# An extension of Świdziński's grammar of Polish

**Maciej Ogrodniczuk**

Chair of Formal Linguistics
Faculty of Modern Languages
Warsaw University

Browarna 8/10, 00-311 Warsaw, Poland
maciej.ogrodniczuk@gmail.com

**Abstract**

The article presents a series of proposed extensions to Świdziński's Formal Grammar of Polish which were introduced in the course of automated syntax verification of a corpus of Polish expressions.

## 1. Introduction

The grammar discussed here is Świdziński's Formal Grammar of Polish[1] (further referenced here as FGP) expressed in a formalism inspired by metamorphosis grammars of Colmerauer (Colmerauer, 1978). FGP has been formulated by Marek Świdziński in his habilitation thesis, later published (with minor modifications) as (Świdziński, 1992) and since then regarded as the largest[2] and most precise formal description of general grammar of Polish.

FGP-based parsing of Polish has a long history with three major milestones resulting in working prototypes of parsers prepared under supervision of Janusz S. Bień: *AMOS-95*[3], *AS*[4] and — the most recent and the only one really usable — *Świgra*, implemented by Marcin Woliński within his doctoral dissertation (Woliński, 2004) and presented in (Woliński, 2005).

None of the previous parsing approaches did intend to modify FGP beyond making corrections necessary to improve parsing and thus they all constituted close computer representation of the original grammar. The idea put forward in Woliński's dissertation and developed in this paper goes in opposite direction: use the most current version of Świgra parsing engine as a testing environment for modifications of the grammar that might result in better understanding of its underlying constructs.

Most of the practical steps to improve FGP described here came into existence in the course of an attempt at automated processing of a treebank of Polish expressions created and marked-up with the results of manual FGP-based parsing within Świdziński's research grant[5]. Though different to Świdziński's approach of competence-based principle of building FGP, such method allowed for immediate verification of proposed changes, therefore all quoted examples of Polish sentences are now parseable with modified version of FGP.

## 2. Elimination of redundant cycles

### 2.1. Recurrence in FGP

FGP capability of representing potentially infinite levels of substructures in parsed sentences is expressed in a peculiar way, which should be regarded as a consequence of Świdziński's conviction[6]:

> (...) to maintain recurrence it is necessary to allow for a lowest-level (simplest) syntax unit to be expressed as a highest-level syntax unit.

As a result[7], FGP contains five cycles, each defined as a loop of clauses containing single non-terminal in clause head and clause body.

For example, the rule

$$\begin{aligned} &zsz(Wf, A, C, T, Rl, O, Neg, I, Z) \\ &\rightarrow s(s1), \\ &\quad zj(Wf, A, C, T, Rl, O, Neg, I, Z, Oz), \\ &\quad rozne(Oz, lub). \end{aligned}$$

is one of the elements of the cycle

$$zr \rightarrow zsz \rightarrow zj \rightarrow zp \rightarrow ze \rightarrow zr$$

Please note that different elements of the cycle might have different conditions attached to the rules (in the example above the condition states that `lub` is not allowed as the value of `Oz`).

Other cycles, all of them identified in Woliński's dissertation, are:

[1] Polish: *Gramatyka Formalna Języka Polskiego*.

[2] It contains 463 rules as compared to 780 rules in Alvey grammar — the English grammar developed within the Alvey Natural Language Tools (Grover et al., 1993), probably the largest existing natural language grammar in Prolog.

[3] See (Bień, 1996a) and (Bień, 1996b).

[4] See (Bień et al., 1999) and (Bień, 2000).

[5] The corpus comprises 5452 expressions corresponding to FGP sentences; each of them contains detailed phrase-level information about its components; see (Świdziński, 1996) for details.

[6] See (Świdziński, 1992), p. 59.

[7] However this idea seems unnatural, it can be justified by the spirit of FGP which means for me the attempt to apply similar mechanisms to language constructs at different — sentence and phrase — levels.

