Marek Łabuzek, Maciej Piasecki,
Computer Science Department, Wrocław University of Technology
({labuzek,piasecki}@ci.pwr.wroc.pl)

Linguistically Annotated Data Sets for the Polish-English Machine Translation System.

1. Introduction

The work presented here originates from the development of the commercial, wide-scale machine translation (MT) system. The system was planned from its very beginning to be fully automated and was designated for wide-market. Its market name is *English Translator* (shorten further to *ET*) and it was created by *Techland* company. Naturally, the work presented here has more technical than scientific character and is primarily performance oriented. However, some experimental techniques being applied in the construction of the system and created data sets, which can be further utilised in the research, makes the subject different from the mere technical report.

During the construction of the system several linguistically well prepared data sets have to be created, in many cases almost from the scratch. The creation of the data-set was facilitated by the number of the software tools built in this purpose. The presentation of the data sets and tools is a main topic of this paper.

Because of the wide scale of *ET*, data sets must cover a significant amount of the 'real' language. Moreover, in order to decrease the time necessary for its development[1], it was assumed to apply Machine Learning methods in as large extent as possible in the system. This assumption influenced strongly the format of many data sets, as will be emphasised further in the paper.

ET has the typical architecture of the *MT* system based on transfer. During the subsequent stages of processing it needs the following linguistic data sets:

- *text segmentation*: context free grammar being a base for *Finite States Automata* performing the segmentation,
- *Part of Speech Tagging* (PST): corpuses + monolingual morpho-syntactic dictionaries,
- *parsing*: monolingual morpho-syntactic dictionaries and subcategorisation dictionary,
- *transfer*: bilingual dictionaries, subcategorisation dictionaries, bilingual subcategorisation dictionaries and even monolingual dictionaries (a part describing derivations e.g. a link between verbs in aspectual pairs during the transfer of constructions of English tenses),
- *word form generation*: monolingual dictionary.

All the data sets, mentioned above, have to be created especially for the needs of *ET*. There was only a chance for the utilisation of some existing morpho-syntactic dictionary of Polish. However, the lack of many features facilitating MT in the existing electronic dictionaries of Polish (e.g. additional syntactic information lexemes or links between Polish verbs and Polish participles, adjectives and adverbs) convinced us to develop our own dictionary. All other data sets had been simply non-available in the time the work on the system had started.

In the following sections, the data sets mentioned above, will be discussed preserving generally the order of the processing. There will be only one exception. The very resource consum-

[1] To be honest, this expectation was not 'fully' met. First results came fast, but qaultiy of translation was low. Improvement of the quality was involving more and more resources and at the same time the progress was constantly slowing down. Anyway, we are still convinced that the methods of Machine Learning can be very useful in the construction of MT systems in the future.

ing process of fully annotated corpus creation has been started as the last one and is still in experimental phase. Thus, the Polish corpus will be presented as the last data set. An additional section covers some of our experiences concerning the utilisation of the large, commercially available English corpus: *Penn Tree Bank*. They were so unexpected, that they may be interesting to be shared with other researchers.

2. Text Segmentation

The text segmentation phase of the translation process consists of two steps. In the first step, the sequence of characters, which forms a text, is divided into blocks. Blocks are words, punctuation marks, numbers, symbols and others. For each kind of block, there is a number of regular expressions, describing allowable sequences of characters that can form a block of a given kind. Each regular expression is compiled into finite state automaton, which accepts, or not, a given sequence of characters.

The blocks which are most difficult to identify are abbreviations with periods and various symbols. The former must be analysed together with period(s) as one block and this means that for each language there should be a separate set of regular expressions describing abbreviations occurring in it. The latter include symbols with periods (the period should not break block) and symbols containing only letters which should be distinguished from words.

In the second step of segmentation, the sequence of blocks obtained in the first step is divided into sentences. This work is done by one hand-crafted finite state automaton. The biggest problem are balanced delimiters (parenthesis, quotes), especially those including whole sentences between them. There is also a problem with abbreviations terminating with period occurring at the end of the sentence because in such situation, period plays two roles: being a part of an abbreviation and terminating a sentence.

3. Monolingual Dictionaries

The monolingual dictionary of our system has two layers: compressed and universal. The compressed layer is used to store and quickly access information, while the universal layer presents the information to other parts of the system in a consistent way. The former is stored on a disc and the latter is created on demand for given words.

