European Language Resource Coordination
Connecting Europe Facility

IFiS PAN

Polszczyzna i inżynieria lingwistyczna

Autor: Marcin Miłkowski (IFiS PAN)

Polszczyzna i jej cechy szczególne

- Polszczyzną posługuje się od 40 do 48 milionów osób:
  - najczęściej używany język zachodniosłowiański na świecie;
  - język urzędowy w Polsce, ale dopuszcza się też języki mniejszości (niemiecki, białoruski, kaszubski i litewski).

- Język trudny — dla obcokrajowców i systemów przetwarzania automatycznego:
  - trudna wymowa dla obcokrajowców;
  - skomplikowana odmiana rzeczowników (9 rodzajów, bardzo dużo wzorów);
  - znaki diakrytyczne, nadal nie zawsze stosowane (Twitter);
  - swobodny szyk zdania;
  - możliwość tworzenia bardzo długich zdań (Maria Dąbrowska w „Nocach i dniach” niejednokrotnie przekracza 500 znaków!).

Polszczyzna – najnowsze tendencje

- Wpływ języka angielskiego:
  - wiele nowych terminów (*smartfon*, *selfie*, *czat*, …);
  - bezpośrednie formy adresatywne, zwłaszcza w reklamie (forma „Ty”, wykorzystanie trybu rozkazującego 2 os. l. poj., „Kupuj!”).

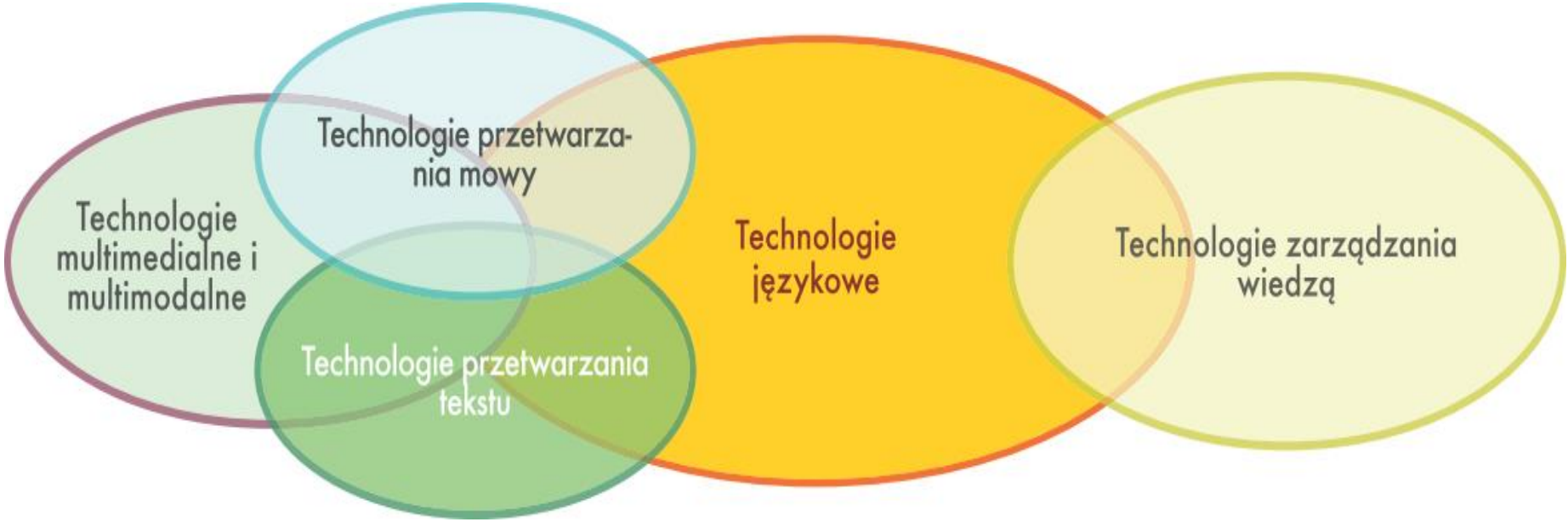
- Wpływ rzeczywistości biurokratyczno-korporacyjnej:
  - nowomowa korporacyjna (*menedżer sprzedaży, target, misja firmy)*;
  - nowomowa europejskiej biurokracji (*interesariusze, …)*

