Maca — a configurable tool to integrate Polish morphological data

Adam Radziszewski
Tomasz Śniatowski
Wrocław University of Technology

Wrocław
University
of Technology

NATIONAL
COHESION STRATEGY

Outline

- Morphological resources for Polish
- Tagset and segmentation differences
- Requirements
- Our solution
- Usage scenarios
- Summary

Introduction

- Morphological analysis: assigning morphological descriptions to tokens
- Token → set of (*MSD tag*, *lemma*) pairs
- MSD — morphosyntactic description tag
  - Part-of-Speech / grammatical class
  - Values of inflectional and syntactic attributes, e.g. case

Example: analysis of the form *myśl*

myśl	subst:sg:nom:f	*thought*
myśleć	impt:sg:imperf	*think!*

Morphological resources for Polish

IPI PAN Corpus tagset

Analyser: Morfeusz SIAT
Large dictionary
Data by recognised Polish linguists
Very restrictive licence

Corpus: IPI PAN (fragment)
660 000 tokens manually annot'd
84 000 different forms
GNU GPL

Morfologik tagset

Analyser: Morfologik
Large dictionary (3.5 mln forms)
Data from ispell/myspell
GNU LGPL or CC BY-SA

*Free: src available

There are more **non-free** analysers & corpora with various tagsets

Morphological resources for Polish (2)

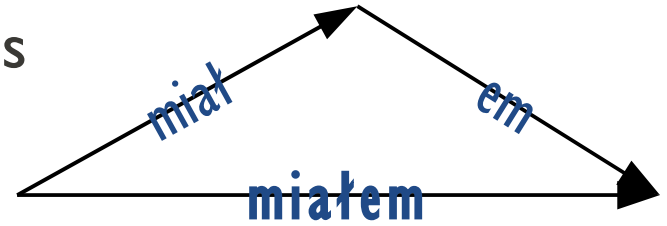
- Important to have corpus and analyser in the same tagset
  - Corpus usually too small to obtain reliable lexical model
  - POS/MSD taggers for Polish rely on external analysers
  - Goal: to integrate corpus morphological data with available analysers
- Important to be able to modify an existing dictionary
  - Correct erroneous entries
  - Extend
  - Supersede entries with domain-specific terminology
  - Integrate multiple dictionaries

Tagset differences

- Traditional Parts-of-Speech (nouns, pronouns, verbs…)
  - Non-free analysers, e.g. POLEX PMDBF
  - Partially Morfologik
- PoS classes based on inflectional properties
  - Morfeusz / IPI PAN Corpus, partially Morfologik
  - Each class assigned a set of attributes whose values must be given
  - If some subset of a PoS not specified for an attribute, should constitute a separate class
  - *Moja* (*my*-fem-sg) inflects as adjective, thus labelled so
  - *Jasno* (*light*) is gradable → adverb; *dziś* (*today*) is not → particle

Segmentation differences

- When attaching MSD tags, we need to know what kind of units (tokens) we want to account for
- Traditionally, strings of letters cut by punctuation and white spaces (Morfologik, POLEX PMDBF)
- Morfeusz: some verb forms are split into parts
  - *Miałem* (*I had* masc) → *miał* (sing. masc.) + *em* (sing. I person)
  - *Miałbym* (I'd have masc) → *miał* (sg. masc.) + *by* (conj.part.) + *m* (sg. I person)
  - Motivation: occasional scrambling *gdyby*+*m miał* (*If I had* masc)
  - Seg. ambiguities: *miałem* is also a noun in instr. case (*dust*)
  - Morfeusz outputs graphs



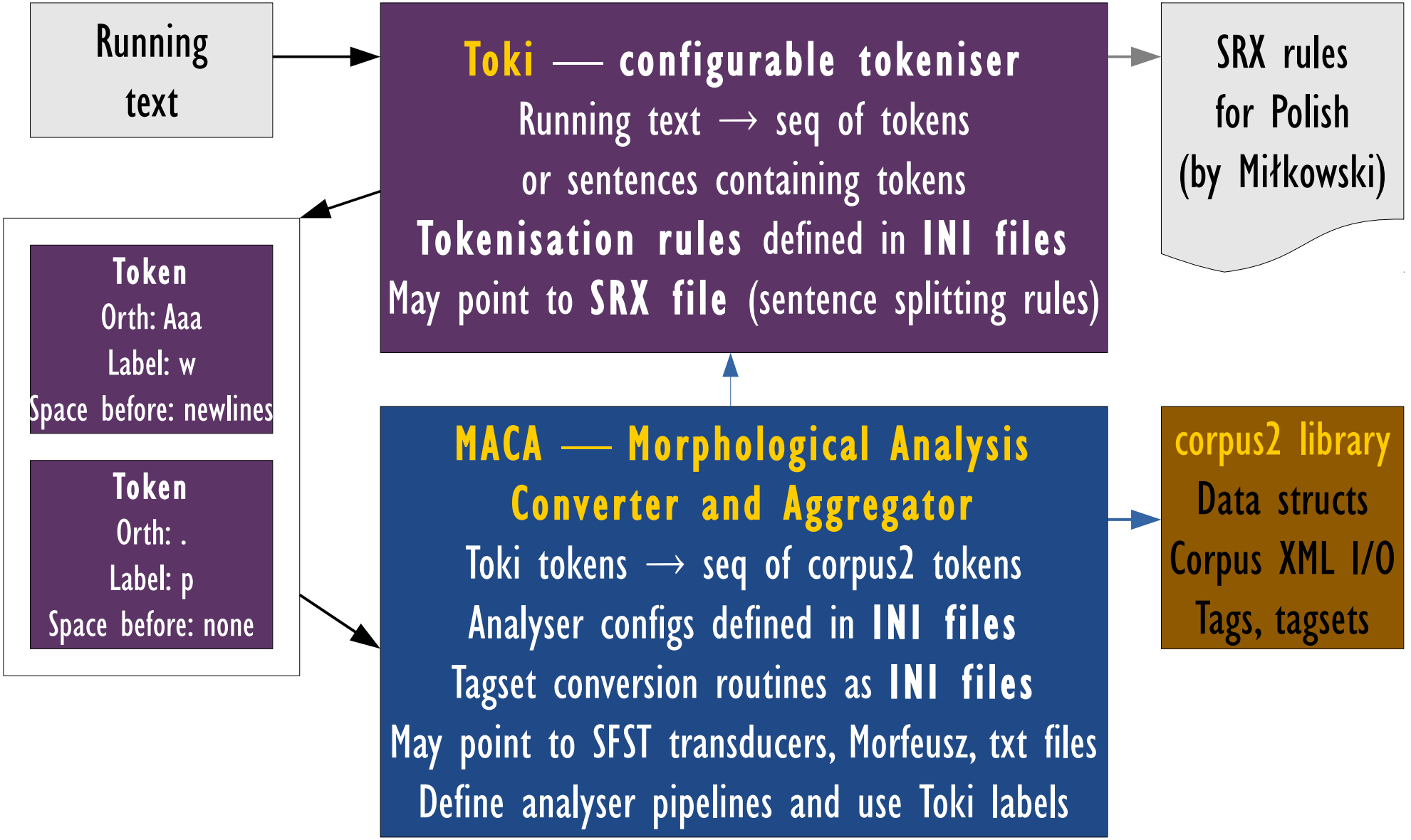
Requirements (functional)

- Integrate available morphological data under different settings, providing multiple configurations
- Select analysers to use at the moment
- Be able to use Morfeusz **until** enough free data available
- Support overriding entries and extending dictionaries
- Tight coupling with tokeniser
  - Take advantage of knowing token type (numbers, words, punct.)
  - Tie different analysis pipelines to different token types
- Handle some differences in tagsets and seg. strategies
- Handle large dictionaries efficiently (transducers)

Requirements (technical)

- Whole functionality as command-line tools and C/C++ library for use in NLP software
  - Performance, low start-up time (no VM)
  - Easy integration with Python and C++
- Re-usability
  - Division into libraries wrt. functionality (I/O, tokeniser, analyser)
  - Useful command-line tools also serving as library API usage examples
- Supporting standards and available resources
  - SRX — segmentation rule exchange format for MT systems
  - Unicode (using ICU library)
  - SFST transducers
  - Support for Morfeusz data (graphs) and XCES XML format (IPI PAN Corpus)

Our solution: MACA system

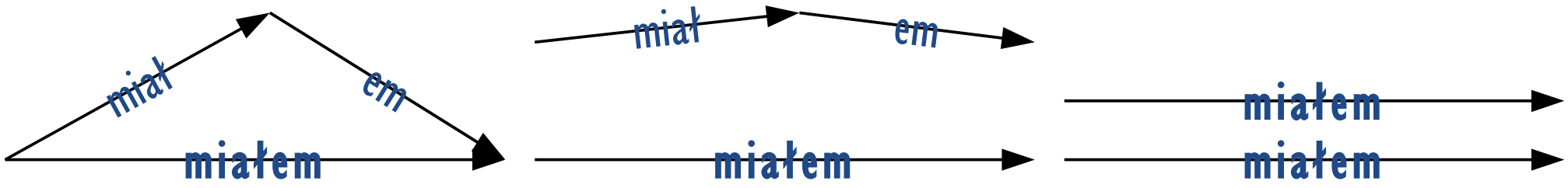


Usage scenarios (1)

- Compiling working analyser from existing data
  - Use one of the provided Toki config or tailor a specific one
  - Compile a text file with dictionary into SFST format
  - Simple Maca config: attaches fixed tags to punctuation and digits, the compiled SFST transducer to the rest
  - Practical usage in another project: converted Morfologik data into the IPIC tagset; resulting in free replacement of Morfeusz
- Using and patching Morfeusz
  - Morfeusz is a library + rudimentary utility to pose queries
  - Morfeusz + Maca is able to analyse running text or XML files
  - When seg. ambiguity encountered, warns and selects shortest path

Usage scenarios (2)

- Simple tag/segmentation conversions
  - Serious tagset conversion is better performed off-line
  - MACA: mapping rules, conditional token joining and splitting
  - Differences in attribute value sets across corpus versions
  - Reducing a tagset to PoS-only tags
  - Reducing ambiguity in Morfeusz output: conversion routines may be applied to graph paths separately before joining



Summary

- A working system, bundled with practical configs & data
- C++ framework to build NLP applications on
- Released under GNU GPL 3.0 at http://nlp.pwr.wroc.pl/redmine/projects/libpltagger
- First open-source C/C++ SRX implementation
- Further work:
  - Python wrappers
  - Support additional corpus formats
  - Support MULTEXT-EAST tag string representation
  - Test for other languages