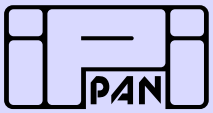
The tools for morphological description of Polish developed & used at IPI PAN

Marcin Woliński



INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. J. K. Ordona 21, 01-237 Warszawa

IPI PAN · April 16, 2007

Outline

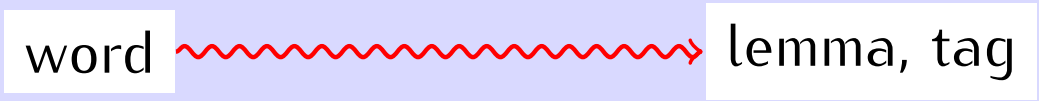
1. Morfeusz – a morphological analyser for Polish
2. Dictionary of proper names developed at IPI PAN
3. Development plans for Morfeusz

The task of morphological analysis

Morphological analysis — interpretation of words as grammatical forms.

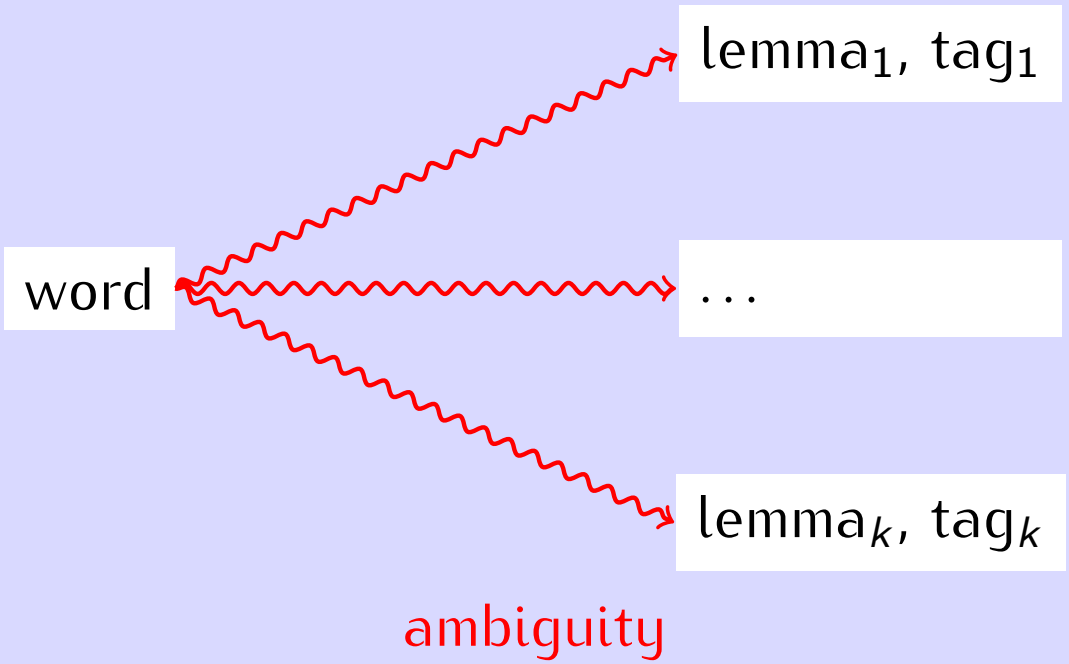
The task of morphological analysis

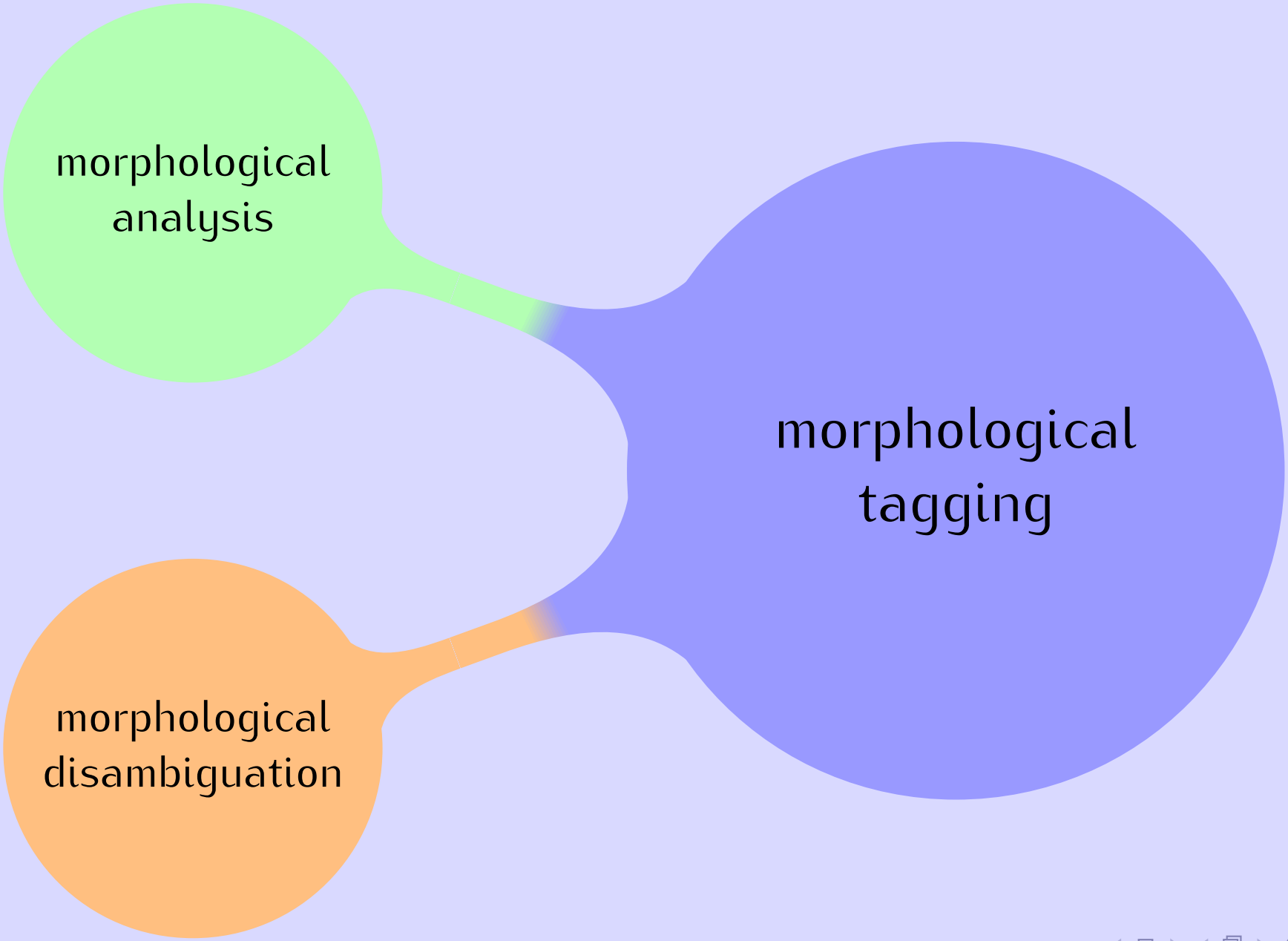
Morphological analysis — interpretation of words as grammatical forms.



The task of morphological analysis

Morphological analysis — interpretation of words as grammatical forms.




morphological analysis
morphological tagging
morphological disambiguation

Morfeusz

Morfeusz is a morphological analyser for Polish developed by Marcin Woliński using linguistic data provided by prof. Zygmunt Saloni.

Morfeusz is available for free for non-commercial and research purposes.

`http://nlp.ipipan.waw.pl/~wolinski/morfeusz/`

The *Morfeusz* library

- *Morfeusz* is provided as Linux shared object file (`.so`) and MS Windows dynamic link library (`.dll`).