The compressed layer groups words into parts of speech and lexemes. It consists of three parts: inflection, lexical information and derivations. For each part of speech, there is a list of inflection forms and a list of lexical information aspects, which mostly describe its syntactic properties such as e.g. reflexiveness of the verb or gradation type of adjectives and adverbs. Each lexeme belongs to only one part of speech. Certainly, one word can be a form of many lexemes. Also, there are lexemes which belong to the same part of speech and have very similar forms, e.g. there are two noun lexemes with base form "*zamek*", which differ only in one form – singular genitive ("*zamku*" i "*zamka*"). There are also lexemes, which have identical forms but differ in values of lexeme attributes, e.g. verbs that have reflexive and non-reflexive version. Lexemes have unique identifiers, which are numbers and are used in other databases of the system and its code to refer to them. They also have a list of values of lexical aspects, which are proper to the part of speech to which they belong.

Whole dictionary is described by the following data: a list of parts of speech, a list of base forms (or more proper: names of lexemes) and a set of inflection patterns. Each part of speech is connected with its inflection and lexical aspects and each base form is annotated with a part of speech, values of lexical information aspects and a name of inflection pattern.

The inflection part of the monolingual dictionary is implemented as a transducer translating between a word form and a pair: an identifier of a lexeme and a code of inflection form. The

source for transducer is generated in a usual way from the list of base forms and a set of inflection patterns. The lexical information part is a simple table of compressed values of lexical aspects.
The derivation part of the monolingual dictionary stores links between lexemes. Currently, we have links between verbs and deverbal nouns, verbs and four participles available in Polish (of course concrete lexemes can have less than five mentioned links) and perfective and imperfective version of a verb. These links are necessary for a proper transfer of tenses and other grammatical structures. Base forms of derived lexemes are described in the same way as inflection forms - in inflection patterns.
The universal layer describes words in a hierarchical way. One syntactic element, describing one word, consists of a word form and a set of syntactical alternatives. A syntactic alternative consists of an identifier of lexeme, a code of basic syntactic category and a set of morphological alternatives. And a morphological alternative is a set of pairs: attribute and value, where attribute can be either inflectional or lexical. Examples of the universal layer representation are presented on Fig. 1.

```
Surface: "chodzenie"
  (
    Syntactic Category: ODS-NOUN
    Semantic Class: 14060
    ( PERSON: F-THIRD-P, CASE: F-NOM, NUMBER: F-SING, GENDER: F-
NEUT, NEG: F-NEG-N )
    ( PERSON: F-THIRD-P, CASE: F-ACC, NUMBER: F-SING, GENDER: F-
NEUT, NEG: F-NEG-N )
    ( PERSON: F-THIRD-P, CASE: F-VOC, NUMBER: F-SING, GENDER: F-
NEUT, NEG: F-NEG-N )
  )

  Surface: "psa"
  (
    Syntactic Category: NOUN
    Semantic Class: 8080
    ( PERSON: F-THIRD-P, CASE: F-GEN, NUMBER: F-SING, GENDER: F-MZYW
)
    ( PERSON: F-THIRD-P, CASE: F-ACC, NUMBER: F-SING, GENDER: F-MZYW
)
  )

Surface: "daj"
  (
    Syntactic Category: VERB
    Semantic Class: 42884
    ( NUMBER: F-SING, PERSON: F-SECOND-P, VB-FORM: F-IMP-FORM,
CONDIT: F-CONDIT-N, ASPECT: F-PF )
  )
```

Fig. 1 Examples of universal layer representation.

A set of *basic syntactic categories* (BSC) was proposed. The set is extended in comparison to a typical list of Polish *parts of speech* and the categories from the set are also a part of the

grammar of the parser. All syntactic categories are organised into hierarchy by explicitly defined *subsumption* relations, e.g.:

```
NOUN: ODS-NOUN, PN, PRON
   PRON: PER-PRON, PRON-NPER, PRON-ZPR, PRON-ZWR, PRON-NEG, PRON-DEM.
```

In the example NOUN has three subcategories: deverbal noun, proper noun and pronoun. PRON, in turn, has six subcategories: personal pronoun, indefinite pronoun, interrogative pronoun and others.

The subsumption relations determine the set of morpho-syntactic attributes assigned to the categories. Some categories of the higher levels are motivated by the correspondence between Polish and English grammar or have a character of semantic subcategories. The subsumption relation is used in machine learning algorithms of our system.

In the Polish monolingual dictionary, we have currently above 180,000 lexemes which gives above 2.5 million forms. The compiled file of transducer has 20MB, which is 30% of a source file.

4. Bilingual Dictionaries

The bilingual dictionary is described in a text file containing words and phrases annotated with part of speech and other information necessary for identifying lexemes and forms of words. The file is compiled into binary files: phrase bases, which store phrases in the form of parse trees and a proper bilingual dictionary, which stores pairs of identifiers of words and phrases.

