

Hungarian word sense disambiguated corpus

Veronika Vincze¹, György Szarvas¹, Attila Almási², Dóra Szauter²,
Róbert Ormándi², Richárd Farkas¹, Csaba Hatvani² and János Csirik¹

¹ MTA-SZTE, Research Group of Artificial Intelligence
6720 Szeged, Aradi Vértanúk tere 1., Hungary
{vinczev, szarvas, rfarkas, csirik}@inf.u-szeged.hu

² University of Szeged, Department of Informatics
6720 Szeged, Árpád tér 2., Hungary
szauter@inf.u-szeged.hu, vizipal@gmail.com, {ormandi, hacso}@inf.u-szeged.hu

Abstract

To create the first Hungarian WSD corpus, 39 suitable word form samples were selected for the purpose of word sense disambiguation. Among others, selection criteria required the given word form to be frequent in Hungarian language usage (frequency rates available in the Hungarian National Corpus (HNC) were used for measurement (Váradi, 2000)), and to have more than one sense considered frequent in usage. HNC and its *Heti Világgazdaság* (HVG) subcorpus provided the basis for corpus text selection. This way, each sample has a relevant context (the whole HVG article), and information on the lemma, POS-tagging and automatic tokenization is also available.

1. Word sense disambiguation

Word Sense Disambiguation (WSD) aims at resolving ambiguities (homonymy, polysemy) in texts. This problem has been present in natural language processing (NLP) since the beginnings, and it is an important intermediate task for most NLP applications (e.g. text comprehension, human-machine interaction, machine translation and information retrieval and extraction).

1.1. Overview of previous research

1.1.1. Word sense disambiguation research in other languages

Word sense disambiguation research concerning, first, English and, later, other languages as well was related in a greater part to SensEval (Kilgariff, 2001), (Mihalcea & Edmonds, 2004) workshops organized by ACL-SIGLex. The book *Word Sense Disambiguation* (Agirre & Edmonds, 2006) published in 2006 and a publication of SemEval workshop (Agirre et al., 2007) organized in 2007 as the next step in SensEval series provide a detailed overview of results up till then.

1.1.2. Word sense disambiguation research on Hungarian language

In relation to the development of Hungarian-English and English-Hungarian machine translation systems, word sense disambiguation tasks in Hungarian have been carried out for a long time (Miháltz, 2005), (Miháltz & Póhl, 2006).

1.2. The task of word sense disambiguation

Word sense disambiguation applications can be divided into two major groups on the basis of the limits of their applicability and the degree of granularity. With regard to scope, distinction can be made between “all-words”

(applied to overall vocabulary) and “lexical sample” (applied to selected word forms only) methods of WSD. As for granularity, fine-grained and coarse-grained levels can be distinguished. The corpus described here is a fine grained lexical sample corpus.

Figure 1 represents different levels of granularity in the case of the verb *jár*. Originally, 16 different senses were selected for annotation. Every box in the chart represents an individual sense denoted by numbers and an approximate English equivalent is also provided. However, the senses can be unified into 3 major groups: movement, relationship and other. Hence, the verb *jár* has got 16 fine-grained but only 3 coarse-grained senses¹. Similar colour represents similar translation to English.

1.3. Guidelines for annotation

In the first phase of the work, possible senses of the selected 39 word forms were defined. In this process, we relied on paper and electronic forms of The Concise Dictionary of the Hungarian Language on the one hand and our linguistic intuition on the other hand. Senses that could be definitely distinguished on the basis of their dictionary definition were considered separate senses².

Following international standards, annotation of corpus samples were carried out by two, independent annotators (qualified linguists), which means that they were not allowed to cooperate. Double annotation made it possible to measure the consistency level of the database and to

¹ For the purpose of a more minute investigation, we have assumed 33 senses altogether on the basis of the Concise Dictionary of the Hungarian Language. We found that independently of the number of senses, they can be unified into 3 major classes: *repetitive movement*, *belonging together* and *other*, which correspond to the 3 main groups in Figure 1.

² Later on, two other criteria were formulated: first, if the senses can be translated with two different vocabulary items in another language, and, second, if word forms have different argument structures, they are considered separate senses.

correct incidental annotation errors after comparing the two.

