

Polszczyzna i inżynieria lingwistyczna

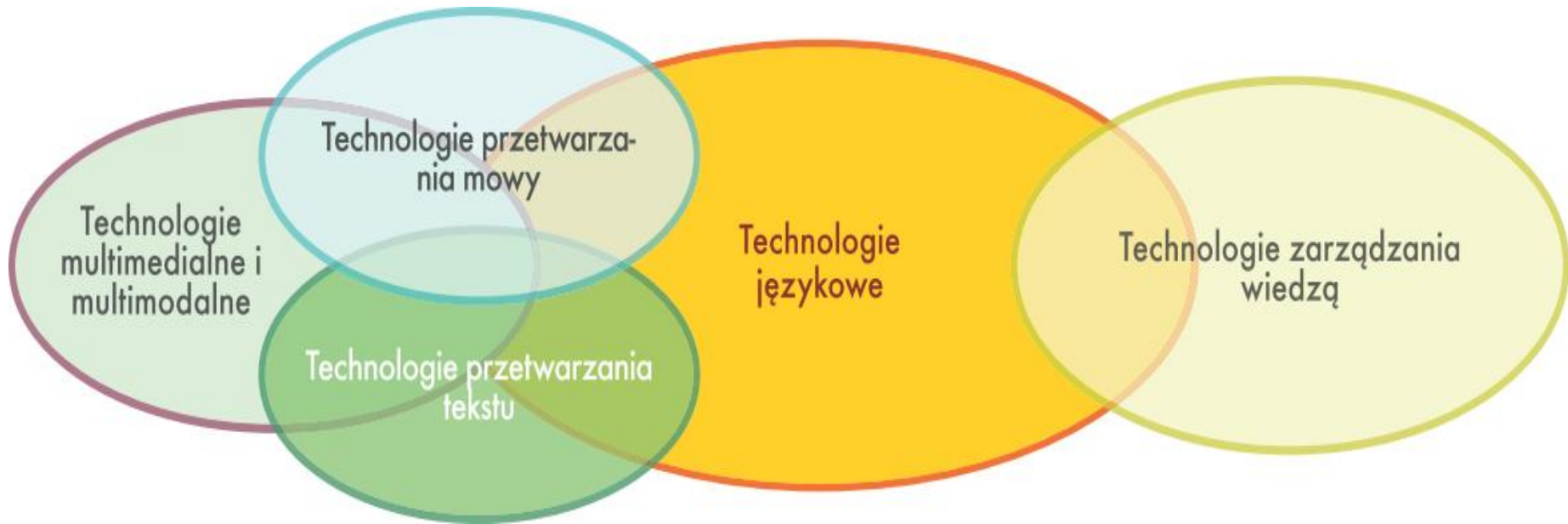
Autor: Marcin Miłkowski (IFiS PAN)

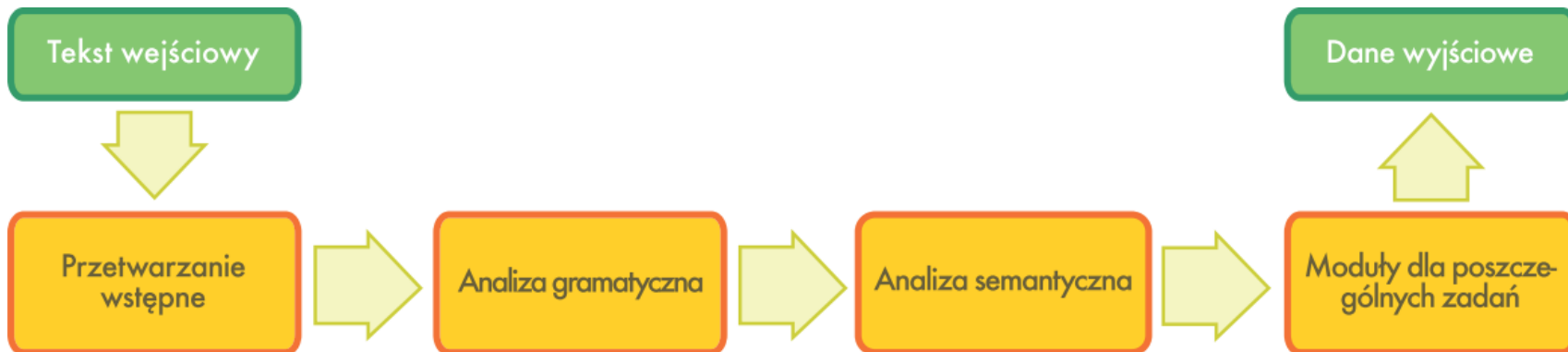


- Polszczyzną posługuje się od 40 do 48 milionów osób:
 - najczęściej używany język zachodniosłowiański na świecie;
 - język urzędowy w Polsce, ale dopuszcza się też języki mniejszości (niemiecki, białoruski, kaszubski i litewski).
- Język trudny — dla obcokrajowców i systemów przetwarzania automatycznego:
 - 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!).



