisfied that only one of the two strings ever occurs in the input and (b) ascertain the meaning of the input string.

The problem is that there may be several correct constructions that never occur in the input to the child. This means that the above strategy can only work if the child has expectations as to what he is going to hear and if these expectations are not met, he or she concludes that the expected construction is not part of the grammar. Where could these expectations come from? Since we are concerned with overgeneralization errors, it seems reasonable to assume that the more frequently the model argument frame (the basis of the overgeneralizaton, e.g., the verb *kell* with dative subject) occurs, the more the child expects to hear the pattern applied to other verbs (e.g., to the verb *kér*). That is, learning is likely to proceed on a statistical basis.

One of our goals is to model this learning mechanism and compare the behaviour of the system to real-world data. As we do not have precise quantitative data on argument frame overgeneralizations in child language, the model curve used here is a typical U-shaped curve observed in the acquisition of past tense morphology, which shows similar over-generalization patterns ([20]). Our results are presented graphically in Figure 2. The curve of the likelihood ratio trial shows a U-shaped curve similar to that observed in child language. (The left graph shows the precision values of likelihood ratio test, the right one shows the children's U-shaped development of the correct past tense of irregular verbs [20].) The horizontal axis of the child data curve represents time: as input sentences accumulate, the initial conservative correct usage of constructions is overgeneralized before further input allows errors to be corrected. The horizontal axis of the machine learning curve shows the size of the corpus, which fulfils a similar function in machine learning. Although the curve rises slowly here, the U-shape is clearly seen. (The recall curve is monotonic increasing, of course.)

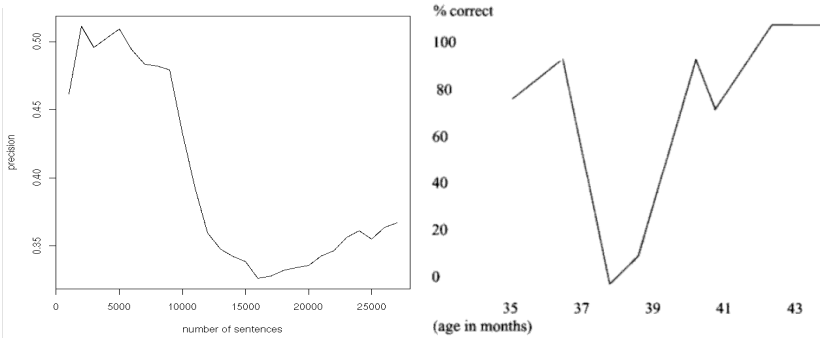

Figure 2
U-shaped curves of machine learning and child language acquisition

6 Conclusion and Future work

One of our aims in constructing a statistical learning model was to model the mechanisms of verb argument frame learning by young children. We have shown that data frequency and the size of the input corpus are important factors in both psycholinguistics and machine learning. Our results reveal that the performance of the system is best when a small number of highly frequent subcategorization frames need to be learnt. We may assume that children start with few errors because they first acquire a few very frequent constructions. At present, however, computational corpus analysis cannot keep up with natural language acquisition: finding large enough corpora is one of the most difficult problems of computational linguistics.

Naturally, the validity of the comparison is marred by the fact that the child language data shown in Figure 2 above and the computer model apply to different linguistic phenomena. In the next phase of our project, argument frame data will be analyzed in Hungarian corpora of child language. Our future plans also include testing the learning algorithm on child directed speech, which will, however, pose even greater problems of data sparseness.

References

[1] Babarczy Anna: The transition to the functional stage in Hungarian language acquisition. In C. de Groot and I. Kenesei (eds.): *Approaches to Hungarian 6*. Szeged: JATE, 1998.

[2] Babarczy Anna: *A path from broader to narrower grammars.* PhD dissertation. The University of Edinburgh, 2002.

[3] B. Boguraev, E. Briscoe, J. Carroll, D. Carter and C. Grover: The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics.* Stanford, CA, 1987, pp. 193-200.

[4] M. Bowerman: Reorganizational processes in lexical and syntactic development. In E. Wanner and L. Gleitman (eds.): *Language Acquisition: The State of the Art*. Cambridge, MA: MIT Press, 1982.

