Extracting Semantic Classes and Morphosyntactic Features for English-Polish Machine Translation

Barbara Gawronska, Björn Erlendsson and Hanna Duczak
Dept. of Languages
University of Skövde
Box 408
541 28 Skövde
Sweden
barbara.gawronska@isp.his.se, bjorn.erlendsson@isp.his.se,
hanna.duczak@isp.his.se

Abstract

This paper describes a procedure aimed at automatic extraction of certain noun and verb categories from Polish texts. The general goal is to construct a lexical database that should be incorporated into a system for machine translation and multilingual generation of summaries. High quality processing of inflectional languages like Polish requires quite elaborated lexical entries, it is therefore highly desirable to automate the process of lexicon construction, at least partially. However, purely statistical methods for languages with less elaborated inflectional systems do not perform especially well on Slavic languages.

As primary cues for automatic subcategorization we used inflectional morphemes expressing the greatest number of semantico-syntactic functions. The crucial semantic category for noun classification was the degree of animacy. Morphosyntactically, this category is expressed by nominal suffixes and subject-verb agreement markers. The procedure for lexical extraction and classification was implemented in Delphi and the system was trained for extraction of so-called superanimate nouns, i.e. nouns denoting male human beings, or groups including both male and female humans. The usability of lexical extraction based on concurrence of morphological features rather than on concurrence of whole word forms is evaluated and discussed.

Keywords: semantic class, morphosyntactic features, gender, animacy, agreement, lexicon construction, machine translation, Polish

1. Introduction

The knowledge about the semantic class and/or semantic features of a word is crucial for many NLP-tasks, such as anaphora resolution, recovering dropped arguments (Han et al. 2000), generation of numerical classifiers (Bond & Paik 2000, Hayashi et al 2001), choice of correct prepositions and verb forms in translation (Gawronska & Duczak 2000, 2001) – just to mention a few examples. These tasks require some kind of semantic analysis (Pulman 1996) in addition to statistical methods. Furthermore, most of the statistical approaches of today (Briscoe and Carroll 1997, Brent 1993, Brown et al. 1991, Charniak 1997, Resnik 1992, Webster & Marcus 1989) well suited for processing English and other languages with a relatively rigid word order, while languages with complicated morphological paradigms and free word order still pose difficulties (Nirenburg 1996, Sheremetyeva & Nirenburg 2000, Niessen & Ney 2000, Thanopoulos et al. 2000, Prescher & al. 2000, Sarkar & Zeman 2000, Farwell and Helmreich 2001). Sarkar & Zeman (2000) show that raw or POS-tagged corpora cannot be utilized for extracting semantic and syntactic classes of Czech lexemes. Niessen & Ney (2000) point out that long-term dependencies in German cannot easily be handled by e.g. Brown's (1993) string translation model and other statistical models where the word is the minimal entry of analysis. Similar problems occur when the input language is Polish or any of the Slavic languages. Thus, in our approach, we employ the internal structure of

words (especially the inflectional affixes) as a starting point for extracting lexical information.

2. Background and Objectives

The goal of our work is to construct a Polish lexicon to be incorporated into a system for multilingual translation and summarization of news reports. The input to the system are CNN online news, and the output languages worked on are for the time being Swedish and Polish.

The lexical resource used for processing English texts is a modified version of WordNet (Miller 1990, 1995). In an early prototype, Polish lexical entries were linked to WordNet "synsets" and subcategorized in the spirit of cognitive semantics (Lakoff 1980, 1993, Jackendoff 1983; Langacker 1991; Talmy 1988). Semantic frames of verbs and prepositions were formulated in terms of "trajectors" (objects that are active or in focus) and "landmarks"(static or passive objects); the trajectors and landmarks were further specified with regard to such conceptual features as dimensions, animacy, patterns of distribution (discrete/continuous, bounded/unbounded), etc. The initial translation experiments proved that this kind of semantic specifications improves the intelligibility and quality of translation and generation; however, manual implementation of the quite elaborated semantic descriptions is time-consuming and inefficient. Therefore, our main objective was to partially automate the process of lexical acquisition and semantic subcategorization. Another objective was to evaluate the hypothesis that, in a language with a rich morphological system, the most marked morphemes, i.e. morphemes expressing the highest number of semantic and syntactic functions, should be used as primary cues in automatic lexical subcategorization.

