Obrady 12. Konferencji na temat Zasobów Językowych i Oceny (LREC 2020), strony 3782-3789
Marsylia, 11-16 maja 2020 r.

OpusTools and Parallel Corpus Diagnostics

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, Jorg Tiedemann
Department of Digital HumanitiesUniversity
of Helsinki, Helsinki/Finlandia{
mikko.aulamo, umut.sulubacak, sami.virpioja, jorg.tiedemann}@helsinki.fi

Streszczenie

Artykułten wprowadza OpusTools, pakiet do pobierania i przetwarzania równoległych ciał włączonych do kolekcji OPUS corpus.Pakiet implementuje narzędzia dostępu do skompresowanych danych w zarchiwizowanym formacie wydania i umożliwia łatwą konwersję między pospolitymi formatami.OpusTools zawiera również narzędzia do identyfikacji języków i filtrowania danych, a także narzędzia do importowania danych z różnych źródeł do formatu OPUS.Pokazujemy wykorzystanie tych narzędzi w równoległym tworzeniu korpusu i diagnostyce danych.Ta ostatnia jest szczególnie przydatna do identyfikacji potencjalnych problemów i błędów w obszernym zbiorze danych.Korzystając z tych narzędzi, możemy teraz monitorować ważność zbiorów danych i poprawić ogólną jakość i spójność gromadzenia danych.

Słowa kluczowe: Korpus (stworzenie, adnotacja itp.);Tłumaczenie maszynowe;Narzędzia, systemy, aplikacje

1. Wprowadzenie

Opus (Tiedemann, 2012) to największa kolekcja powszechnie dostępnych korpusów równoległych.Kolekcja stale się rozwija z biegiem lat i jest szeroko wykorzystywana w pracach nad tłumaczeniami maszynowymi i badaniami międzylingwistycznymi.Obecnie zawiera 57 wydanych korpusów obejmujących ponad 700 języków i wariantów językowych tworzących ponad 70 000 bitekstów w sensie dopasowanych par językowych we wszystkich korpusachw kolekcji.Rozmiar i popularność OPUS sprawia, że konieczne jest zbudowanie efektywnej infrastruktury umożliwiającej różnym użytkownikom uzyskanie i dostęp do danych, a niniejszy dokument wprowadza dwa pakiety, które dostarczają narzędzi do tego celu.Celem tych pakietów jest ułatwienie pobierania, konwersji i przetwarzania danych zawartych w OPUS z linii poleceń lub z aplikacji za pomocą biblioteki wdrażającej te narzędzia.Oba pakiety odnoszą siędo biblioteki Pythona z narzędziami wiersza poleceń i uzupełniającym-modułem Perla, zarówno udostępnianym jako open source, jak i z pobłażliwymi licencjami.

W poniższych sekcjach przedstawiamy narzędzia i ich podstawowe zastosowanie, a także omówimy, w jaki sposób zastosowaliśmy te narzędzia do tworzenia nowych zbiorów danych i prowadzenia systematycznej diagnostyki całej bazy danych.Dzięki dostępności OpusTools można teraz przeprowadzić dokładne kontrole rozległych zbiorów danych w celu weryfikacji ważności kodowania, znalezienia uszkodzonych linków i struktur oraz identyfikacji innych problemów z danymi.

2. Charakterystyka OPUS

Opus zawiera korpus równoległy z wielu różnych źródeł.Każdy z nich ma swoje cechy charakterystycznei właściwości mogą się znacznie różnić w zależności od oryginalnych danych i ich dystrybucji.Filozofią OPUS jest utrzymywanie marży i adnotacji w jak największym stopniu,-ale ujednolicenie niezbędnego formatu danych, aby dostęp do równoległych danych był jak najbardziej przejrzysty.Oznacza to, że dane Corpus są konwertowane do samodzielnego (wolnego od schematów) XML, który zachowuje oryginalne znaczniki, ale konsekwentnie dodaje niezbędne znaczniki, które są niezbędne do wyrównania i dalszegoprzetwarzania językowego.Wyrównanie jest zapisywane jako adnotacja standoff w formacie XCES Align (dla wyrównania zdania) i „Format Mojżesza" (dla wyrównania słów).Stosując tę zasadę, dane mogą

być przechowywane niezależnie od adnotacji dostosowania, co pozwala na skuteczne wdrażanie i przechowywanie masowo równoległych danych, a także, w razie potrzeby, umożliwia alternatywne dostosowanie.Rysunek 1 pokazuje przykład adnotacji standoff używanej w OPUS do określania powiązań między zdaniami.Każdy plikdopasowania zdania może zawierać dowolną liczbę elementów linkGrp wcelu wyrównania dokumentów z gromadzenia danych.Dokumenty są określane za pomocą ścieżki w stosunku do pierwiastka XML podkopasu OPUS ielementy linku zapewniają wyrównanie zdania zestawami identyfikatorów zdania, które są oddzielone średnikiem.Tworzenie alternatywnego wyrównania odbywa się po prostu poprzez stworzenie nowegopliku wyrównania zdań i nie ma konieczności dokonywania dalszych modyfikacji z oryginalnymi danymi corpus.Zauważ, że wyrównanie zdania jest dwujęzyczne, jak pokazano w przykładzie.Jednakże adnotacja standoff- umożliwia wyrównanie masywnie równoległych zbiorów danych we wszystkich parach językowych bez powielania żadnego z powiązanych plików danych.Ponadto, mogą istnieć alternatywne pliki corpus o różnych poziomach adnotacji bez konieczności ponownego dopasowania tych alternatywnych plików.Rysunek 2 pokazuje przykłady takich notowanych plików, wszystkie wyrównane w ten sam sposób z wyrównaniem zdania standoff przechowywanego w plikach zewnętrznych.Więcej szczegółów na temat struktury danych w OPUS można znaleźć na OPUS Wiki.[1]

Inną zasadą w OPUS jest dostarczanie danych w innych wspólnych formatach, tak aby były one łatwo dostępne dla szerokiego zakresu zastosowań.Te formaty danych są jednak generowane z bazowego kodowania opartego na XML, które służy jako kopia nadrzędna każdego korpusu.Użytkownicy danych OPUS zazwyczaj nie są świadomi tych zasad i pobierają format danych, który najbardziej odpowiada ich potrzebom.

Ideą OpusTools jest teraz ujednolicenie dostępu do danych podstawowych w XML i do innych generowanych formatów poprzez dostarczaniepodstawowych bibliotek i narzędzi wiersza poleceń do pobierania i konwersji danych corpus.Zapewniają one również wygodne narzędzia do podstawowego filtrowania i losowego dostępu do archiwizowanych danych w ich skompresowanej formie, która jest wykorzystywana do dystrybucji danych.To ostatnie jest szczególnie ważne, ponieważ rozmiary niektórych cor—

[1] Http:// opus.nlpl.eu/trac/wiki/ Dataformats

```
<?xml version=„1.0” kodowanie=„utf-8”?>
<!DOCTYPE cesAlign PUBLIC
                    „-//CES//DTD XML cesAlign//EN” „”> <cesAlign
version=„1.0”>
<linkGrp targType=”
          odDoc=„en/0/1089124/4995691.xml.gz”
          toDoc=„fr/0/1089124/4588599.xml.gz”>
<link id=„SL0” xcelets=„1;1” nakładanie się=„0.331”/>
<link id=„SL1” xcelets=„2 3;2” nakłada się=„0.560”/> <link id=„SL2”
xcelets=„4;”/>
<link id=„SL3” xcelets=„5 6;3” nakładanie się=„0,854”/> <link id=„SL4”
xcelets=„7 8 9;4” nakładanie się=„0.699”/> <link id=„SL5” xcelets=„10
11;5” nakładanie się =„0.776”/>
```

Rysunek 1:Przykład wyrównania zdania w formacie XCES Align.Element linkGrp określa pary dokumentów, które są wyrównane i linki pomiędzy poszczególnymizdaniami są podane welementach linku.Opcjonalne nakładanie się atrybutów w tym przykładzie odnosi się do współczynników nakładania się czasu, które są używane jako funkcja w wyrównaniu napisów.