An important criterion was that word forms were annotated only within the frame of a given POS. E.g. only the nominal senses of the word *pont* 'point' were annotated. In an adverbial sense – meaning 'exactly' – it

was not annotated, that is, it is polysemy that was taken into consideration, while homonymy not. For the same reason, the "századrész" 'one hundredth' sense of the word form *század* (nominal senses: 'century' or 'company' (military unit)) was not included as it is a fraction number.

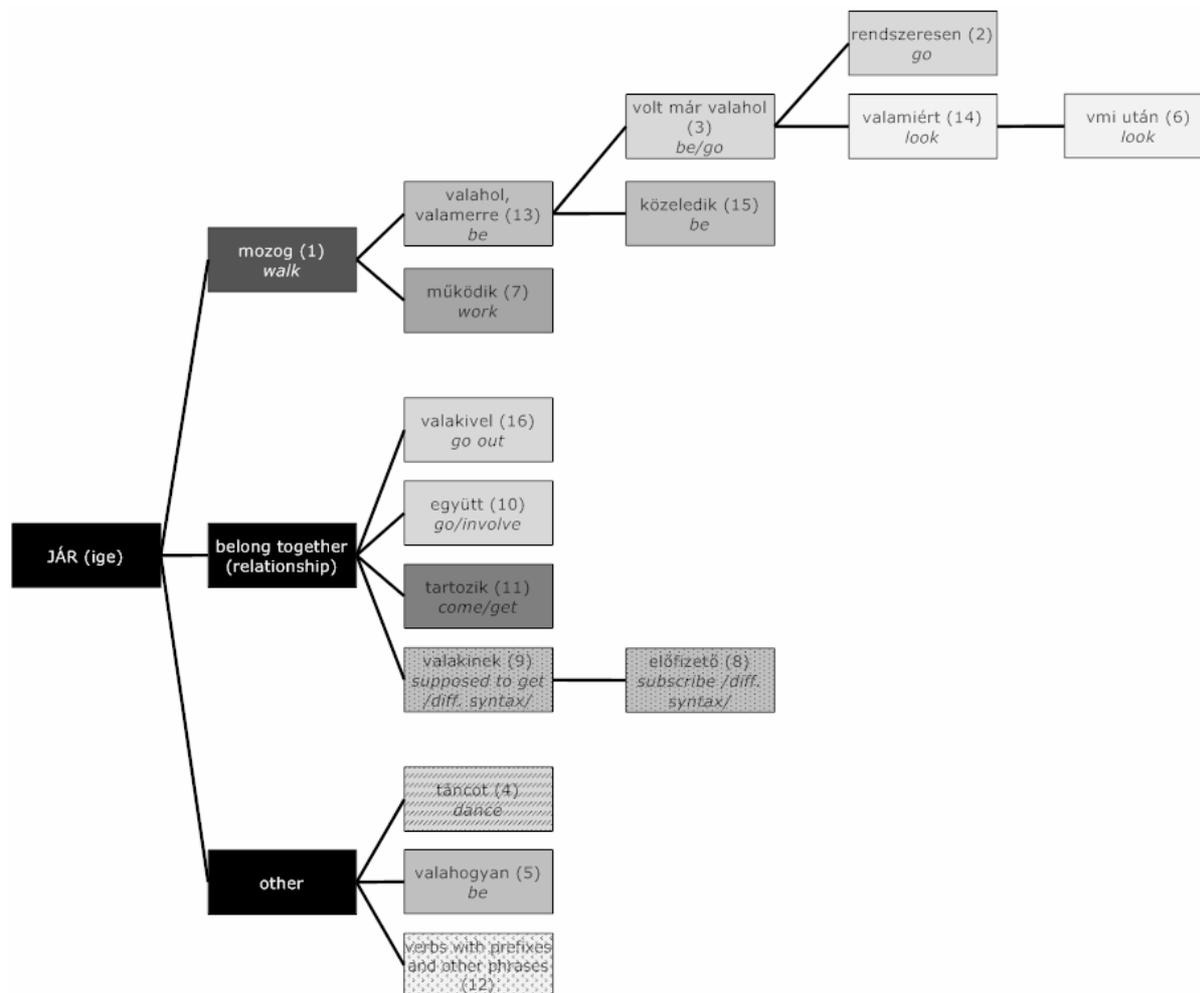


Figure 1: Sense distinctions for the verb *jár* 'go/be/look'.