Currently, we have 110,000 translations in the Polish-English dictionary (not counting virtual entries for derivations, which are translated through verbs) and 125,000 translations in the English-Polish dictionary.

5. Subcategorisation Dictionaries

The subcategorisation dictionary has been constructed mainly to support the parsing (this dictionary is also used during transfer, what is discussed later in this section). The parser is based on Machine Learning techniques[2] what influenced strongly the format of the data in the subcategorisation dictionary. All entries have taken a form of a tree structure describing some 'situation' being met during parsing. Each tree has a distinguished leaf node called 'head'. This node, or more precisely, the identifier of a lexeme kept in it, is used as an index for retrieval of the tree from the dictionary. The notion of the 'head' has been used here rather awkwardly: it collapses with *head* understood as syntactic relation. That is why, we use PRED as name for syntactic relation corresponding to HEAD.

```
SNT { SUBJ NP_N}{ PRED VP {strzec PRED VERB }{się MOD PART }( OBJ
                                                          NP_G)}
SNT { SUBJ NP_N}{ PRED VP {błogosławić PRED VERB }{ IOBJ NP_A}}
SNT { SUBJ NP _N}{ PRED VP {błogosławić PRED VERB }{ IOBJ NP_D}}
SNT { SUBJ NP _N}{ PRED VP {dawać PRED VERB }{ IOBJ NP_D}{ OBJ
                                                          NP_A}}}
```

Fig. 2 Examples of the trees from Subcategorisation Dictionary of Polish

Besides the structure, each tree describes several requested elements (see Fig. 2):

- *subtrees* – specified as compound category,

[2] Parser works according to *shift-reduce* scheme and its logic is based on modified *Decission Tree Learning* technique.

- *syntactic/semantic roles* – only syntactic ones are presented in Fig. 2,
- *values of morphological attributes* – e.g. _G annotating the case,
- *specific lexemes* (exactly their identifiers)- they enrich the key for retrieval.

Moreover, the Polish subcategorisation dictionary codes additional information concerning the optional elements and groups of optional elements (where at least one element of the group must be realised) and sequences with fixed order. Here, we are following the notation proposed by Polański (Polański 1984).

Presently, the subcategorisation dictionary contains mainly the verb trees (probably the most important between all other types for the parsing). But also, there are some trees describing adverbial constructions, some compound adverbs, prepositions and conjunctions. An important part of the dictionary, constantly growing, is formed by trees describing *multiword idioms*. The present size of the dictionary is about 18.000 trees.

Translation of the lexeme only on the base of a bilingual dictionary is very often ambiguous or simply wrong. The results can be significantly improved when we use the subcategorisation context during translation. A bilingual subcategorisation dictionary suits these needs. Moreover, during the transfer we should not only translate a lexeme on the base of its subcategorisation but also we should transform, during the transfer, the whole structure described by the tree from the dictionary.

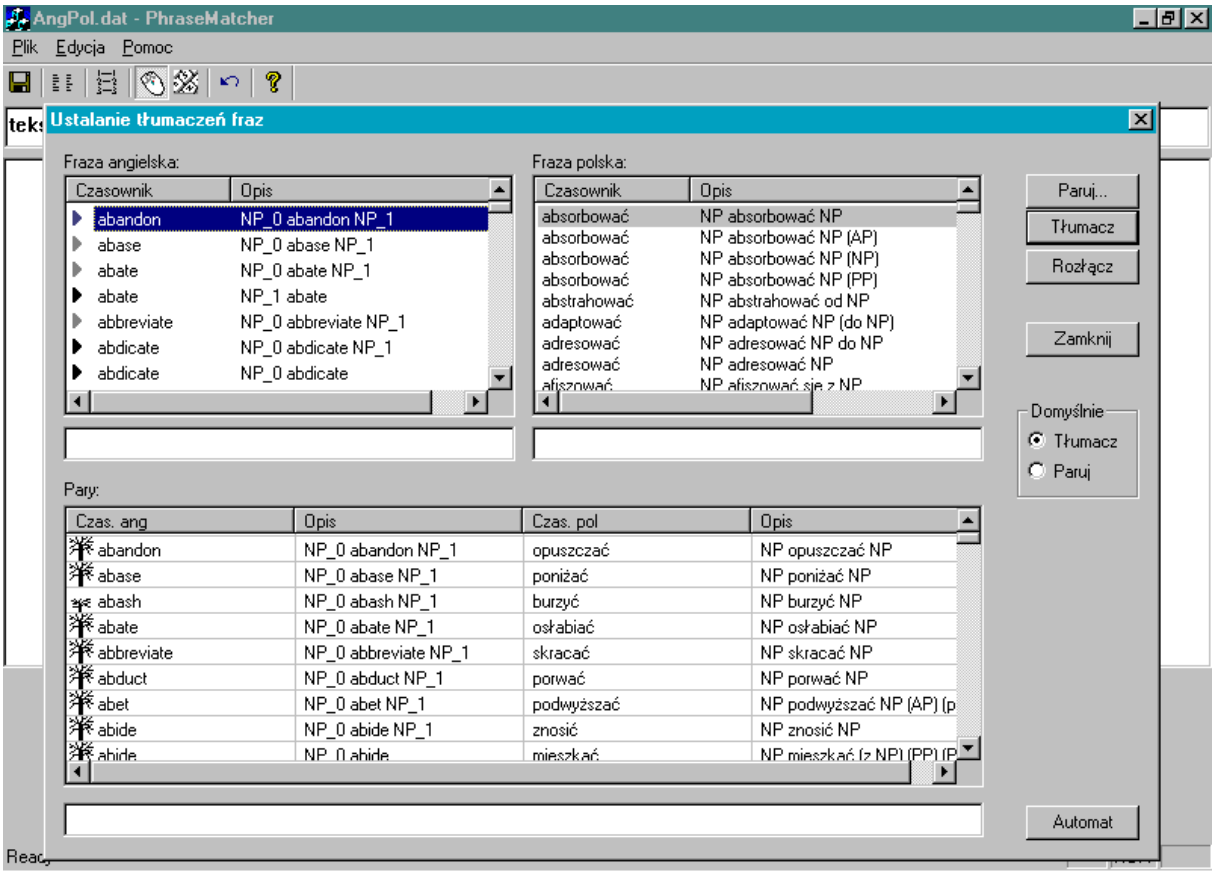

Fig. 3 *PhraseMatcher* – a list browser view enabling to pair the subcategorisation trees.

The bilingual subcategorisation dictionary associates pairs of the trees and delivers additional information controlling the transfer e.g. the information defining the corresponding pairs of arguments, identifying arguments to be deleted or controlling the process of transformation of the source argument into the target in case when their categories differ significantly.

A special tool with a graphical interface has been created in order to facilitate the process of definition of the entries in the bilingual subcategorisation dictionary. It allows for graphical

browsing of the source and target trees in two columns and observing the resulting pairs in the list below (Fig. 3). Next, the whole controlling information can be graphically defined (Fig. 4). In both trees the nodes are uniquely numbered. Human operator using fast access by accelerator keys, or slower but more comprehensible access by mouse, is able to pair the corresponding nodes. The corresponding nodes can be located in both trees on different levels of the structure and can be of different categories. The automatic transfer procedures deal with these differences. The information about pairing is stored with the link of the trees.
The design of user interface was oriented on the achievement of the efficiency of hand movements. Several version of the design have been evaluated and the finally chosen version was the one with the best estimation of the time consumed. The analysis was performed by the application of analytical usability assessment methods (e.g. *GOMS*, *Fitz law*).

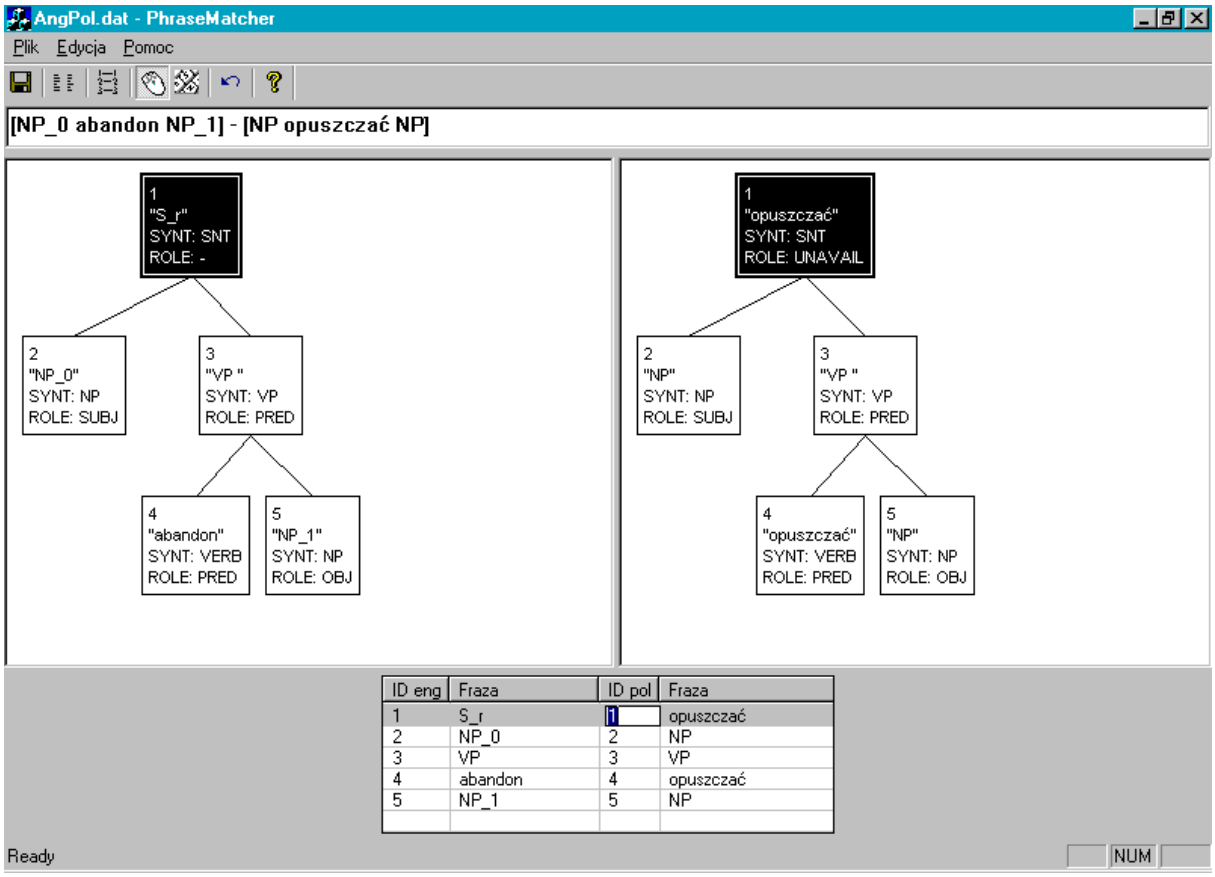

Fig. 4 *PhraseMatcher* – a tree browser view enabling to pair the nodes of two linked trees.

6. Adaptation of English Corpus

When we decided to purchase a large, professional corpus of English, we thought that the biggest problem would be to persuade the managers of the company to spend not a small sum of money for the commercial licence. We were plainly wrong: the worst was before us: *Penn Tree Bank* (*PTB*) has appeared to be not a tool of 'walk up and use' type.
The first big mistake was partially our own: we defined the system of syntactic categories for the parser before purchasing *PTB*. However, using the system of categories close to the one assumed in a large grammar of English (XTAG 1999), we did not expect any more serious problems. But, we had been confronted with the necessity of conversion of syntactic categories. The problem was so serious that a sophisticated expert system had to be constructed to

perform the task (but still not completely – see below the remark on subcategorisation of multiword verbs, below).
Other mistakes were not obviously ours[3]:

1. We found quite a big number of ambiguities left in *PTB* tags (e.g. doubled markings for: gerund/noun, adjective/noun, past form/past participle etc.).
2. The patterns of subcategorisation of multiword verbs and phrasal verbs concerning the tagging the subsequent words as prepositions and adverbs are not explicitly given. Moreover, they seem to be very unstable across the corpus.
3. *PTB* has been expanded during the subsequent stages of development. At least two of them were connected with change in the system of tags. However, in the present version of *PTB* both systems can be met! The differences are not very big but e.g. word "to" had been tagged in the previous version of *PTB* with a tag TO, the present version of *PTB* has two different tags for "to" used as preposition and 'modal' (before infinitive form). One can still meet in *PTB* quite a big number of TO tags.
4. Finally, we have encountered many simple mistakes (!) e.g. pronouns tagged as determiners and vice versa.

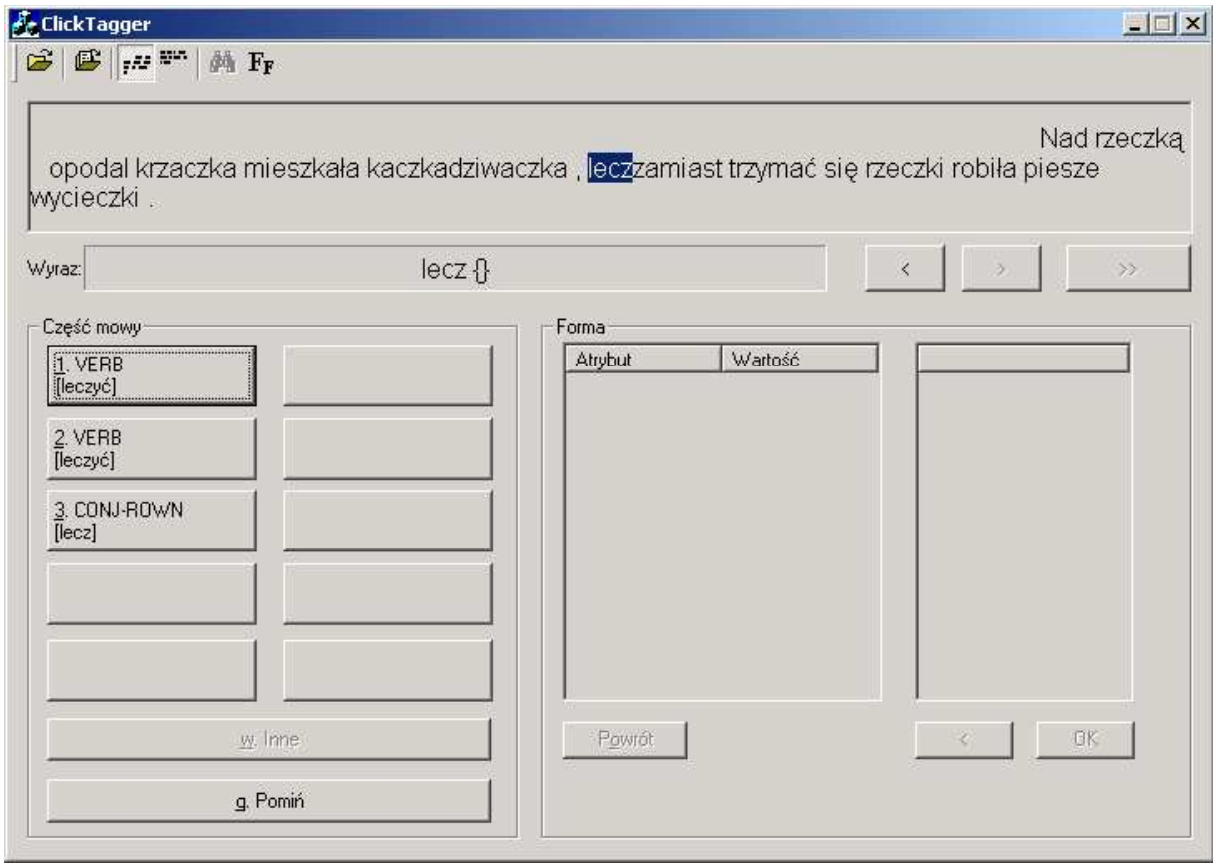

Fig. 5 *ClickTagger* – disambiguating BSC of words in the corpus.

Besides construction of the expert system, in order to solve the problems mentioned in points 1-4, we had to make a lot of manual disambiguations and corrections (e.g. change of tag COMP for PREP, and vice versa, in the case of word "for"). The number of 'touched' words

[3] Our remarks are related to the version of *PTB* being described as « preliminary », but anyway being commercially sold and delivered.

is up to 1.5% of all words of *PTB*, even, still leaving the problem of subcategorisation translation (or more exactly: correction) unresolved. Anyway, after all this work had been done, we were still able only to use at most a half of *PTB*.

7. Polish Corpus

In Polish, there are morpho-syntactic ambiguities of three kinds – a word: is a form of different lexemes of the same BSC, is a form of lexemes of different BSC, represents different forms of the same lexemes. The first two kinds are significantly less frequent than in English but the third one is very frequent. To solve a problem by construction of a tagger, we needed a large fully annotated corpus of Polish. In the time the *ET* project was started, there was no available (even only commercially) such corpus. We decided to create our own one.
The corpus has been collected from all publicly available sources of electronic texts, mainly from web pages. About ~2GB of rough data has been stored. Obviously the quality of the data gathered in this way is very low. However, the quality of the final, annotated form of the corpus is reasonably high due to process of preparation of the annotated data.

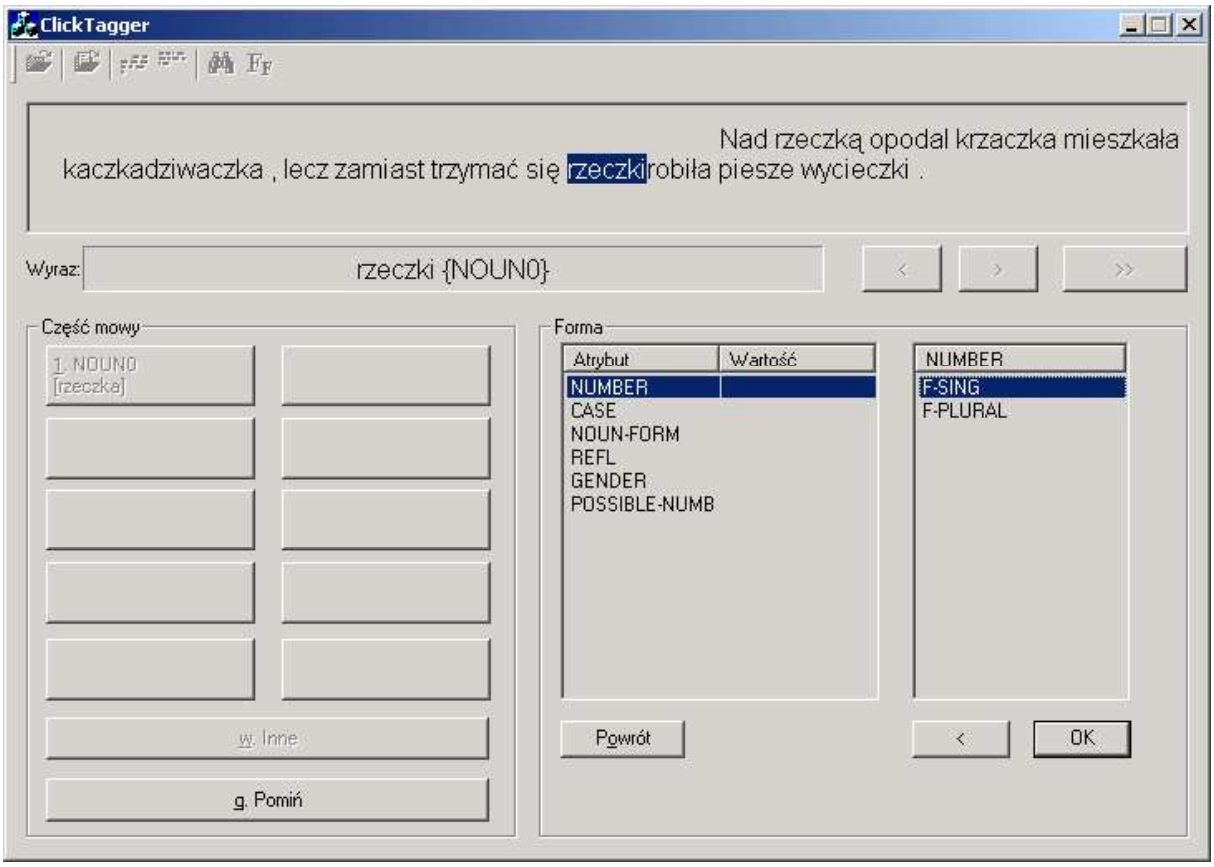

Fig. 6 *ClickTagger* – disambiguating values of morpho-syntactic attributes.

All words in texts have been manually described with a help special tool called *ClickTagger* (Fig. 5). *ClickTagger* is a tool with a graphical interface, collaborating with the monolingual dictionary of *ET*. The tool automatically assigns tags to all unambiguous words and presents only ambiguous ones to the human operator showing each such word centred in the context of the surrounding words (only a limited number of words is visible, but there is still a possibility of backward and forward scrolling) – see Fig. 5. First, the basic syntactic category is disambiguated, next subsequent attributes. In Fig. 5, the situation is shown, in which BSC of word *lecz* is ambiguous between three possibilities. The human operator is forced first to disam-

biguate BSC. Next, *ClickTagger* always shows the list of possible values of morpho-syntactic attributes, from the dictionary, see Fig. 6, and goes further as quickly as some decision made clarifies the given word. Unknown words are marked with special markers.
Results of manual tagging are next assessed by a human supervisor. The supervisor works on a printed form. He has also a possibility for introduction of appropriate changes into the monolingual dictionary in case of any mistakes had been identified in the corpus. The same, the work on corpus annotation gives also the good opportunity for elimination of mistakes from the dictionary.
Finally, the operators perform the computer aided correcting of the corpus (based on new extended dictionary).
The 'side effect' of this process is introduction of many new lexemes of 'internet jargon' into the dictionary (e.g. "host", "web"). However, having in mind that *ET* is being commonly used for the translation of web pages, this side effect has a quite positive influence on the quality perceived by the users.
The present size of the corpus is about 65.000 tagged and corrected words. Next 60.000 tagged words still waits to be corrected.
There are two tag sets: a full tag set and reduced tag set. Each full tag consists of name of a basic syntactic category followed by a list of morpho-syntactic attribute – value pairs. The full tag facilitates the process of assessment, mentioned above. An example of a part of the corpus with full tags (the primary form) is presented on Fig. 7.

```
Dzięki {PREP} temu {PRON CASE=F-DAT PERSON=F-THIRD-P NUMBER=F-SING
GENDER=F-NEUT} ograniczeniu {ODS-NOUN PERSON=F-THIRD-P CASE=F-DAT
NUMBER=F-SING GENDER=F-NEUT NEG=F-NEG-N} zmieszczą {VERB TENSE=F-
PRESENT NUMBER=F-PLURAL PERSON=F-THIRD-P VB-FORM=F-BASE-FORM
CONDIT=F-CONDIT-N ASPECT=F-PF} się {PART} zapewne {ADV GRADE=F-GR-
REG} naraz {ADV GRADE=F-GR-REG} na {PREP} ekranie {NOUN PERSON=F-
THIRD-P CASE=F-LOC NUMBER=F-SING GENDER=F-MNZYW} ; {SEP-DELIM}
```

Fig. 7 A fragment of the annotated part of Polish Corpus.

The full tag set consists of about 1600 different tags. Having in mind that the number of words presently in the corpus is not very big (the number was given earlier), the full tag set is much to big to be useful during teaching of POS automatic tagger with the help of *Machine Learning or Statistical Learning* methods. We have tried to reduce the number of tags. But, the natural limitation is here the usefulness of information conveyed by tags according to the reduction of the number of ambiguities during parsing. Reducing the number of tags we get more and more general tags. More general tags 'cut' less number of possibilities from the dictionary. We have been hardly able to reach a lower bound of about 270 different tags. Paying here some trade-off with usefulness. We believe that the further reduction is not possible and that the optimal number of tags (besides the full set) is above 500.
On such limited corpus, and limited number of tags, we have been able to perform only some initial experiments with POS tagger learning. The achieved 86% performance of the tagger counted in a standard way (according to all words), did not give any noticeable improvement during the parsing.

8. Further Development

The work is planned in the following areas. The monolingual dictionary will be enlarged by new words, especially specialized ones and the inflection will be corrected if any errors will be encountered. New derivation links will also be added. The first kind of a planned link is a

unidirectional link from perfective version of the verb to the imperfective one, since bidirectional links are not proper when imperfective verb has more than one perfective counterpart e.g. *chorować - zachorować, rozchorować się*. Other planned link are adjective - adverb and noun - adjective (so called relational adjective i.e. describing the feature of being related with some concept e.g. *grupa - grupowy*). Bilingual dictionaries will be enlarged, especially by specialized words and various phrases. Subcategorization dictionary is planned to cover all verbs from the monolingual dictionary. The English corpus will be further adapted to our dictionary and parsing methods and the Polish one - enlarged and tried to be utilized especially in a tagging phase of the translation process.

References

(Bień 1991) Bień J.S. "Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji", Wyd. UW, Warszawa 1991.

(Bień & Szafran 2002) Bień J. S., Szafran K. *An Experimental Parser of Polish.* To appear in proceedings of Formal Description of Slavic Languages Leipzig'99.

(Daciuk 1998) Daciuk J. Incremental Construction of Finite-State Automata and Transducers, and their Use in the Natural Language Processing. PhD thesis, Technical University of Gdańsk, 1998.

(Markowski 1999) "Nowy słownik poprawnej polszczyzny", red. Markowski A., PWN, Warszawa, 1999.

(Polański 1984)„Słownik syntaktyczno-generatywny czasowników polskich", red. Polański Kazimierz, Instytut Języka Polskiego PAN, 1984.

(Saloni & Świdziński 1998) Saloni Zygmunt, Świdziński Marek. Składnia współczesnego języka polskiego. PWN, Warszawa 1998.

(Szpakowicz 1983) Szpakowicz Stanisław. Formalny opis składniowy zdań polskich. Wyd. UW, Warszawa, 1983.

(Świdziński 1992) Świdziński Marek. Gramatyka formalna języka polskiego. Wyd. UW, Warszawa, 1992.

(Vetulani et al. 1998) Vetulani Z., Walczak B., Obrębski T., and Vetulani G. Unambiguous coding of the inflection of Polish nouns and its application in electronic dictionaries – format POLEX. Wydawnictwo Naukowe UAM, 1998.

(XTAG 1999) The XTag Group of Institute for Research in Cognitive Science, University of Pennsylvania. A Lexicalized Tree Adjoining Grammar for English. www.cis.upenn.edu/~xtag (1999)