2.1. Highly inflected languages – a challenge for lexicon building procedures

Difficulties in statistically based recognition and classification of word forms increase with the number of functions that can be expressed by a single morpheme. For isolating languages, lexicons can be successfully constructed from large corpora by statistical methods which regard the word as the basic unit. Agglutinative languages, like Turkish or Finnish, require morphological analysis, but distinguishing the word stem from inflectional affixes in these languages is a relatively straightforward task.

Inflectional languages, like the Slavic and the Caucasian languages, make use of so-called 'portemanteau'-morphemes, marking several features at the same time and often melting with the stem as a result of morphophonological assimilation. This makes many word forms highly ambiguous, and poses difficulties to stem identification algorithms. E.g. a Polish word form like *stali* has the following interpretations:

- noun, female, dative, singular: *stal* + *i* → *stali* ("steel" + dative)
- adjective, male, human, plural, nominative: *stał* + *i* → *stali* ("steady, permanent, constant" – with regard to male human individuals)
- verb, male, human, past tense, 3rd person, plural: *sta* + *li* → *stali* ("they (men) stood")

Similar examples are legio. The free word order makes the disambiguation of homonymous forms yet more problematic. Verbs do not occupy a predictable position in the sentence, and long distance dependencies between adjective attributes and head nouns are not unusual.
The most natural starting point in a procedure aimed at classification of inflected word forms should be to extract the most specific entities, i.e. entities marked for the maximal number of features, or for the most specific features. As shown above, the verbal suffix *-li* expresses six syntactico-semantic categories; this seems to be the maximal number of features that a grammatical affix in Polish may comprise. This suffix specifies the subject of the verb with respect to semantic and syntactic features; in particular, it defines the animacy status of the subject. This feature must be taken into account in any NLP system for Polish. Our assumption was that the *-li* suffix could be used as a cue for automated extraction of the animacy feature, and other morphosyntactic features as well. We also assumed that, once the most specific verb forms and the most specific semantic class of nouns (+male, + human) are extracted, it will be easier to identify the less specific noun categories.

2.2. The animacy hierarchy in Polish

The category of animacy is grammaticalized to different extent in different languages. In several Asian languages, the distinction between animate and non-animate objects is expressed by numeral classifiers. Bond and Paik (2000) and Bond et al. (2001) show the usefulness of these markers for automatic noun classification in Malaysian, and our approach is inspired by this work, although the object of our study is a typologically different language.

All Slavic languages mark the distinction between animate and non-animate referents by means of case suffixes. In most Slavic languages, the animacy distinction is binary, but in Polish three main animacy degrees are distinguished:

- inanimate – ex: *stół* "table", *samochód* "car". In this group, the accusative case has the same form as the nominative in singular and plural;
- animate – traditionally including human females, and living non humans (animals, fantasy figures), e.g. *dziewczynka* "girl", *pies* "dog", *ptak* "bird", *wampir* "vampire". The accusative form of grammatically male nouns is identical with the genitive in singular, and with the nominative in plural;
- so-called 'superanimate', or 'male-animate' – the group incorporates nouns denoting human males in singular, e.g. *żołnierz* "soldier", *mąż* "husband", and grammatically male nouns denoting only males or males and females together in plural, e.g. *przywódcy* "leaders", *naukowcy* "scientists". The genitive and accusative forms are equal both in singular and in plural; additional morphological exponents, not shared by the other groups, are the suffixes *-li* on verbs in plural and *-i* on adjective attributes in plural.

The examples below show some morphosyntactic differences between the three groups:

1a. inanimate:

Młod-**e**	drzewa	sta-**ły**
young-PL- NOM	trees	stand-PAST-3P-PL

"Young trees stood/were standing (there)"

1b. animate:
Młod-**e** psy sta-**ły**
young-PL- NOM dogs stand-PAST-3P-PL
"Young dogs stood/were standing (there)"

1c. animate:
Młod-**e** dziewczyny sta-**ły**
young-PL- NOM girls stand-PAST-3P-PL
"Young girls stood/were standing (there)"

1d. superanimate:
Młodz-**i** chłopcy sta **-li**
young-PL-MA-HUM-NOM boys stand-PAST-MA-HUM-3P-PL
"Young boys stood/were standing (there)"