- Żeńskie nazwy zawodów:
  - *profesorka*, a nie *pani profesor;* lecz *pani biskup Helsinek*, a nie !*biskupka Helsinek*

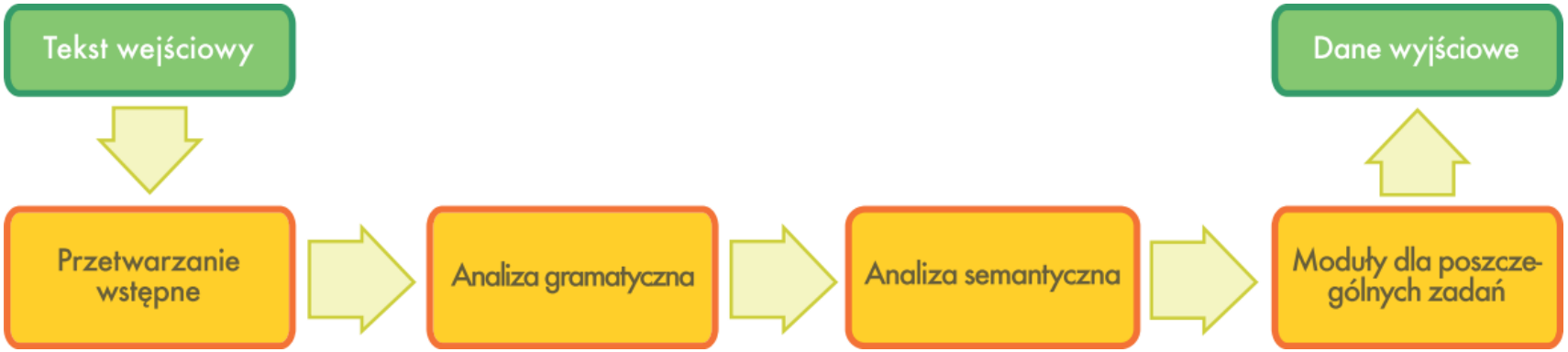
- Wpływ sieci społecznościowych i mediów:
  - szybkie rozprzestrzenianie się neologizmów, zwłaszcza ironicznych

European Language Resource Coordination
Connecting Europe Facility

Inżynieria lingwistyczna



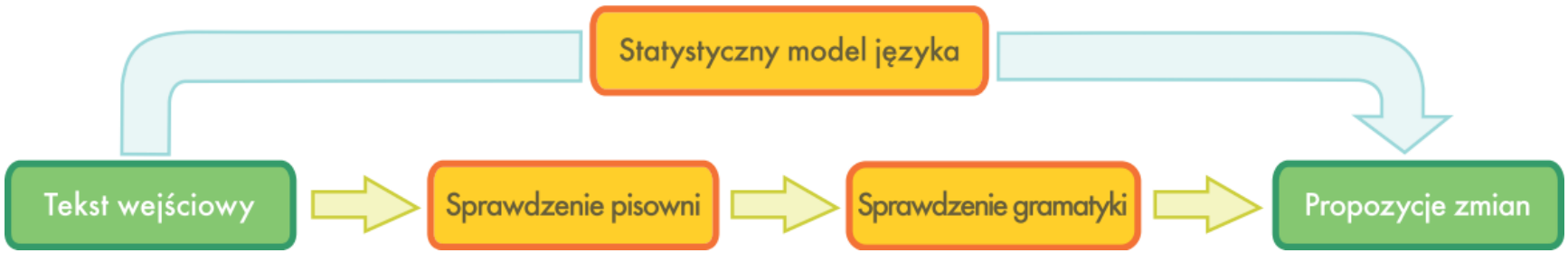
Inżynieria lingwistyczna



Przetwarzanie wstępne: normalizacja danych, usuwanie formatowania, wykrywanie języka i kodowania znaków itd.
Analiza gramatyczna: wykrywanie orzeczenia i jego dopełnień, okoliczników itd.; określanie struktury zdania.
Analiza semantyczna: ujednoznaczenie (które ze znaczeń słowa *wina* jest odpowiednie w danym kontekście?), identyfikacja okazjonalizmów (takich jak *ona, ten samochód* itp.); przedstawianie znaczenia zdania w postaci czytelnej dla komputera.

