Marek Świdziński
Uniwersytet Warszawski

A new version of the formal grammar of Polish: corpus-backed improvements and corrections

Marek Świdziński*, Marcin Woliński^, Katarzyna Głowińska^
*Institute of Polish, Warsaw University
^Institute of Computer Science, Polish Academy of Sciences

Abstract

The paper reports on a research project aimed at half-automatic construction of a treebank for the Polish language that was just completed. The tool was the parser Świgra, based upon the formal grammar of Polish (FGP). It generated sets of constituency trees for 10,000 utterances taken from the subcorpus of the National Corpus of Polish (NCP). The sets of trees were examined by a team of annotators expected to select the correct interpretation for each utterance. Apart from the treebank itself, the work resulted in a refinement of the grammar.

Résumé

L'article rend compte d'un projet de recherche qui vient d'aboutir et qui avait pour but la construction semi-automatisée d'une banque d'arbres pour le polonais. L'outil utilisé était l'analyseur syntaxique Świgra, lui-même fondé sur une grammaire formelle de la langue polonaise (FGP). Cet outil a généré des ensembles d'arbres de constituants pour 10 000 énoncés extraits d'un sous-corpus du Corpus National du Polonais. Ces ensembles d'arbres ont été examinés par une équipe d'annotateurs qui devaient choisir le seul arbre approprié pour chaque énoncé. En dehors de la banque d'arbres elle-même, le travail a abouti à la redéfinition de la grammaire.

Introduction

The paper gives a sketchy report on a linguistic project that was a part or perhaps a by-product of another, fairly larger project aimed at half-automatic construction of a treebank for Polish. The latter, completed in October 2011, made use, *inter alia*, of a formal grammar of Polish compiled in the early 1990s. The former project, in turn, consisted of parallel testing and verification of the grammar both from an empirical and a technical perspective, which resulted in a number of improvements to the grammar and in plans of its further development. With all this, the paper is technical, rather than purely linguistic.

Formal grammar of Polish

The formal definition of the Polish language referred to in this paper was given in « Gramatyka formalna języka polskiego » [Formal grammar of Polish; henceforth FGP] by Marek Świdziński (Świdziński, 1992). FGP was designed as a theoretical model of the ideal Polish native speaker's competence with no corpus background, as practised by Saussure-style structuralists. Unlike their models, the one presented here is formal, i.e., it defines a certain set of expressions – the formal language – to be compared with another language: the set of correct utterances called *Polish*. Most structures of Polish were accounted for by FGP. Only several types of structures (not many) were omitted: non-finite sentences//clauses, coordinate phrases, nominal phrases with numerals, discontinuous constructions and some other idiosyncrasies.

The definition of the Polish language given by FGP putsa particular emphasis on various agreement phenomena typical of highly inflected languages. FGP is a context-free phrase structure grammar applying the formalism of metamorphosis grammar (cf. Colmerauer, 1978, for the original article and Abramson and Dahl, 1989, for developments). Syntactic units are represented by terms, each carrying an appropriate set of attributes that formalize grammatical features of those units. Rules of the grammar, fairly numerous and complicated, define particular units as sequences of other units, establishing correspondences between grammatical features. The syntactic hierarchy is multi-level; trees defined by the rules of the grammar, with non-binary branching at some levels and very long unbranched branches, hardly resemble those the generativists operate with.

It should be admitted that FGP has little in common (if any) with the generative tradition, be it GB, HPSG, or LFG. FGP is simply a sort of calculus invented to account for nothing but one natural language (Polish) or, to be more exact, on its surface. Rather, it is comparable with such formal grammars, as, e.g., [Dać tu coś metamorficznego angolskiego!!! Janusz coś o tym mówił!]

The grammar represents a syntactic, not lexical formalism. All agreements are accounted for via rules, mainly by parameter matching. A great number of attributes is proposed to cope with agreements of various kinds. FGP, too, operates with a large repertoire of nonterminals.

Almost every syntactic unit has got three types of realizations:

- typical (= coordinate),
- simple (= subordinate), and
- degenerated (= elementary).

Three syntactic mechanisms are thoroughly examined:

- syntactic implication (valency),
- agreement, and
- linear ordering.

Polish is treated as if it were a fixed word order language. All possible permutations are given explicitly.

