Polish Lexical Analyser
Version 1.1

Stanisław Galus

1 November 2008

1 Running the program

PLA performs lexical analysis of Polish text and compiles dictionaries used during analysis. The command line for PLA is

pla analyse [word–file]... input–file output–file
pla compile source–word–file word–file
pla decompile word–file source–word–file
pla list–symbols

The first form performs lexical analysis of *input–file* using given dictionaries *word–files* and sending output to *output–file*. The second form compiles text form of the dictionary *source–word–file* into its binary form *word–file*. The third form decompiles binary form of the dictionary *word–file* to its text form *source–word–file*. The fourth form lists the set of all lexical symbols. For example, the command

pla compile basic.swf basic.wf

compiles the source word file *basic.swf* into dictionary *basic.wf*, and the command

pla analyse basic.wf wojski.txt wojski.out

analyses text from the file *wojski.txt* using the dictionary *basic.wf* and sends the result to the file *wojski.out*.

Program exit code 0 is returned on success. In case of a failure, 1 is returned and an error message is printed on the standard error. If a word file is seriously corrupted, the program may be aborted.

2 The set of lexical symbols.

The basic subset of lexical symbols used by the program is outlined in subsection 2.1, which is an excerpt from *Parsing Polish as a Context-Free Language*

(see Kłopotek M. A., Wierzchoń S. T., Trojanowski K. (editors), *Intelligent Information Processing and Web Mining. Proceedings of the International IIS: IIPWM'06 Conference*, Springer-Verlag, Berlin 2006, p. 329–333). The list of all lexical symbols is described in subsection 2.2.

2.1 The basic subset of lexical symbols

We use the following grammatical categories:

case $C = \{nom, gen, dat, acc, instr, loc, voc\}$,

number $N = \{sg, pl\}$,

gender $G = \{m, m1, m2, m3, f, n, m-3, m-1, -m1, -f\}$,

person $P = \{1, 2, 3\}$,

degree $D = \{pos, comp, sup\}$,

aspect $A = \{impf, pf\}$.

The set of cases has seven members: nominative, genitive, dative, accusative, instrumental, locative, vocative. There are two numbers: singular and plural. The genders are masculine, masculine-personal, masculine-animal, masculine-inanimate, feminine, neuter, masculine-animate, masculine-impersonal, non-masculine-personal and non-feminine genders, respectively. There are three persons: first, second and third. The degrees are positive, comparative and superlative. The two aspects are imperfect and perfect.

We define the following inflexional patterns:

substantival declension $Ps = N \times C$,

adjectival declension $Pa = \{sg\} \times \{m-3, m3, f, n\} \times C \cup \{pl\} \times \{m1, -m1\} \times C$,

numeral declension $Pn = \{m1, -m1\} \times C$,

declension of numeral 1 $P1 = \{m-3, m3, f, n\} \times C$,

declension of numeral 2 $P2 = \{m-1, m1, f, n\} \times C$,

singular declension by case $Pc = C$,

singular declension by gender $Pg = \{-f, f\}$,

conjugation $Pj = \{sg\} \times \{m, f, n\} \times P \cup \{pl\} \times \{m1, -m1\} \times P$,

genderless conjugation $Pk = N \times P$,

simple forms of imperative $Pi = \{sg\} \times \{2\} \cup \{pl\} \times \{1, 2\}$.

Pronouns fall into seven types: personal *pers*, reflexive *refl*, indefinite *ind*, negative *neg*, demonstrative *dem*, possessive *poss*, interrogative-relative *intr* pronouns. Numerals can be arranged into four types: cardinal *card*, collective *coll*, fractional *frac* and ordinal *ord* numerals. The seven forms of verbs are infinitive *inf*, present for imperfect verbs, future for perfect verbs *simp*, past tense *past*, imperative mood *imp*, subjunctive mood *subj*,

impersonal form of past tense *ifpt* and simultaneous adverbial participle for imperfect verbs, anticipatory adverbial participle for perfect verbs *advp*.

The set of genders which may be assigned to a noun may be defined as $GN = \{0, m1, m2, m3, f, n\}$, where *0* denotes no gender. Types of pronouns inflected by a pattern are denoted as: $T0 = \{ind, neg, dem, intr\}$, $TPs = \{pers\}$, $TPa = \{pers, ind, neg, dem, poss, intr\}$, $TPn = \{ind, dem, intr\}$, $TPc = \{refl, ind, neg, intr\}$. We also define the set of punctuation marks $PM = \{., ;, ,, :, –, \ldots, ?, !, (,), [,], /, \{, \}, <, >, „, ”, ≪, ≫, ‘, ’, "\}$ and the set of graphic signs $GS = \{=, +, \%, §, \#, \$, \&, *, @, \backslash, \hat{}, \_, |, \tilde{}\}$.

The set of lexical symbols may be classified as follows.