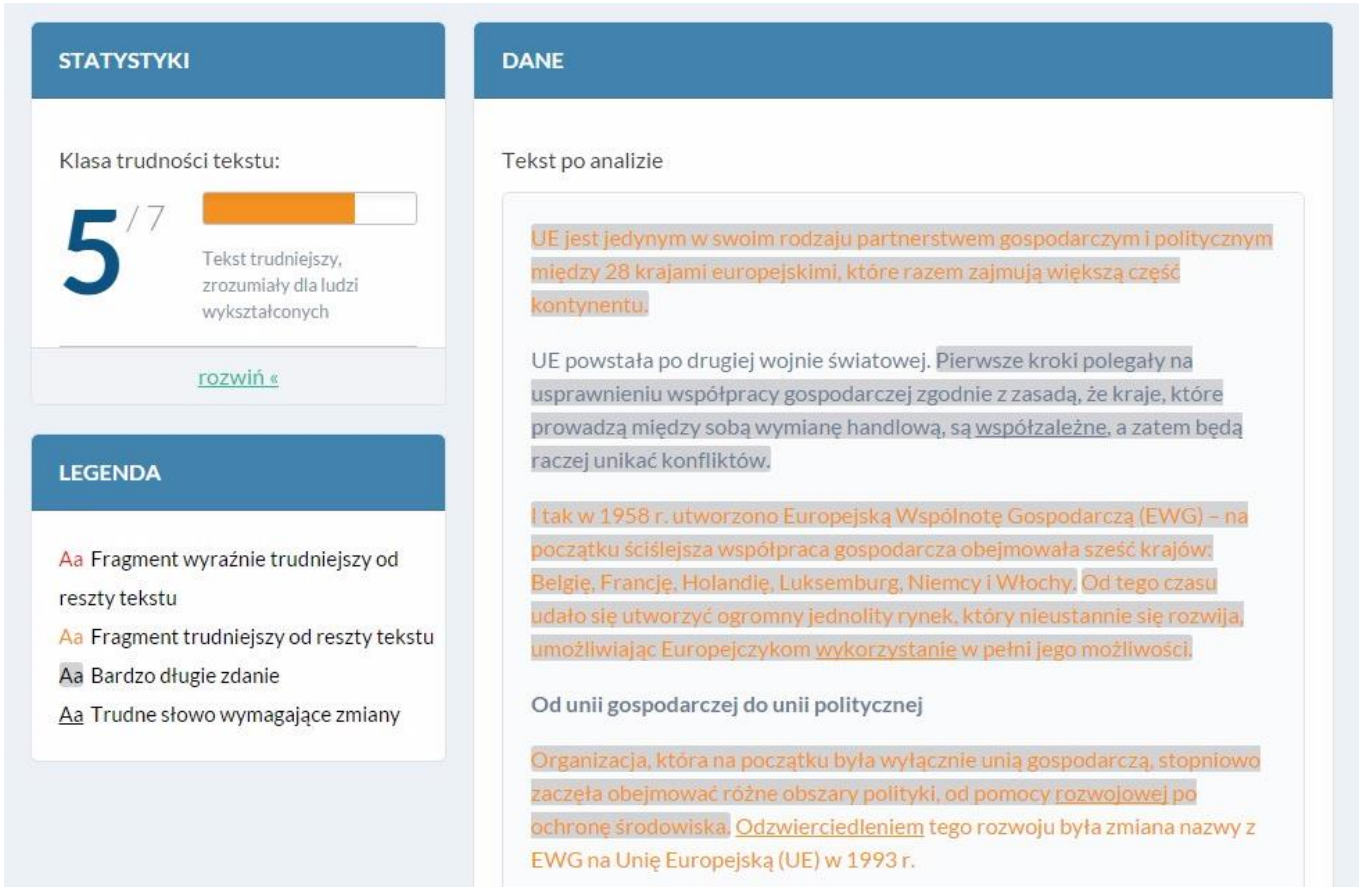
Zastosowania inżynierii w polszczyźnie

- **Korekta pisowni i korektory gramatyczne**
(komercyjne w pakiecie Microsoft Office, Google Docs; otwarte – program LanguageTool)
- **Systemy sporządzania dokumentacji technicznej i prawnej**
- **Jasnopis**