An interesting distinctive case is a category that can be labelled as 'semianimate'. These are grammatically male nouns, which - under certain conditions - adapt morphological markers that are typical of animates, although their referents are neither human beings nor animals. Several semantic subgroups can be distinguished here, e.g. certain (mostly spherical) vegetables – *ziemniak* "potato", *pomidor* "tomato"); mushrooms; car and aeroplane marks – *Ford , Cadillac, Boeing*; nouns denoting dances – *walc* "waltz"*, polonez* "polonaise" etc. A common semantic denominator seems to be the association with the features '+mobile' or '+spherical'. Another peculiar phenomenon in Polish is the fact that a coordinated noun phrase comprising two "plain" animate nouns becomes superanimate (+male, +human) as a whole if one of the nouns is +male, and the other one +human, as in *Dziewczyna i pies biega**li*** ("A girl and a dog were running").

Table 1 is an attempt to summarise the main characteristics of the Polish animacy system.

Animacy degree	**Grammatical gender**	**Semantic features**	**Accusative form**	**Adjective ending in plural**	**Verb ending in plural, past tense**
inanimate	+ma/+fe	-alive	acc=nom	-e	-ły
	+ne	+/- alive			
semianimate	+ma	- alive, + mobile or + spherical	sg: acc=gen or acc=nom, pl: acc=nom	-e	-ły
animate	+ma/+fe	+ alive	sg: acc=gen, pl: acc=nom	-e	-ły
superanimate	+ma	+ human	acc=gen	-i/-y	-li

Table 1: The grammatical and semantic characteristics of Polish nouns.

3. Extracting superanimate nouns from corpora

3.1. Challenges

As shown above, the superanimate nouns are the most plausible candidates for automatic extraction from Polish text corpora, as they form a semantically homogenous group and – when in plural – they occur as subjects of past tense verbs with the highly marked *-li* suffix.

Furthermore, there is a practical aspect: reference to male human individuals is generally frequent in world news; thus, via Internet, it was easy to access a training corpus which was supposed to contain a considerable number of superanimate nouns. The main idea was to extract all sentences containing verbs with the *-li*-suffix from the corpus, then identify their subjects, and finally – to store the head nouns of the subject phrases in the lexicon with automatically attached appropriate specification (+noun, +male, +human, + nominative, +plural, declension pattern). The declension pattern was supposed to be inferred from the plural endings of the nouns.

The principle is quite straightforward, but, for Polish, there were certain challenges to face:

- The ambiguity of the final *-li*: all words ending in *-li* cannot be automatically classified as verbs. As indicated in section 1.1., this suffix may coincide with other endings, e.g. with the dative or genitive suffix *-i* of nouns with stems ending in [l] (*sta**li*** "steel+gen/dat", *cywi**li*** "civilians+gen"). This problem could be solved by establishing a stop-list containing about 50 word forms.
- The ambiguity of the superanimate nouns' plural suffixes: traditional grammars list about 36-38 declension patterns of Polish nouns. Some of the plural suffixes are restricted to the superanimate nouns, but many of them occur in other categories (cf. *Ameryka**nie*** "Americans", *mieszka**nie*** "apartment".). Proper names increase the problem (e.g. *Rabba**ni***-proper name vs. *rabi**ni*** "rabbins").
- Relative clauses: the logical subject of a *-li* verb may be an antecedent to a relative pronoun; as a consequence this noun can get accusative, dative or instrumental case ending and will not be discovered by a procedure searching for nominative forms (e.g. *Policja aresztowała terrorys**tów**, którzy…*- "The police arrested the terrorists+gen, who…").
- Coordinated noun phrases: The verb may end in *-li* if the subject is a coordinated NP containing two or more superanimate nouns (*prezydent i premier powiedzie**li**…*- "the President and the Prime Minister said…"), but also when only one of the nouns belongs to the superanimate category (*prezydent i jego żona powiedzie**li**…* — "the President and his wife said…"). For more intricate, but less frequent problems concerning coordinated NPs, see section 2.2. Negative conjunctions, like Polish equivalents of "neither…nor" cause similar complications.