Pora jest rozległa w taki sposób, że wymaga, aby wspólne systemy plików przetwarzały dane w surowej, nieskompresowanejformie.Na przykład, najnowszy opensubtitles corpus zawiera około 3,7 miliona pojedynczych dokumentów w 67 językach z dopasowaniem w ponad 3 600 bitextów.Jeden z najnowszych dodatków, JW300 obejmuje 380 języków w ponad 46 bitextów.Łącznie istnieje ponad 9,2 mlndokumentów indywidualnych tylko w ostatnich wydaniach wszystkich korpusów i liczba ta jest podwojona przez różne rodzaje wstępnego przetwarzania, które są dostarczane, surowego tekstu i korpusu tokenized, które są częściowo opatrzone dodatkowymi informacjami językowymi.Ponadto, bitexts są wydawane w natywnym formacie XML(patrz Rysunek 2), zwykłym formacie tekstowym i formacie wymiany pamięci tłumaczeniowej-(TMX).Obecnie wydawnictwa zajmują łącznie 5,9 TB miejsca w skompresowanym formacie.

Powyższe liczby ilustrują potrzebę odpowiedniej infrastrukturyi skutecznych narzędzi do zarządzania różnymi zbiorami danych.Jest to motywacja do wdrożenia swobodnie dostępnych narzędzi OPUS opisanych poniżej.Tworzą one wygodną bibliotekęi skrzynkę narzędziową do pobierania, ekstrakcji i konwersji danychz kolekcji OPUS.Dodatkowo pomagają one w prowadzeniu systematycznej diagnostyki zbiorów w celu identyfikacji błędów i problemów w zbiorach danych.Poniżej przedstawimy najpierw oba pakiety i ich funkcjonalność.Następnie przekazujemy informacje na temat ich wykorzystania w tworzeniu nowych zbiorów danych i wreszcie informujemy o stosowaniu narzędzi OPUS do badań diagnostycznych i kontroli stanu zdrowia psychicznego.

3. Pakiet OpusTools

Pakiet OpusTools jest zestawem narzędzi do pobierania i zarządzania danymi z korpusu równoległego z OPUS.Pakiet-składa się z biblioteki Pythona i powiązanych skryptów wiersza poleceń.Dodatkowo istnieje pakiet Perl do tworzenia nowych zbiorów danych i dostępu do danych równoległych.

3.1. Narzędzia linii poleceń

Pakiet OpusTools zawiera pięć skryptów linii poleceń opartych na Pythonie 3:Opus_read, opus_express, opus_cat, opus_get i opus_langid.2[Skrypty] umożliwiają pobieranie danych [2]

OPUS, przesyłanie danych w określonych formatach, ekstrahowanie zestawówszkoleniowych, programistycznych i testowych z danych i innych.Rysunek 3 przedstawia przegląd scenariuszy.

opus_read jest skryptem do pobierania równoległych corpora i konwersji ich na pożądane formaty.Opus corpora zawiera pliki wyrównania formatu XCES wskazujące na dwa pliki zdań XML w różnych językach.Format wyrównania XCES łączy zdania w plikach źródłowych z zdaniami w plikach docelowych przy użyciu ID zdania. Pliki zdania w OPUS corpora są skompresowane do archiwów ZIP i opus_read ułatwia odczytanie danych bezpośrednio z skompresowanych plików. opus_read przetwarza dany plik wyrównania i tworzy wyjście w jednym z czterech formatów:normalne, moze, TMX lub XCES linki. opus_read najpierw próbuje odczytać pliki OPUS z lokalnych katalogów.Jeśli wymagane pliki nie zostaną znalezione, narzędzie oferuje możliwość ich pobrania.Pliki zdania można pobrać w formacie surowym, tokenizowanym lub parsowanym.

opus_read zawiera podstawowe filtry do usuwania niechcianych par zdań przed utworzeniem pliku wyjściowego.Nonalignments, gdzie segment źródłowy lub docelowy jest pusty, można pominąć.Alternatywnie, można określić pewną liczbę segmentów źródłowych i docelowych, np. możliwe jest włączenie tylko jednego do jednego wyrównania w wyjściu.Niektóre korpusy zawierają wynik atrybutu dla każdej pary zdań.Na przykład, pary zdań w opensubtitles corpus nakładają się na siebie wyniki, które wskazują, w jakim stopniu znaczniki czasowe obu segmentów się pokrywają.opus_read jest w stanie odfiltrować pary zdań, które nie przekraczają danego progu punktów atrybutu.Ponadto pary segmentów mogą być usuwane napodstawie wyników identyfikacji językowej.Etykiety językowe i wyniki ufności mogą być dodawane do zdania plikówXML za pomocą skryptu opus_langid.

opus_express jest skryptem zbudowanym na opus_read, który może wyodrębnićgotowe do użycia zestawy treningowe, rozwojowe i testowe dla pary językowej z jednej lub więcej corpora OPUS.Proceduranajpierw wypełnia określoną ilość zdań dla zestawu testowego, następnie kontynuuje to samo dla zestawu programistycznego, a resztę wyrzuca do zestawu treningowego.Skrypt może opus_express opcjonalniewstępnie przetasowywać dane przed podziałem lub odwrotnie oznaczać i zachować granice dokumentów przez podziały dla modeli na poziomie dokumentu. opus_express zawiera równieżopcję wykorzystania wyników atrybutów, takich jak nakładanie się wartości, które są ekstrahowane przez opus_read w jego przełączaniu na jakość-świadomość, którapriorytetowo traktuje pary zdań wyższej wiarygodności, przekraczające konfigurowalny prógdo sortowania w zestawach testu i rozwoju.

opus_cat jest używany do odczytu jednojęzycznego korpusu z OPUS lub pojedynczych plików w tych korpusach.Pliki mogą być drukowane w formacie XML lub mogą być konwertowane na zwykły tekst. opus_cat jest przydatny do ręcznego sprawdzania domeny lub jakości pojedynczego korpusu, ponieważ jest w stanie odczytać pliki bezpośrednio z archiwów ZIP w corpora OPUS.

opus_get jest skryptem do pobierania równoległych plików corpus z OPUS.Przed pobraniem, corpora może być przeszukiwana i wymieniana według nazwy, języka źródłowego i języka docelowego.Na przykład, można pobrać pliki dla określonej pary językowej w jednym korpusie, wszystkie pliki par językowych w

[2] https: //github.com/Helsinki-NLP/OpusTools

Surowy format XML:

```
<?xml version=„1.0” kodowanie=„utf-8”?>
<dokument>
  <CHAPTER ID="1">
    <P id="1">
      <s id="1">Wznawianie sesji </s>
    </P>
    <SPEAKER ID="1"Nazwa=„Prezes”>
      <P id="2">
        <s id="2">Deklaruję wznowienie sesji Parlamentu Europejskiego odroczonej w czwartek, 14 czerwca 2001 r. </s> </P>
```

