Recent Advances in MULTIFLEX, a Morphological Generator of Multi-Word Units

Agata Savary

November 17, 2008

Morfeusz/MULTIFLEX Platform

- Aim: description of **morphology and variation of compounds**
- Information on the language level: classes, categories and values of the **IPIPAN tagset**
- Generating inflected forms of single words with **Morfeusz**
- Combining (by a **graph**) inflected forms of single words to create inflected forms of compounds
- **Unification** and value **inheritance** for a compact description

Example 1: *Maria Skłodowska-Curie*

Maria Skłodowska-Curie, Marii Skłodowskiej-Curie, . . .

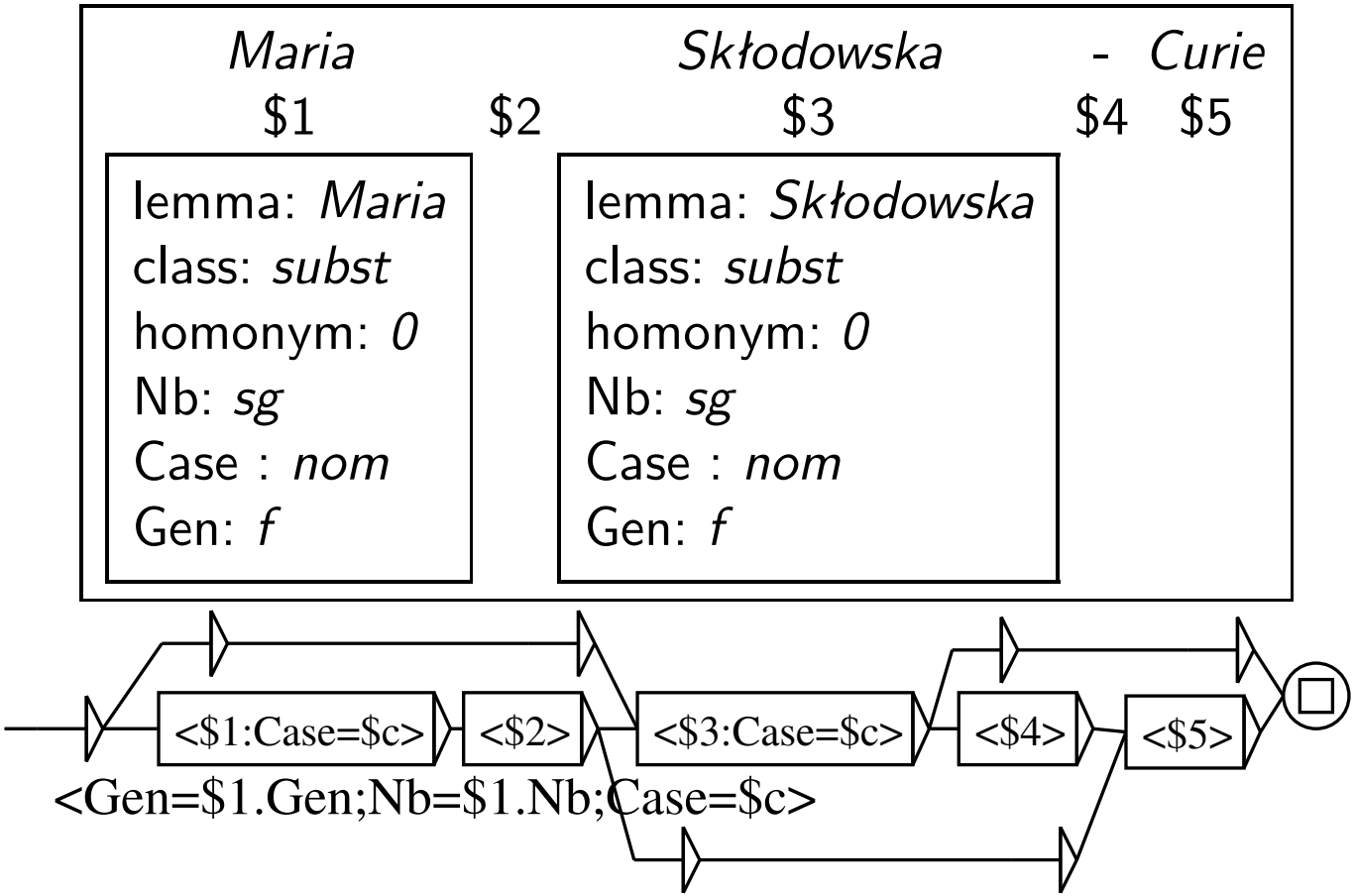
Skłodowska-Curie, Skłodowskiej-Curie, . . .