3.2. Training

A procedure taking these difficulties into account was implemented in the Delphi programming language (an object-oriented version of Pascal) along the lines shown in Figure 1. The procedure looks for sentences with words ending in *-li*. These word are checked against the stop-list containing non-verbs ending in the same characters. If the *-li* word is not in the stop-list, it is considered to be a verb, and the search for superanimate nouns begins. Words that end in character combinations listed as male plural suffixes are stored in the database with a full sematic and morphological specification (+ noun, + male, + human, + nominative, + plural) and the declension number of the plural suffix is attached to the lexical entry. Most of the superanimate nouns can be identified this way; often, more than one noun is extracted from a sentence, due to the frequent occurrence of appositional constructions, as in *Afganistan opanowa**li*** *talib**owie*** – *uczni**owie*** *szkół religijnych* – "The Talibans – students of religious schools – got Afganistan under their control".

If no words ending in male plural suffixes are found in the sentence, the procedure looks for conjunctions and ascribes partial specifications to the words found on both sides of the conjunction, again looking on the endings of the words. Proper names are disregarded, words ending in *-a* are classified as female animate nouns (specification 4 in Figure 1b), and other words not ending in vowels are assumed to be singular superanimates in nominative (specification 3). In a sentence like *Minister obrony Donald Rumsfeld* ***i*** *generał Henry Shelton potwierdzi****li*** *te informacje* – "The defense secretary D.R. and general H.S. confirmed this information", the nouns *minister* and *generał* are correctly identified as +male, + human, +singular. Another clue for identification of superanimates is the presence of the +male, +human, +nominative form of the relative pronoun ***którzy***. Words directly preceding the relative pronoun are assumed to be plural superanimates, as in *Walki zostały zainicjowane przez talib****ów****,* ***którzy*** *ponownie zaatakowa****li*** *Kabul* – "The strikes had been initiated by the Talibs, who attacked Kabul for the second time". Nouns identified this way get specification 2 (+ma,+hum,+pl), where the case value and the declension number have to be added during post-editing.