- 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





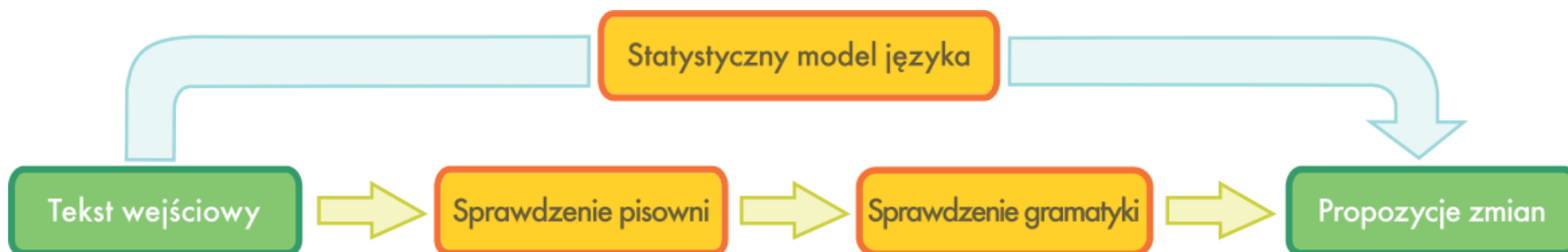
Przetwarzanie wstępne: normalizacja danych, usuwanie formatowania, wykrywanie języka i kodowania znaków itd.

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.




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



STATYSTYKI

Klasa trudności tekstu:

5 / 7 

Tekst trudniejszy, zrozumiały dla ludzi wykształconych

[rozwiń »](#)

LEGENDA

- Aa Fragment wyraźnie trudniejszy od reszty tekstu
- Aa Fragment trudniejszy od reszty tekstu
- Aa Bardzo długie zdanie
- Aa Trudne słowo wymagające zmiany

DANE

Tekst po analizie

UE jest jedynym w swoim rodzaju partnerstwem gospodarczym i politycznym między 28 krajami europejskimi, które razem zajmują większą część kontynentu.

UE powstała po drugiej wojnie światowej. Pierwsze kroki polegały na usprawnieniu współpracy gospodarczej zgodnie z zasadą, że kraje, które prowadzą między sobą wymianę handlową, są współzależne, a zatem będą raczej unikać konfliktów.

I tak w 1958 r. utworzono Europejską Wspólnotę Gospodarczą (EWG) – na początku ściślejsza współpraca gospodarcza obejmowała sześć krajów: Belgię, Francję, Holandię, Luksemburg, Niemcy i Włochy. Od tego czasu udało się utworzyć ogromny jednolity rynek, który nieustannie się rozwija, umożliwiając Europejczykom wykorzystanie w pełni jego możliwości.

Od unii gospodarczej do unii politycznej

Organizacja, która na początku była wyłącznie unią gospodarczą, stopniowo zaczęła obejmować różne obszary polityki, od pomocy rozwojowej po ochronę środowiska. Odzwierciedleniem tego rozwoju była zmiana nazwy z EWG na Unię Europejską (UE) w 1993 r.

Analiza trudności tekstu z podstawowymi informacjami o Unii Europejskiej

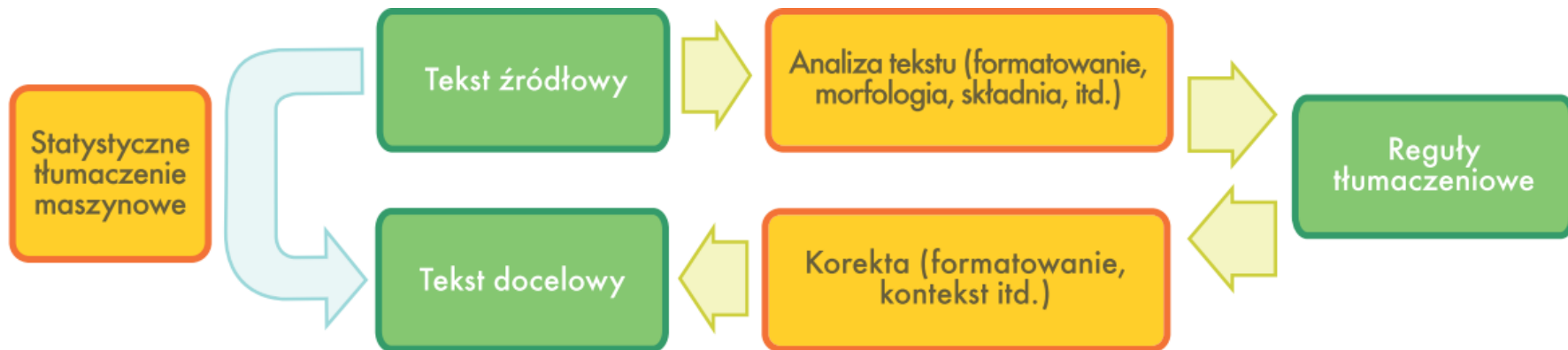


- **Polskie wyszukiwarki Szukacz i Nesprint** przegrały w konkurencji z wyszukiwarką Google, która stosuje statystyczne analizy języka (i autokorektę)
- **Grupowania wyników w wielu językach dostarcza polska wyszukiwarka Carrot Search**

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



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





- **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)



- **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