ZRK Zintegrowany Rejestr Kwalifikacji

Instytut Badań Edukacyjnych

Andrzej Walczak

Projekt ZRK, ekspert, tłumacz

Warszawa, 12 grudnia 2019

Wykorzystanie narzędzi korpusowych do tłumaczenia i inne wyzwania związane z tłumaczeniem tekstów o kompetencjach

European Funds Knowledge Education Development

Republic of Poland

ZRK Zintegrowany Rejestr Kwalifikacji

European Union European Social Fund

ZAMIAST WSTĘPU

Rem tene, verba sequentur

Opanuj treść, a słowa same się znajdą

SPECYFIKA TŁUMACZENIA TEKSTÓW O KOMPETENCJACH

Specjalistyczne słownictwo i frazeologia daleko wykraczające poza zakres dostępnych słowników dwujęzycznych

Konieczność zrozumienia tematu, którego ramy często wykraczają poza kompetencje tłumacza

Utrzymanie precyzji i jednoznaczności tłumaczenia

Zachowanie naturalnego brzmienia tłumaczonego tekstu

FRAGMENT

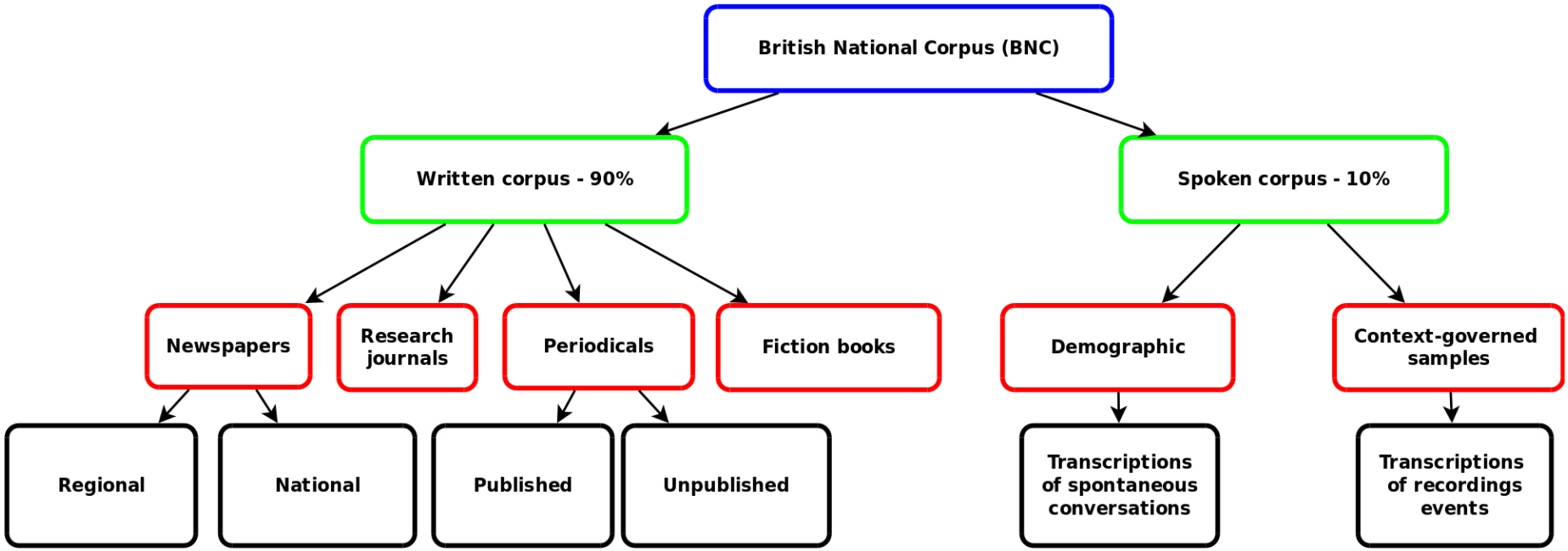
"Absolwent technikum kształcącego w zawodzie technik górnictwa odkrywkowego powinien być przygotowany do wykonywania następujących zadań zawodowych: 1) wykonywania robót związanych z odwadnianiem górotworu i zwałowisk, 2) wykonywania robót związanych z udostępnianiem i urabianiem złoża, 3) wykonywania robót związanych z transportem nakładu i kopaliny."

KORPUSY JĘZYKOWE

Korpus (ang. corpus, z łac. corpus „ciało”) – zbiór tekstów służący badaniom lingwistycznym, np. określaniu częstości występowania form wyrazowych, konstrukcji składniowych, kontekstów, w jakich pojawiają się dane wyrazy. [wikipedia]

Przemyślany zbiór autentycznych tekstów w postaci cyfrowej.

STRUKTURA BNC – KORPUS ZRÓWNOWAŻONY



ANGLOJĘZYCZNE KORPUSY ONLINE

WebCorp Live - concordance the web in real-time.

http://www.webcorp.org.uk/live/

British National Corpus (BNC)

https://www.english-corpora.org/bnc/

Corpus of Contemporary American English (COCA)

https://www.english-corpora.org/coca/

NARZĘDZIA SŁOWNIKOWE DZIAŁAJĄCE W OPARCIU O KORPUSY

- diki.pl
- linguee.pl
- glosbe.com
- bab.la
- context.reverso.net
- ludwig.guru

NARZĘDZIA DO TŁUMACZENIA MASZYNOWEGO DZIAŁAJĄCE W OPARCIU O KORPUSY

- deepl.com
- translate.google.com
- bing.com
- worldlingo.com

PROGRAMY WSPOMAGAJĄCE TŁUMACZENIE (CAT)

Programy "stacjonarne":

- Trados Studio
- Wordfast
- memoQ
- OmegaT (darmowy) ...

Programy online:

- MateCat
- Wordfast
- Smartcat

Wszystkie wykorzystują pamięci tłumaczeń TMX, listy terminologiczne

WADY KORPUSÓW Z PUNKTU WIDZENIA TŁUMACZA

Informacje mają charakter ilościowy - brak komentarzy, definicji - wnioski trzeba wyciągać samodzielnie lub/i weryfikować w innych źródłach

Pokazują jak język wykorzystywany jest w praktyce, ale nie gwarantują poprawności – również rodzimi mówcy popełniają błędy - konieczne jest zachowanie krytycyzmu

Wiarygodność zależy w dużym stopniu od obszerności, jakości technicznej i merytorycznej zgromadzonych danych

JAK ZBUDOWAĆ WŁASNY SPECJALISTYCZNY KORPUS

Zebranie możliwie wielu autentycznych tekstów związanych z danym tematem (najlepiej w postaci "czystych" plików tekstowych, zakodowanych w systemie UTF-8) - teksty mogą pochodzić z elektronicznych publikacji (np. pdf), baz danych, arkuszy, z serwisów internetowych etc.

Krytyczna, techniczna analiza zebranych tekstów (tekst wyświetlany jest poprawnie, słowa nie są podzielone w przypadkowych miejscach), czy nazwy plików mają przemyślane nazwy

Opcjonalne tagowanie tekstów - spowalnia korzystanie z korpusu, ale podnosi jego użyteczność

PRAWNE ASPEKTY (WEB-SCRAPING)

Analiza pliku robots.txt

```
User-Agent: *
Disallow: /search/quicksearch.html
Disallow: /search/find.html
Disallow: /search/text/academy.html
Allow: /
```

Uważna lektura "Terms and conditions"

- wykorzystanie niekomercyjne
- nie wolno udostępniać osobom trzecim

ŹRÓDŁA TEKSTÓW NA TEMAT KOMPETENCJI - I

Strony rządowe / samorządowe / stanowe itp. na temat zatrudnienia, rynku pracy, ścieżek kariery, doradztwa zawodowego:

- O*NET (USA) (1100 opisów)
- Jobs and Skills (Australia) (564 opisy)

Strony firm/organizacji zajmujących się rekrutacją personelu:

- America's Job Exchange (USA) (737 opisów)
- Betterteam (USA) (934 opisy)

ŹRÓDŁA TEKSTÓW NA TEMAT KOMPETENCJI - II

Strony dla pracowników pomagające przygotować CV

- Great Sample Resume (USA) (643 opisy)

Strony oferujące doradztwo w zakresie kariery

- Inputyouth (UK) (958 opisów)
- Prospects (UK) (473 opisy)
- Target Jobs (UK) (348 opisów)

ŹRÓDŁA TEKSTÓW NA TEMAT KOMPETENCJI - III

Specjalistyczne strony branżowe:

- Media Match (USA) (208 opisów)
- Top Accounting (USA) (226 opisów)

Strony oferujące pomoc w poszukiwaniu pracowników

- Workable (USA) (958 opisów)

DARMOWE NARZĘDZIA KORPUSOWE

Lawrence Anthony software:

https://www.laurenceanthony.net/software.html

AntConc – corpus analysis toolkit

AntGram – n-gram and p-frame generation tool

TagAnt – Part-Of-Speech (POS) tagger

DO CZEGO PRZYDAJE SIĘ KORPUS W PRACY TŁUMACZA

Etap wstępny - eksploracja korpusu, poznanie jego specyfiki (słownictwo, frazeologia, styl, specyfika języka)

Właściwe tłumaczenie – sprawdzanie słów, związków frazeologicznych w kontekście, szukanie synonimów, analiza specyfiki wykorzystania związków frazeologicznych w różnych odmianach języka, analiza przypadków wątpliwych, weryfikacja przyjętych rozwiązan

WYKORZYSTANIE KORPUSU DO EKSPLORACJI TEKSTU

- Jakie słowa najczęściej występują w korpusie? (wordlists, stopwords)
- Jakie związki frazeologiczne często występują w korpusie? (n-grams)
- Jakie słowa wyróżniają dany tekst na tle innych, podobnych w grupie? (keywords)

WYKORZYSTANIE KORPUSU W TŁUMACZENIU

- "Sanitary installation" - czy takie połączenie w ogóle istnieje?
- Czy we względnie formalnych tekstach mogę wykorzystać czasownik frazowy, np. *carry out*?
- Czy, jeśli istnieje rzeczownik dobrze opisujący daną czynność, mogę wykorzystać formę gerundialną (pasuje do wyliczenia), np. zamiast *assembly -> assembling*, zamiast *construction -> constructing, design -> designing*?
- Wiem, że mogę napisać *drawing up project plans,* ale nie przychodzi mi do głowy jakieś inne, bardziej formalne słowo.
- W tłumaczeniach pojęcia *kopalnia odkrywkowa* mam m.in. *open-pit*, *opencast* – czy rzeczywiście nie ma między nimi różnicy?
- Szukam czasowników, które brzmiałyby naturalnie ze słówem *budget*

WILDCARDS – SYMBOLE WIELOZNACZNE

- Gwiazdka (*) - brak znaku, jeden lub więcej znaków
- Plus (+) - brak znaku lub jeden znak
- Znak zapytania (?) - jeden dowolny znak
- Krzyżyk (#) - jakiekolwiek słowo
- "Małpa" (@) - jedno dowolne słowo lub brak słowa
- Pionowa kreska (|) - alternatywa, albo-albo

refer*	referred, reference, referee, referring ...
refer+	refer, refers
communicate # with	communicate directly with, communicate clearly with etc.
work* # with	

ZRK Zintegrowany Rejestr Kwalifikacji

"Operating and developing the Integrated Qualifications Register"

Project co-financed by the European Social Fund of the European Union

Instytut Badań Edukacyjnych/Educational Research Institute
IQS Project Office
Górczewska 8, 01-180 Warsaw, Poland
phone: +48 22 24 17 100, +48 22 24 17 111
e-mail: rejestr@ibe.edu.pl

http://rejestr.kwalifikacje.edu.pl | http://www.ibe.edu.pl

European Funds Knowledge Education Development
Republic of Poland
ZRK Zintegrowany Rejestr Kwalifikacji
European Union European Social Fund