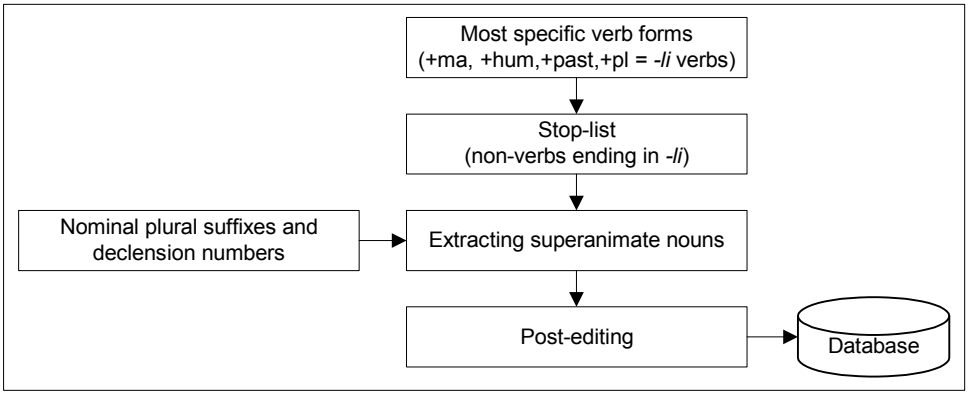

Figure 1a: The general outline of the procedure for lexical extraction

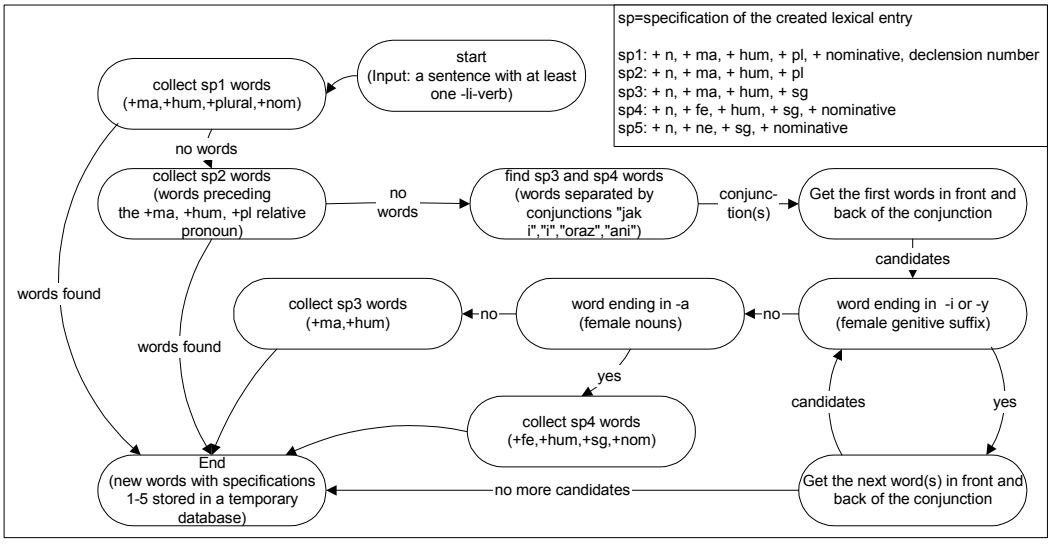

Figure 1b: The extraction and classification module

The training set consisted of four weeks' world news reports taken from the Polish web-site www.onet.pl, collected between September 25th and October 20th 2001. One-week text amount comprised about 11 000 words. According to a preliminary estimation, 42% of the words in each corpus were nouns and 22% of the nouns referred to male persons. The database was empty at the beginning of the training. The results of extracting nouns from the training set are shown in figures 2 a-b. The term "unknow forms" in 2a-b refers to forms that are recognized as nouns, but not present in the database.

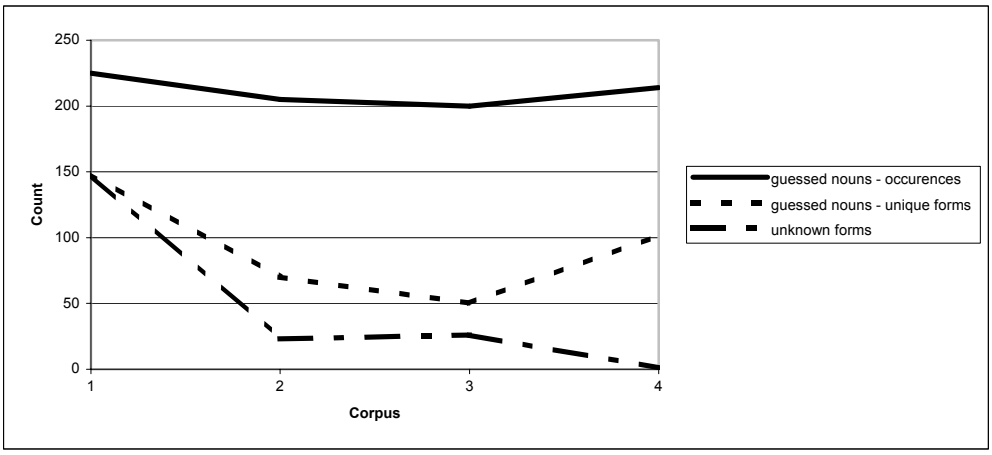

Figure 2a. The decrease of unknown superanimate noun forms during the training phase

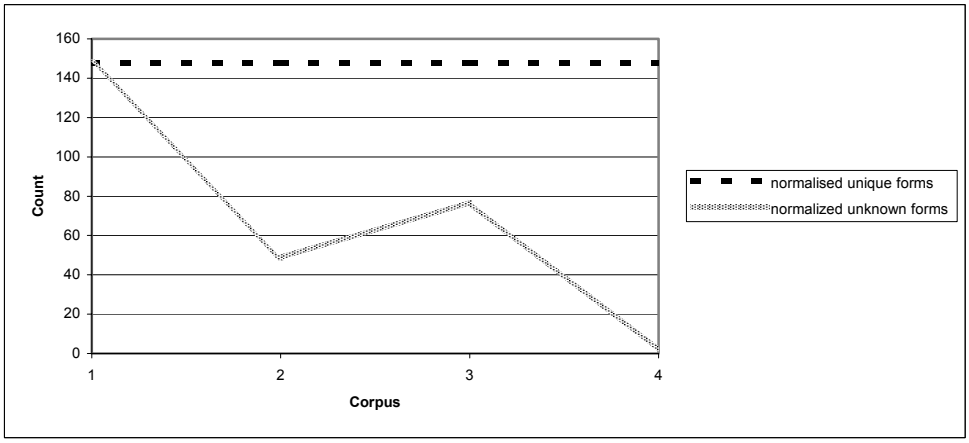

Figure 2b. The decrease of unknown superanimate noun forms during the training phase – normalized data

3.3. Post-editing and Testing

During the training phase, the words that had been added to the data base were post-edited after each of the three first sessions. The results are shown in Figure 3. The human editor had five options: no change (the first category in Figure 3), animacy classification correct, but case and declension have to be specified (category 2), change the specification automatically into +ne,+sg,+nom (this is the case of the ambiguous and highly productive suffix *–nie*; category 3 in the figure), change the specification manually (categories 4 and 5). It can be

seen that a significant majority of the processed words has been classified correctly by the system.