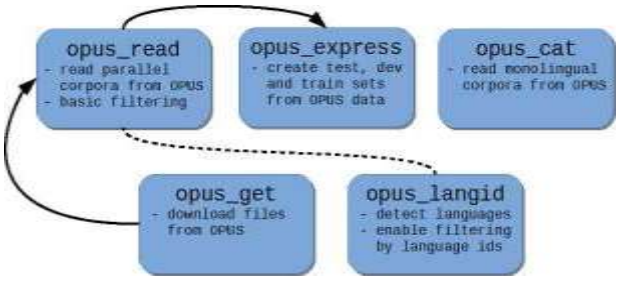
Tokenizowany (notowany) format XML:

```
<?xml version=„1.0” kodowanie=„utf-8”?>
<dokument><CHAPTER ID="1"><P id="1">
<s id="1">
<chunk type=„NP” id=„c-1”>
  <w hun= „NN” tree= „NN” lem= „resumption” pos= „NN” id=„w1.1”>Wrócenie</w>
</chunk>
<chunk type=„PP” id=„c-2”>
  <w hun= „IN” tree=„IN” lem=„of” pos=„IN” id=„w1.2”>z</w>
</chunk>
<chunk type=„NP” id=„c-3”>
  <w hun= „DT” tree=„DT” lem=„the” pos=„DT” id=„w1.3”>the</w>
  <w hun= „NN” tree= „NN” lem= „sesja” pos= „NN” id=„w1.4”>sesja</w>
</chunk>
</s>
```

UD Parsed XML format:

```
<?xml version=„1.0” kodowanie=„utf-8”?>
<dokument>
<CHAPTER ID="1">
  <P id="1">
    <s id="1">
      <w xpos=„NOUN” head="0" feats=„Number=Sing” UPOS=„NOUN” lemma=„Resumption” id=„1,1” deprel=„Root”>Resumption </w> <w xpos=„ADP”
      head=„1,4”UPOS=„ADP” lemma=„w” id=„1,2”
      <w xpos= "DET" head="1.4" feats="Definite=Def|PronType=Art" UPOS="DET" lemma="Number="1,3" deprel="det"> </w> <w xpos=„NOUN” head=„1,1”
      feats=„Number=”Sing„UPOS=”NOsc>
    </s>
```

Rysunek 2:Przykłady danych XML zakodowanych w OPUS.Różne rodzaje adnotacji mogą być dodawane bez niszczenia wyrównania zdania, który jest przechowywany jako adnotacja standoff linków pomiędzy identyfikatorami zdania.



Rysunek 3:Pięć skryptów opartych na Pythonie OpusTools.Każdy ze skryptów może być używany indywidualnie. opus_express jest zbudowany na opus_read, a opus_read wykorzystuje opus_get do pobierania plików OPUS. opus_langid musi być stosowany doplików zdania, aby umożliwić filtrowanie id języka dla opus_read.

jeden corpus lub wszystkie pliki dla określonego języka w całym OPUS. opus_read używa opus_get do automatycznego pobieraniażądanych plików corpus.

opusJangid jest używany do dodawania etykiet identyfikujących języki punktów ufności dla każdego zdania w danym pliku zdania XML.Identyfikacja językowa odbywa się za pomocą dwóch narzędzi:Pycld2, wiązania[3]Pythona dla detektora języka kompaktowego 2 i[4] langid.py (Lui i Baldwin, 2012).opus_langid musi być stosowany do zdania plików XML zanim opus_read może filtrować paryzdań przez etykiety językowe.Rysunek 4 pokazuje przykład pliku zdania, który został przetworzony z opus_langid.

3.2. Biblioteka Pythona OpusTools

Oprócz tego, że są skryptami linii poleceń, opus_read, opus_cat, opus_get i opus_langid są powiązane z modułami Pythona, które mogą być importowane i używane we własnych skryptach.Moduły zapewniają taką samą funkcjonalnośćjak narzędzia linii poleceń, a także bardziej szczegółowe zarządzanie danymi poprzez wykorzystanie podmodułów i funkcji.Wszystkie kody Pythona są napisane w Pythonie 3.

Moduł OpusRead można inicjować parametrami odpowiadającymi flagom podanym do opus_readisłuży do pobierania i konwersji plików corpus z OPUS.WEwnętrznie OpusRead wykorzystuje moduły analizujące XMLz podbiblioteki parse zawartej w pakiecie Opus- Tools Python.Podbiblioteka zawiera moduły do przetwarzania plików wyrównania XCES i plików zdań.W

[3]https://github.com/aboSamoor/pycld2
[4]https://github.com/CLD2Owners/cld2

```
<?xml version=„1.0” kodowanie=„utf-8”?>
<tekst>
<p id="1">
    <s cld2= „en” cld2conf=„0.99” id=„s1.1” LangID=„en” langidconf=„1.0”>
     Oświadczenie premiera Ingvara Carlssona o polityce rządu podczas otwarcia szwedzkiego parlamentu we wtorek, 4 października 1988 r.
    </s>
</p>

<p id="2">
    <s cld2= „en” cld2conf=„0.98” id=„s2.1” LangID=„en” langidconf=„1.0”>
     Wasza Wysokość, Wasze Królewskie Wysokości, Panie Przewodniczący, Posłowie do szwedzkiego parlamentu.
    </s>
</p>
```

Rysunek 4:Przykład pliku zdań, w którym etykiety językowe i wyniki pewności siebie zostały dodane do znaczników zdania.

Moduł AlignmentParser analizuje dany plik łącza XCES i inicjuje moduły SentenceParserdo przetwarzania plików zdań.AlignmentParser wypuszcza segmenty jednozdanie, podczas gdy SentenceParser wypuszcza pojedynczezdania z obu stron wyrównania.LinksAlignmentParser może być użyty w przypadku, gdy potrzebne są tylko linki XCES i można pominąć parsowanie pliku zdania.Dla parsowania zdania istnieje również alternatywnymoduł ExhaustiveSentenceParser,który jest bardziej wytrzymały niż SentenceParser,ale nieco wolniejszy przy przetwarzaniu tylko niewielkiej części dużego korpusu.Każdy z modułów w parse sublibrary może być indywidualnie zaimportowanydo skryptu Pythona i używany do ekstrakcji pojedynczych zdań,par zdań lub linków XCES.

OpusCat jest modułem Pythona używanym przez skryptopus_cat i oba mają taką samą funkcjonalność odczytywania jednojęzycznych plików zdań z OPUS.OpusCat wykorzystuje zmodyfikowaną wersję modułu SentenceParser:Podczas odczytywania plików z pojedynczymi zdaniami proces analizowania zdania nie musi przebiegać zgodnie z kolejnością określoną w pliku wyrównania, a SentenceParser w OpusCat po prostu wypuszcza każde zdanie w pliku.Zarówno OpusCat, jaki SentenceParser mogą być importowane jako moduły Pythona, aby mieć szczegółową kontrolę nad czytaniem plików jednojęzycznych.**Moduł** OpusGet uruchamia skryptopus_get z możliwościąpobierania corpora.Importując moduł w kodzie Pythona, można uzyskać szczegółowe informacje o corpora OPUS w strukturach danych Pythona.Informacje te obejmują liczbę par wyrównania, liczbę dokumentów, liczbę żetonów i rozmiar w kilobajtach wśród innych pozycji.