Maria Skłodowska, Marii Skłodowskiej, . . .

Maria Curie, Marii Curie, . . .

Skłodowska, Skłodowskiej, . . .

Annotation of components and inflection graph



Annotated forms

Maria Skłodowska-Curie	*Maria Skłodowska-Curie:subst:sg:***nom***:f*
Marii Skłodowskiej-Curie	*Maria Skłodowska-Curie:subst:sg:***gen***:f*
Marii Skłodowskiej-Curie	*Maria Skłodowska-Curie:subst:sg:***dat***:f*
Skłodowską-Curie	*Maria Skłodowska-Curie:subst:sg:***inst***:f*
Skłodowskiej	*Maria Skłodowska-Curie:subst:sg:***loc***:f*

Example 2: *ulica Marii Skłodowskiej-Curie*

ulica Marii Skłodowskiej-Curie, ulicy Marii Skłodowskiej-Curie, . . .;

ulica Marii Skłodowskiej, ulicy Marii Skłodowskiej, . . .;

ulica Marii Curie, ulicy Marii Curie, . . .;

ulica Skłodowskiej-Curie, ulicy Skłodowskiej-Curie, . . .;

ulica Skłodowskiej, ulicy Skłodowskiej, . . .;

Marii Skłodowskiej-Curie;

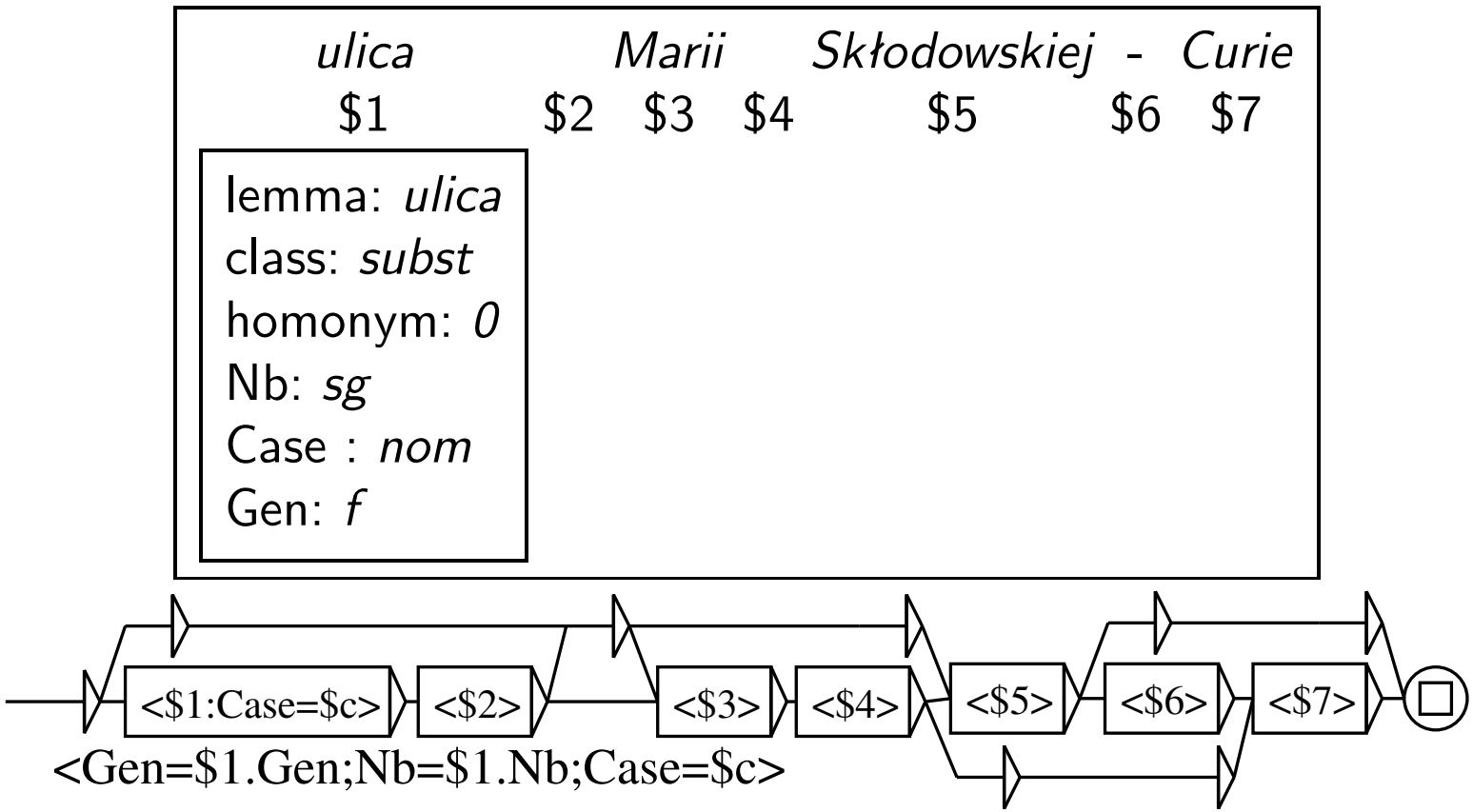
Marii Skłodowskiej;