The usability of the lexical extraction procedure was then tested on unseen world news and on texts belonging to different domains (financial news, sport, and science).

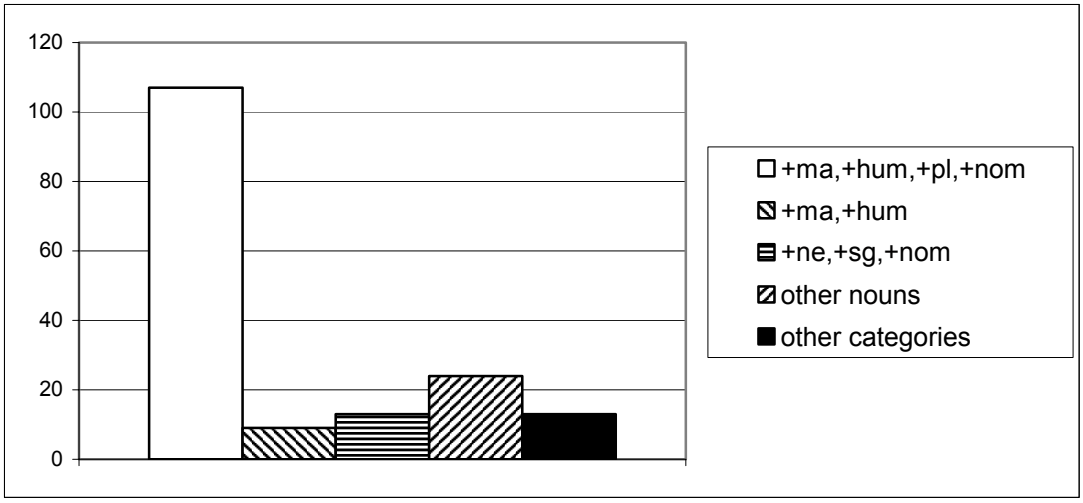

Figure 3. The results of post-editing of the training set

	World news	Sport	Science	Business
Nouns (types) found in the database	166	47	64	48
Nouns (types) added to the database	24	121	98	78
Total	190	168	162	126

Table 2: The lexical coverage of different text domains

As expected, most of the superanimate nouns in *-li*-sentences from unseen world news (87%) were present in the database. Most new superanimate nouns have been discovered in sport reports due to the high frequency of nouns denoting male sportsmen, nationalities and origin. The results of post-editing displayed no significant domain-related differences in percentage terms: the proportions among the five categories shown in Figure 3 remained in principle unchanged – what changed was the number of items to be postedited.

A pilot study of contrastive lexical coverage, conducted on English news reports, showed that the nouns stored in the database after the four training sessions enable translation of 82% English nouns referring to human beings, i.e. about 22% of all nouns in the texts. The corpus used in this study was not a parallel English-Polish corpus, but the texts dealt with the same kind events (mainly the war against terror). It was worth noting that our database enabled translation of nouns that are frequent in the news of today, but not present in conventional lexicons, like *muhadżedini* "muhajedeens" or *Pasztuni* "Pashtuns". These nouns are not encoded in WordNet, but translation can be achieved by simple translitteration rules (*sh* → *sz*, *ee* → *i* etc.)

4. Extracting adjectives, quantifiers and female animate nouns

Although our work concentrated on nouns, a side effect of noun classification was the acquirement of a considerable number of verb entries, including certain information about the agent of the verb. A verb that can occur with the *-li* suffix is obviously able to take +human agents; an overwhelming majority of the extracted verbs (ca 70%) denoted typically human activities, and only less than 1% of the verbs could take non-living subjects.This information, in combination with the knowledge about superanimate nouns, could be used for extracting further noun classes. The next step was therefore extraction of female animate nouns. From the verb forms stored in the database and classified as human activities, we generated female past tense singular forms with the suffix *-ła* **(**the plural suffix *-ły* is not marked for grammatical gender and thus not useful for gender classification). Female nouns in the nominative form that occurred together with the -ła forms of previously extracted verbs were assumed to be +animate. This assumption turned out to be correct for 84% of extracted nouns.

