Rokovania 12. konferencie o jazykových zdrojoch a hodnotení (LREC 2020), strany 3782 – 3789
Marseille, 11. – 16. mája 2020

OpusTools a paralelná diagnostika korpusu

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

Abstraktné

Tento papier predstavuje OpusTools, balík na stiahnutie a spracovanie paralelných korpusov zahrnutých v kolekcii OPUS corpus.Balík implementuje nástroje na prístup ku komprimovaným dátam v ich archivovanom formáte vydania a umožňuje jednoduchý prevod medzi bežnými formátmi.OpusTools obsahuje aj nástroje na identifikáciu jazyka a filtrovanie dát, ako aj nástroje na import dát z rôznych zdrojov do OPUS formátu.Použitie týchto nástrojov ukazujeme v paralelnej korpusovej tvorbe a v dátovej diagnostike.Druhá možnosť je obzvlášť užitočná pri identifikácii potenciálnych problémov a chýb v rozsiahlom súbore údajov.Pomocou týchto nástrojov môžeme teraz monitorovať platnosť súborov údajov a zlepšiť celkovú kvalitu a konzistentnosť zberu údajov.

Kľúčové slová: Korpus (vytvorenie, poznámky atď.);Strojový preklad;Nástroje, systémy, aplikácie

1. Úvod

Opus (Tiedemann, 2012) je najväčšia zbierka voľne dostupných paralelných korpusov.Kolekcia v priebehu rokov neustále rastie a je široko využívaná v práci na strojovom preklade a krížovom lingvistickom výskume.V súčasnosti obsahuje 57 vydaných korpusov pokrývajúcich viac ako 700 jazykových a jazykových variantov, ktoré vytvárajú viac ako 70 000 bitextov v zmysle zladených jazykových párov vo všetkých korpusochv kolekcii.Veľkosť a popularita OPUS si vyžaduje vybudovanie efektívnej infraštruktúry, ktorá umožňuje rôznym užívateľom získať a pristupovať k dátam a tento papier zavádza dva balíky, ktoré poskytujú nástroje na tento účel.Cieľom týchto balíkov je uľahčiť sťahovanie, prevod a spracovanie dát zahrnutých v OPUS z príkazového riadku alebo z aplikácií využívajúcich knižnicu implementujúcu tieto nástroje.Oba balíkyodkazujú na knižnicu Python s nástrojmi príkazového riadku a doplnkovým Perlovýmmodulom, oba poskytované ako open source a s povoľnými licenciami.

V nižšie uvedených častiach uvádzame nástroje a ich základné použitie a tiež diskutujeme o tom, ako sme tieto nástroje aplikovali na vytváranie nových súborov dát a na spustenie systematickej diagnostiky celej databázy.S dostupnosťou OpusTools je teraz možné vykonávať starostlivé hygienické kontroly rozsiahlych dátových súborov s cieľom overiť platnosť kódovania, nájsť rozbité odkazy a štruktúry a identifikovať ďalšie problémy s dátami.

2. Charakteristiky OPUS

Opus zahŕňa paralelné korpusy zo širokej škály zdrojov.Každý z nich má svoje vlastné zvláštnostia vlastnosti sa môžu podstatne líšiť v závislosti od pôvodných dát a ich distribúcie.Filozofiou v OPUS je udržať značenie a anotáciu čo najviac,ale zjednotiť základný dátový formát, aby bol prístup k paralelným dátam čo najtransparentnejší.To znamená, že korpusové dáta sú konvertované na samostatný XML, ktorý zachováva originálnu značku, ale dôsledne pridáva základné značenie, ktoré je potrebné pre zosúlaďovanie a ďalšie jazykovéspracovanie.Nastavenie je uložené ako standoff anotácia vo formáte XCES Align (pre zarovnanie vety) a „Moses formát“ (pre zarovnanie slov).Na základe tejto zásady sa údaje môžu uchovávať oddelene od anotácie zarovnania, ktorá umožňuje efektívnu implementáciu a uchovávanie masívne paralelných údajov a v prípade potreby umožňuje aj alternatívne zosúlaďovanie.Na obrázku 1 je znázornený príklad anotácie pozastavenej pozície, ktorá sa používa v OPUS na špecifikovanie prepojení medzi vetami.Každý súbor nazosúladenie vety môže obsahovať ľubovoľný počet prvkov prepojeniaGrp s cieľom zosúladiť dokumenty zo zberu údajov.Dokumenty sú špecifikované pomocou cesty vo vzťahu ku koreňu XML podkorpusu OPUS aprvky prepojenia poskytujú zarovnanie vety súbormi vety ID, ktoré sú oddelené bodkočiarkou.Vytvorenie alternatívneho zarovnania sa jednoducho robí vytvorením novéhosúboru zarovnania vety a nie je potrebné vykonať žiadne ďalšie úpravy s pôvodnými korpusovými dátami.Všimnite si, že zarovnanie vety je dvojjazyčné, ako je znázornené v príklade.Avšak anotácia typustandoff umožňuje zladiť masívne paralelné súbory dát vo všetkých jazykových pároch bez duplikácie ktoréhokoľvek z prepojených dátových súborov.Okrem toho môžu existovať alternatívne korpusové súbory s rôznymi úrovňami anotácie bez potreby opätovného zosúlaďovania týchto alternatívnych súborov.Obrázok 2 znázorňuje príklady takýchto anotovaných súborov, všetky sú zosúladené rovnakým spôsobom s nastavením stavu vety uloženého v externých súboroch.Viac informácií o dátových štruktúrach v OPUS nájdete na stránke OPUS Wiki.[1]

Ďalšou zásadou v OPUS je poskytovať dáta v iných spoločných formátoch, aby boli ľahko prístupné pre širokú škálu aplikácií.Tieto dátové formáty sú však vygenerované zo základného kódovania založeného na XML, ktoré slúži ako hlavná kópia každého korpusu.Používatelia údajov OPUS zvyčajne nie sú informovaní o týchto princípoch a sťahujú dátový formát, ktorý najviac vyhovuje ich potrebám.

Myšlienka OpusTools je teraz zjednotiť prístup k hlavným dátam v XML a k iným generovaným formátom poskytnutím-základných knižníc a nástrojov príkazového riadku na získavanie a prevod korpusových dát.Poskytujú tiež vhodné nástroje na základné filtrovanie a náhodný prístup v archivovaných dátach v ich komprimovanej forme, ktorá sa používa na distribúciu dát.To druhé je obzvlášť dôležité, pretože veľkosť niektorých kor—

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

```
<?xml verzia=„1.0“ kódovanie=„utf-8“?>
<!DOCTYPE cesAlign PUBLIC
                          „-//CES//DTD XML cesAlign//EN“ ""><linkGrp targType=“
            od Doc=„en/0/1089124/4995691.xml.gz“
            toDoc = „fr/0/1089124/4588599.xml.gz“>
<link id=„SL0“ xtargets=„1;1“ prekrývanie=„0.331“/>
<link id=„SL1“ xtargets=„2 3;2“ prekrývanie=„0,560“/> <link id=„SL2“
xtargets=„4;“/>
<link id=„SL3“ xtargets=„5 6;3“ prekrývanie=„0,854“/> <link id=„SL4“
xtargets=„7 8 9;4“ prekrývanie=„0,699“/> <link id=„SL5“ xtargets=„10
11;5“ prekrývanie=„0,776“/>
```

Obrázok 1:Príklad zarovnania trestu vo formáte XCES Zoradiť.Prvok linkGrp špecifikuje páry dokumentov, ktoré sú zosúladené a prepojenia medzi jednotlivými vetami sú uvedené vprvkoch odkazu.Voliteľnéatribúty prekrývania v tomto príklade sa vzťahujú na pomery časových prekrývaní, ktoré sa používajú ako prvok v zarovnaní titulkov.