2. Presentation of the corpus

In this chapter, the main characteristics of the corpus are introduced.

2.1. Corpus components

When planning the corpus, 300-500 samples of each word form were to be annotated. This size makes it possible that the subcorpora prepared for the individual word forms can be compared to data available for other languages. However, the finalized database also contains unannotated samples and samples with single annotation, that were annotated only by one of the linguists.

Ambiguous word forms that are positioned high in a word frequency list based on Hungarian corpora were chosen for annotation. We planned to disambiguate nominal, verbal and adjectival senses as well. The selected word forms are the following:

Adjectives: *anyagi* 'material', *élő* 'living, live', *erős* 'strong, hot, etc.', *képes* 'capable, illustrated', *pontos* 'precise, etc.', *szociális* 'social'.

Nouns: *család* 'family', *élet* 'life', *ház* 'house', *helyzet* 'situation', *intézmény* 'institute, institution', *iskola* 'school', *kép* 'picture, photo', *képviselő* 'representative, vanilla cream bun', *kormány* 'government, steering-wheel', *nap* 'Sun, day, light', *oldal* 'side, page, aspect, website', *ország* 'country, kingdom', *perc* 'minute, knuckle', *pont* 'dot, point, full stop', *program* 'program, programme', *század* 'century, company', *személy* 'person, slow train', *szervezet* 'organization, organism', *tanár* 'teacher', *világ* 'world, light', *víz* 'water'.

Verbs: *függ* 'hang, depend', *hat* 'affect', *jár* 'go, be, look', *kap* 'get', *kerül* 'cost, go round', *marad* 'stay, remain', *rendelkezik* 'have, command', *szerepel* 'act, occur', *tart* 'hold, go toward, last', *tartozik* 'belong, owe', *tud* 'know, be able', *válik* 'become, get divorced'.

The average number of senses is rather high (6 senses per word form), out of which – on an average – only 5 occur in the corpus texts. But if negligible senses are not considered (occurring in 1-2% of the cases), then the average number of senses present is even lower, 3.7 (moderate ambiguity). The word form *tanár* ‘teacher’ is of special importance, because it proved to be completely unambiguous in the texts examined in spite of the criterion that word forms have more than one, frequent senses in the given language. Some words, however, despite homogeneous language usage, occur in several senses (e.g. in the case of the verb *jár* ‘go, be, look’, 14 out of its possible 16 senses occurred in the text).

2.2. Corpus format

When building the corpus, we followed the format designed for corpora prepared for WSD tasks of *SensEval/SemEval* international conference workshops

organized by the Association for Computational Linguistics. This choice enabled us to adapt the existing XML format, and, hopefully, standard storage of data will also help in distributing the corpus.

An example from the corpus:

```
<instance id="jár.V.mnsh.01" docsrc="press-hvg.1">
  <answer instance="jár.V.mnsh.01" senseid="jár_v_5_valahogyan"/>
  <context>
    ...
    Ezért is tűnik e pillanatban irrelevánsnak annak felvetése , hogy nem
    <head>jártak</head> volna -e jobban a szocialisták , ha Horn Gyulát
    időben katapultálták a pártelnökségből , de legalábbis a
    miniszterelnök-jelöltségből , és mondjuk a külügyminiszeri
    tevékenységével és személyes karakterével , amellet fiatalabb
    korával a választóközönség számára esetleg vonzóbb Kovács László
    vezényletével vágnak bele a választási kampányba .
    ...
  </context>
</instance>
```

	Word forms	Samples with double annotation	Samples with single annotation	Samples without annotation
Adjectives	6	2087	462	688
Nouns	21	6853	2714	11459
Verbs	12	3537	1898	13501
Total	39	12477	5074	25648

Table 1: Corpus statistics.

2.3. Basic statistical attributes of the corpus

Table 1. represents the basic statistical attributes of the corpus. For each part-of-speech, it contains the numbers of samples that were doubly annotated and later finalized by the third annotator, samples that were annotated only by one of the linguists and samples without annotation as well. For a more detailed analysis see our website³.

2.4. Availability

The first version of the corpus was developed within the scope of the project titled *The construction Hungarian WordNet Ontology and its application in Information Extraction Systems* (Hatvani et al., 2007). The corpus – for research and educational purposes – is available and can be downloaded free of charge.