Moduł OpusLangid ma taką samą funkcjonalność jak skrypt opus_langid:dodawanie etykiet językowych i identyfikacja językowa punktów pewności do plików zdań XML.Dodatkowo OpusLangid zawiera klasęLanguageIdAdder, która może być używana do uzyskiwania etykietjęzykowych i identyfikowania punktów pewności zarówno z pycld2 jak i langid.py dla zwykłego zdania tekstowego z pojedynczą funkcją wywołania.

3.3. Moduł OpusTools Perl

Pakiet uzupełniający narzędzia OPUS jest dostępny jako moduł Perla z dopuszczalną licencją MIT.5

[5]https: //github.com/Helsinki-NLP/ OpusTools—perl

zawiera narzędzia wiersza poleceń, które są przydatne zwłaszcza do tworzenia nowych zbiorów danych, ale także ogólnie do szybkiego dostępu do danych w różnych formatach.Niektóre funkcje sąobecnie zastąpione implementacjami w bibliotece Pythona opisanymi powyżej, a my skupimy się tutaj na narzędziach, które obsługują dodatkowe przypadki użycia.Narzędzia te należą głównie do następujących trzech kategorii:

Narzędzia do konwersji: Narzędzia, które mogą być wykorzystane do importu i eksportu danych w różnych formatach plików i znacznikach danych.Głównym celem jest import nowych zbiorów danych w OPUS i tworzenie plików danych, które są wydawane w różnych formatach.

Narzędzia do osiowania: Wyrównywanie zdań i wyrazów może być używane na różne sposoby, a te narzędzia zapewniają pewne wygodne operacje na górze wyrównanych bitextów.

Inne narzędzia do przetwarzania: Kategoria ta obejmuje narzędzia do adnotacji i indeksowania.

W pierwszej kategorii mamy narzędzia importowe, takie jak Moza2opus, tmx2opus i xml2opus.Ex
Skrypty portowe to opus2moses, tmx2moses, opus2text i opus2multi.

xml2opus jest prostym skryptem, który dodaje granice zdań-do dowolnych danych XML.Wykrycie granic zdania odbywa się przy użyciu narzędzi wydanych za pomocąkorpusu równoległego Europarl (Koehn, 2005) i zapakowanych w module Perl Lingua::Sentence.W przyszłości zostaną zintegrowane dodatkowe narzędzia oparte na klasyfikacjach banków drzew UD.Znaczniki inline, które dodają znaczniki w zdaniach, nie są obecnie obsługiwane.

Moses2opus odczytuje wyrównane zwykłe pliki tekstowe jako powszechnie używane w tłumaczeniu maszynowym z wyrównanymi zdaniami w tej samej linii.[5]Narzędzie konwertuje dane do prostegoautonomicznego XML dla danych corpus i XCES Align format dla wyrównania zdania standoff tak jak jest używany w OPUS.Obecnie obsługiwane jest tylko dwujęzyczne wejście.Zwykłe pliki tekstowe nie zawierają granic zdania, ale nadal mogą zawierać wyrównania zdań, które nie są jedno do jednego.Dlatego moza2opus dodaje znacznik zdania za pomocąLingua::Wysłanie i -odpowiedniodostosowuje ustawienia zdania standoff.Skrypt obsługuje również dzieleniebitextów na mniejsze części.Puste

[5]Nazwa pochodzi od pakietu Mojżeszowego, który spopularyzował format.

linie w źródle i

językdocelowy może być używany do określania granic dokumentów.Ponadto korpus można podzielić na części o jednakowej wielkości, stosując próg długości dla maksymalnej liczby jednostek tłumaczeniowych włączonych do jednej części.

tmx2opus konwertuje pamięci tłumaczeniowe w formacie TMX na OPUS XML.Narzędzie dodaje granice zdania w taki sam sposób jak moses2opus.Pozwala również przepuścić kilka plików TMX przez narzędzie konwersji i jest w stanie scalić informacje w przypadku nakładania się zdań, które są objęte w kilku jednostkach tłumaczeń.Jest to przydatne podczas przetwarzania danych, które pochodzą jako różne bitexts, ale obejmujące tę samą zawartość.Tak więc, tylko unikalne zdania są przechowywane w wynikowym OPUS XML dla każdego języka, nawet jeśli pojawiają się one w różnych jednostkach tłumaczeń z dopasowaniem do różnych języków. tmx2opus może również przetwarzać pamięci tłumaczeniowez więcej niż dwoma językami w jednostce tłumaczeń,i będzie produkować dwujęzyczne pliki wyrównania zdań dla wszystkich par językowych, jak są one konieczne w OPUS.Ponadto możliwe jest również podzielenie danych na mniejsze części podobne do tego, co robi mozai2opus.Właściwości plików TMX można również skopiować do przekonwertowanych danych w celu zachowania dodatkowych metadanych.Zastosowanie tmx2opus do utworzenia przywożonego korpusu ParaCrawlw OPUS jest opisane w sekcji 4.
Skrypty eksportują głównie konwertowanie danych w przeciwnym kierunku.opus2moses i opus2textkonwertują daneOPUS XML na zwykły tekst i są one w większości przestarzałe i zastąpione implementacją pakietu Python wprowadzonego wcześniej. tmx2mosesjest wygodnym skryptem do wyodrębnienia wyrównanych zdań z dowolnych plików TMX i nie ogranicza się do danych OPUS.

opus2multi jest narzędziem, które może tworzyć wielorakie zbiory danych z OPUS corpora.W OPUS wszystkie zbiory danych są wyrównane dwujęzycznie, ale w niektórych przypadkach chcemy mieć wyrównanie obejmujące więcej niż dwa języki.W tym celu opus2multi może pomóc połączyć dwujęzyczne dopasowanie zdaniai wyodrębnić linki w większej liczbie języków.Narzędzie działa na plikach wyrównania zdań standoff i korzysta z języka obrotowego do tworzeniajednostek tłumaczeniowych we wszystkich językach.W tym celu rozszerza sięczęściowo pokrywające się ze sobą wyrównanie zdań do czasu, gdy wszystkie języki zostaną pokryte bez dalszych konfliktów w wynikającej z tegojednostce tłumaczeń (tj. żadne pozostałe nie pokrywają się z innymi jednostkami).Wynikiem tego procesu sąpliki wyrównania zdań, które są (dla wygody) drukowane dwujęzycznie przy użyciu formatu XCES Align, które mogą być dalej przetwarzane za pomocą OpusTools w celu ekstrakcji rzeczywistych par wyrównania.Istnieje również możliwość kontrolowania maksymalnej wielkości jednostki tłumaczeniowej (w liczbie zdań w jednym języku), ponieważ rozmiar może wzrastać bez ograniczeń w procesie rozszerzania.Uwzględniono również eksperymentalną cechę włączania linków międzyjęzykowych do dalszego odwzorowywania tranzytywnych map.Jest to przydatne dla zbiorów danych, takich jak opensubtitles, w których alternatywne plikinapisów mogą być używane do łączenia pomiędzy różnymi językami.