Marii Curie;

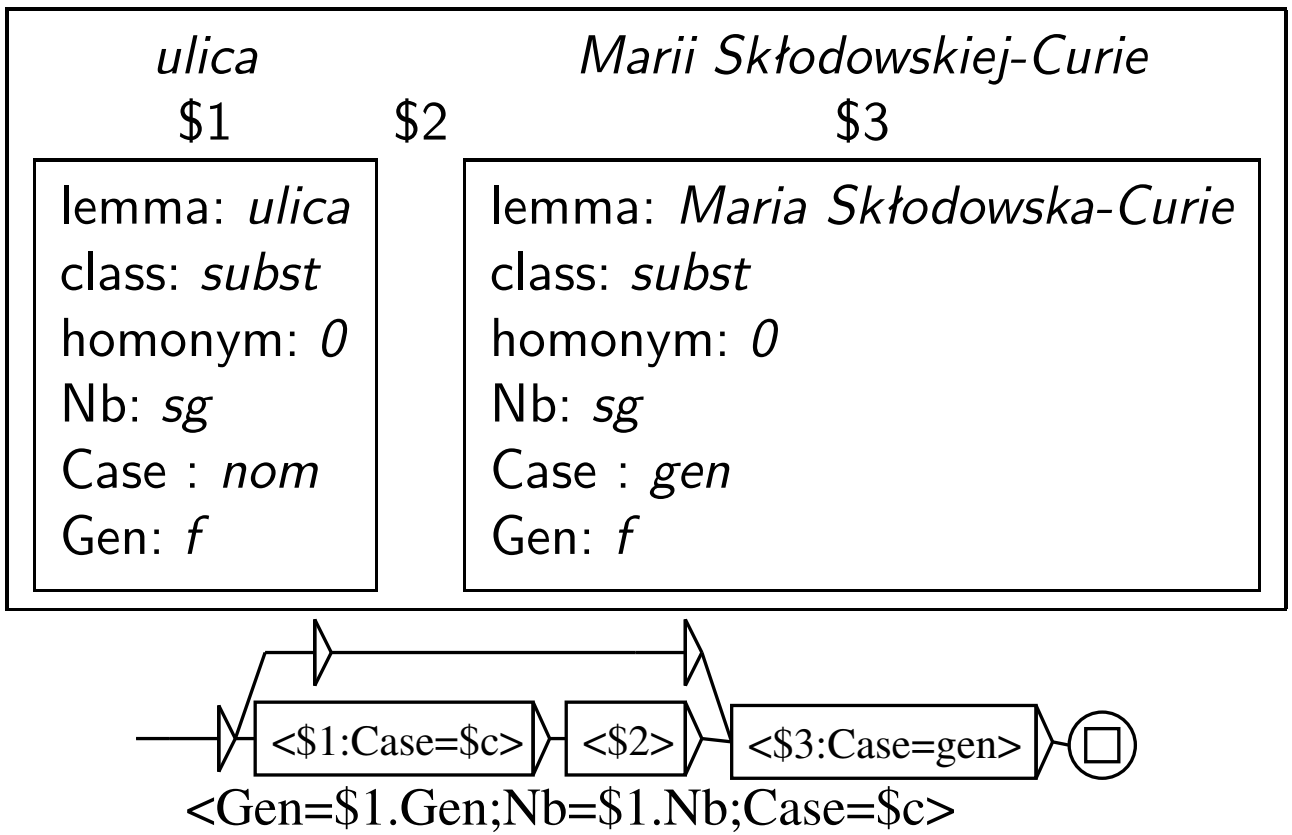
Skłodowskiej-Curie;