Four layers of the structure are distinguished in FGP, which results in a multi-level hierarchy:

- word (token),
- phrase,
- clause (= elementary sentence), and
- sentence.

FGP comprises no less that 500 rules that look like that given below:

```
ZE (wf,a,c,t,rl,o,wa,wb,wc,neg,i,z,ow)
= FL (a,c,rl,o,neg,i,NP)
< FF (wf,a,c,t,rl,o,wa,wb,wc,k,neg,NI,z,ow)
```

```
    FW (wa,k,a,c,rl,o,neg,NI,NP)
    FW (wb,k,a,c,rl,o,neg,NI,NP)
    FW (wc,k,a,c,rl,o,neg,NI,NP)
  $ RÓŻNE(z,BY".CHOĆBY.CO.CZYŻBY.GDYBY.JAKBY.
        JAKI.JAKOBY.KTO.KTÓRY.P.P'.P".PZ.ŻEBY).
```

where `ZE` stands for an elementary clause, `FF` for a finite phrase, `FW` for a complement, `FL` for an adjunct, attributes in brackets formalize various morphological or syntactic features such as aspect (`a`), tense (`c`), gender and number (`rl`), negativity (`neg`), valency (`wa`, `wb`, `wc`) and other. The rule above defines clauses with an initial adjunct, the whole set of rules for finite sentences containing almost twenty.

It is worth noticing that the rules of FGP do not introduce terminals, i.e., the lowest level to reach is that of preterminals. FGP does not provide the lexicon: only the so-called function words (conjuntions, complementizers, relative and interrogative pronouns, punctuation marks) are introduced by rules that define some preterminals.

As a model of linguistic competence, FGP carefully accounts for syntactic ambiguity. This means that a number of different structures is often assigned to a given expression, which are either technical variants of one invariant, or reveal the « meaningfulness » of the difference. Finally, there are empty constituents and terminals in FGP, not unfamiliar to an average native speaker's linguistic intuition.

First verifications and developments

As not based upon any corpus data, FGP had to be empirically verified. In 1993, a research project was undertaken to check the accuracy of the definition of Polish given in FGP. The linguistic material was obtained from the half-million-word corpus of the frequency dictionary of Polish (SFPW, 1990); interestingly, the corpus that dates from the late 1960's is the first Polish corpus both founded and stored digitally.

The team members used to take every 10th utterance from the corpus. Their analysis consisted of a sort of manual and provisional parsing of those utterances according to the rules of FGP whenever possible. « Terminals » were not real tokens; rather, the analysis used to stop at the clausal or phrasal level. Grammatical features revealed in the course of structuring were also attributed to clauses and phrases, not to words. The results founded a database of about 10,000 utterances. In the database no less than 30 features are introduced, including clause type, head with its lemma and morphological characteristics, length of the clause and its components, word order and many other.

The report was published as « Własności składniowe wypowiedników polskich » [Syntactic features of Polish clauses and sentences] (Świdziński, 1996). What followed were a number of improvements and developments of FGP: redefinition of negativity (Przepiórkowski, Świdziński, 1997), formal description of Polish numeral expressions (Świdziński, 2005), and a new definition of the nominal phrase (Świdziński, Woliński, 2009).

Automatic syntactic analyser

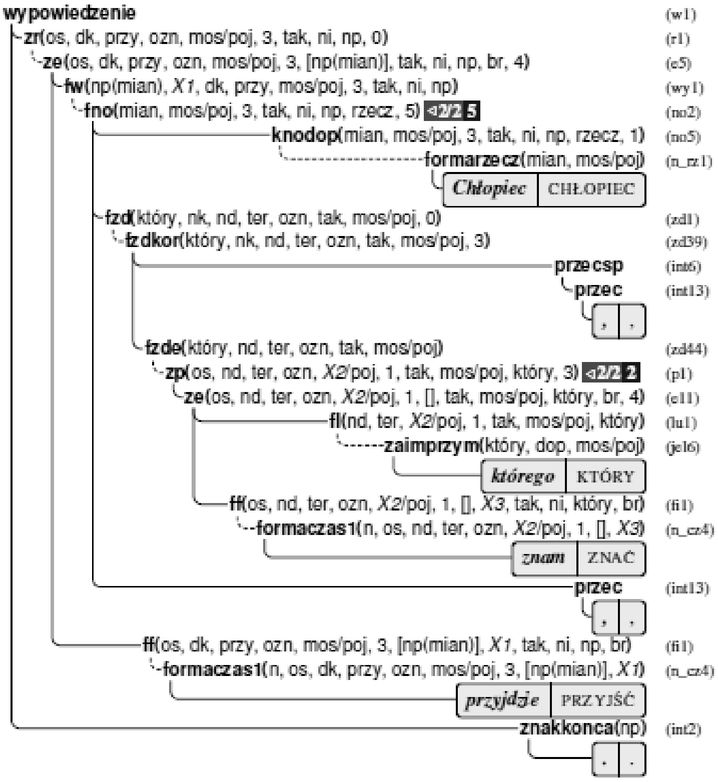
After some local endeavours to implement fragments of FGP in the 1990's the grammar was successfully implemented by Marcin Woliński who designed an automatic syntactic analyser Świgra in his Ph.D. dissertation « Komputerowa weryfikacja gramatyki Świdzińskiego » [A computational verification of Świdziński's grammar] (Woliński, 2004).