3. Evaluation

In this chapter, consistency of the annotation is evaluated, the protocol followed in the course of unification of differences of the two parallel annotations is presented, and finally, using a classifier based on a simple vector space model we demonstrate that better results can easily be obtained than when using the most frequent senses.

3.1. Consistency rate of annotators (consistency control)

In the first step, the set of senses to be distinguished was

defined for each word form and provided with a short description (definition). Hungarian WordNet was also extended with synsets of senses used in the corpus, but missing from HuWN. Following international standards, annotation was carried out by two independent linguists. Consistency rate of annotators, that is, the accuracy of annotation performed by experts is lower when the proportion of the most frequent sense is not too high. In these cases, disambiguation is a difficult task (since consistency was measured between qualified linguists). On the other hand, these are the cases when applications like machine translation or information retrieval systems, could benefit most from an effective WSD solution, instead of choosing always the most frequent sense. The consistency rate of annotators was 84.78% for the whole corpus.

One of the most difficult tasks for the annotator was to keep his consistency throughout the whole process of annotation. If, in a problematic case, he decided on a given sense, then he had to tag the word form in the same way in a later, similar occurrence. E.g. if he tagged the ‘point’ of the ‘Hungarian BUX index’ with the sense *pont_2: unit of evaluation* in the first instance, then he had to do likewise in all the following cases, even if there was “great distance” between the individual occurrences⁴.

⁴ It is difficult for an annotator to keep his own consistency, but it is even more difficult for two annotators to reach consistency. This is why inconsistencies might occur in the corpus. Keeping certain „distant cases” in mind is one great difficulty of manual annotation.

³ www.inf.u-szeged.hu/hlt

In practice, senses were not always given the most precise definitions, they didn't always reflect theoretical or lexical differences of meaning. Many a times, overly fine-grained shades of meaning have also been distinguished, which made consistency rate decrease and even the annotator's own consistency could deteriorate. In order to reach higher consistency levels the system can be further improved.

In certain cases, the morphological analyzer did not categorize the word forms properly. E.g. the present tense verb form *vált* 'change' was analyzed as the past tense of *válik* 'become', and so the program offered it for annotation among the senses of *válik* 'become'. In this case, the word form was not annotated.

From the similarity of text domains followed that certain senses occurred a lot more frequently than others. As the corpus is based on HVG texts, the word form *kormány* occurs exclusively in its political sense 'government', but if the corpus was to contain texts on automobiles, the number of occurrences of its other sense 'steering wheel' would grow immediately.

Collocations, idioms and proverbs pose a special problem because in many of the cases it is impossible to know which sense the given word form assumes within the phrase. E.g. in the proverb *sok víz lefolyik a Dunán addig* 'it'll be a long time' (lit. 'much **water** will flow on the Danube till that time'), the sense 'water' can be identified precisely: *víz_2: a mass of water covering an area of the Earth's surface*; however, it is still a question whether it can be given this tag or the so-called *other* tag, since it constitutes a part of a fixed expression.

3.2. Finalization of corpus annotation

In correction phase, a third, independent annotator checked the cases when annotations were dissimilar and finalized the tags of these samples. In this way, annotation of the doubly annotated subcorpus is more or less consistent. Samples with single annotation were not checked in this phase.

Most of the inconsistencies were due to the fact that the annotators interpreted certain overlapping senses in a different way. For instance, the senses *jár_6: look after something* and *jár_14: go to several places for something* were usually confronted, that is, one annotator voted for *jár_6*, while the other selected *jár_14*, which suggests that the distinction between the two senses is questionable. Cases when one annotator tagged a given word form in a given context as *other*, while the other annotator felt that this particular occurrence of the word could be tagged as another, more specifically defined sense were responsible for most of the inconsistencies. E.g. in the case of *kap* 'get', it was very frequent that one annotator chose the tag *kap_1: something is given to him* in expressions such as *jogdíjat kap* 'to receive royalty payment' while the other one tagged it as *other*.