Skłodowskiej

Flat description of variants: *ulica Marii Skłodowskiej-Curie*



New solution : embedded description of *ulica Marii Skłodowskiej-Curie*



Morphology of numerals in the IPIPAN tagset

Two kinds of "traditional" numerals :

- Cardinal numerals (**num**) - a class on its own
  - have a fixed number
  - inflect in case, gender, and accomodability
  - complex morpho-syntactic behaviour
- Ordinal numerals - behave morphologically as adjectives
  - have a fixed number
  - inflect in case, gender, and accomodability

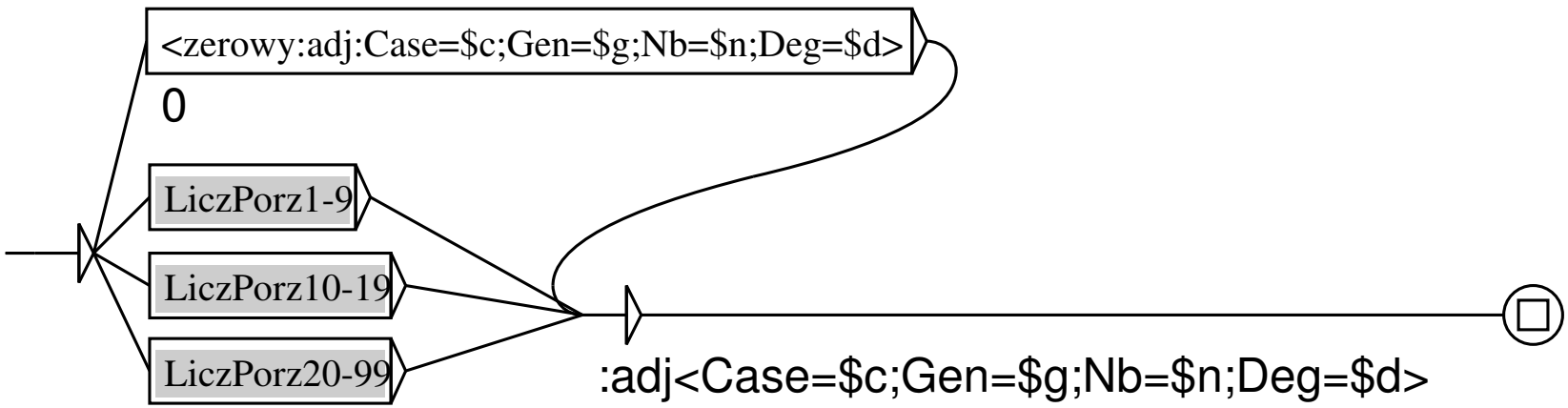
Problem: how to describe multi-word numerals ?

- \- their number is infinite
- \- their are spelled with letters or digits (dwudziesty vs. 20.)
- \+ their vocabulary is small
- \+ the rules of their creation are very regular

Ordinal numerals - what we wish

- A canonical form for each numeral:
  {dziesięcio-tysięczny, 10-cio-tysięczny, 10-tysięczny, 10000., 10 000.} $\Rightarrow$ 10000
- A complete annotation: 10000:adj:sg:m1:nom, etc. (?)
- Morphological analysis:
  dziesięcio-tysięczny $\Rightarrow$ 10000:adj:sg:m1.m2.m3:nom
- Morphological generation:
  10000:adj:sg:m1:nom $\Rightarrow$ {dziesięcio-tysięczny,
  10-cio-tysięczny, 10-tysięczny, 10000., 10 000.}

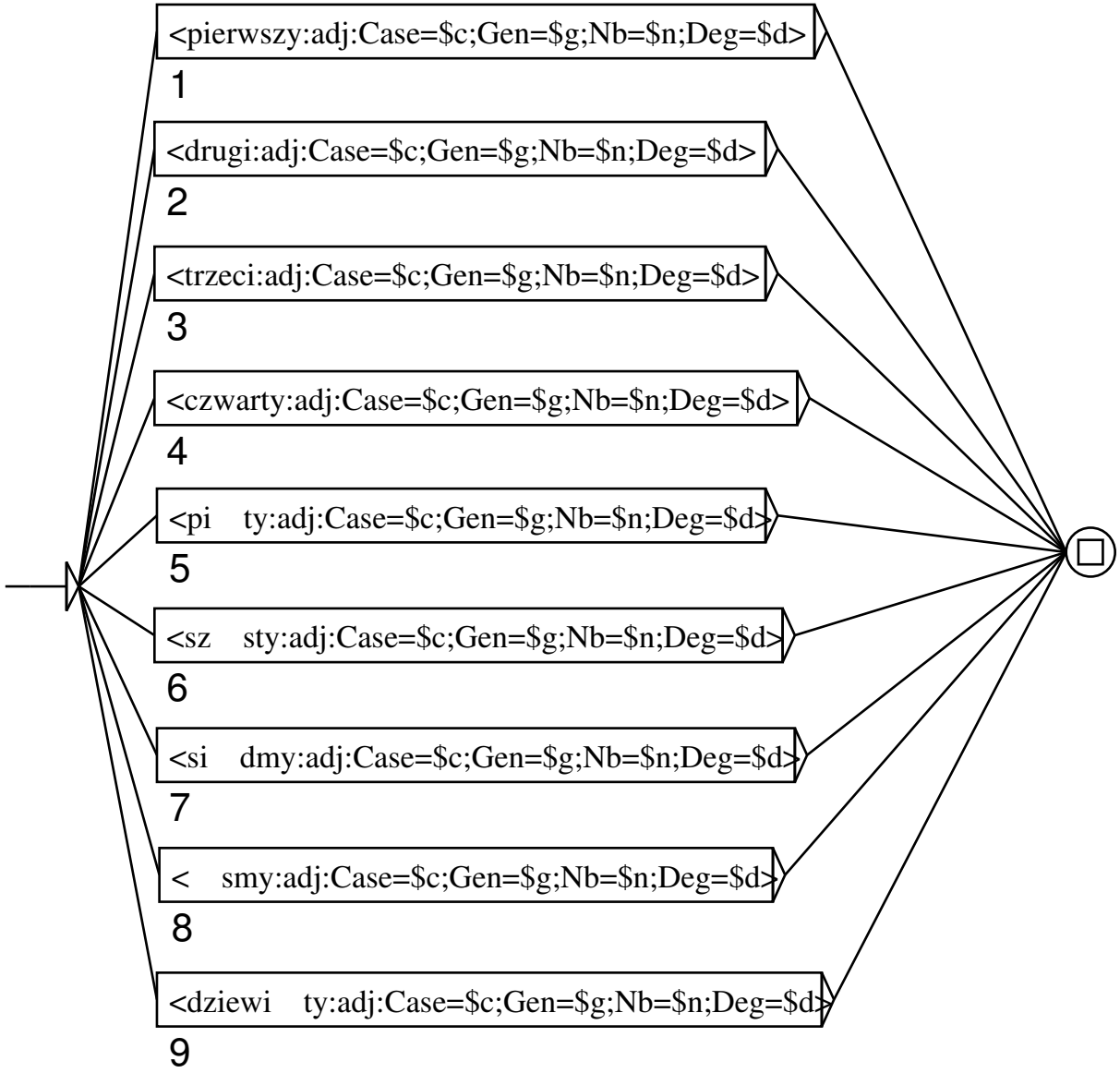
Graph-based description: ordinal numerals 0-99 (simplified)



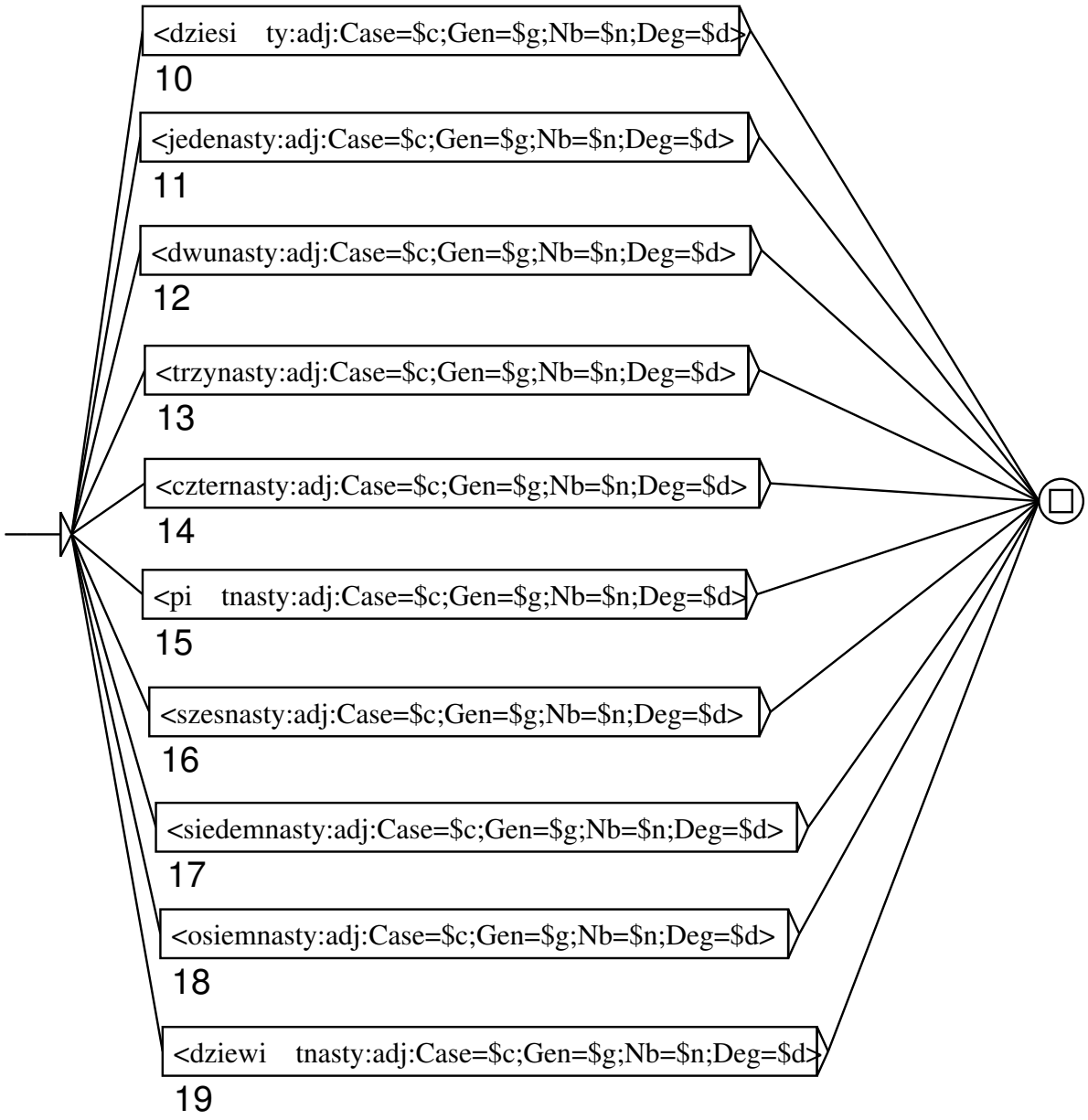
Info **inside** the boxes: textual variants

Info **under** the boxes: canonical forms

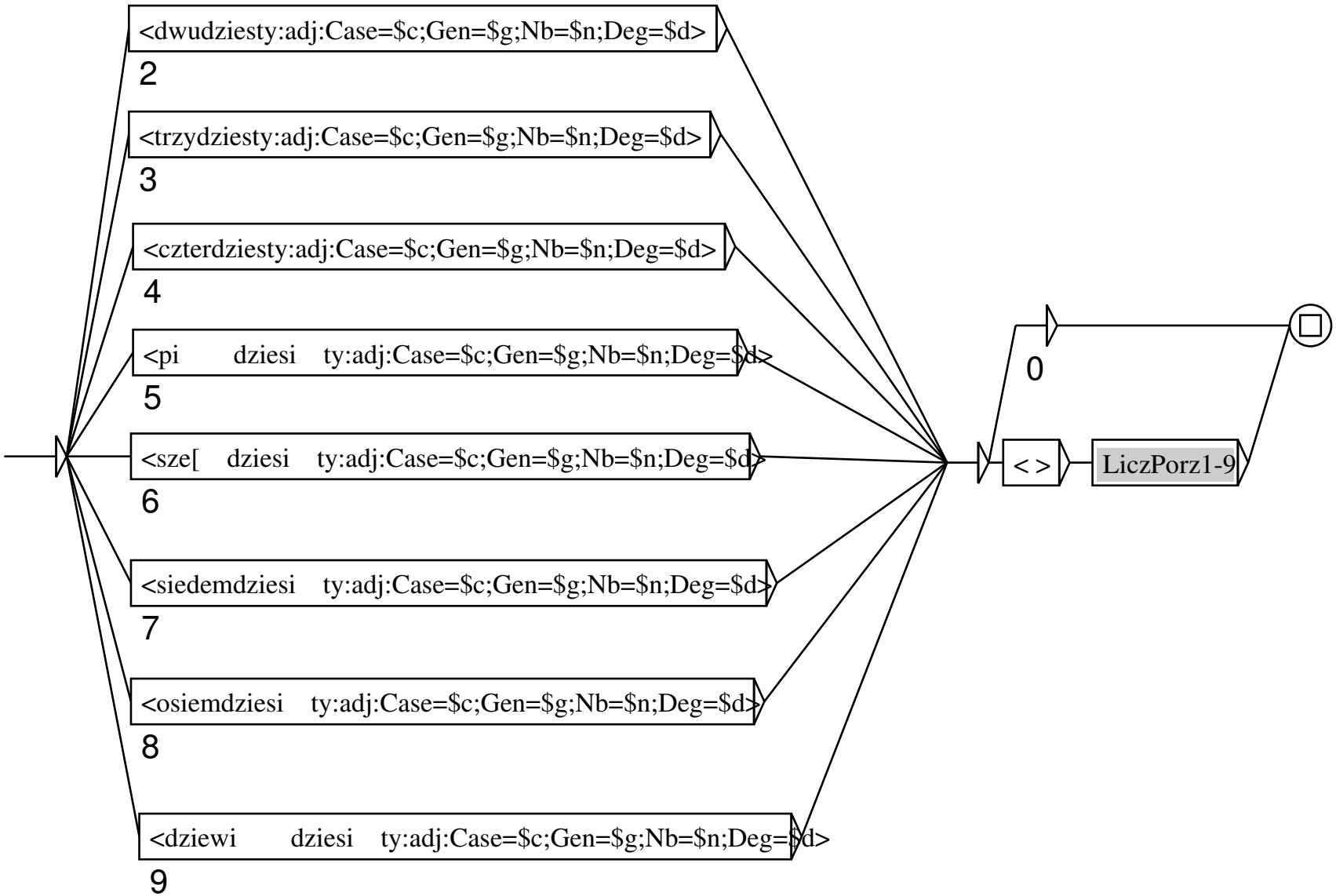
Sub-graph: **LiczPorz1-9**



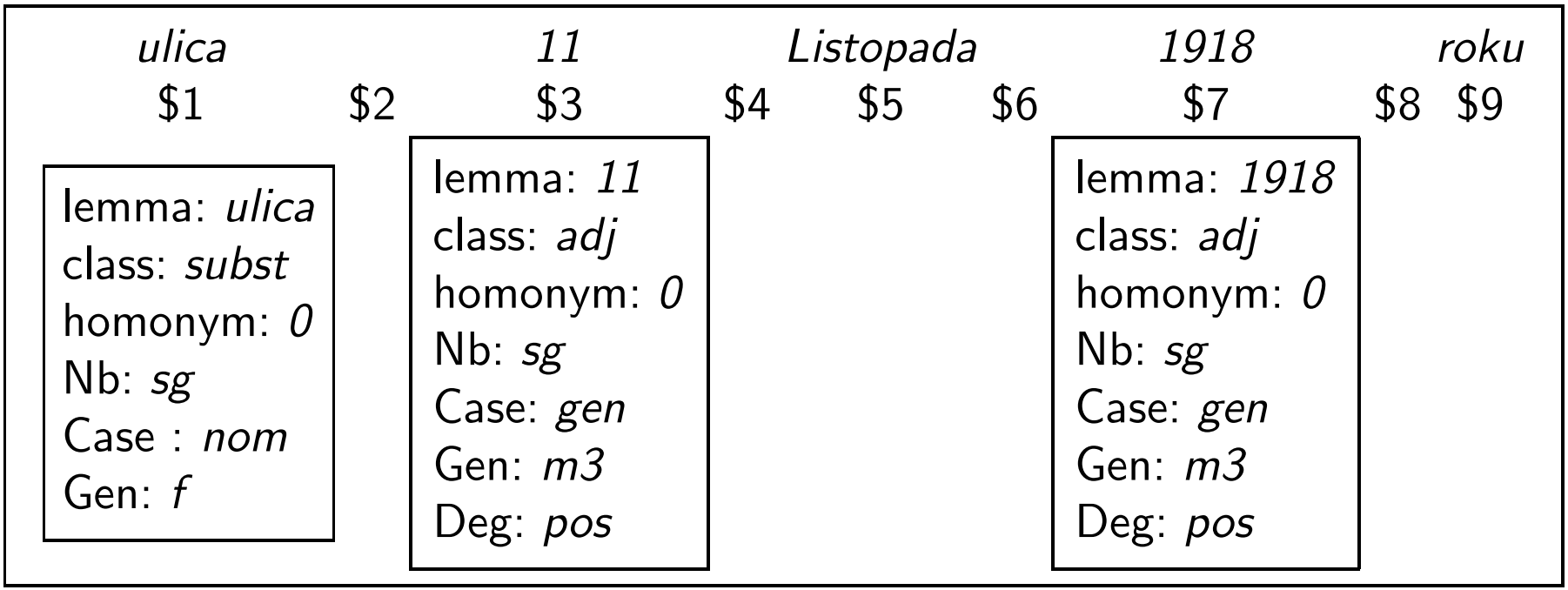
Sub-graph: **LiczPorz10-19**



Sub-graph: **LiczPorz20-99**



Perspective: using graphs for “inflection” of compound numerals in MULTIFLEX



“Inflected forms” to be generated

ulica 11 Listopada 1918 roku

ulica 11 Listopada 1918

11 Listopada 1918 roku

ulica Jednastego Listopada 1918 roku

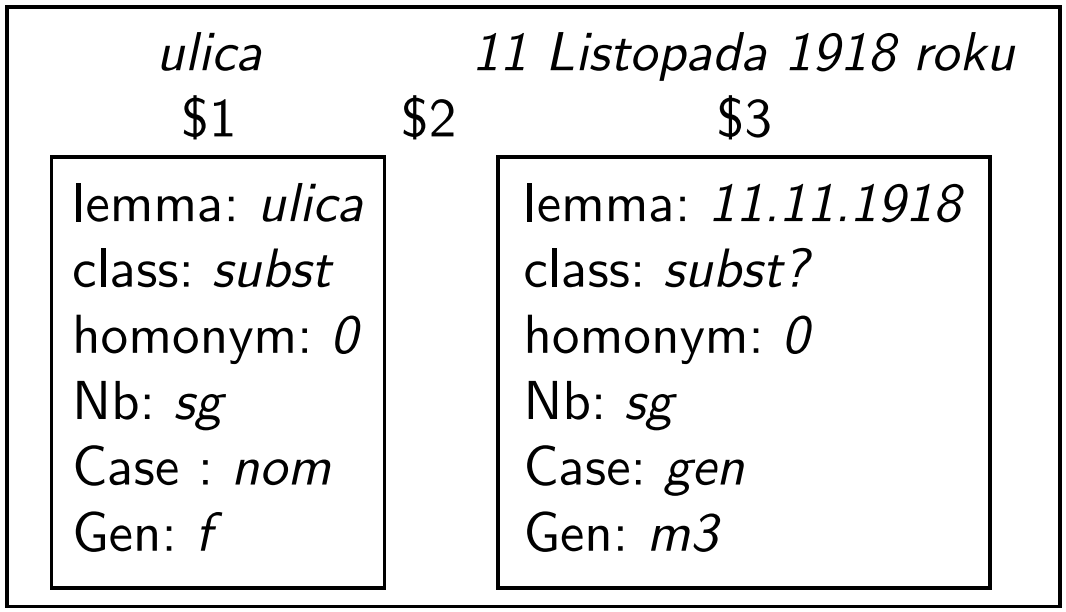
ulica 11-ego Listopada 1918 roku

ulica 11 Listopada Tysiąc Dziewięćset Osiemnastego roku

? *ulica 11 Listopada Osiemnastego roku*

etc.

One step further



Graphs describe possible dates and their variants:

- 11.11.1918
- 11/11/1918
- 11 listopada 1918
- 11 listopada osiemnastego roku
- jedenasty listopada osiemnastego roku
- etc.