Narzędzia do osiowania w pakiecie OpusTools pomagają przetwarzać dopasowaniezdania w ich formacie standoff. opus—swap —align po prostu zamienia kierunek osiowania.Opus zapewnia tylko wyrównanie w jednym kierunku

(ponieważ są one symetryczne), ale czasami jest to wygodne, aby mieć dostęp do linków w innym kierunku, jak również.-opus-merge —align łączypliki wyrównania zdań iusuwa duplikaty, jeśli istnieją. opus-split-align dzieli pliki wyrównania zdań na osobne pliki z jednym na grupę wyrównania, tzn. zrównany dokument.W końcu opus-pivoting umożliwia stworzenie przechodniego wyrównania zdań pomiędzy dwoma językamiużywającymi języka przegubowego i linkami do języka przegubowego.Jest to wygodne dla korpusów, które pochodzą z bitextów, które nie obejmują wszystkich par językowych, ale tylko dopasowują się do określonego języka, takiego jak angielski.Zakładając, że istnieje znaczne nakładanie się pomiędzy bitextami, powiedzmy A^ P i B^ P, opus-pivoting ekstraktujepowiązania pomiędzy zdaniamiw A i B, tworząc nowy bitext A^ B. Sekcja 4.ilustruje użycie przez przykład stworzenia MultiParaCrawl.Wreszcie, inne narzędzie do wyrównania, opus—pt2dice,wydobywa surowe probabilistyczne słowniki dwujęzyczne z fraz-translation-tables stworzonych z wyrównania słów i przy użyciu narzędzi SMT wychodzących z zestawu narzędzi Mojżesza.Słowniki te wykorzystują pewne heurystyki do filtrowania danych, a narzędzie tworzy również dodatkowe wyniki Dice jako symetryzowaną wartość wyrównania z warunkowych prawdopodobieństwtłumaczenia zawartych w oryginalnych tabelach frazy, co jest przydatne do dwujęzycznego ekstrakcji leksykonu (Smadja et al., 1996).

Inne narzędzia: Ostatnia kategoria narzędzi zawiera dodatkowe narzędzia do przetwarzaniadanych, takie jak opus —udpipe i opus —index.Pierwszy z nich realizuje opakowanie wokół UDPipe (Straka i Strakova, 2017) do adnotacji danych OPUS i przechowywania wyników w zgodnych z OPUS XML.OpusTools może korzystać z wcześniej przeszkolonych modeli pochodzących z LIN-DAT.[6]Wreszcie, opus-index jest narzędziem do indeksowania corpora OPUS za pomocą Corpus Work Bench (CWB) (Evert i Hardie, 2011).Tworzy wszystkie pliki importu i uruchamia koder, jeśli jest dostępny, aby utworzyć wieloaspektową corpora, która ma być zapytana za pomocą wyszukiwarki CWB.

4. ParaCrawl i MultiParaCrawl

W tej sekcji chcielibyśmy przedstawić import danych ParaCrawl, aby zademonstrować wykorzystanie OpusTools.ParaCrawl Corpus[7] zostałwydobyty poprzez przeglądanie sieci i stosowanie skomplikowanego rurociągu dopasowania dokumentów i zdań woparciu o pakiet Bitextor (Espla-Gomis, 2009).Obecna wersja v5.0 obejmuje 24 języków europejskich,a projekt zapewnia automatycznie czyszczone bitexty dlajęzyków dopasowanych do angielskiego.Wielkość wynosi od 100 000 jednostek tłumaczeniowych (maltański-angielski) do ponad 50 milionówjednostek (francusko-angielski), a pliki danych są dystrybuowane w formacie tekstowym lub TMX.Chociaż istnieje kilka bonusowych par językowych, które zawierają również dwa bitexty nie wliczając angielskiego, większość kolekcji jest dwujęzycznie dopasowana do treści angielskiej.

Celem integracji ParaCrawl w OPUS jest udostępnienie danych w natywnym formacie OPUS, a także pełne pokrycie wszystkich par językowych zawartych w kolekcji.W tym celu, uprzednio wprowadzone

[6]https://lindat.mff.cuni.cz

[7]https://paracrawl.eu

język	pliki	żetony	zdania	BG	CS	da	de	El	ES	et	Fi	FR	GA	cześć.	HU	to jest	To	IV	szczur	NL	Pi	PT	r	SK	SI	SV
BG	1	57.7	2.6M		0.5M	0.4M	0.7M	0.4M	0.7M	0.3M	0.4M	0.8M	96.6k	0.3M	03M	0.5M	0.3M	0.2M	68.0k	0.4M	0.4M	0.5M	0.4M	0.4M	03M	0.4M
CS	1	119.0M	5.3M	0.5M		0.8M	1.4M	0.6M	13M	0.4M	0.6M	1.3M	0.1M	0.4M	0.6M	1.2M	0.3M	0.3M	79.1k	0.9M	1.0M	1.0M	0.6M	0.8M	0.3M	0.7M
da	1	108.3M	4.7M	0.4M	0.8M		1.4M	0.6M	1.4M	0.4M	0.8M	1.4M	0.1M	0.4M	03M	1.3M	0.3M	0.3M	88.3k	1.3M	0.9M	1.2M	03M	0.5M	0.3M	1.3M
de	1	909.7M	38.3M	0.7M	1.4M	1.4M		0.8M	7.0M	0.4M	0.8M	8.1M	0.1M	0.5M	0.8M	6.0M	0.4M	0.3M	82.8k	3.1M	1.8M	3.6M	0.7M	0.6M	0.4M	1.4M
El	1	94.9M	3.8M	0.4M	0.6M	0.6M	0.8M		1.0M	0.2M	03M	1.0M	0.1M	0.3M	0.4M	0.9M	0.3M	0.2M	76.1k	0.7M	0.6M	0.9M	03M	0.4M	0.3M	0.6M
ES	1	961.5M	38.7M	0.7M	13M	13M	7,1 M	1.0M		0.4M	0.9M	9.9M	0.1M	0.5M	0.8M	6.8M	0.4M	0.3M	78.2k	2.9M	1.8M	6.0M	0.9M	0.6M	0.3M	1.4M
et	1	26.5M	1.4M	0.3M	0.4M	0.4M	0.4M	0.2M	0.4M		0.4M	0.4M	95.1k	0.2M	03M	0.3M	0.3M	0.2M	81.2k	0.3M	0.3M	0.3M	03M	0.2M	03M	0.4M
Fi	1	54.4M	3.2M	0.4M	0.6M	0.8M	0.8M	0.5M	0.9M	0.4M		1.0M	0.1M	0.3M	03M	0.8M	0.3M	0.3M	80.7k	0.8M	0.7M	0.8M	03M	0.4M	03M	1.2M
FR	1	1.3G	51.1M	0.8M	1.4M	1.4M	83M	1.0M	10.1	0.4M	1.0M		0.1M	0.5M	0.8M	7.1M	0.4M	0.3M	82.3k	3.4M	1.8M	4.6M	0.9M	0.6M	0.4M	1.4M
GA	1	24.8M	0.8M	97.6k	0.1M	0.1M	0.1M	0.1M	0.1M	96.4k	0,1 M	0.1M		67.5k	0,1 M	0.1M	78.0k	75.7k	54.7k	99.4k	983k	0.1M	75.6k	0.1M	76.7k	96.5k
HR	1	43.2M	1.9M	0.3M	03M	0.4M	0.5M	03M	0.5M	0.2M	03M	0.5M	68.2k		0JM	0.6M	03M	0.2M	50.6k	0.4M	0.4M	0.4M	03M	0.3M	0.3M	0.3M
HU	1	107.0M	4.1M	0.3M	0.7M	03M	0.8M	0.4M	0.8M	0.3M	03M	0.9M	0.1M	0.3M		0.8M	0.3M	0.3M	76.4k	0.6M	0.7M	0.6M	0.6M	0.5M	0.3M	0.5M
to jest to	1	562.3M	22.0M	0.5M	13M	13M	6.1M	1.0M	7.0M	0.4M	0.8M	7.2M	0.1M	0.6M	0.8M		0.4M	0.3M	91.4k	2.6M	1.7M	3.9M	0.9M	0.6M	0.4M	1.3M
To jest to	1	25.6M	1.3M	0.3M	03M	03M	0.4M	03M	0.4M	0.3M	0.4M	0.4M	79.0k	0.2M	03M	0.4M		0.3M	73.4k	0.4M	0.4M	0.4M	03M	0.3M	03M	0.4M
IV	1	22.5M	1.1M	0.2M	03M	03M	0.3M	0.2M	0.3M	0.2M	03M	0.3M	763k	0.2M	03M	0.3M	0.3M		66.9k	0.3M	0.3M	0.3M	03M	0.3M	03M	0.3M
ULT	1	4.2M	0.2M	68.4k	79.5k	88.8k	83.3k	76.5k	78.7k	81.7k	81.1k	82.9k	55.0k	50.8k	76.8k	92.0k	73.8k	67.2k		85.7k	863k	87.2k	68.7k	82.6k	713k	86.5k
NL	1	237.9M	10.6M	0.4M	0.9M	13M	3.1M	0.8M	3.0M	0.3M	0.8M	3.5M	0.1M	0.4M	0.6M	2.7M	0.4M	0.3M	86.4k		13M	2.2M	0.6M	0.5M	0.3M	1.3M
Pi	1	144.8M	6.7M	0.4M	1.1M	0.9M	1.9M	0.6M	1.9M	0.3M	0.7M	1.8M	99.3k	0.4M	0.7M	1.8M	0.4M	0.3M	86.8k	1.2M		1.5M	0.7M	0.6M	0.3M	1.0M
PT	1	320.	13.5M	0.5M	1.0M	12M	3.6M	0.9M	6.1M	0.3M	0.8M	4.7M	0.1M	0.4M	0.7M	4.0M	0.4M	0.3M	87.9k	2.2M	L6M		0.7M	0.5M	0.3M	1.2M
r	1	65.7M	2.9M	0.4M	0.6M	03M	0.7M	0.5M	0.9M	0.3M	03M	0.9M	76.4k	0.3M	0.6M	0.9M	0.3M	0.2M	69.2k	0.6M	0.7M	0.7M		0.4M	0.3M	0.6M
SK	1	41.6M	2.1M	0.4M	0.8M	03M	0.6M	0.4M	0.6M	0.2M	0.4M	0.6M	0.1M	0.3M	03M	0.6M	0.3M	0.3M	83.1k	0.5M	0.6M	0.5M	0.4M		0.4M	0.5M
SI	1	31.8M	1.5M	0.2M	03M	03M	0.4M	03M	0.3M	0.2M	02M	0.4M	773k	0.3M	03M	0.4M	03M	0.2M	71.9k	0.3M	0.4M	0.3M	03M	0.4M		0.3M
SV	1	131.5M	6.1M	0.4M	0.7M	13M	1.4M	0.6M	1.5M	0.4M	12M	1.4M	973k	0.3M	03M	1.4M	0.4M	0.3M	87.1k	1.3M	1.0M	1.2M	0.6M	0.5M	0.3M	