Two other classes that by now have been extracted from the training corpus are adjectives and quantifiers co-ocurring with superanimate nouns. The common principle for extracting different word classes is shown in Figure 4. As can be seen, the main cues for automatic word classification are grammatical and semantic markedness, syntactic agreement and frequency of a particular inflectional form.

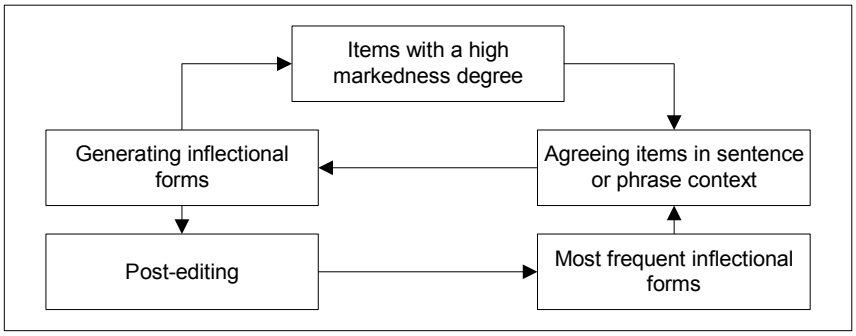

Figure 4: The general procedure for extracting and classifying different word classes.

5. Conclusions

The database built from both the training and the test corpus used in the experiment contains for the time being about 1600 open-class entries: ca 1000 nouns (92% of them +animate), 470 verbs and 180 adjectives. Those items cover ca 75% of the corpus vocabulary (excluding proper nouns). Among the unclassified words in the corpus we found only 2.9% superanimate nouns, which shows that the extraction has been quite effective

Among the unclassified words, the largest group (40%) are inanimate nouns. Extracting the category “inanimate” should therefore not be especially difficult. However, the inanimate nouns require a more fine-grained classification and thus some enrichment of the procedure shown in Figure 4 will probably be necessary. Our goal is a system based on conceptual features and able to handle such intricate problems as aspect choice and anaphora processing; thus, we do not want to refrain from detailed semantic classification of lexemes.

Gawronska and Duczak (2001) argue for the need to encode semantic features like +/- container, +/-uniplex and +/- solid in machine translation-oriented Polish lexicon; it seems to be possible to extract (at least to a certain extent) these features from prepositions and verbal prefixes co-occurring with the inanimate nouns.

The major conclusion from our experiment is that it is possible to partially automate the process of lexicon acquisition for a highly inflected language without diminishing quality requirements. This process can be made more efficient if both contextual cues and morphological cues, especially the features of inflectional prefixes, are employed. Highly marked grammatical affixes proved to be appropriate cues in the first phase of automating.

6. References

Bond, Francis, Ruhaida Binti Sulong, Takefumi Yamazaki & Kentaro Ogura: 2001, 'Design and Construction of a machine-tractable Japanese-Malay Dictionary', in Maegaard, B. (ed.): *MT Summit VIII. Machine Translation in the Information Age,* Santiago de Compostela, Spain, pp. 53-58.

Bond, Francis & Kyonghee Paik: 2000, 'Re-using an ontology to generate numeral classifiers', in *18th International Conference on Computational Linguistics: Coling 2000,* Saarbruecken, Germany, pp. 90-96.

Briscoe, Ted, & John Carroll: 1993, 'Generalized probabilistic LR parsing of natural language (corpora) with unification-based methods', *Computational Linguistics* 19:25-59.

Brent, Michael R: 1993, From grammar to lexicon: Unsupervised learning of lexical syntax, *Computational Linguistics* 19, pp. 243-262.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer: 1991, 'Word-sense disambiguation using statistical methods', in *ACL* 29, pp. 264-270.

Charniak, Eugene: 1997, 'Statistical parsing with context-free grammar and word statistics', in *Proceedings of 14th National Conference on Artificial Intelligence (AAAI '97),* pp.598-603.

Farwell, David. & Steve Helmreich: 2001, 'Towards Pragmatics-based Machine Translation', in *Proceedings of the Workshop MT Summitt VIII: MT 2010 – Towards a Road Map for MT*, Santiago de Compostela, Spain, pp. 22-25.