There occurred some cases when the third annotator did

not agree with either of the two previous annotations. In these cases – if possible – the finalization of the particular tag was discussed by the three annotators. E.g.: in the context *a víz felforralása vagy – erre szolgáló filtereken való – átszűrése* 'the boiling of water and straining it through filters designed for this purpose' the two annotators gave the following senses: *víz_1: liquid essential for life* and *víz_2: a mass of water covering an area of the Earth's surface*; however, the third annotator voted for the sense *víz_3: drinking water, bathwater*. (The final tag was *víz_3*.)

A negligible part of inconsistencies was due to obvious mistakes: one of the annotators clicked on the adjacent tag or forgot to tag a sample etc.

3.3. Baseline measurements, C4.5, naïve Bayes classifiers

When evaluating word sense disambiguation procedures, the rate of the most frequent sense (MFS) is usually considered as baseline accuracy since this is the precision that can be obtained trivially. A system output (disambiguated occurrences of word forms) can be considered evaluable if it assigns proper senses to word forms in proportion higher than the rate of the MFS.

To build supervised learning models it is necessary to convert the examples of the task to a format that can be easily handled by the algorithm. During our experiments we used only a short global context of the tagged word form (one paragraph) and we represented it by using the well-known token level *vector space model (VSM)*. We also used local contextual features, describing the surrounding words in a window of size 3. We considered only nouns, verbs, adjectives and adverbs as contextual features and used lemmatized word forms. These features are useful as position information is lost in traditional VSM representation and local context often contain useful disambiguation cues. In the case of word sense disambiguation, this representation is obviously too simple, as widely used morpho-syntactic, topic and other features are not considered here. Our results are intended as a comparative baseline for distributing the corpus.

In our experiments we applied a token unigram Vector Space Model as feature representation and local contextual features in a window of size +/-3. We performed a leave-one-out evaluation of C4.5 classifiers with default parameters, as of the Weka package and 5 instances per leaf. Our results show that statistical models outperform the *most frequent sense (MFS)* heuristic currently used by applications that perform no WSD (MFS yields 69.39% accuracy on the whole corpus, while Naïve Bayes model yields 70.27 and C4.5 yields 72.70% performance using our simplistic feature representation). The difference in accuracy is salient for more complicated target words in the corpus, where the *most frequent sense* is relatively low.