Rysunek 5:Statystyki z MultiParaCrawl corpus - wielojęzyczne rozszerzenie ParaCrawl poprzez osiowanie osi przez angielski.Trójkąt górno-prawny daje rozmiar pod względem wyrównywania zdań w zwykłym formacie tekstowym, a trójkąt dolny-lewy pokazuje wielkość wyekstrahowanych plików TMX pod względem unikalnych jednostek tłumaczeniowych na parę językową.

narzędzia tmx2opus i opus-pivoting stają się przydatne. tmx2opus jest nie tylko użyteczny do ekstrakcji wyrównań z oryginalnego źródła TMX, ale także zapewnia funkcjonalnośćdodawania znacznika granic zdania i zmniejszenia nadmiarowości pomiędzy różnymi bitextami.Użycie unikalnej opcji tmx2opus zmniejsza rozmiar angielskiejczęści korpusu (tj. 252 milionów oddzielnie wyrównanych angielskich zdańw 23 bitextach) do mniej niż 60 % oryginalnych danych.Jednocześnie unikatowość umożliwia równieżzbudowanie wieloprzykładowego korpusu poprzez obracanie linków do angielskiego w nowo utworzonym, unikalnym zestawie zdań.W tym celumożna wykorzystać opus-pivoting, jak wyjaśniono wcześniej.Korzystając z tej procedury, można by stworzyć 253 dodatkowych bitextów orozmiarach do 10 milionów jednostek tłumaczeniowych.Rysunek 5 podsumowuje nieangielskie bitexty w Mul- tiParaCrawl.

5. Diagnostyka równoległego Korpusu

Nasza procedura diagnostyczna dla kolekcji OPUS wykorzystuje narzędzie opus_read wierszpoleceń (opisane w sekcji3.1.) do pobierania wyrównanych danych tekstowych dla danejpary języków w danym korpusie.W tym celu opus_read przetwarzanatywne dane sformatowaneprzez XML do wygenerowania żądanegopodzbioru danych, a następnie wykonuje konwersję do zwykłegoformatu tekstowego.Podczas tego procesu,rutyna diagnostyczna słucha wszelkich błędów, które mogą się pojawić, i zapisuje je do opracowania raportu diagnostycznego do późniejszej analizy.Systematycznie wykonujemy tę procedurę dla każdej pary języków dostępnych pod każdym korpusem OPUS.[8]Dla naszej diagnostyki wykorzystujemy pełną granulaarność dostarczaną przez OPUS, zbierając oddzielne odczyty dla różnych korpusów,które komponują bitexty, a także trzymając regionalne wariantyjęzyków osobno, a nie je sprzęgając.Aby przeprowadzić tego rodzaju wyczerpującą analizę, przeprowadziliśmy łącznie 87 948 zadań tablicy procesora równolegle, z czasem pracy wahającym się od

[8]Nie wykonaliśmy diagnostyki dwóch najnowszych dodatkówdo OPUS:Infopankki i MultiParaCrawl.

1 sekundy do 5,2 godziny, a każde zadanie wykorzystujące od 4 do128 GB pamięci.W sumie całaanaliza diagnostyczna trwała około 1000 godzin,średnio 18,2 godziny na korpus.Podczas gdy ziarnistośćnaszej analizy będzie przydatna wewnętrznie do wykrywania anomalii w OPUS w celu ułatwienia naprawy, zbieramy również nasze dane w celu generowania danych dotyczących całego korpusu, które zgłaszamy i omawiamy w tej sekcji.