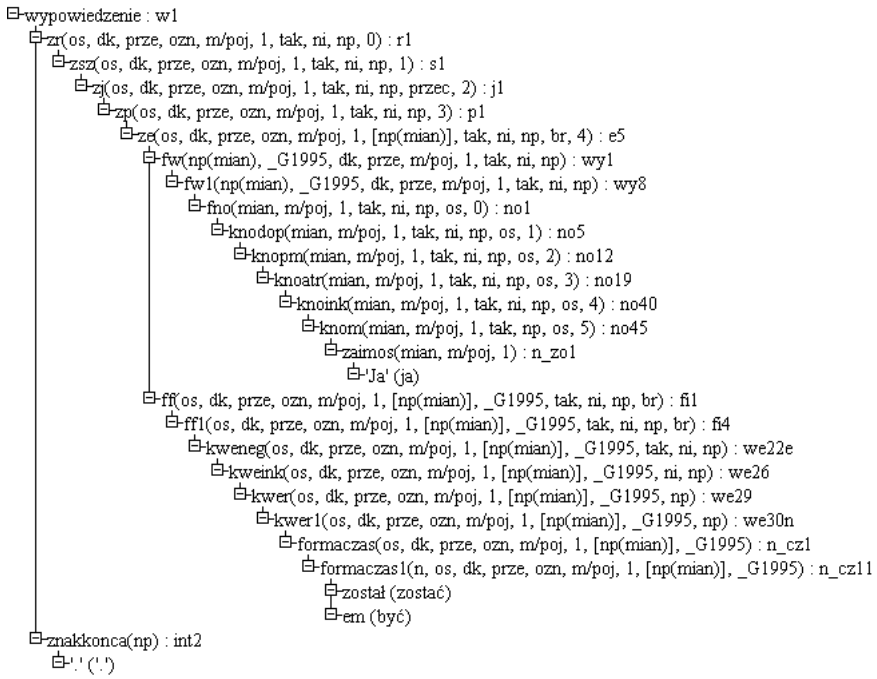


Figure 1. A result parsing tree generated with the original grammar

$$fno \rightarrow knodop \rightarrow knopm \rightarrow knoatr \rightarrow knoink \rightarrow knom \rightarrow fno$$

$$fps \rightarrow kpspm \rightarrow kpsps \rightarrow kpsink \rightarrow kprzysl \rightarrow fps$$

$$fpt \rightarrow kptno \rightarrow kptpm \rightarrow kptps \rightarrow kptink \rightarrow kprzym \rightarrow fpt$$

$$fzd \rightarrow fzdsz \rightarrow fzdj \rightarrow fzdkor \rightarrow fzd$$

Such approach causes substantial problems for implementators of FGP parsers because presence of single non-terminal on both sides of the cycled clauses creates idle loops which result in infinite number of unproductive analyses (for a sentence analysed with a clause taking part in the cycle) containing repeated cycle path.

To minimize interference with the grammar but eliminate idle cycles, Woliński blocks them by introducing additional parameter for each of the cycled clauses which controls cycle length to avoid idle paths while parsing — without having to restructure the grammar.

The opposite idea is to test Woliński's assumption that all non-terminals involved in each individual cycle are distributionally equivalent and thus they can be replaced with (or merged into) single non-terminal.

### 2.2. Merging non-terminals

At first, one common non-terminal has been introduced for each cycle to replace all cycled non-terminals. For instance, one single non-terminal `zd` has been introduced to replace all sentence-level non-terminals (`zr` $\rightarrow \ldots \rightarrow$ `ze`) etc. After applying this change, cycled clauses could have been (in most cases) turned off since clause head and clause body carried the same non-terminal with equivalent parameters.

One of the most apparent results of this procedure is flattening the parsing trees — e.g. all types of sentences (from complex sentences with co-ordinate clauses to elementary sentence) are now expressed with a single non-terminal `zd` which, if applicable, enables the parser to reach the phrase level in one step.

During the process of rewriting the grammar it appeared that for some clauses additional conditions (added for blocking certain parsing paths, e.g. to disallow occurrences of *lub* conjunction at the beginning of a sentence) make the translation of grammar rules far more than a trivial task. Consequently, an attempt has been made to apply all the conditions to the new grammar by adopting such techniques as shifting them to a higher-level clause, redistributing them over equivalent other clauses or applying additional conditions. In a few cases, mutually exclusive conditions appeared in two separate clauses which could then be merged into single one not carrying any conditions.