[5] M. Bowerman: The non-negative evidence problem: How do children avoid constructing an overly general grammar. In J. A. Hawkins (ed.): *Explaining Language Universals*. Oxford: Blackwell., 1988.

[6] M. R. Brent: From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics* 19. 2, 1993, pp. 243–262.

[7] Ted Briscoe, Anne Copestake and Bran Boguraev: Enjoy the paper: lexical semantics via lexicology. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90).* Helsinki, 1990, pp. 42–47.

[8] Ted Briscoe and John Carroll: Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*. Washington, DC, USA, 1997.

[9] CoNLL: Conference on Natural Language Learning. http://ifarm.nl/signll/conll/

[10] Csendes D., Csirik J., Gyimóthy T.: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. In *Proceedings of TSD 2004*. Brno, Czech Republic, 2004, vol. 3206.

[11] Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor: Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*. 2004.

[12] Halácsy P., Kornai A., Oravecz Cs.: Hunpos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 209–212.

[13] D. Ienco, S. Villata and C. Bosco: Automatic extraction of subcategorization frames for Italian. In *Proceedings of the Sixth Language Resources and Evaluation (LREC '08)*. Marrakech, Morocco, 2008.

[14] A. Korhonen, G. Gorrell and D. McCarthy: Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong, 2000, pp. 199–206.

[15] Kornai A., Halácsy P., Nagy V., Oravecz Cs., Trón V., and Varga D.: Web-based frequency dictionaries for medium density languages. In Adam Kilgarriff and Marco Baroni (eds.): *Proceedings of the 2nd International Workshop on Web as Corpus*. 2006, pp. 1–9.

[16] Kornai A., Rebrus P., Vajda P., Halácsy P., Rung A., and Trón V.: Általános célú morfológiai elemző kimeneti formalizmusa. In Alexin Z. and Csendes D. (eds.): *II. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 2004, pp. 172–176.

[17] B. MacWhinney: *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription format and programs. Volume 2: The Database*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[18] Ch. Manning: Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL '93)*. Columbus, Oh., 1993, pp. 235–242.

[19] M. Maragoudakis, K. L. Kermanidis and G. Kokkinakis: *Learning subcategorization frames from corpora: A case study for modern Greek.*

Department of Electrical and Computer Engineering, University of Patras, 2000.

[20] W. O'Grady: *How Children Learn Language*. Cambridge University Press, 2005.

[21] Fernando Pereira, N. Tishby and L. Lee: Distributional Clustering of English Words. In *Proceedings of the 31st Annual Meeting of the ACL*. 1993, pp. 183-190.

[22] S. Pinker: *Language Learnability and Language Development.* Cambridge, MA: Harvard University Press, 1984.

[23] Réger Zita: The functions of imitation in child language. *Applied Psycholinguistics* 7, 1986, pp. 323–352.

[24] C. J. van Rijsbergen (ed.): *Information retrieval*. London, 1979.

[25] Andrea Sanfilippo and Victor Poznanski: The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Resources. In *Proceedings of the Applied Natural Language Processing Conference.* Trento, Italy, 1992, pp. 80—87.

[26] Sass Bálint: Extracting Idiomatic Hungarian Verb Frames. In T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (eds.): *Advances in Natural Language Processing*. 5th International Conference on NLP, FinTAL 2006, Turku, Finnország, pp. 303-309.

[27] S. Schulte im Walde: The induction of verb frames and verb classes from corpora. In A. Lüdeling and M. Kytö (eds.): *Corpus Linguistics. An International Handbook*. Berlin, Mouton de Gruyter, 2008.

[28] Trón V., Halácsy P., Rebrus P., Rung A., Vajda P. and Simon E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of 5th International Conference on Language Resources and Evaluation*. ELRA, 2006, pp. 1670–1673.

[29] K. Wexler and P. W. Culicover: *Formal Principles of Language Acquisition.* Cambridge, MA: MIT Press, 1980.

[30] D. Zeman and A. Sarkar: Automatic extraction of subcategorization frames for Czech. In *Proceedings of the International Conference on Computational Linguistics (COLING '00).* 2000, pp. 691—697.