- Simple programming interface: a function that takes a piece of text as an argument and returns a table of interpretations.
- The programming interface is in C for portability between compilers.
- Additional modules for using *Morfeusz* in programs written in Perl, SWI Prolog, SICStus Prolog, Java (by Dawid Weiss), and PHP (by Piotr Wilkin).
- Input/output possible in UTF-8, ISO 8859-2, CP1250, and CP852.

Tokarski's description of Polish inflection

Jan Tokarski

Schematyczny indeks a tergo polskich form wyrazowych,
edited by Zygmunt Saloni,
second edition, PWN, Warszawa 2002.

Tokarski's description of Polish inflection

Jan Tokarski

Schematyczny indeks a tergo polskich form wyrazowych,
edited by Zygmunt Saloni,
second edition, PWN, Warszawa 2002.

Sample lemmatisation rules:

-kście	*mIV LV*	-kst	kontekście, tekście, mikście	(6)
-kście	*żIV D*	-ksta	sekście	

An example analysis

Lemmatisation rules matching the word **czystkom**:

-kom	*mIII lD*	-ek	czystek
-kom	*mIII lD*	-k	czystk
-kom	*mIV N*	-kom	czystkom
-kom	*nII lD*	-ko	czystko
-kom	*nVI lD*	-kum	czystkum
-kom	*żIII lD*	-ka	czystka
-kom	*żIII(m) lD*	-ko	czystko
kom	*żIV lG*	koma	czystkoma
-kom	*blp lD*	-ka	czystka
-kom	*blp lD*	-ki	czystki
-kom	*Vla i*	-komić	czystkomić

An example analysis

Lemmatisation rules matching the word **czystkom**:

-kom	*mIII lD*	-ek	czystek
-kom	*mIII lD*	-k	czystk
-kom	*mIV N*	-kom	czystkom
-kom	*nII lD*	-ko	czystko
-kom	*nVI lD*	-kum	czystkum
-kom	*żIII lD*	-ka	czystka
-kom	*żIII(m) lD*	-ko	czystko
kom	*żIV lG*	koma	czystkoma
-kom	*blp lD*	-ka	czystka
-kom	*blp lD*	-ki	czystki
-kom	*Vla i*	-komić	czystkomić

An example analysis

Lemmatisation rules matching the word **czystkom**:

-kom	*mIII lD*	-ek	czystek
-kom	*mIII lD*	-k	czystk
-kom	*mIV N*	-kom	czystkom
-kom	*nII lD*	-ko	czystko
-kom	*nVI lD*	-kum	czystkum
-kom	*żIII lD*	-ka	czystka
-kom	*żIII(m) lD*	-ko	czystko
kom	*żIV lG*	koma	czystkoma
-kom	*blp lD*	-ka	czystka
-kom	*blp lD*	-ki	czystki
-kom	*Vla i*	-komić	czystkomić

An example analysis

Lemmatisation rules matching the word **czystkom**:

-kom	*mIII lD*	-ek	czystek
-kom	*mIII lD*	-k	czystk
-kom	*mIV N*	-kom	czystkom
-kom	*nII lD*	-ko	czystko
-kom	*nVI lD*	-kum	czystkum
-kom	*żIII lD*	-ka	czystka
-kom	*żIII(m) lD*	-ko	czystko
kom	*żIV lG*	koma	czystkoma
-kom	*blp lD*	-ka	czystka
-kom	*blp lD*	-ki	czystki
-kom	*Vla i*	-komić	czystkomić

Tokarski's index in *Morfeusz*

Lemmatisation rules matching the lemma **czystka** *żIII*:

czystka	-tka	*żIII*	*N*	-tka
czystki	-tki	*żIII*	*GlNTV*	-tka
czystce	-tce	*żIII*	*DL*	-tka
czystkę	-kę	*żIII*	*T*	-ka
czystką	-ką	*żIII*	*I*	-ka
czystko	-tko	*żIII*	*V*	-tka
czystek	-tek	*żIII*	*lG*	-tka
czystkom	-kom	*żIII*	*lD*	-ka
czystkami	.ami	*S lI*	⇒ .om	*S lD*

The core dictionary of *Morfeusz*

...		
kontekstu	1	subst:sg:gen:m3
konteksty	1	subst:pl:nom.acc:m3
kontekstów	2	subst:pl:gen:m3
kontekście	4st	subst:sg:loc.voc:m3
kontem	2o	subst:sg:inst:n1.n2
...		

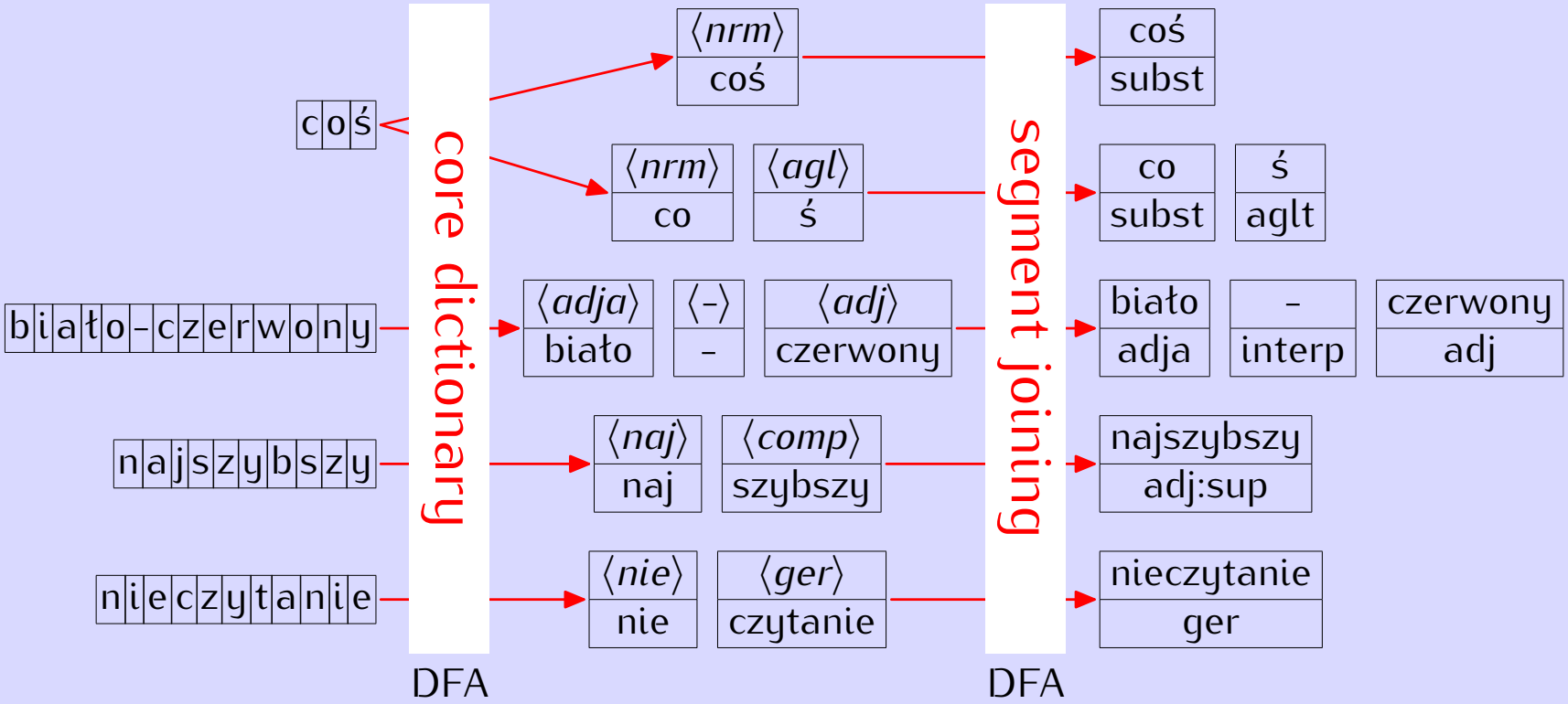
The lemmas in the list are provided implicitly.
E.g., *4st* above means: to get the lemma for *kontekście* strip the last 4 letters and replace them with *st*.

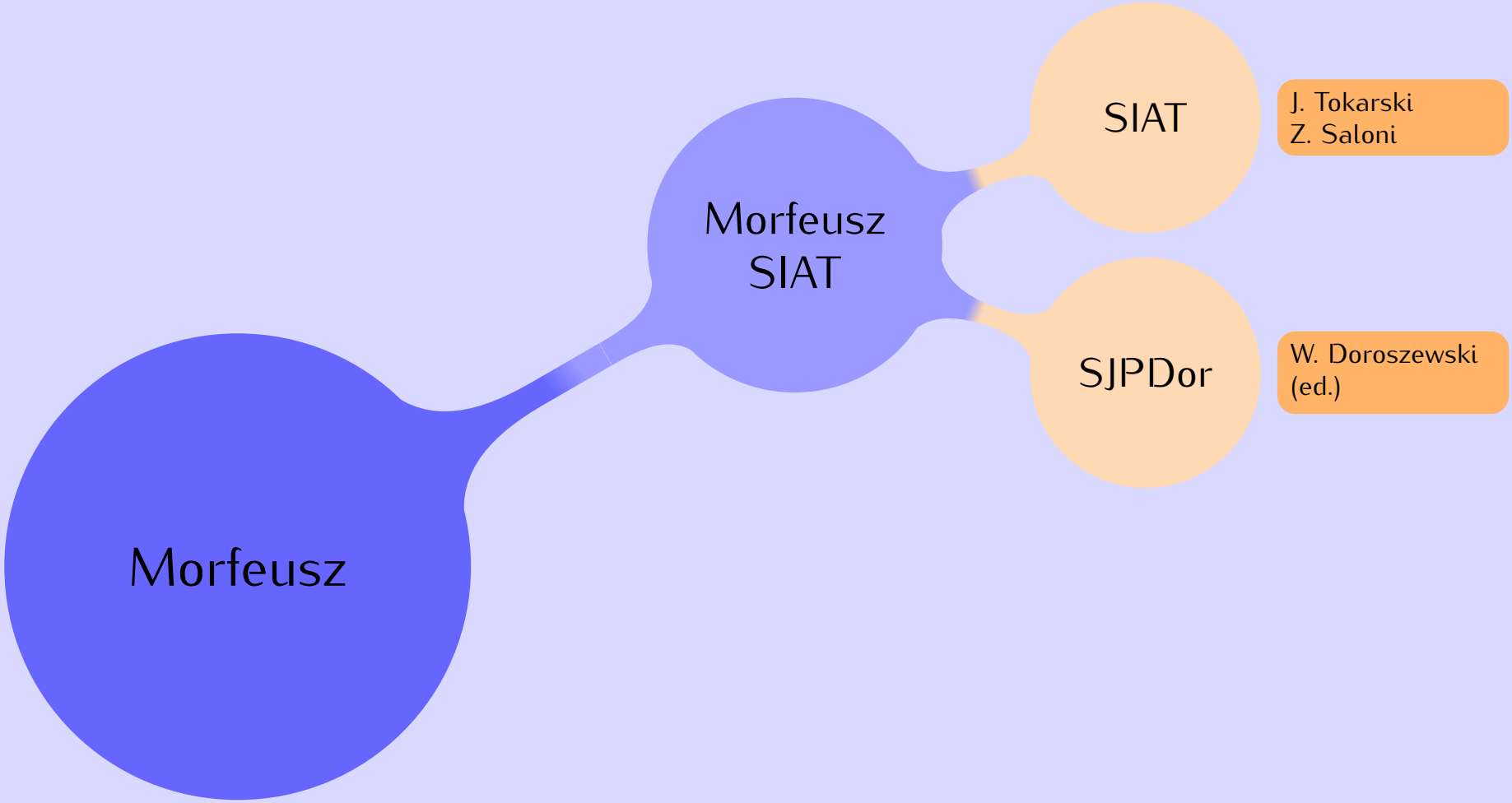
The core dictionary of *Morfeusz*

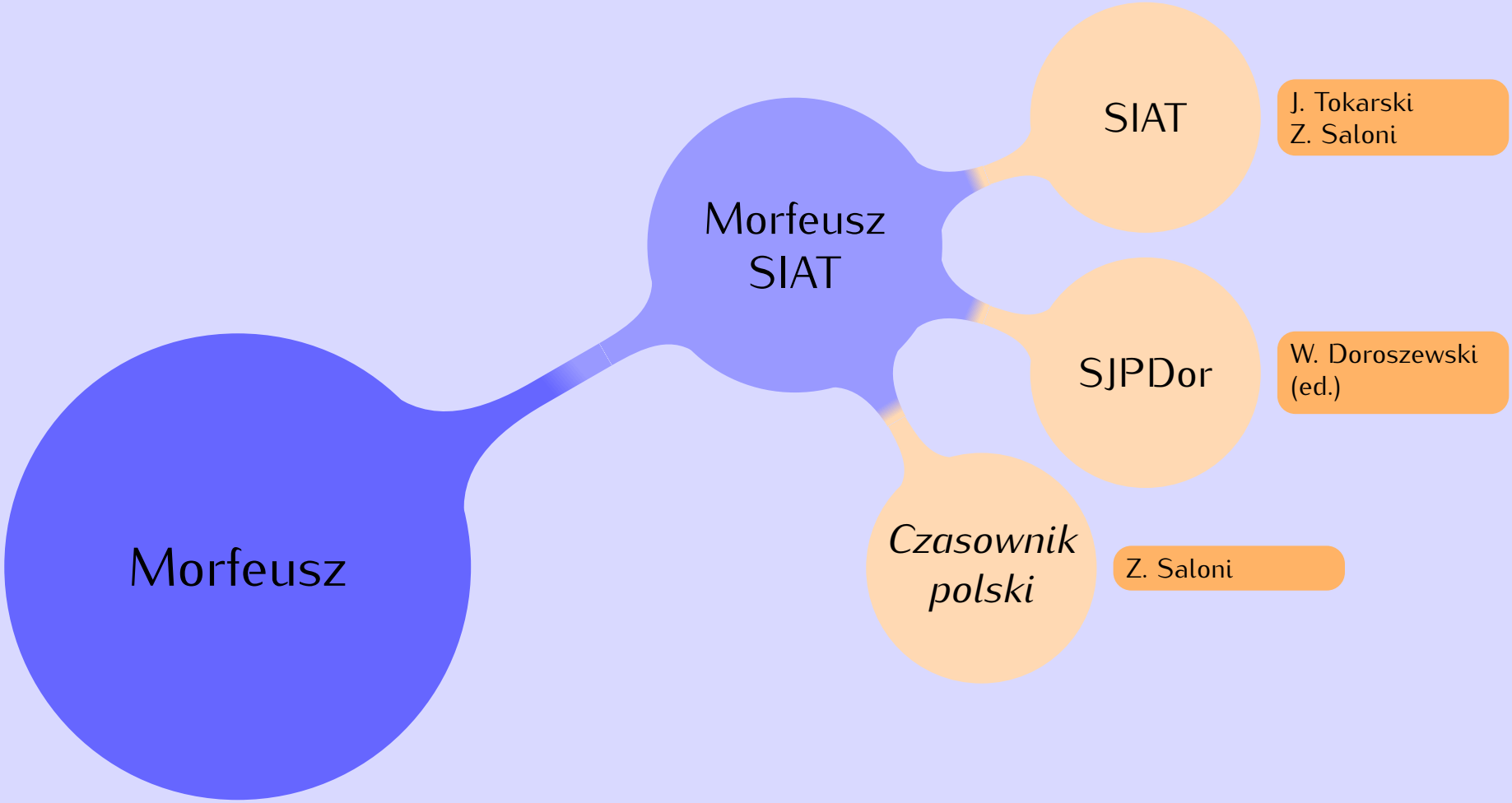
...		
kontekstu	1	subst:sg:gen:m3
konteksty	1	subst:pl:nom.acc:m3
kontekstów	2	subst:pl:gen:m3
kontekście	4st	subst:sg:loc.voc:m3
kontem	2o	subst:sg:inst:n1.n2
...		

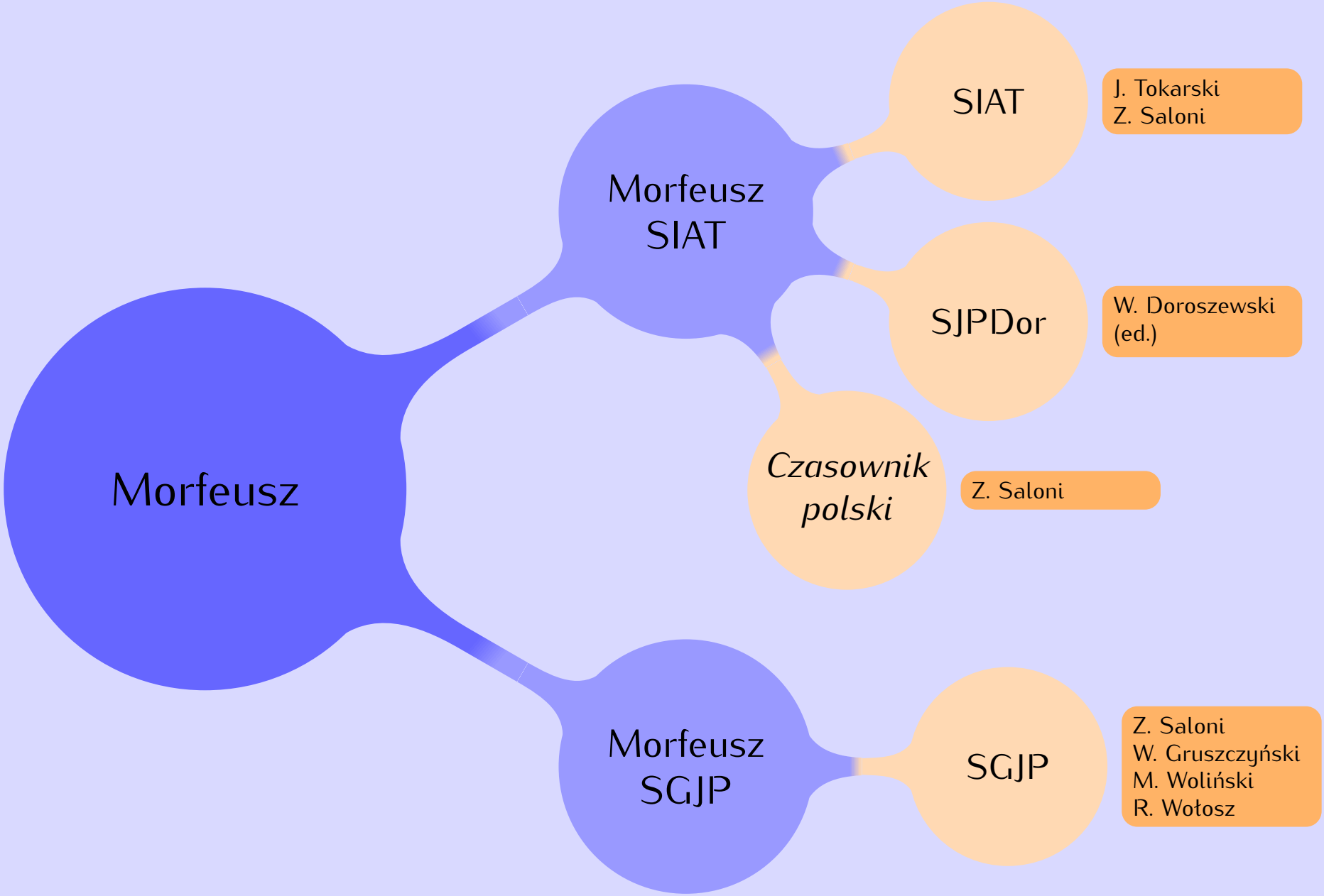
Tags in the dictionary are in accordance with The IPI PAN Tagset developed for annotation of The IPI PAN Corpus.

The representation of the linguistic data in *Morfeusz*




Morfeusz
Morfeusz SIAT
SIAT
J. Tokarski
Z. Saloni
SJPDor
W. Doroszewski
(ed.)


Morfeusz
Morfeusz SIAT
SIAT
J. Tokarski
Z. Saloni
SJPDor
W. Doroszewski (ed.)
Czasownik polski
Z. Saloni

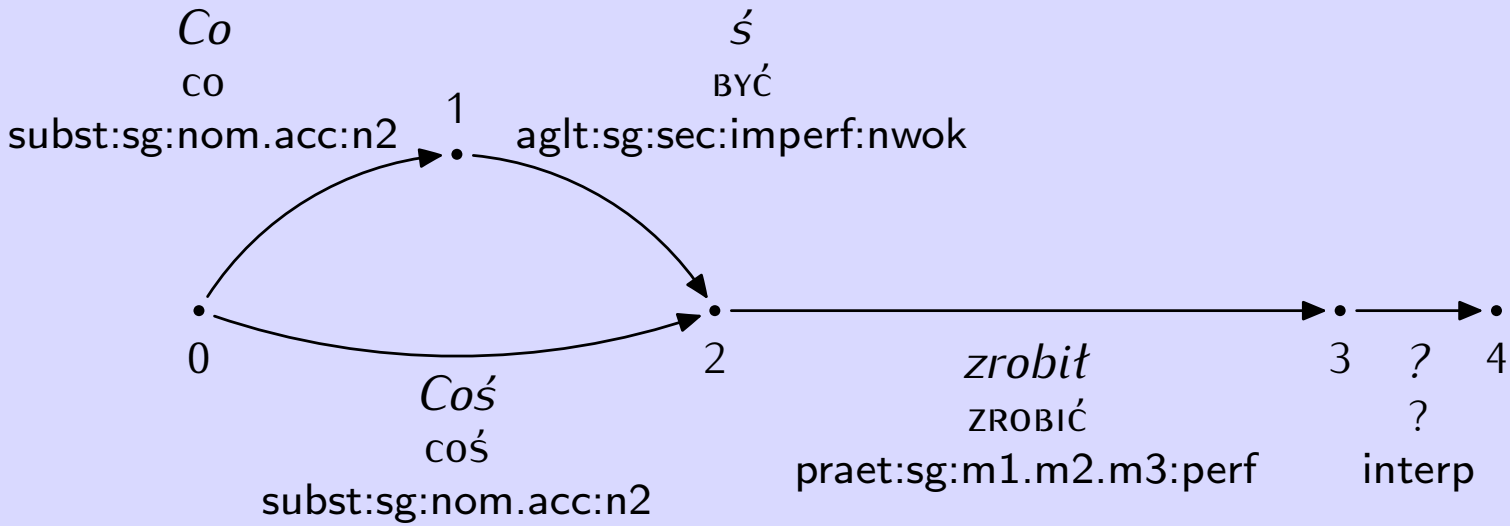

Morfeusz
Morfeusz SIAT
SIAT
J. Tokarski
Z. Saloni
SJPDor
W. Doroszewski (ed.)
Czasownik polski
Z. Saloni
Morfeusz SGJP
SGJP
Z. Saloni
W. Gruszczyński
M. Woliński
R. Wołosz

Morfeusz SIAT in numbers

- about 115,000 lexemes in the dictionary,
- about 1,700,000 different Polish words,
- about 35,000 words per second (on Pentium IV 2GHz),
- recognises

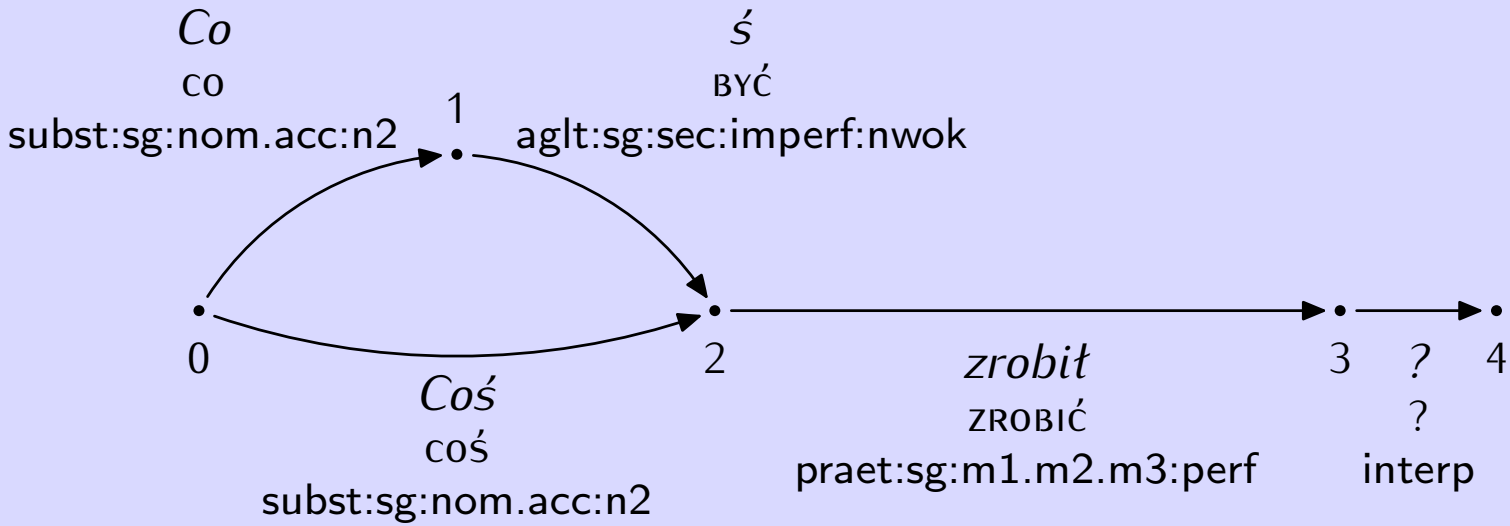
running words	word types	corpus
96.6%	87.0%	Frequency Dictionary of Polish (0.5 Mw)
95.7%	69.0%	The IPI PAN Corpus (v. 1.0, 85 Mw)

The representation of the results of an analysis



- **nodes** — positions in the text (between tokens)
- **edges** labelled with possible token interpretations

The representation of the results of an analysis



0	1	*Co*	CO	subst:sg:nom.acc:n2
1	2	*ś*	BYĆ	aglt:sg:sec:imperf:nwok
0	2	*Coś*	COŚ	subst:sg:nom.acc:n2
2	3	*zrobił*	ZROBIĆ	praet:sg:m1.m2.m3:perf
3	4	*?*	?	interp

A dictionary of proper names developed at IPI PAN

- Contains words from the IPI PAN Corpus which were unknown to *Morfeusz SIAT*. Most of them are proper names (about 90%).
- For those words complete lexemes are created by a group of lexicographers.
- For cross-validation each word is processed (at least) by two persons.
- About 5000 lexemes already in the dictionary.

Features contained in the dictionary

- grammatical class

Features contained in the dictionary

- grammatical class
- gender for nouns, aspect for verbs

Features contained in the dictionary

- grammatical class
- gender for nouns, aspect for verbs
- inflection

Features contained in the dictionary

- grammatical class
- gender for nouns, aspect for verbs
- inflection
- common or proper
  - common
  - first name
  - last name
  - geographical name
  - name of an institution or organisation
  - other proper name

Features contained in the dictionary

- grammatical class
- gender for nouns, aspect for verbs
- inflection
- common or proper
  - common
  - first name
  - last name
  - geographical name
  - name of an institution or organisation
  - other proper name
- kind of language
  - general
  - archaic
  - local dialect

The tool used

- Web-based application named *Kuźnia leksemów* (*Lexeme forge*).
- Coded in PHP by Daniel Janus.
- Data gathered in an SQL database.
- Inflectional patterns for Polish nouns developed by Włodzimierz Gruszczyński are used to ease entering of inflectional forms.
- Lexicographers answer questions asked by the program.

Development plans for *Morfeusz*

Long needed changes:

- case-sensitivity,
- abbreviations,
- distinguishing of homonyms,
- morphological guesser,
- features of proper names,
- small changes in the tagset: parts of multi-word idioms, a more detailed classification for non-inflecting lexemes, ordinal numerals (?),
- ability to attach additional (special purpose) dictionaries.

Homonyms

- Same lemma, different inflection:
  - BAL (ball, dance): `subst:sg:gen:m3` *balu*
  - BAL (log of wood): `subst:sg:gen:m3` *bala*
- Same lemma, same inflection, different meaning:
  - PARA (pair)
  - PARA (steam)

Homonyms

- Same lemma, different inflection:
  - BAL (ball, dance): `subst:sg:gen:m3` *balu*
  - BAL (log of wood): `subst:sg:gen:m3` *bala*
- Same lemma, same inflection, different meaning:
  - PARA (pair)
  - PARA (steam)

Only the first case needs to be distinguished by Morfeusz with homonym numbers: BAL 1 and BAL 2, or perhaps BAL;-U and BAL;-A.

Do we need information on the case of letters?

- Przyszedł Jan Biały.
- Przyszedł Jan biały jak śnieg.

Do we need information on the case of letters?

- Przyszedł Jan Biały.
- Przyszedł Jan biały jak śnieg.

Proposed values for case: lower, required capital, facultative capital, all caps(?), lower-that-should-be-capital(?), mixed(?).

Features in *Morfeusz*'s output

from	to	word	lemma	tag
0	1	*Co*	CO	subst:sg:nom.acc:n2
1	2	*ś*	BYĆ	aglt:sg:sec:imperf:nwok
0	2	*Coś*	coś	subst:sg:nom.acc:n2
2	3	*zrobił*	ZROBIĆ	praet:sg:m1.m2.m3:perf
3	4	*?*	?	interp

Proposed new set of features in *Morfeusz*'s results

- from
- to
- word
- lemma
- homonym identifier
- tag
- is_proper
- case
- source (main dictionary, guesser, domain dictionary, dialect, archaic, ...)

Bibliography

Marcin Woliński.
System znaczników morfosyntaktycznych w korpusie IPI PAN.
Polonica, XXII–XXIII:39–55, 2003.

Marcin Woliński.
Morfeusz — a practical tool for the morphological analysis of Polish.
In Mieczysław Kłopotek, Sławomir Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, pages 503–512. Springer, 2006.