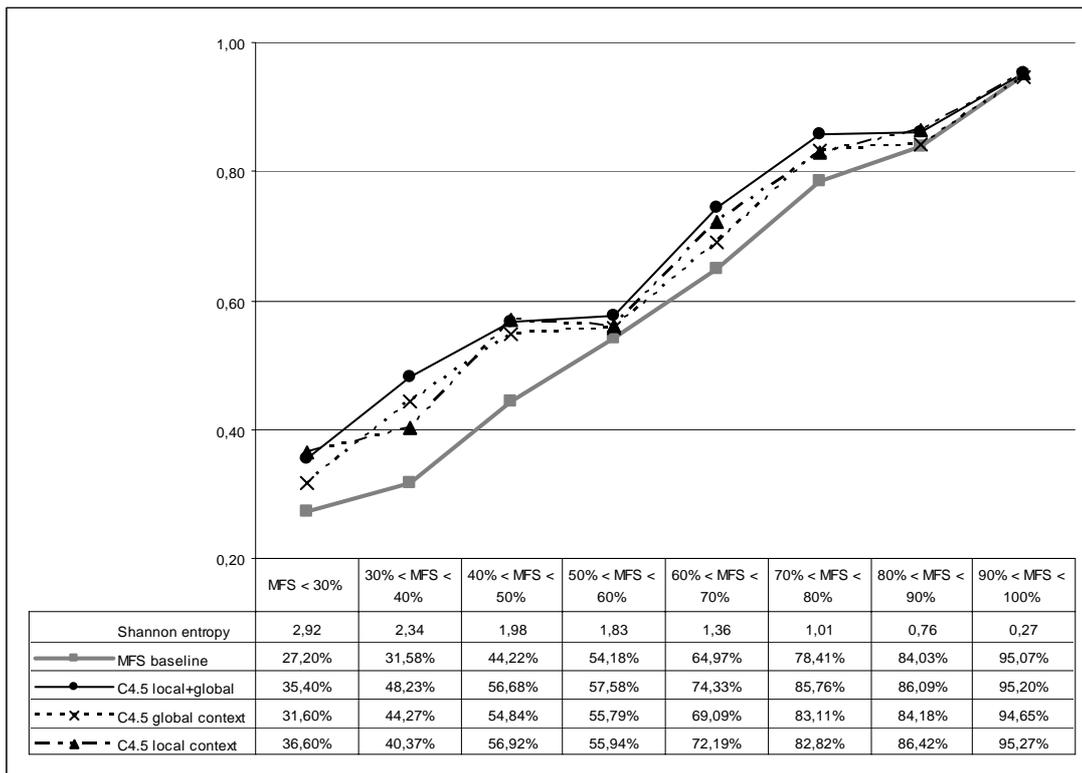


Figure 2: Average accuracies of MFS, and C4.5 classifier with different feature types for target words of different complexity. Complexity is measured here by the accuracy of MFS heuristic and the average Shannon entropy of sense labels is also presented in the table.

In our experiments we applied a token unigram Vector Space Model as feature representation and local contextual features in a window of size +/-3. We performed a leave-one-out evaluation of C4.5 classifiers with default parameters, as of the Weka package (Witten and Frank, 2005) and 5 instances per leaf. Our results show that statistical models outperform the *most frequent sense (MFS)* heuristic currently used by applications that perform no WSD (MFS yields 69.39% accuracy on the whole corpus, while C4.5 yields 74.86% performance using our simplistic feature representation). The difference in accuracy is salient for more complicated target words in the corpus, where the *most frequent sense* is relatively low. Figure 2. shows the average accuracies of MFS and C4.5 (using only local, only global and both contextual feature types, along with the average Shannon entropy of sense labels for target words of varying complexity (MFS falling between a certain interval). The figure confirms that a significant improvement can be achieved even with simple models on complex target words (~10% improvement on MFS, for MFS < 80%), while on target words where the MFS is predominant, simple classifiers converge to baseline performance. C4.5 yields consistently higher accuracies than MFS; demonstrating that WSD is feasible for Hungarian with relatively simple models and small labeled sample size. Figure 3. on next page shows the average accuracies of MFS and C4.5 classifiers for different levels of polysemy. Only senses with frequency higher than 5% are counted here, i.e. the first column shows average accuracy for

target words that have only 1 sense with frequency higher than 5%, while the last column shows average accuracies for target words with 6 such senses. The figure demonstrates that simple classifiers are most successful for words that have 3-5 frequent senses. The two figures also demonstrate that polysemy level and accuracy of MFS highly correlate with each other.

4. Suggestions for correction and further development

In order to reach a higher level of consistency, the system can be further ameliorated by the review of the cases of individual senses as a set:

- All word forms having a given sense must be reconsidered and compared to one another in order to check if annotation is consistent within the given sense domain.
- The above step is worth performing in the case of each sense.
- Samples that somehow mismatch the others must be retagged.

In some cases, it can be fruitful to reconsider the distinction of certain senses (see e.g. the case of *jár*). The more senses are assumed, the lower the agreement rate among annotators will be. Such is the case with ambiguous or fuzzy senses. Hence, from the perspective of computational linguistics, well-defined and clearly distinguishable senses are of primary importance. On the basis of our experiences, 3-5 coarse-grained senses on an average seem to meet the above-mentioned requirements. In order to get that “ideal number” of senses, some further

lexicographic or syntactic research should be carried out such as the description of the agreement structure of different senses, substitution with synonyms, identification of the domain etc. Annotation of texts from other domains can also be useful when determining the frequency of different senses.

The development of an “all-words” WDS is a demanding task since possible senses of all vocabulary items of the Hungarian language should be elaborated. The task is further impeded by the fact that sometimes it is the phrase that has a certain sense in the given context and it can hardly be determined whether the word form itself has this meaning: e.g. in the sentence, *A védőoltások*

növekedése terén pedig erős képzelőerő kell a tényleges kormányzati cselekvési mező megtalálásához. ‘With respect to the increase of vaccinations **strong imagination** is needed to find the real governmental field of action.’, it is the whole context that carries the ironic sense and not only the word form *erős* ‘strong’.