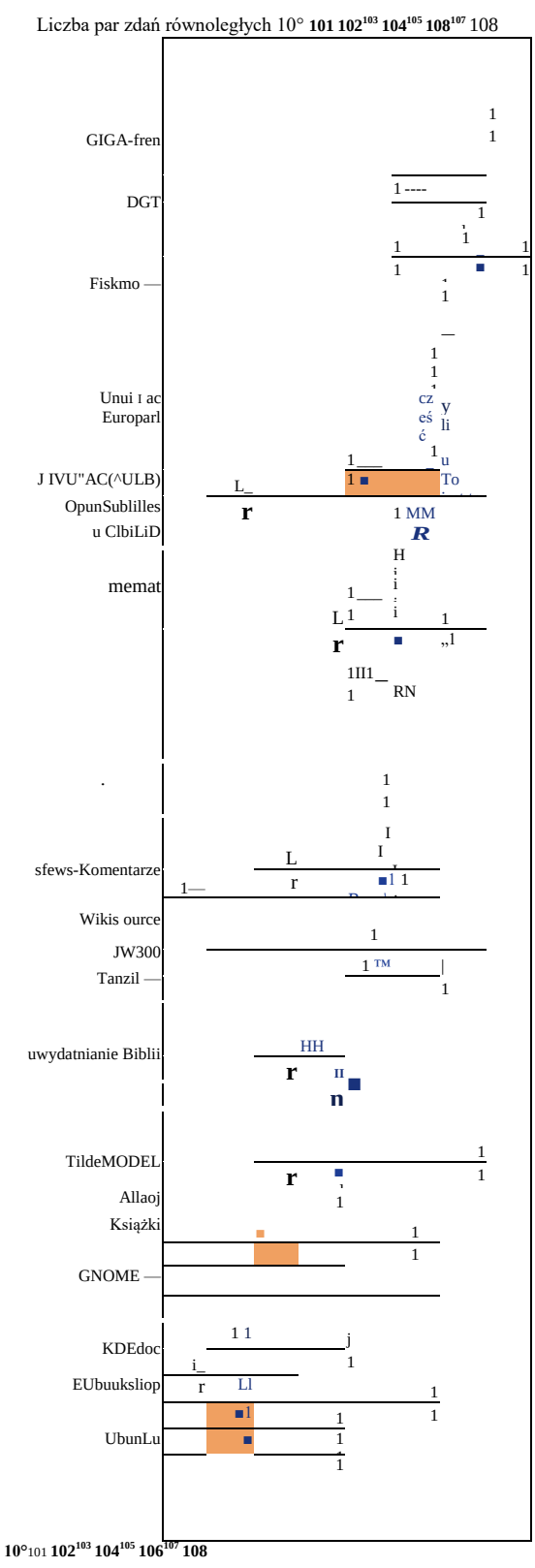
5.1. Analiza błędów

Zalogowane w naszym raporcie „diagnozy" wymieniają przyczyny każdego błędu wyszukiwania, co daje nam środki do niezawodnego ich zlokalizowania i naprawienia.Zestawienie wszystkich diagnoz, wyniki pokazują, że podczas gdy 37 korpusów jest całkowicie bezbłędnych, pozyskiwanie danych zablokowało się dla co najmniej jednej pary językowej dla pozostałych 18 korpusów.Obfitośćbłędów odzyskiwania w tych korpusach różni się od maleńkiego ułamka do całego korpusu (pokazane na rysunku 8).Zdecydowanawiększość tych błędów wynika z nieuformowanych danych XML z nieprawidłowymi tokenami (96,2 %) lub niedopasowanymi tagami (3,5 %).Nasze dotychczasowe częściowe kontrole sugerują, że można je przypisać drobnym błędom konwersji, takim jak nieuniknione znaki specjalnei jednostki XML występujące w oryginalnych danych przed importem do OPUS.Kolejna bardzo mała część błędów (0,3 %) wskazuje brakujące pliki danych w głównym systemie plików, w którym OPUS jest hostowany, co prawdopodobnie wskazuje na błędy kopii i pozostaje do zbadania.

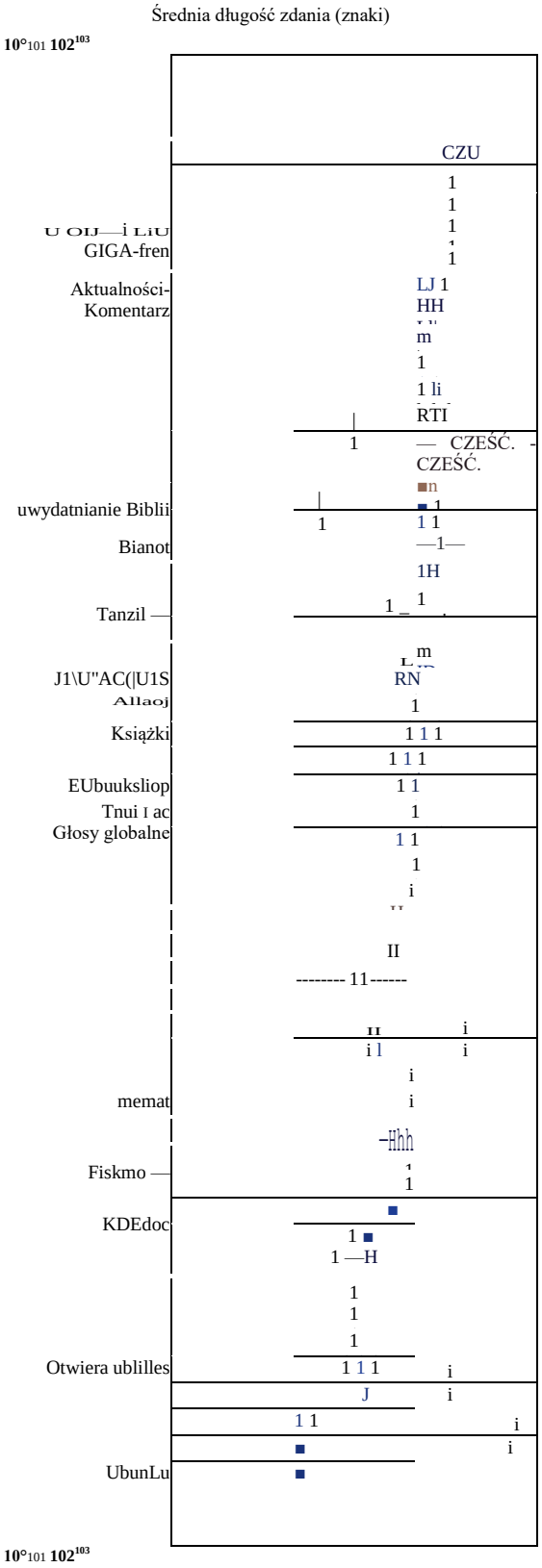
5.2. Statystyka Corpus-Wide

Oprócz katalogowania kwestii odzyskiwania danych, nasza procedura diagnostyczna oblicza również niektóre podstawowestatystyki ilościowe, takie jak zgłaszane koszty obliczeniowe dla odzyskiwania danych, i różne pomiary na pobranych danych dla każdegokorpusu, języka i pary językowej.Nasze analizystatystyczne w większości nie wykazały godnych uwagi trendów ani odchyleń, z wyjątkiem niektórych środków, którewskazywały na względne wariancje i poziom hałasu w danych dotyczącychcałego korpusu.Na rysunkach 6 i 7 podajemy

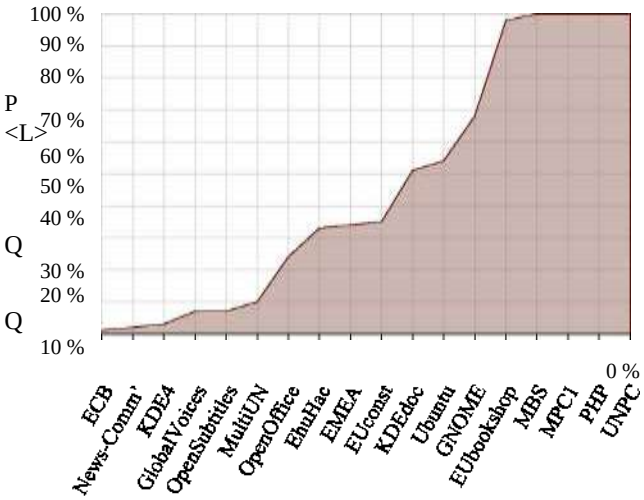
rozkład dwóch miar na zestaw dostępny lan- średnia długość zdania, odpowiednio.Obie pary guage dla każdego korpusu:średnia liczba pomiarów sen została zwizualizowana za pomocą pole-i-whisker działki do par tence (lub, dokładniej, jednostek tłumaczeniowych) i podkreślenie różnic dystrybucyjnych, gdzie punkty końcowe



Rysunek 6:Rozkład liczby par zdań równoległych na zestaw dostępnych par językowych dla każdego korpusu.