Although the formalism of FGP is a variation of metamorphosis grammar, which makes it natural to attempt an implementation in Prolog, M. Woliński decided to build a parser from scratch. What should be emphasized is the fact that FGP had to be modified. First of all, the analysis starts from the bottom:

> The bottom-up parsing strategy is hardwired in the parser. For highly inflected languages like Polish it seems to be the best choice since inflectional features of particular words when going bottom-up come into play early in the process, blocking therefore many impossible paths in the search space.
> The result of a parse is a shared parse forest [...]. Since sets of parse trees are sometimes quite large even for sentences that are neither long, nor particularly complicated it is an interesting challenge to navigate in the set and process it as a whole without generating all possible trees.

(Świdziński, Woliński, 2009: p. 145). For implementation purposes FGP was coupled with the morphological analyser Morfeusz SIaT (Wolinski, 2006). A number of rules were added to the grammar to make the parser work that way. Some tricks had to be used to limit void recursion in the rules, to get rid of rules introducing empty realizations of several syntactic units, same as five context-sensitive rules, impermissible in the formalism. The account of free word order of Polish clauses, half-formal in FGP, was designed more accurately. One can say that what entered the parser was just FGP', the first update of FGP.

Below a tree generated by Świgra is given:



wypowiedzenie (w1)
zr(os, dk, przy, ozn, mos/poj, 3, tak, ni, np, 0) (r1)
ze(os, dk, przy, ozn, mos/poj, 3, [np(mian)], tak, ni, np, br, 4) (e5)
fw(np(mian), X1, dk, przy, mos/poj, 3, tak, ni, np) (wy1)
fno(mian, mos/poj, 3, tak, ni, np, rzecz, 5) ◁2/2 5 (no2)
knodop(mian, mos/poj, 3, tak, ni, np, rzecz, 1) (no5)
formarzecz(mian, mos/poj) (n_rz1)
Chłopiec CHŁOPIEC
fzd(który, nk, nd, ter, ozn, tak, mos/poj, 0) (zd1)
fzdkor(który, nk, nd, ter, ozn, tak, mos/poj, 3) (zd39)
przecsp (int6)
przec (int13)
, ,
fzde(który, nd, ter, ozn, tak, mos/poj) (zd44)
zp(os, nd, ter, ozn, X2/poj, 1, tak, mos/poj, który, 3) ◁2/2 2 (p1)
ze(os, nd, ter, ozn, X2/poj, 1, [], tak, mos/poj, który, br, 4) (e11)
fl(nd, ter, X2/poj, 1, tak, mos/poj, który) (lu1)
zaimprzym(który, dop, mos/poj) (je16)
którego KTÓRY
ff(os, nd, ter, ozn, X2/poj, 1, [], X3, tak, ni, który, br) (fi1)
formaczas1(n, os, nd, ter, ozn, X2/poj, 1, [], X3) (n_cz4)
znam ZNAĆ
przec (int13)
, ,
ff(os, dk, przy, ozn, mos/poj, 3, [np(mian)], X1, tak, ni, np, br) (fi1)
formaczas1(n, os, dk, przy, ozn, mos/poj, 3, [np(mian)], X1) (n_cz4)
przyjdzie PRZYJŚĆ
znakkonca(np) (int2)
. .

'The boy whom I know will come.'

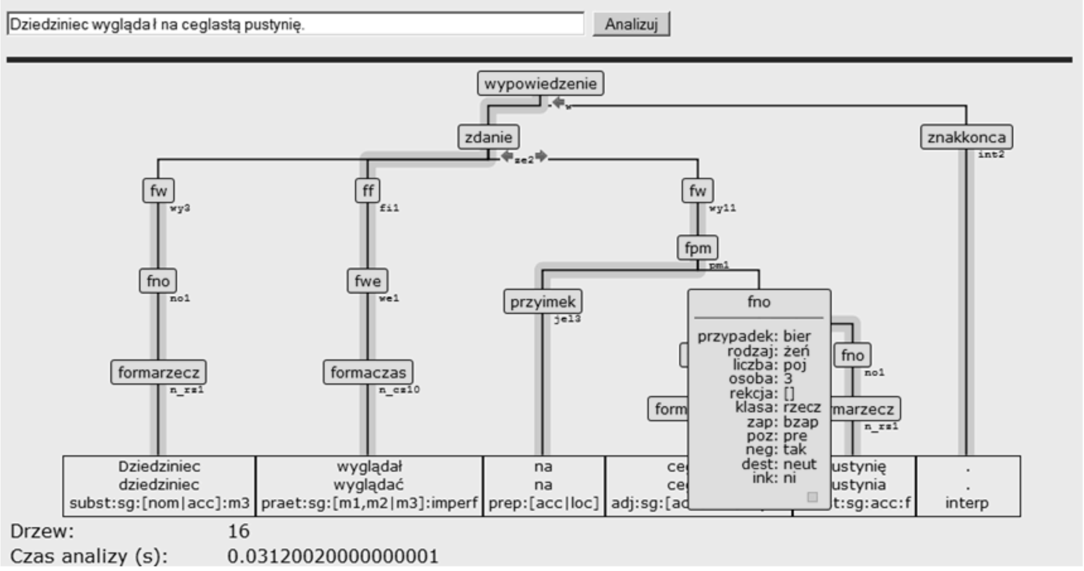
The tree is one of the three trees obtained by the analiser for this utterance.

The treebank project

In 2008, a 3-year project was launched by M. Woliński. Its aim was to build a treebank of 20,000 utterances on the basis of the National Corpus of Polish (NKJP) (Świdziński, Woliński, 2010). Utterances taken therefrom were processed by Świgra which, as a rule, provided sets of trees for each. The resulting forests, visualised by means of the application Dendrarium, were subject to manual verification by a team of linguists composed of 15 people. Their duty was to choose the only proper interpretation for the analyzed utterance, i.e., to syntactically disambiguate it. The idea sounds theoretically justified (in fact, human communication consists in an interchange of unequivocal pieces of information) but is alien to the original, ambiguity-revealing philosophy of FGP.

The empirical basis of the project was a one-million-word balanced subcorpus of NKJP. This subcorpus was manually annotated with morphological features, which means that the results of a morphological analyser were disambiguated and descriptions were added for words unknown to the analyser (mainly proper names). Consequently, every word in this corpus used to meet one morphological interpretation, which made syntactic disambiguation feasible (to say nothing of blocking unwanted paths in the process of parsing).

Dendrarium is a web-based tool for treebank development (Woliński, 2010). The tool provides an environment for disambiguation and validation of parse forests generated by the grammar. The tree in this environment looks like this:



Dziedziniec wyglądał na ceglastą pustynię.

fno

przypadek: bier
rodzaj: żeń
liczba: poj
osoba: 3
rekcja: []
klasa: rzecz
zap: bzap
poz: pre
neg: tak
dest: neut
ink: ni

```
Drzew:                 16
Czas analizy (s):      0.03120020000000001
```

Dziedziniec wyglądał na ceglastą pustynię.
'The courtyard looked like a bricky desert.'

As it is easy to see, not only are structures visualized but also values of morphological and syntactic attributes of each node (each syntactic unit) are revealed. It is worth noticing that the tree above is one of the 16 trees provided by the parser.

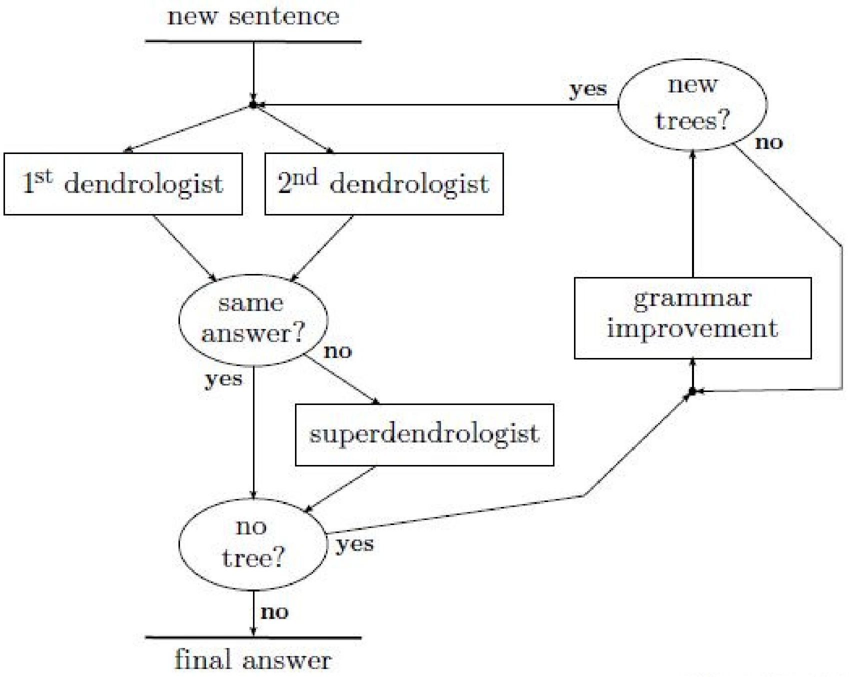
Each set of trees was analyzed by two linguists independently. They either rejected some of the trees as improper or vague, signalled cases of the lack of the proper tree (empirically deviant utterance or correct but unaccounted for), or chose the only one – even in cases of more variant trees. The results, if different, were then subject to adjudication by the supervisor (superdendrologist). This role was played by Katarzyna Głowińska. She prepared a very detailed instruction whose purpose was to help the team members make unequivocal choices and, consequently, used to take care of solving possible conflicts between those working on the same pack of utterances.

The annotators (dendrologists, in our jargon) were expected to check interpretations provided by Świgra concerning (a) segmentation, (b) arguments of the finite verb, (c) centers of clauses and phrases, and (d) morphological characteristics of each token. Their decisions, same as their hesitations, happen to have been source of improvements to the grammar.

Of particular importance were cases when no correct tree was found. Special answers were to be chosen:

- utterance is really incorrect,
- error in morphology or segmentation,
- utterance is not a finite sentence (subordinate or coordinate),
- sentence is correct but too difficult to describe,
- sentence is correct, grammar improvement requested.

The latter three cases used to imply adjustments to the grammar. The whole work of the team can be illustrated by the diagram below:



The project is reported in Woliński, Głowińska, and Świdziński, 2011.

A new version of the grammar: FGP2

Unlike FGP, which was, in a sense, an impulse to design a parsing device for Polish, the new grammar can be said to have originated from Świgra. The closing of the treebank project notwithstanding, FGP2 by M. Woliński and M. Świdziński is still under construction. Below, modifications to the grammar are mentioned no matter whether introduced in the course of the project, or just prompted for the future work. The refinements to the grammar used to be made within the project simultaneously and iteratively.

1. Formal apparatus

The adjustments concerned first of all the formal apparatus. The rules of FGP2 look now and work different:

```
zdanie(Wf,A,C,T,Rl,O,Neg,Dest,I,Pk,Sub) -->
 s(ze1),
 ff(Wf,A,C,T,Rl,O,Wym,Neg,Dest1,I,Pk1,na),
 sequence_of([
          fw(W2,A,C,Rl,O,post,Neg,Dest2,ni,Pk2,po)
              ^[oblwym_iter,Wym,W2,ResztaWym],
          fl(A,C,Rl,O,Neg,Dest2,ni,Pk2,po)
              ^[najwyżej3,0,_,_]
          ]
          ^[obldest_iter,Dest1,Dest2,Dest]
          ^[sprawdz_pk_iter,Pk1,Pk2,Pk] ),
  ResztaWym = wym([],OW),resztawym(OW)   .
```

The above rule substitutes the definition of ZE mentioned at the beginning of this article. The number of rules was significantly reduced. More values of arguments are calculated in the process of parsing, rather than given explicitly.

2. Flat hierarchy

One of the most troublesome aspects of building the treebank is the phenomenon of tree overgeneration. Instead of tricks applied when implementing FGP for Świgra to avoid vicious recursion in defining various syntactic units FGP2 drastically flattens the hierarchy. For a given syntactic unit, all types of dependents are introduced by one rule which provides us with a « comb-like » structure instead of binary branching; in other words, head and dependents are constituents of the same level. Compare the two structures:

(ten) (nowy) (kolega) (Piotra) (z Krakowa) (FGP2)
(ten (nowy (((kolega) (Piotra)) (z Krakowa)) (FGP)
‘this new colleague of Peter from Cracow’

The « comb » philosophy was originally applied by FGP in the definition of finite sentences, and only there. Its application to other syntactic units is technically profitable as reducing overgeneration of trees for a given utterance. However, it provokes some linguistic objections since it may effect in overlooking real syntactic ambiguities.

3. Lexical background

FGP did not provide any lexical-syntactic information. Although finite sentences, same as verb(al) phrase and verbal preterminals, had valency parameters, no data of that type was stored as FGP contained lexical rules for function words only. On the contrary, Świgra, whose part FGP2 is, has got an access to the lexicon and operates with sets of valency frames for a great number of verbal units. Moreover, non-verbal syntactic units in the grammar and in the lexicon are also assigned valency information. The syntactic information of that type is indispensable for the complement vs. adjunct dichotomy purposes.

4. System-typical constructions complemented

FGP was intended to account on a majority of syntactic structures of Polish, leaving some types, typical of Polish and frequent in the corpus, untouched for various reasons. FGP2 supplements its predecessor by addition of rules defining structures that were omitted by FGP. Let us mention some of those constructions:

- coordinate phrases of all types (including exocentric nominal phrases),
- numeral-nominal phrases,
- nominal phrases with a dependent in apposition,
- prepositional-nominal phrases with dependents,
- subordinate clauses used as full utterances.

The examples below illustrate the types, respectively:

- *małemu, ale ciekawemu* 'small$_{\text{dat,sing,masc}}$ but interesting$_{\text{dat,sing,masc}}$'

  (Mam) nie tylko przyjść, ale i zostać. '(I am) not only to come, but also to stay.'

  Albo tutaj, albo w Warszawie (jest praca.) 'There are jobs (either here, or in Warsaw.'

  (Zaproś) Piotra, Marię, kuzynów, każdego! '(Do invite) Peter, Marie, cousins, everybody else!
  On i ona (spali.) // (Spał//Spali) on i ona. 'He and she were sleeping.'

  Przyszli chłopiec z dziewczyną. 'The boy with a girl have come.'
- *pięć kobiet (przyszło.) // cztery kobiety (przyszły.)* 'Five//Four women have come.'
- *Wolfgangiem Amadeuszem* 'Wolfgang Amadé$_{\text{instr}}$'
- *dokładnie o piątej* 'exactly at five'
- *Albo on przyjdzie.* 'Or perhaps he will come.'
  Ponieważ przyjdę. 'Since I'll come.'
  Że przyjdę. 'That I'll come.'

Special rules are introduced to define utterances beginning with a conjunction or a complementizer. Such utterances, quite common in corpora, are nothing but clauses cut out from sentences and used as an independent utterance.

5. Commaity

A sophisticated system of rules and parameters has been designed to cope the issue of comma placement within utterances, sentences, clauses and phrases of various types. In particular, the problem of comma haplology, i.e. of zeroing of the repeated comma in the neighbourhood of another, has been solved without introducing the empty comma, as practised in FGP. The parameter of « commaity », typical of most syntactic units, is a pair of subparameters that store the information of whether a given unit can appear in the left or right neighbourhood of either a comma (be it a punctuation mark or a conjunction), or another punctuation mark. In the sentence below:

Wiem, Mario, że jesteś tutaj. 'I know, Mary, that you are here.'

the second comma works as both the vocative phrase's obligatory separator (final) and the subordinate clause separator (initial), also obligatory. Bifunctional commas of that type are accounted for by imposing restrictions on possible combinations of the values of those subparameters.

6. Individual idiosyncrasies

The corpus analysis usually allows the researcher to discover idiosyncratic constructions. They are to be accounted for mainly by introduction of extra rules.