5. Acknowledgements

This work was supported in part by the NKTH grant of Jedlik Ányos Programme. Grant no. NKFP/07/A2, title: *Hatékony tudásmenedzsment eszköz kialakítása nyelvészeti és gráfelméleti eszközök integrációjával a hazai és közösségi vámpathóságok részére.*

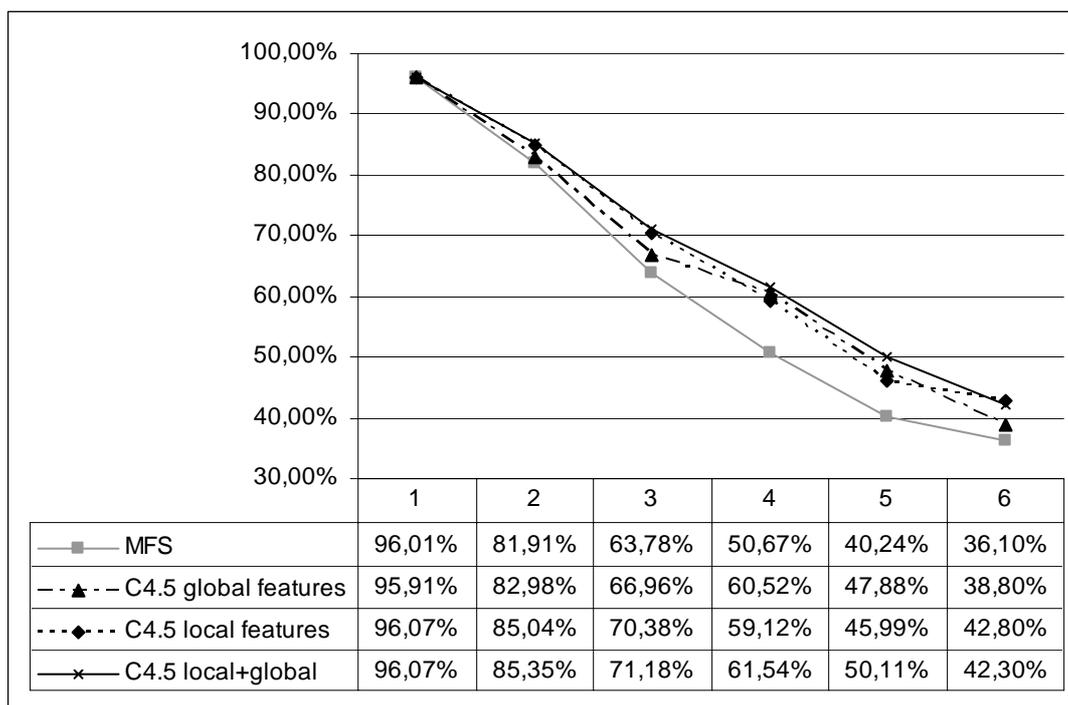


Figure 3: Average accuracies of MFS and C4.5 classifiers with different features for target words of different polysemy level (different number of senses having 5% or higher frequency).

6. References

- Agirre, E., Edmonds, P. (2006). Word Sense Disambiguation – Algorithms and Applications. In Ide, N., Véronis J. (Eds.), *Text, Speech and Language Technology Series*. Volume 33. Dordrecht: Springer.
- Agirre, E., Márquez, L., Wicentowski, R. (2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague: Association for Computational Linguistics.
- Hatvani, Cs., Kuti, J., Miháltz, M., Szarvas, Gy. (2007). *GVOP 3.1.1-2004-05-0191/3.0 – Magyar ontológia építése és alkalmazása információkinyerő rendszerekben, projektzáró összefoglaló jelentés*. Technical Report. Szeged.
- Kilgariff, A. (2001). *Proceedings of Senseval 2: Second International Workshop on the Evaluating Word Sense Disambiguation Systems*. Toulouse: Association for Computational Linguistics.
- Mihalcea, R., Edmonds, P. (2004). *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona: Association for Computational Linguistics.
- Miháltz, M. (2005). Towards A Hybrid Approach To Word-Sense Disambiguation In Machine Translation. In *Proceedings Modern Approaches in Translation Technologies Workshop at RANLP-2005*. Borovets.
- Miháltz, M., Póhl, G. (2006). Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból. In *Proceedings of III. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged.
- Váradi, T. (2000). Szótár, Korpusz – Magyar Nemzeti Szövegtár. In Gecső, T. (Ed.), *Lexikális jelentés, aktuális jelentés*. Segédkönyvek a nyelvészet tanulmányozásához IV. Budapest: Tinta Kiadó.
- Witten, I. H., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann.