„KORBA”
Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)

Pracownia Historii Języka Polskiego
XVII i XVIII wieku IJP PAN,

Instytut Podstaw Informatyki PAN

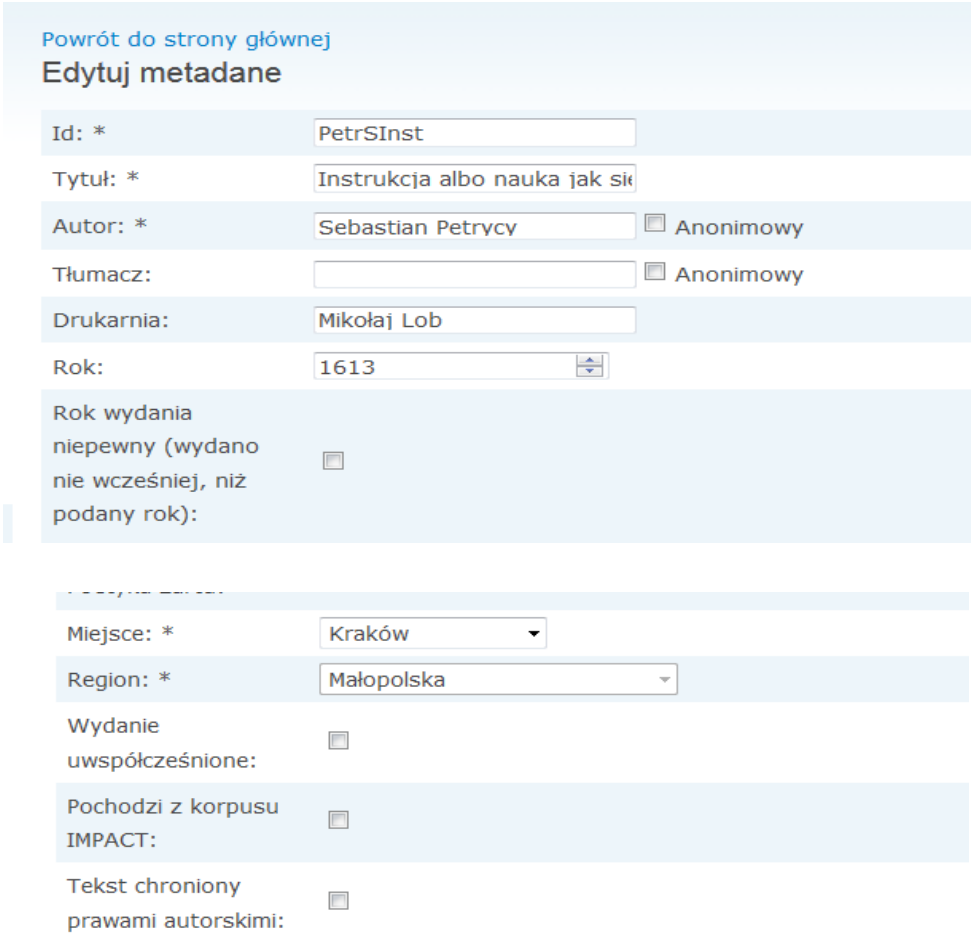
Podstawowe informacje o projekcie

- Projekt realizowany przez IJP PAN we współpracy z IPI PAN, finansowany w ramach Narodowego Programu Rozwoju Humanistyki na lata: 2013-2017
- Kierownik projektu: prof. dr hab. Włodzimierz Gruszczyński
- Planowana objętość korpusu to 12 mln segmentów
- Korpus historyczny rozszerzający Narodowy Korpus Języka Polskiego
- Kryptonim: KORBA (KORpus BArokowy)
- http://korba.nlp.ipipan.waw.pl/login/?next=/ (obecnie dostępny dla członków zespołu pracującego nad korpusem oraz w Pracowni Historii Języka Polskiego XVII i XVIII w. IJP PAN w Warszawie)

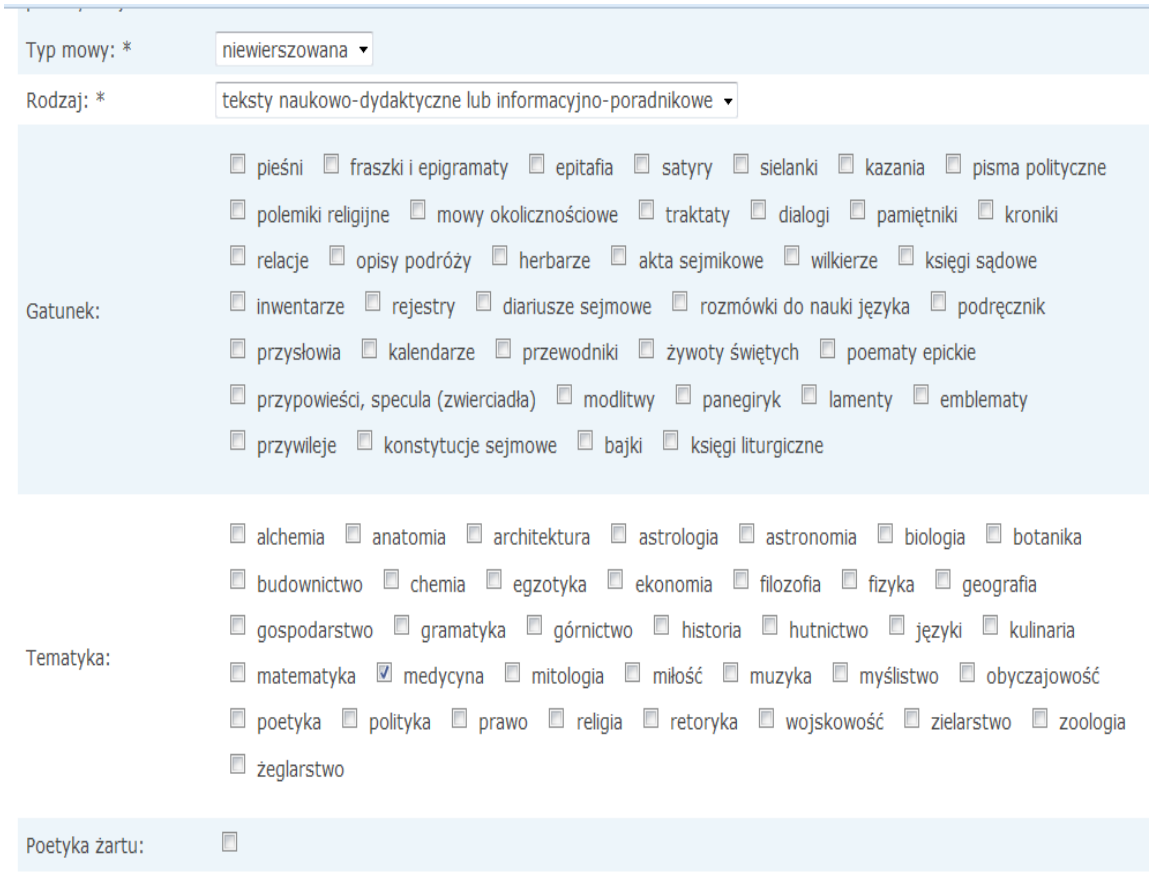
Zakres znakowania

- Znakowanie socjolingwistyczne (informacje o autorze, wydawcy, tłumaczu) – umieszczone w „metryczce” każdego tekstu.
- Znakowanie stylistyczno-genologiczne – umieszczone w „metryczce” każdego tekstu.
- Znakowanie strukturalne i tekstowe, umożliwiające dokładną lokalizację cytatu w tekście (paginacja; rozdziały, części, księgi itp.; notki marginesowe, przypisy itp.; podpisy pod ilustracjami; także oznaczenie wszystkich wtrętów obcych, z podziałem na języki).
- Znakowanie morfosyntaktyczne
  - 0,5 mln segmentów – ręczne
  - reszta – automatyczne

Znakowanie socjolingwistyczne



Znakowanie stylistyczno-genologiczne



Znakowanie strukturalne i tekstowe

- Oznaczenie istotnych elementów struktury tekstu, np. elementów strony tytułowej, fragmentów obcojęzycznych itp.
- Znakowanie Korby odbywa się w szablonie dokumentu Word umożliwiającym wprowadzanie znaczników za pomocą skrótów klawiaturowych
- Na potrzeby projektu został stworzony program do konwersji plików Worda na format XML zgodny ze standardem TEI

Znaczniki strukturalne

[POCZĄTEK DOKUMENTU] - [KONIEC DOKUMENTU]
[POCZĄTEK STRONY] - [KONIEC STRONY]
[POCZĄTEK POZIOMU 1] - [KONIEC POZIOMU 1]
[POCZĄTEK STRONY TYTUŁOWEJ] -[KONIEC STRONY TYTUŁOWEJ]
[POCZĄTEK STRONY POTYTUŁOWEJ] - [KONIEC STRONY POTYTUŁOWEJ]
[POCZĄTEK MOTTA] - [KONIEC MOTTA]
[POCZĄTEK PRE-TEKSTU] -[KONIEC PRE-TEKSTU]
[POCZĄTEK POST-TEKSTU] - [KONIEC POST-TEKST]
[POCZĄTEK LISTY] - [KONIEC LISTY]
[POCZĄTEK INDEKSU] - [KONIEC INDEKSU]
[POCZĄTEK SPISU TREŚCI] - [KONIEC SPISU TREŚCI]
[POCZĄTEK NOTKI] - [KONIEC NOTKI]
[ILUSTRACJA]
[WZÓR MATEMATYCZNY]
[ZAPIS NUTOWY]
[INNA PRZERWA W TEKŚCIE]
[FRAGMENT NIECZYTELNY]
[TEKST W JĘZYKU OBCYM]
[ALFABET NIEŁACIŃSKI]

Znaczniki tekstowe

Tytuł publikacji	Podtytuł części
Powtórzony tytuł	*Streszczenie części*
Podtytuł publikacji	Podpis pod rysunkiem
Autor	**Oznaczenie składki**
Miejsce wydania	***Numer strony***
Drukarnia	*Kustosz*
Rok wydania	*Żywa pagina*
Inny element strony tytułowej	*Numer rozdziału*
Tłumacz	*Numer wersetu*
Język oryginału	*Element dodany w Biblii Gdańskiej*
Numer wydania	Tekst odredakcyjny
Tytuł części	Tekst współczesny

Oznaczenia języków obcych

Arabski	Niemiecki
Czeski	Południowosłowiański
Francuski	Wschodniosłowiański
Grecki	Skandynawski
Hebrajski	Turecko-tatarski
Hiszpański	Węgierski
Litewski	Włoski
Łacina	Inny

Komentarze i poprawki skryptorów

- Informacja o liczbie nieczytelnych znaków
  - COMED<..>
- Informacja o możliwości błędnego odczytania słowa
  - gospodarz <?>
- Informacja o zaskakującym zapisie
  - gorzski <!>
- Rozwinięcie skrótów oznaczonych tyldą
  - atramẽt <atrament>
- Poprawka oczywistej literówki
  - wsziędze <wszędzie>

Próbka tekstu – Word

[POCZĄTEK·DOKUMENTU]¶

[POCZĄTEK·STRONY·TYTUŁOWEJ]¶

PIECHOTNE·CWICZENIE¶

Albo· WOIENNOSC· PIESZA· Ktorą· Láćinnicy· Pedestrem· Militiam· názywáią.· Wodzom,· Pułkownikom,· y· wszelkiey· Woienney· Stárszynie.· Lubo· ktoszkolwiek· iest· Woienney· správy·Miłośnikiem.¶

Do·wiadomośći·podána.¶

Przez·BLAZEIA·LIPOWSKIEGO¶

Nakładem·Ierzego·Forstera,·I.K.M.·Bibli·opoli.<?>¶

W·KRAKOWIE¶

W·Drukárni·v·Wdowy·Lukaszá·Kupiszá,·I.K.M.¶

Typogr.·*Roku·Panskiego·1660.¶*

[KONIEC·STRONY·TYTUŁOWEJ]¶

Próbka tekstu – XML

```
- <teiCorpus>
    <xi:include href="KORBA_header.xml"/>
  - <TEI>
      <xi:include href="header.xml"/>
    - <text>
      - <front>
          <pb/>
        - <titlePage xml:id="txt_1-titlePage">
          - <docTitle xml:id="txt_1.1-docTitle">
              <titlePart type="main" xml:id="txt_1.1.1-titlePart">PIECHOTNE CWICZENIE</titlePart>
            - <titlePart type="sub" xml:id="txt_1.1.2-titlePart">
                Albo WOIENNOSC PIESZA Ktorą Láćinnicy
                <foreign xml:lang="la" xml:id="txt_1.1.2.1-foreign">Pedestrem Militiam</foreign>
                názywáią. Wodzom, Pułkownikom, y wszelkiey Woienney Stárszynie. Lubo ktoszkolwiek iest Woienney správy Miłośnikiem.
              </titlePart>
            </docTitle>
            <note xml:id="txt_1.2-note">Do wiadomośći podána.</note>
            <docAuthor xml:id="txt_1.3-docAuthor">Przez BLAZEIA LIPOWSKIEGO</docAuthor>
          - <note xml:id="txt_1.4-note">
              Nakładem Ierzego Forstera, I.K.M. Bibli
              <uncertain reason="unreadable" xml:id="txt_1.4.1-uncertain">opoli.</uncertain>
            </note>
          - <docImprint xml:id="txt_1.5-docImprint">
              <pubPlace xml:id="txt_1.5.1-pubPlace">W KRAKOWIE</pubPlace>
              <publisher xml:id="txt_1.5.2-publisher">W Drukárni v Wdowy Lukaszá Kupiszá, I.K.M. Typogr.</publisher>
              <docDate xml:id="txt_1.5.3-docDate"> Roku Panskiego 1660.</docDate>
            </docImprint>
          </titlePage>
```

Znakowanie morfosyntaktyczne

- Przypisanie każdemu segmentowi informacji gramatycznej:
  - o jego przynależności do określonej klasy gramatycznej
  - o wartościach kategorii gramatycznych
- Umożliwia zaawansowane przeszukiwanie korpusu, np. wyszukanie wszystkich form fleksyjnych leksemu
- Znakowanie Korby będzie się odbywać na wzór znakowania NKJP
- Tagset (zbiór znaczników) będzie oparty na tagsecie NKJP, który zostanie dostosowany do potrzeb opisu polszczyzny barokowej
- Znakowanie będzie się odbywać na tekstach transkrybowanych, a nie transliterowanych

Konwersja do wersji transkrybowanej

transliteracja	transkrypcja
B. Nie záda mi żaden tey rzeczy, ktoraby **bárziey** owemu niżeli mnie należeć nie miáłá. A ták śmierć y Kupidyn pozárszy się iáko **bestie**, posnęli w Kośćiele Bacchusowym, gdźie niewyszumiawszy z przepićia śmierć Cupidinow, á Cupido śmierći należący przypasawszy sáydak do boku szły, swych należących odpráwowáć powinnośći.	B. Nie zada mi żaden tej rzeczy, któraby **barzyej** owemu niżeli mnie należeć nie miała. A tak śmierć i Kupidyn pozarszy się jako **beztie**, posnęli w Kościele Bacchusowym, gdzie niewyszumiawszy z przepicia śmierć Cupidynów, a Cupido śmierci należący przypasawszy sajdak do boku szły, swych należących odprawować powinności.

https://bitbucket.org/jsbien/pol

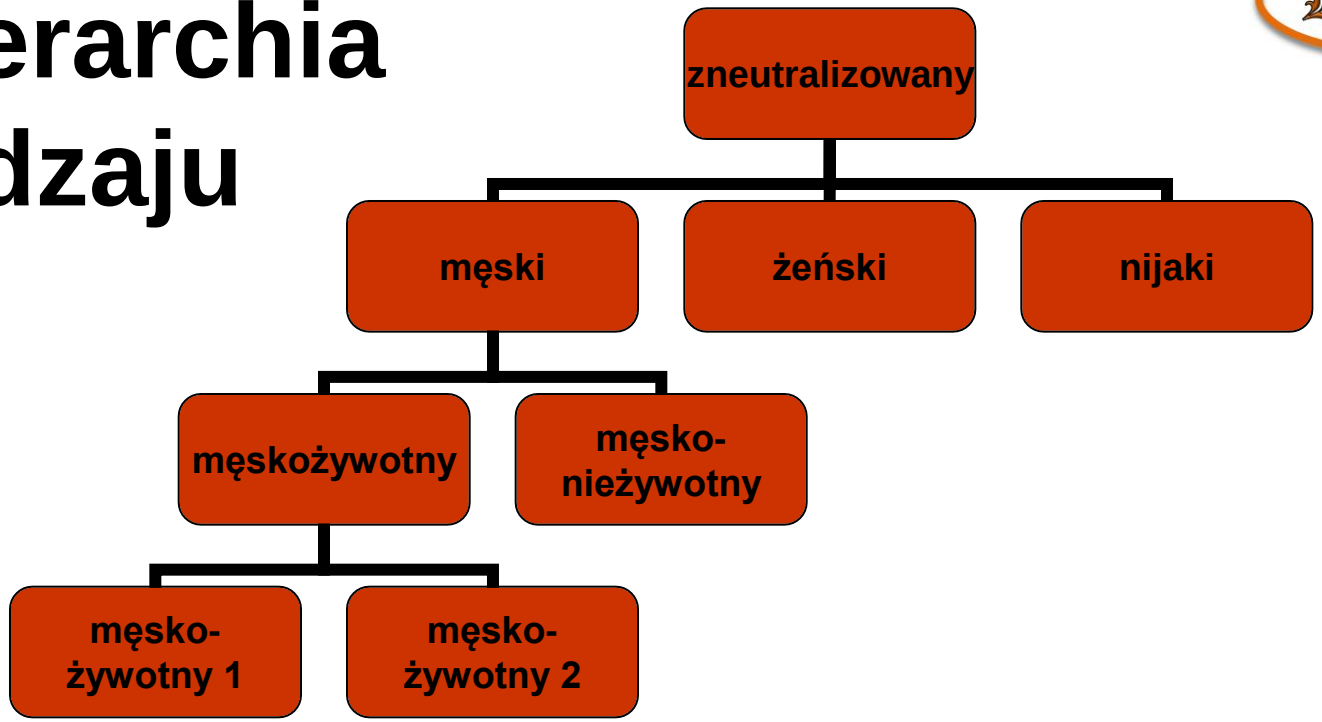
Narzędzia wspomagające znakowanie morfosyntaktyczne

- Analizator morfologiczny – przypisuje formie gramatycznej wszystkie możliwe interpretacje
- Tager – ujednoznacznia interpretację gramatyczną
- Słownik gramatyczny – zawiera paradygmaty leksemów danego języka, stanowi podstawę działania analizatora morfologicznego
- Podstawą Morfeusza – analizatora morfologicznego przeznaczonego do analizy tekstów współczesnych (używanego przy znakowaniu NKJP) – jest *Słownik gramatyczny języka polskiego* (SGJP – http://sgjp.pl)
- Podstawą Morfeusza XVII-wiecznego będzie „postarzony” SGJP

„Postarzanie” SGJP

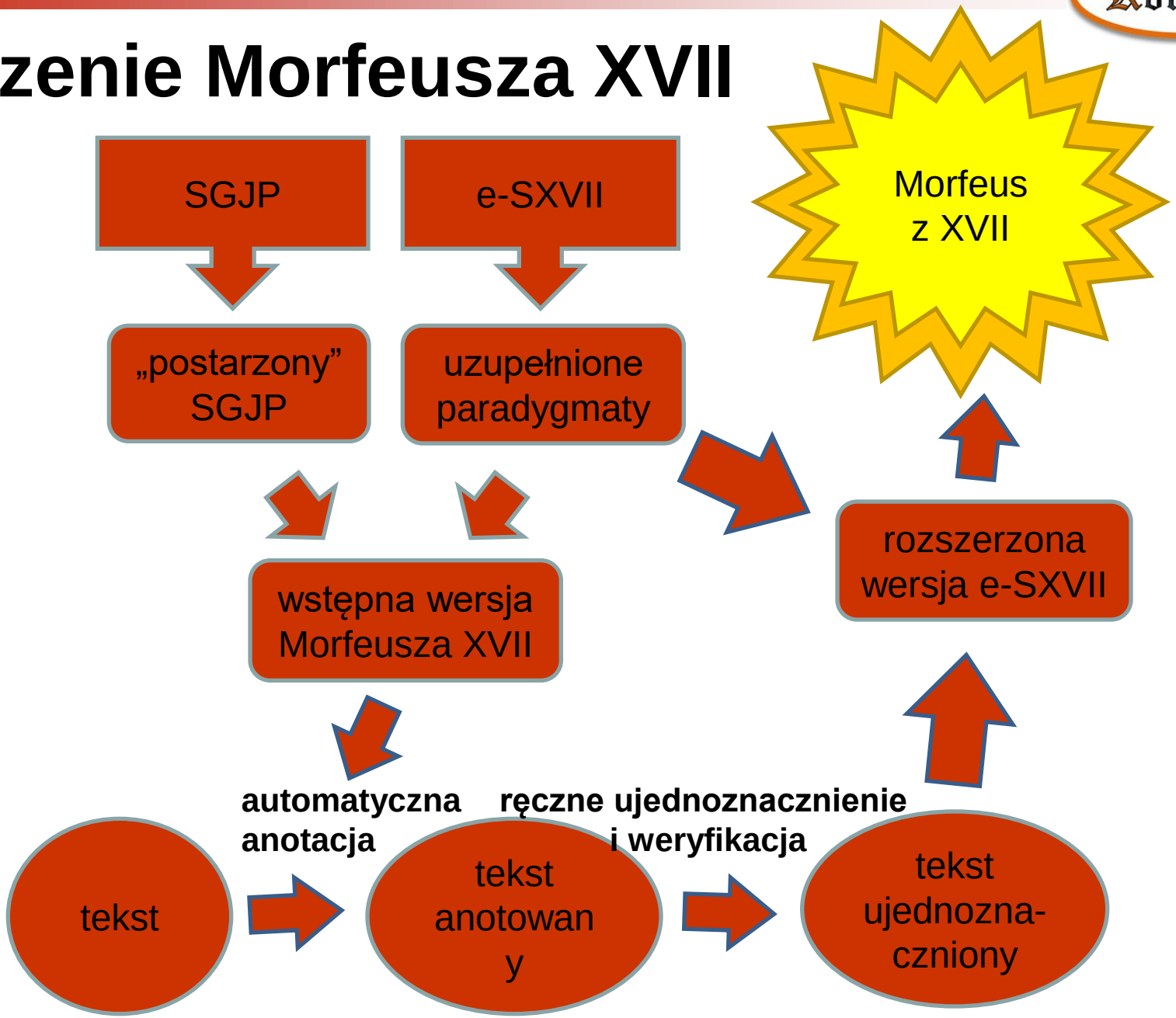
- Uzupełnienie SGJP o paradygmaty leksemów „wyciągniętych” z e-SXVII (niepełne paradygmaty zostaną automatycznie uzupełnione)
- Modyfikacja paradygmatów SGJP
  - Dodanie form historycznych do paradygmatów współczesnych, np. imiesłów przysłówkowy uprzedni zostanie dodany do paradygmatów czasowników niedokonanych (*widziawszy*)
  - Dodanie dodatkowych wartości do niektórych kategorii gramatycznych, np. liczby podwójnej (*żabie*)
  - Dodanie nowych, regularnie tworzonych derywatów, np. imiesłowów przymiotnikowych w stopniu wyższym (*bolątszy*)
  - Modyfikacja kategorii rodzaju

Hierarchia rodzaju



forma	wartości gramatyczne	lemat	rodzaj
abrysom	Dat. pl	ABRYS (?) ABRYSA (?) ABRYSO (?)	zneutralizowany
abrysu	Gen. sg	ABRYS	męski
abrys	Acc. sg	ABRYS	męskonieżywotny
abrysę	Acc. sg	ABRYSA	żeński

Tworzenie Morfeusza XVII



Dziękujemy za uwagę