A distributional subclassification of particles is proposed. Sentential (general), phrasal, and adnumeral particles are distinguished, stored in the lexicon and introduced by separate rules:

> *już* 'already' – general (attachable everywhere)
> *nazbyt* 'too much', as in *On jest nazbyt mądry.* 'He is too smart.' – adadjectival (cf. **On kupił nazbyt książkę.*
> *z* 'about', as in *Będzie ich z pięciu.* 'There will be about five of them.' – adnumeral

One of the most interesting findings of the project is that of adjectival phrases which are dependents of the adjectival phrase:

> *On jest jakiś głupi.* 'He's sort of stupid.'

The constituent *jakiś*, embedded in the whole phrase, agrees with its head in case, number, and gender. Such a structure constitutes a strange type of the adjectival phrase in which one adjective inherits its grammatical characteristics from another adjective.

Conclusions

Text corpora analyses intended as a testbed of a given NLP device often result in redefinition of that device. FGP2 seems to be more adequate than FGP. Its corpus coverage looks better.

Hovever, it does not mean that the game is over. The by-product of the Treebank project, FGP2, is expected to develop independently. Importantly, the FGP2 development is corpus-prompted, but not only. Another source of further modifications is the analysis of the annotators' work, their hesitations, decisions, and objections. It would be nice to get a grammar located somewhat between the Saussurian competence account and a definition of the language understood as a set of expressions.

A number of loose ends still remain at hand. First of all, the complement vs. adjunct distinction is very difficult to decide, either theoretically or empirically. Discontinuous constructions, typical of highly inflected languages and not easy to cope with, wait for a thorough analysis. Finally, there is a great number of non-finite utterances in text corpora which badly need a rigorous description.

References

NKJP, 2008-2011, *Narodowy Korpus Języka Polskiego*, http://nkjp.pl, Warszawa.

Przepiórkowski Adam, Świdziński Marek, 1997, *Polish verbal negation revisited: a metamorphosis vs. HPSG account.* Prace IPIPAN * ICS PAS Reports, Vol. 929. Warszawa, p. 36.

SFPW, 1990, Kurcz Ida, Lewicki Andrzej, Sambor Jadwiga, Woronczak Jerzy, Szafran Krzysztof, *Słownik frekwencyjny polszczyzny współczesnej*, Kraków 1990.

Świdziński Marek, 1992, *Gramatyka formalna języka polskiego*. Wydawnistwa Uniwersytetu Warszawskiego: Warszawa, p. 420.

Świdziński Marek, 1996, *Własności składniowe wypowiedników polskich,* ELIPSA: Warszawa 1996, p. 168.

Świdziński Marek, Woliński Marcin, 2009, « A new formal definition of Polish nominal phrases », dans *Aspects of Natural Language Processing. Essays dedicated to Leonard Bolc on the Occasion of His 75th Birthday*, *LNCS 5070*, M. Marciniak, A. Mykowiecka (eds.), Springer: Berlin–Heidelberg–New York, p. 143-162.

Świdziński Marek, Woliński Marcin, 2010, « Towards a bank of constituent parse trees for Polish », dans *Text, Speech and Dialogue, 13th International Conference, TSD 2010, Brno, September 2010, Proceedings*, P. Sojka et al (eds.), *LNAI 6231*, Springer: Heidelberg, p. 197–204.

Woliński Marcin, *Komputerowa weryfikacja gramatyki Świdzińskiego*, unpublished Ph.D. dissertation, ICS, PAS, Warszawa, p. 142.

Woliński Marcin, 2006, « Morfeusz—a practical tool for the morphological analysis of Polish », dans *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, M. Kłopotek, S. Wierzchoń, K. Trojanowski (eds.), Springer: Heidelberg, p. 503–512.

Woliński Marcin, 2010, « Dendrarium—an open source tool for treebank building », dans *Intelligent Information Systems*, M. Kłopotek, M. Marciniak, A. Mykowiecka, W. Penczek, S. Wierzchoń (eds.), Siedlce, Poland, p. 193–204.

Woliński Marcin, Głowińska Katarzyna, and Świdziński Marek, 2011, « A Preliminary Version of Składnica—a Treebank of Polish », dans *Proceedings of the 5th Language & Technology Conference*, Z. Vetulani (ed.), Poznań, Poland, p. 299-303.