European Language Resource Coordination
Connecting Europe Facility

Jasnopis: analiza trudności tekstu



Analiza trudności tekstu z podstawowymi informacjami o Unii Europejskiej

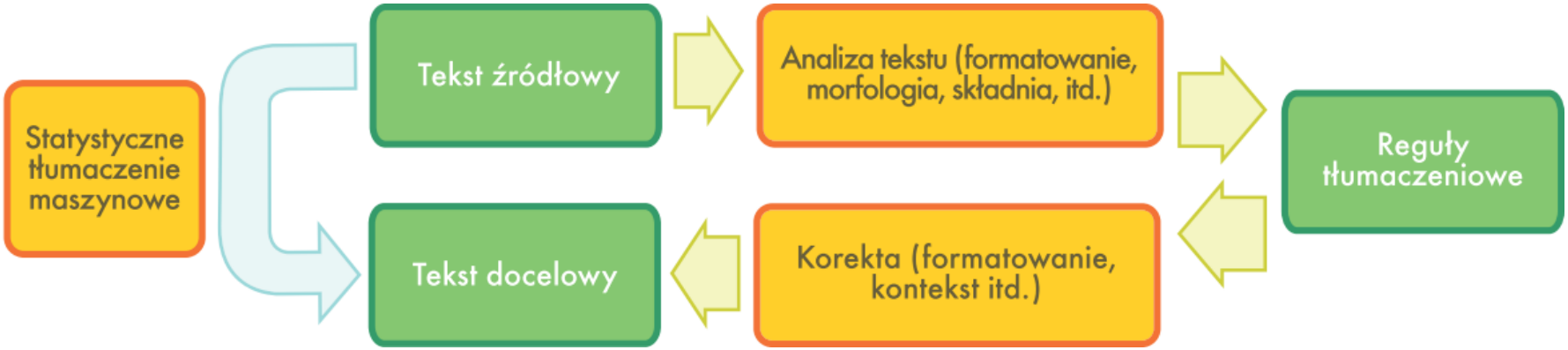
Wyszukiwarki i interakcja głosowa

- **Polskie wyszukiwarki Szukacz i Nesprint** przegrały w konkurencji z wyszukiwarką Google, która stosuje statystyczne analizy języka (i autokorektę)
- **Grupowania wyników w wielu językach dostarcza polska wyszukiwarka Carrot Search**

- **Interakcja głosowa** stosowana w smartfonach (Siri i Cortana nie obsługują polszczyzny)
- **Syntezator mowy** Ivona jest bardzo popularny
- **Systemy dialogowe** używane do obsługi klienta
- **Rozpoznawanie mowy**: Skrybot i PrimeSpeech

Tłumaczenie maszynowe

- **Translatica (system regułowy)**
- **Google Translate**
- **Bing Translate**
- **Moses (stosowany w biurach tłumaczeń)**



Inne zastosowania

- **Automatyczne streszczanie dokumentów**
  - prototyp: system Lakon
- **Boty dialogowe na stronach internetowych przedsiębiorstw**
  - Także rozrywkowe: PZU (mocnopomocni.pl)
- **Boty generujące spam**
- **Wykrywanie plagiatów, zwłaszcza prac naukowych**
  - System plagiat.pl (ekspansja na rynki poza Polską)

Bariery i możliwości rozwoju

- **Spektakularne sukcesy inżynierów lingwistycznych:**
  - polska Słowosieć, największy na świecie słownik typu WordNet, zbudowany najszybciej i dostępny na swobodnej licencji;
  - Narodowy Korpus Języka Polskiego prowadzi do powstawania nowych podstawowych narzędzi przetwarzania języka i słowników (parsery, słowniki walencyjne, systemy rozpoznawania anafory, formalne modele gramatyki)
  - wyszukiwarka NEKST