Gawronska, Barbara & Hanna Duczak: 2000, 'Understanding Politics by Studying Weather', in White, J.S. (ed.): *Envisioning Machine Translation in the Information Future*, Berlin/ New York: Springer Verlag, pp. 147-157.

Gawronska, Barbara & Hanna Duczak 2001. 'Image-schemata in machine translation: a cognitive analysis of Polish prepositions and prefixes'. In: *Proceedings of COMPLEX 20011. 6th Conference on Computational Lexicography and Corpus Research: "Computational Lexicography and New EU Languages"*. Centre for Corpus Linguistics, Dept. of English, University of Birmingham, pp. 21-33.

Gawronska, Barbara: 2001, 'PolVerbNet: an experimental database for Polish verbs', in Maegaard, B. (ed.): *MT Summit VIII. Machine Translation in the Information Age,* Santiago de Compostela, Spain, pp. 121-126.

Han, Chung-hye, Benoit Lavoie, Martha Palmer, Owen Rambow, Richard Kittredge, Tanya Korelsky, Nari Kim & Myunghee Kim: 2000, 'Handling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System', in *Proceedings of the 4th Conference*

of the Association for Machine Translation in the Americas, AMTA 2000 , Springer-Verlag, pp. 40-53.

Hayashi, Minoru, Setsuo Yamada, Akira Kataoka & Akio Yokoo: 2001, in Maegaard, B. (ed.): *MT Ssummit VIII. Machine Translation in the Information Age,* Santiago de Compostela, Spain, pp. 157-161.

Jackendoff, Ray: 1983, *Semantics and Cognition*, Cambridge, MA: MIT Press.

Lakoff, George: 1993, *The Contemporary Theory of Metaphor*, in Ortony, A. (Ed.), Metaphor and Thought*,* 2d ed. Cambridge: Cambridge University Press.

Langacker, Ronald: 1991, *Concept, Image, and Symbol. The Cognitive Basis of Grammar*, Berlin/New York: Mouton de Gruyter.

Miller, George, ed.: 1990, *WordNet: An On-Line Lexical Database*, Volume 3(4) of the International Journal of Lexicography. Oxford University Press.

Miller, George:1995, 'WordNet: An on-line lexical database', Communications of ACM, 38(11).

Niessen, Sonja & Hermann Ney: 2000, 'Improving SMT quality with morpho-syntactic analysis', in *Proceedings of Coling in Europe,* Saarbruecken, Germany, pp. 1081-1085.

Nirenburg, Sergei: 1996, 'Supply-side and demand side of lexical semantics', Introduction to the *Workshop on Breadth and Depth of Semantic Lexicons at ACL'96*.

Prescher, Detlef, Stefan Riezler & Mats Rooth:2000, 'Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution', in *Proceedings of Coling in Europe*, Saarbruecken, Germany, pp. 649-655.

Pulman, Stephen G.: 1996, Semantics, in Cole Ronald A., Joseph Mariani, Hans Uszkoreit, Annie Zaenen, & Victor Zue (Eds): *Survey of the State of the Art in Human Language Technology* http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

Resnik, Philip: 1992, 'WordNet and distributional analysis: A class-based approach to lexical discovery', in *Workshop on Statistically-Based Natural-Language-Processing Techniques*, San José.

Sarkar, Anoop & Daniel Zeman: 2000, 'Automatic Extraction of Subcategorization Frames for Czech', in *Proceedings of Coling in Europe*,Saarbruecken, Germany, pp. 691-697.

Sheremetyeva, Svetlana & Sergei Nirenburg: 2000, 'Acquisition of a Language Computational Model for NLP', in *18th International Conference on Computational Linguistics: Coling 2000,* Saarbruecken, Germany, pp. 1111-1115.

Talmy, Leonard: 1988, 'The Relation of Grammar to Cognition', in Rudzka-Ostyn, B. (Ed.), *Topics in Cognitive Linguistics* Amsterdam/Philadelphia: John Benjamins, pp. 165-205.

Thanopoulos, Aristomenis, Nikos Fakotakis & Georg Kokkinakis: 2000, 'Automatic Extraction of Semantic Relations from Specialized Corpora', in *18th International Conference on Computational Linguistics: Coling 2000,* Saarbruecken, Germany, pp. 836-842.

Webster, Mort & Mitch Marcus: 1989, 'Automatic acquisition of the lexical semantics of verbs from sentence frames', in *ACL* 27, pp. 177-184.