Rysunek 7:Rozkład średnich długości zdania (w znakach) na zestawy dostępnych par językowych dla każdego korpusu.



Rysunek 8:Procent par językowych w corpora OPUS, dla których pobieranie danych z narzędziami OPUS zwraca błędy.Na wykresie pominięto bezbłędną korpusę.

Pokaż najniższe i najwyższe wartości,[9]a dwie połówki pola reprezentują drugi i trzeci kwartyl wartości, oddzielony medianą.

Jednym z najbardziej uderzających szczegółów z rysunku 6 jest kontrastmiędzy wariancjami.Ponad jedna trzecia korpusu wykazuje bardzo niewiele wariancji między parami językowymi, co oznacza, że dane są w pełni wielorakie, podczas gdy inne, jak JW300 i otwarte napisy pokazują odwrotnie bardzo dużą wariancję, gdzie różnica w wielkości dostępnych danych może obejmować kilka rzędów wielkości.Patrząc konkretnie na pierwsze kwartyle, niektóre korpusy, takie jak QED i Tatoeba wydają się mieć znaczną część par językowych zawierającychbardzo niewiele jednostek tłumaczeniowych, co prawdopodobnie wskazuje na wysoki poziom detekcji językalub wyrównywania zdań.Na rysunku 7, pierwszy kwartyl wydaje się mieć podobny zakres względny dla niektórych ciał, co oznacza, że zdania zawierają średnio tylko kilka znaków dla niektórych dostępnych par językowych.Prawdopodobnie nie jest przypadkiem, że te przypadki w większości odpowiadają korpusowi, które zostały zebrane z naturalnie hałaśliwych danych.Ponadto najmniejsze i największe mediany wartości na rysunku 7 wskazują na wyjątkowo krótkie i wyjątkowodługie „typowe" zdania w odpowiedniej korpusie,które mogą wskazywać na silny kontrast w segmentacji tekstu lubwyraźnie różne domeny danych.Na przykład, trzy korpusy o najniższych medianach zawierają tłumaczenia oprogramowania-komputerowego, podczas gdy dokumenty z Organizacji Narodów Zjednoczonych dają najwyższą medianę długości.

6. Wnioski i przyszłe prace

W artykule tym wprowadzamy OpusTools, pakiet otwartych bibliotek i narzędzi linii poleceń dla efektywnego i wygodnego dostępu do korpusu równoległego w szerokim zbiorze danych OPUS.Pakiet wdraża narzędzia do pobierania, konwersji, filtrowania i przetwarzania równoległych zbiorów-danych oraz ułatwia dostęp do skompresowanych i zarchiwizowanych plików z kolekcji.Zapewnia również bibliotekę pythonową umożliwiającą programowy dostęp do danych, dzięki czemu łatwe jest włączenie przetwarzania danych do rozwoju innych narzędzi

[9]Nieograniczone punkty końcowe wskazują na ekstremę wykraczającą poza granice osi x.

.Ponadto prezentujemy narzędzia do konwersji i wyrównania danych, które mogą być stosowane przy przygotowywaniu nowych zbiorów danych z różnych źródeł.Demonstrujemy ich użycie na przykładzie niedawno dodanego MultiParaCrawl corpus, który rozszerza oryginalny zbiór danych o wyrównania oparte na przesunięciachpomiędzy wszystkimi parami językowymi, przyczyniając się dorosnącego zasięgu bazy danych OPUS.

Chociaż utrzymywanie kolekcji tak dużej, jak OPUS jest bardzo wytrzymałe jest dość trudne, rozwiązywanie problemów będzie łatwiejsze i szybsze przy diagnostyce w pełni nakreślonej.Ogólnie rzecz biorąc, podczas gdy błędy odzyskiwania danych obejmują wyraźne punkty działania, analizy statystyczne wydają się raczej sugerować znaczną różnorodność jakościowąi ilościową wśród corpora OPUS, z tendencjami pozornie w obrębie oczekiwań i przypadkami granicznymi, które można przypisać hałasowi w oryginalnych danych.Naszym zamiarem jest rozwiązanie wszystkich problemów związanych z pobieraniem danych, tak aby korzystanie z narzędzi OPUS było płynnym doświadczeniem dla wszystkich użytkowników, a także usprawnienie naszej rutyny jako narzędzia diagnostycznego, które stałoby się standardową częścią procesu rozszerzania OPUS o nowe korpusy.

Potwierdzenia

ERC Niniejsza praca jest częścią projektu FoTran, finansowanego przez Europejską Radę ds. Badań Naukowych (ERC) un—

H Program Unii Europejskiej w zakresie badań naukowychi innowacji w ramach programu „Horyzont 2020” (porozumienie w sprawie dotacji Na 771113)' jak również projekt MeMAD, finansowany z unijnego programu badań i innowacji w ramach programu „Horyzont 2020” (-porozumienie w sprawie dotacji Na 780069).

7. Odniesienia bibliograficzne

Evert, S. i Hardie, A. (2011).Ławka robocza z XXI-wieku:Aktualizacja architektury zapytań na nowe tysiąclecie.In *Proceedings of the Corpus Linguistics2011 Conference, University of* Birmingham, UK.Smadja, F., McKeown, K. R. i Hatzivassiloglou, V. (1996).Tłumaczenie kolokacji dwujęzycznych leksykonów:Podejście statystyczne.*Językoznawstwo*obliczeniowe, 22(1):1-38.

8. Odniesienia do zasobów językowych

Espla-Gomis, Miquel.(2009).*Bitextor:darmowe/Otwarte źródło oprogramowania do zbioru Wspomnienia Tłumaczeń z wielojęzycznych stron internetowych.*

Koehn, Philipp.(2005).*Europarl:Równoległy korpus dla statystycznego tłumaczenia maszynowego*.AAMT.

Lui, Marco i Baldwin, Timothy.(2012). *langid.py:Narzędzie do identyfikacji języka poza półką.*Stowarzyszenie-lingwistyki obliczeniowej.

Straka, Mediolan i Strakova, Jana.(2017).*Tokenizing, POS Tagging, Lemmatizing i Parsing UD 2.0 z UDPipe*.Stowarzyszenie lingwistyki obliczeniowej.

Tiedemann, Jorg.(2012).*Równoległe dane, narzędzia i interfejsyw OPUS.*Europejskie Stowarzyszenie Zasobów Językowych(ELRA).