The detailed test results from elimination of redundant cycles and adding the nominal group definition (described below) were published in (Ogrodniczuk, 2005).

## 3. Adding syntax group definitions to FGP

Another modification extends FGP definition of nominal phrase with formal description of co-ordinate nominal group adapted from (Szpakowicz and Świdziński, 1983) to make possible parsing of sentences with complex nominal expressions.

Definition of nominal phrase in FGP, however sophisticated, is not intended to cover co-ordinate relationship between nominal constructs, not even as simple as in *kot i pies* (two simple nominal phrases joined with co-ordinate conjunction). However, a formal description of such constructs is contained in above-mentioned article, where the hierarchy of co-ordinate nominal groups is expressed using similar formalism as in (Świdziński, 1992)

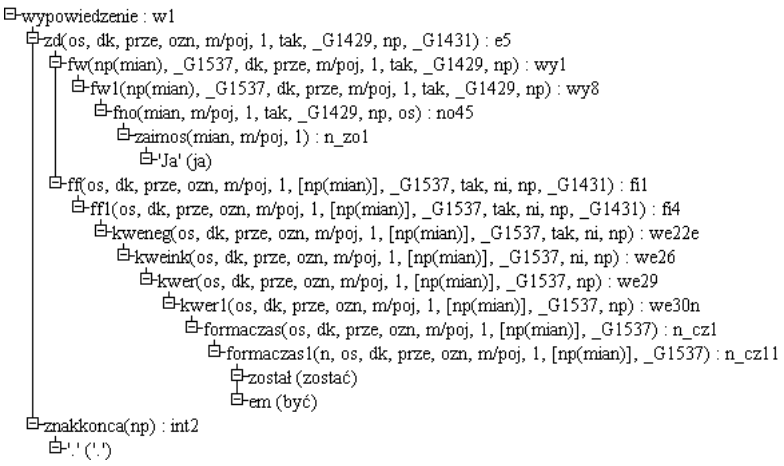


Figure 2. Flattened parsing tree: the grammar without cycles

which facilitates their direct inclusion in FGP:

| | |
|---|---|
| `rgn` | co-ordinate nominal group |
| `sgn` | sequential nominal group |
| `jgn` | homogeneous nominal group |
| `pgn` | single nominal group |

To test this definition, it has been included into FGP and the general test suite for nominal groups has been prepared. Two differences have been made to the description from Szpakowicz and Świdziński's article: the ellipsis parameter (present in the description „for future extension of the definition") has been omitted and the values of class parameter (introduced to group together nominal constructs representing the same grammatical class) have been limited to negative pronoun[8]. The definition has been embedded in FGP by replacing all occurrences of nominal phrase (`fno`) in FGP clause bodies with newly defined general nominal group (`gno`), which in turn included definition of the highest-level unit — co-ordinate nominal group (in only clause).

Single nominal group, left in the article for future specification, has been defined as FGP nominal phrase (although it should be noted that its description is significantly simpler than authors' intentions for single nominal group). To make both descriptions compatible, existing nominal phrase definition had to be extended with additional class and selective negation parameters.

The following extension has been made to the original definition of the nominal group: contrary to *bądź* conjunction, the article does not mention sequential constructs with *czy* (functional *or*, as in *„Inni chwalą się przed zdumioną publiką przebijaniem policzków czy rąk."* `[3949]`[9]). To allow for parsing them, I extended the body of the rule defining sequential conjunctions of particular type (`5`) with *czy*.

FGP extended in this way can be now used to parse real-life nominal expressions such as quite common compound subjects.

### 3.1. The position parameter

One of the most interesting components of the nominal group description is by far the position parameter. It is defined as the place of nominative nominal group with respect to verbal group in the sentence and can be expressed as `post`, `pre` or `undefined`.