$NOUN(g, n, c)$	$(n, c) \in Ps$	noun of gender $g \in GN$	(84)
$ADJ(d, n, g, c)$	$(n, g, c) \in Pa$	adjective of degree $d \in D$	(126)
$PRON(t, 0)$		pronoun, $t \in T0$	(4)
$PRON(t, Ps, n, c)$	$(n, c) \in Ps$	pronoun, $t \in TPs$	(14)
$PRON(t, Pa, n, g, c)$	$(n, g, c) \in Pa$	pronoun, $t \in TPa$	(252)
$PRON(t, Pn, g, c)$	$(g, c) \in Pn$	pronoun, $t \in TPn$	(42)
$PRON(t, Pc, c)$	$c \in Pc$	pronoun, $t \in TPc$	(28)
$NUM(card, Ps, g, n, c)$	$(n, c) \in Ps$	numeral, *card*, $g \in \{m3, n\}$	(28)
$NUM(card, Pn, g, c)$	$(g, c) \in Pn$	numeral, *card*	(14)
$NUM(card, P1, g, c)$	$(g, c) \in P1$	numeral, *card*	(28)
$NUM(card, P2, g, c)$	$(g, c) \in P2$	numeral, *card*	(28)
$NUM(coll, c)$	$c \in Pc$	numeral, *coll*	(7)
$NUM(frac, 0)$		numeral, *frac*	(1)
$NUM(frac, g)$	$g \in Pg$	numeral, *frac*	(2)
$NUM(ord, n, g, c)$	$(n, g, c) \in Pa$	numeral, *ord*	(42)
$V(a, inf)$		verb of aspect $a \in A$, *inf*	(2)
$V(a, simp, n, p)$	$(n, p) \in Pk$	verb of aspect $a \in A$, *simp*	(12)
$V(a, past, n, g, p)$	$(n, g, p) \in Pj$	verb of aspect $a \in A$, *past*	(30)
$V(a, imp, n, p)$	$(n, p) \in Pi$	verb of aspect $a \in A$, *imp*	(6)
$V(a, subj, n, g, p)$	$(n, g, p) \in Pj$	verb of aspect $a \in A$, *subj*	(30)
$V(a, ifpt)$		verb of aspect $a \in A$, *ifpt*	(2)
$V(a, advp)$		verb of aspect $a \in A$, *advp*	(2)
$PASSP(n, g, c)$	$(n, g, c) \in Pa$	adjectival passive participle	(42)
$ACTP(n, g, c)$	$(n, g, c) \in Pa$	adjectival active participle	(42)
$VN(n, c)$	$(n, c) \in Ps$	verbal noun	(14)
$PASTP(n, g, c)$	$(n, g, c) \in Pa$	adjectival past participle	(42)
$ADV(d)$		adverb of degree $d \in D$	(3)
PREP		preposition	(1)
CONJ		conjunction	(1)
PART		particle	(1)
INTERJ		interjection	(1)
NONE		word of none part of speech	(1)
UNWORD		unknown word	(1)
$PMARK(m)$	$m \in PM$	punctuation mark	(24)
HYPHEN		hyphen	(1)
NUMBER		a sequence of digits	(1)
$GSIGN(s)$	$s \in GS$	graphic sign	(14)
UNCHAR		unknown character	(1)

The number of lexical symbols presented in each row is given in parentheses. The total number of symbols is 974. However, since nominative and vocative of reflexive pronouns do not exist, the tags $PRON(refl, Pc, nom)$ and $PRON(refl, Pc, voc)$ are not used. All other tags are non-empty.

2.2 The full list of lexical symbols

The full list contains 8670 lexical symbols. The first 972 symbols are those named in subsection 2.1. The remaining symbols are their variants with either the particles *-że*, *-ż* or the endings *-(e)m*, *-(e)ś*, *-(e)śmy*, *-(e)ście* or both added (see *Nowy słownik ortograficzny PWN wraz z zasadami pisowni i interpunkcji*, PWN, Warszawa 1996, p. LVI–LVII). The numbers of symbols in the main classes of lexical symbols are as follows.

Class	Basic	Particle *-że*	Endings *-(e)m* etc.	Total
NOUN	84	84	672	840
ADJ	126	126	1008	1260
PRON	338	338	2704	3380
NUM	150	150	1200	1500
V	84	84	–	168
PASSP	42	42	336	420
ACTP	42	42	336	420
VN	14	14	112	140
PASTP	42	42	336	420
ADV	3	3	24	30
PREP	1	1	8	10
CONJ	1	1	8	10
PART	1	1	8	10
INTERJ	1	1	8	10
NONE	1	1	8	10
UNWORD	1	–	–	1
PMARK	24	–	–	24
HYPHEN	1	–	–	1
NUMBER	1	–	–	1
GSIGN	14	–	–	14
UNCHAR	1	–	–	1
Total	972	930	6768	8670

For example, the positive adverb of the basic form $ADV(pos)$ *daleko*, has the following variants:

- $ADV(pos, ze)$ *dalekoż*,
- $ADV(pos, 1sg)$ *dalekom*,
- $ADV(pos, 2sg)$ *dalekoś*,
- $ADV(pos, 1pl)$ *dalekośmy*,

- *ADV*(*pos*, *2pl*) *dalekoście*,
- *ADV*(*pos*, *1sg*, *ze*) *dalekożem*, *dalekomże*,
- *ADV*(*pos*, *2sg*, *ze*) *dalekożeś*, *dalekośże*,
- *ADV*(*pos*, *1pl*, *ze*) *dalekożeśmy*, *dalekośmyż*,
- *ADV*(*pos*, *2pl*, *ze*) *dalekożeście*, *dalekoścież*.

The manner in which the notation of lexical symbols is broadened is clear. The total number of lexical symbols can also be expressed as

$$972 + 930 + (930 - 84) \times 8 = 8670.$$

3 Preparing input files for analysis

The input file for analysis should be prepared as a text file using the following characters:

lower case letters abcdefghijklmnopqrstuvwxyz
ąćęłńóśźż
ľšşťžŕáâăäĺçčéëěíîďðňôőöřůúűüýţß

upper case letters ABCDEFGHIJKLMNOPQRSTUVWXYZ
ĄĆĘŁŃÓŚŹŻ
ĽŠŞŤŽŔÁÂĂÄĹÇČÉËĚÍÎĎĐŇÔŐÖŘŮÚŰÜÝŢ

digits 0123456789

hyphen -

punctuation marks .:,;–- . . . ?!()[]/⟨⟩„”«»‘’'"

graphic signs =+%§#$&*@\ˆ_|˜

white space space, horizontal tab, line feed, vertical tab, form feed, carriage return

All other characters are assigned the symbol *UNCHAR*. The above characters should be represented with the ISO_8859–2:1987 character encoding. Missing characters should be represented as follows:

–	--	(two hyphens)
. . .	. . .	(three dots)
⟨	<	(less than sign)
⟩	>	(greater than sign)
„	,,	(two commas)
”	''	(two apostrophes)
«	<<	(two less than signs)
»	>>	(two greater than signs)

4 Preparing dictionaries

A source word file is a sequence of entries. Each entry may be preceded by a number of blank lines and must be followed by one or more blank lines or the end of the file. An entry consists of a heading followed by a number of forms, each of which occupies one line. If a form of an entry has an alternative, the entry should be compiled as two entries. Missing forms should be denoted by single hyphens. Each entry may be preceded by a number of comments and each line of an entry may be closed with a comment. A comment starts with a hash and goes up to the end of the line.

The heading determines kind of the entry. The numbers of forms of entries of each kind as well as the order and meaning of forms are fully explained in a template source word file *template.swf*.

5 Output files from analysis

An output file from lexical analysis is a text file which lists all alternatives for each lexical symbol found. Punctuation marks, hyphens, graphic signs, numbers, words of none part of speech, unknown words and unknown characters always have exactly one explanation. The remaining symbols can have any number of explanations. Each lexical symbol is printed on the first position on its own line. An unknown character is not printed at all and its line remains empty. This line is followed by one or more explaining lines, each of which starts on the sixth column following five spaces. On these lines, an assigned lexical symbol (which never contains white space) is printed and if it is not a *PMARK*, *HYPHEN*, *GSIGN*, *NUMBER*, *NONE*, *UNWORD*, or a *UNCHAR*, the main form is printed after one space until the end of the line.

6 Additional files

The program is accompanied by three additional files. The file *wojski.txt* is an example of input file for analysis. The file *template.swf* is a template for source form of dictionaries, containing all possible kind of entries explained.

The file *basic.swf* is the source of the main dictionary, containing well over 120000 entries. To build this file, I initially used the list of words for the program Ispell, compiled by Piotr Gackiewicz, Włodzimierz Macewicz and Mirosław Prywata in 2000. Since then, *Słownik języka polskiego PWN* and *Nowy słownik poprawnej polszczyzny PWN* became the main source.

7 Bugs

The following numerals: *zero*, *tysiąc*, *milion*, *miliard*, *bilion*, *trylion*, *kwadrylion*, *kwintylion*, *sekstylion*, *septylion*, *oktylion*, *nonylion* have been compiled as if they were inflected according to substantival declension, possessing forms of both singular and plural. These numerals have also been compiled as nouns. The class of cardinal numerals inflected substantivally $NUM(card, Ps, g, n, c)$ seems to be a mistake (cf. Alicja Nagórko, *Zarys gramatyki polskiej (ze słowotwórstwem)*, PWN, Warszawa 2000, p. 153).

Verbal nouns have been most often compiled both in singular and plural to avoid lengthy semantic considerations.

Within the scope of accepted classification, the fraction of incorrect entries in the main dictionary is so small that it is difficult to estimate.