According to relationships between position, gender and number the sentence containing such nominal group is regarded correct by the authors:

1. when gender and number of the verbal group agree with the parameters of the nominal group as a whole[10] — in such case the nominal group can occur before or after the verbal group:

   masculine and plural, `undefined` position
   *(Nie przyszli) ani dziecko, ani ojciec.*
   *Ani dziecko, ani ojciec (nie przyszli).*

2. when gender and number of the verbal group agree with the parameters of the first component of the nominal group — in such case the nominal group can occur only after the verbal group:

   neuter and singular, `post` position
   *(Nie przyszło) ani dziecko, ani ojciec.*

3. when gender and number of the verbal group agree with the parameters of the last component of the nominal group — in such case the nominal group can occur only before the verbal group:

   masculine and singular, `pre` position
   *Ani dziecko, ani ojciec (nie przyszedł).*

[8]The original description contains five other classes; four of them had been left for future development by article authors while the remaining one — genitive numeral, referenced explicitly, had to be removed from the grammar since current version of FGP lacks formal description of Polish numerals.

[9]Record `3949` from the corpus described in (Świdziński, 1996); I will use this notation throughout the document to identify the source of examples that initiated extension of the grammar.

[10]This implies plural number; gender category is calculated using separate rules, the main of which is domination of masculine gender, see (Kallas, 1976; Świdziński, 1978) for details.

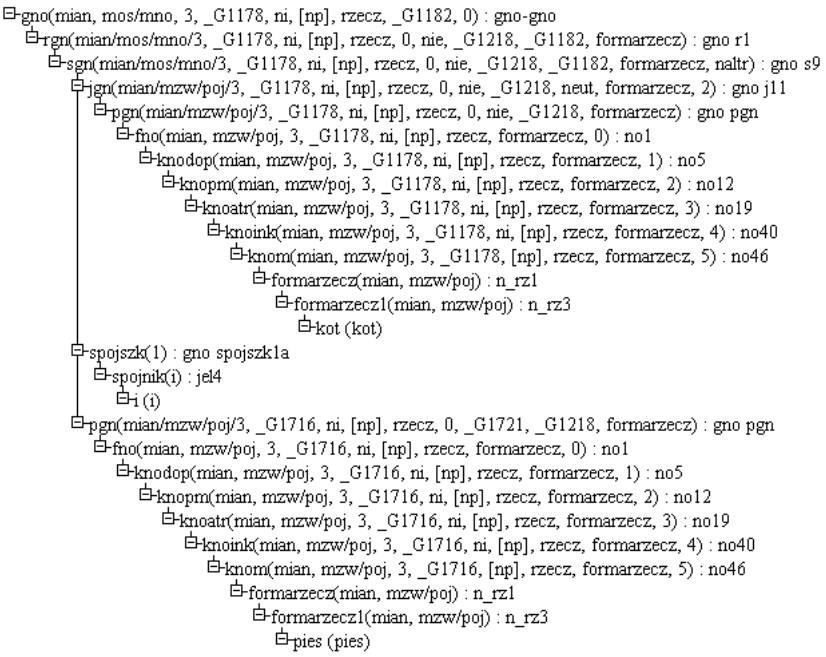


Figure 3. Fragment of a parsing tree with nominal group

Nevertheless, certain sentences formed according to these rules seem contrary to linguistic intuition (e.g. *Powiedziała mu ona i on.*)[11]. Basing on this ambiguity, unorthodox decision has been made to make use of the parameter outside the grammar. The underlying idea is to mark the result parsing trees containing wrong position as incorrect without changing the grammar significantly (since the article deals only with the calculation of position parameter, not with how it further intervenes with the verbal group).

Elimination of the parsing trees with incompatible position, gender and number has been achieved by changing the way result trees are post-processed in the parser. The tree fragments containing calculated position of the verbal and nominal groups inconsistent with the position parameter resulting from nominal group definition are removed from the result set and are therefore blocked to be used as valid results.

### 3.2. Other syntax groups

Basing on the idea described above three other types of grouping constructs were introduced to FGP: groups of adjectives, adverbs and nominal phrases with prepositions.

Here are the example sentences which are now parseable with FGP:

- adjectival group: *„Jestem kobietą czułą i łagodną."* `[1367]`,
- adverbial group: *„Tu i ówdzie legły, wiły się ciała."* `[5618]`,
- nominal-prepositional group: *„Guanajuato powstało ze srebra i dla srebra."* `[3595]`,
- mixed adverbial and nominal-prepositional group: *„Przyjeżdżali konno lub w powozach."* `[3670]`.

## 4. Modification of four negative constructs

Different types of commonly used negative constructs seems to have been neglected in FGP. Analysis of corpus expressions resulted in slight modification of the grammar on many different levels, starting from correction of the agreement of negation parameters to adding particular types of negated language constructs.

### 4.1. Negated verb and infinitive

Failure to parse with FGP sentences containing negated verb and required sentence phrase built around infinitive such as *„Nie mogę spać."* `[2000]`, *„Czy nie powinieneś się poradzić lekarza?"* `[2887]` or *„Dlaczego nie pozwala mi spokojnie odpocząć?"* `[4703]` required analysis of the negation agreement between verbal phrase and required phrase. Świdziński writes[12]:

> Among (...) thirteen parameters of elementary sentence five agrees with all constituent phrases, namely aspect, tense, gender-number, person and negation.

which is true for all requirements except for infinitive and sentence phrase. As for sentence phrase, the lack of agreement has already been included in FGP[13] while the rule for verbal phrase in the infinitive had to be corrected to unbind negation parameters.

### 4.2. Negated conditional

The number of sentences from the processed corpus contained broken conditional forms (as in *„Ty byś nie pił."*

[11] An ordinary „language user" could say that the more distant the verb in the sentence is from the subject, the more likely the sentence seems correct, but it definitely require more studies.

[12] (Świdziński, 1992), section 6.2.1.

[13] See rule `wy19`.

[3270]) which by definition are turned down by FGP as uncontinuous.

General solution to this common problem in not trivial since expression between conditional agglutinant *byś* and pseudoparticle *pił* can have arbitrary structure. Woliński presents a simple solution that restricts FGP to constructs when conditional agglutinant appears in direct surrounding of pseudoparticle form. However, his definition allows for parsing sentences with phrases such as *„jedlibyście"*, *„byście jedli"* and *„nie jedlibyście"*, yet it is impossible to accept negated phrase *„byście nie jedli"*.

Simple extension of this definition makes use of the mechanism developed for *się* pronoun appearing between conditional agglutinant and pseudoparticle — a variant of a grammar rule has been introduced to allow for parsing of constructs containing *nie* particle directly before pseudoparticle form.

### 4.3. Negated prepositional phrase

Prepositional phrases can also carry negation particle directly before preposition and yet distributional context of such modified construct remains unchanged. FGP does not contain mechanisms for parsing negated prepositional phrases (as in *„Nie o to chodzi."* [2950] or *„Gadali o Broni i nie o Broni."* [2956]), so it has been extended with a simple rule introducing a variant of prepositional phrase carrying negation particle before preposition and the nominal group.

### 4.4. Negated superlative

In definition of adverbial phrase FGP limits the use of negation to constructs with negative pronoun, which rejects sentences containing negated superlative (as in *„Poczynasz sobie nie najgorzej, chłopcze."* [2053]). To allow for using them, a new rule for adverbial constructs containing negated superlative adverbs has been added.

## 5. Other minor extensions

Several minor extensions have been included in the grammar to allow for parsing groups of undoubtedly correct sentences which were not accepted by FGP. Below I present four examples of such changes that illustrate the nature of modifications:

- compound subordinate conjunctions — to allow for parsing sentences such as *„Będą mogły być krzywoliniowe, a więc będą dyskami."* [3868] containing compound conjunction *a więc* new rule has been added to FGP; it allows for using compound conjunction in place of „standard" conjunction of *więc*-type group (*więc*, *zatem* and *przeto*), but apart from its incorporating context,
- gerund with *się* — constructs containing gerund forms with reflexive pronoun (as in *„Obrazują one zachowanie się organizmu kosmonauty."* [3563]) are now parsed with a new rule for basic nouns with *się*,
- adverbial modifier with *po* — constructs built with a preposition and particular form of an adverb (as *„Mówimy po polsku.* [3159], *„Po prostu nawaliła winda."* [1985]) are treated as compound adverbs,
- free phrase sequences — parsing sentences containing comma-separated sequences of free phrases such as *„Wydostali się w dolinę, na uprawne suche pola."* [4902] requires extension of free phrase definition since FGP limits the list of realizations of such phrase to certain types[14]; the extension adds new rule for free phrases that allow for stacking them in sequences.

## 6. Conclusions and perspectives

The most evident benefit resulting from the changes in FGP (although they should only be considered as the training ground for further modifications) is noticeable increase of parseable sentences in the processed corpus — which should result in broader acceptance of real-life sentences. Apart from ongoing work concerning serious extension of the scope of the grammar such as integration of the description of numerals[15]), such minor changes are of similar importance since they improve the power of expression of FGP with language facts far less obvious than essential syntax properties. Both levels of changes do not seem possible without using corpora, which proves that the approach of verification of FGP against small, but well-annotated corpus was a good starting point for similar experiments.

Another direction of work should definitely be improvement of parsing environment to include representation of phraseological properties of Polish. At first glance this does not seem possible without rebuilding the grammar to make use of such information, but the issue definitely requires more studies.

## 7. References


Bień, Janusz S., 1996a. *Computer validation of a description of Polish syntax*. [In Polish]. TR 96-06 (227), Institute of Informatics, Warsaw University.

Bień, Janusz S., 1996b. *Computer validation of Świdziński's formal grammar*. [In Polish]. PTJ Bulletin, vol. LII, pp. 147–164.

Bień, Janusz S., 2000. *Test suite for verification and assessment of parsers of Polish. Final (slightly modified) report of KBN 8 T11C C 002 13 grant*. [In Polish]. Electronic version: ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/tajp/.

Bień, Janusz S., K. Szafran, and M. Woliński, 1999. *Experimental Parsers of Polish*. [in:] *3. Europäische Konferenz "Formale Beschreibung slavischer Sprachen, Leipzig 1999"*, vol. 75 of *Linguistische ArbeitsBerichte,* series, pp. 185–190, Institut für Linguistik, Universität Leipzig.

Colmerauer, A., 1978. *Metamorphosis grammar*. [in:] L. Bolc (ed.), *Natural Language Communication with Computers*, Lecture Notes in Computer Science 63, pp. 133–189, Springer-Verlag.


[14] See (Świdziński, 1992), section 6.5.2.

[15] See e.g. (Gruszczyński and Saloni, 1978) and (Derwojedowa et al., 2003).

Derwojedowa, M., M. Rudolf, and M. Świdziński, 2003. *Two formal approaches to Polish numeral phrases implemented*. [in:] *Studia z gramatyki i leksykologii języka polskiego*, pp. 93–108, Uniwersytet Mikołaja Kopernika, Toruń.

Grover, C., J. Carroll, and E. Briscoe, 1993. *The Alvey Natural Language Tools grammar*. (4th release). Computer Laboratory, Cambridge University, UK, Technical Report 284. Electronic version: `ftp://ftp.cl.cam.ac.uk/nltools/reports/grammar.ps`.

Gruszczyński, W. and Z. Saloni, 1978. *Syntax of numeral phrases in modern Polish*. [in:] Studia gramatyczne, Vol. II. Wrocław, pp. 17–42.

Kallas, K., 1976. *On sentences such as* Pachniał wiatr i morze, Andrzej i Amelia milczeli. [In Polish]. Studia z filologii polskiej i słowiańskiej, vol. 4, Warsaw.

Ogrodniczuk, M., 2005. *Restructuring Świdziński's grammar of Polish*. [w:] Proceedings of 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Wydawnictwo Poznańskie i Fundacja Uniwersytetu im. A. Mickiewicza. Poznań 2005, s. 177-181.

Szpakowicz, S. and M. Świdziński, 1983. *Formal definition of co-ordinate nominal group in present written Polish*. [In Polish]. Studia gramatyczne IX, ISBN 83-04-03303-8.

Świdziński, M., 1978. *Sample transformations in Polish*. [In Polish]. "Polonica" IV.

Świdziński, M., 1992. *Formal grammar of Polish*. [In Polish]. Warsaw University Dissertations, Warsaw.

Świdziński, M., 1996. *Syntax properties of Polish expressions*. [In Polish]. Institute of Polish Philology, Warsaw University, Warsaw.

Woliński, M., 2004. *Computer-aided verification of Świdziński's grammar*. [In Polish]. Ph.D. thesis. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Woliński, M., 2005. *An efficient implementation of a large grammar of Polish*. [w:] Proceedings of 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Wydawnictwo Poznańskie i Fundacja Uniwersytetu im. A. Mickiewicza. Poznań 2005, s. 177-181.