Pora je rozsiahla tak, že vyžaduje, aby spoločné súborové systémy spracovávali dáta v nespracovanej, nekomprimovanejforme.Napríklad, posledný opensubtitles korpus obsahuje približne 3,7 milióna individuálnych dokumentov v 67 jazykoch so zarovnaním vo viac ako 3 600 bitextoch.Jeden z posledných doplnkov, JW300 pokrýva 380 jazykov vo viac ako 46 000 bitextov.Spolu je viac ako 9,2 milióna individuálnychdokumentov len v posledných vydaniach všetkých korpusov a toto číslo je zdvojnásobené rôznymi typmi predbežného spracovania, ktoré sú poskytované, surový text a tokenizované korpusy, ktoré sú čiastočne označené dodatočnými jazykovými informáciami.-Okrem toho, bitexty sa uvoľňujú vo natívnom formáte XML-(pozri obrázok 2), jednoduchom textovom formáte a výmene prekladovej pamäte(TMX).V súčasnej dobe zaberajú celkom 5.9 TB priestoru v komprimovanom formáte.

Uvedené čísla ilustrujú potrebu vhodných infraštruktúra účinných nástrojov na riadenie rôznych súborov údajov.To je motivácia pre implementáciu voľne dostupných nástrojov OPUS popísaných nižšie.Vytvárajú pohodlnú knižnicua skrinku nástrojov na sťahovanie, extrahovanie a konverziu dátzo zbierky OPUS.Okrem toho pomáhajú vykonávať systematickú diagnostiku zberu s cieľom identifikovať chyby a problémy v súboroch údajov.Nižšie budeme najprv prezentovať dva balíky a ich funkčnosť.Následne poskytujeme informácie o ich využití pri vytváraní nových súborov dát a nakoniec informujeme o aplikácii nástrojov OPUS pre diagnostické štúdie a hygienické kontroly.

3. Balík OpusTools

Balík OpusTools je súbor nástrojov na sťahovanie a správu paralelných korpusových dát z OPUS.Balíkpozostáva z knižnice Python a súvisiacich skriptov príkazového riadku.Okrem toho existuje balík Perl pre vytváranie nových dátových súborov a prístup k paralelným dátam.

3.1. Nástroje príkazu-Line

Balík OpusTools obsahuje päť skriptov založených na príkazovom riadku Python 3:Opus_read, opus_express, opus_cat, opus_get a opus_langid.2[skripty] umožňujú sťahovanie

OPUS dát, výstup údajov v špecifických formátoch,-extrahovanie školení, vývojových a testovacích súborov z údajov a ďalšie.Obrázok 3 zobrazuje prehľad scenárov.

opus_read je skript pre sťahovanie paralelných korpusov a ich prevod do požadovaných formátov.Opus korpus obsahuje súbory formátu XCES, ktoré poukazujú na dva súbory XML vety v rôznych jazykoch.Formát nastavenia XCES spája vety v zdrojových súboroch na vety v cieľových súboroch pomocou vety ID vety. Vetné súbory v OPUS korpusoch sú komprimované do ZIP archívov a opus_read umožňuje čítať údaje priamo z komprimovaných súborov. opus_read analyzuje danýsúbor nastavenia a vytvára výstup v jednom zo štyroch formátov:normálne, mach, TMX alebo XCES odkazy. opus_read first sa snaží čítať OPUS súbory z lokálnych adresárov.Ak nie sú nájdené požadované súbory, nástroj ponúka možnosť ich stiahnutia.Súbory vety môžu byť stiahnuté v surovom, tokenizovanom alebo paredizovanom formáte.

opus_read obsahuje základné filtre pre odstránenie nechcených párov viet pred vytvorením výstupného súboru.Nonalignments, kde zdroj alebo cieľový segment je prázdny, môžu byť vynechané.Alternatívne možno špecifikovať určitý počet zdrojových a cieľových segmentov, napr. je možné zahrnúť do výstupu iba jedno zarovnanie.Niektoré korpusy obsahujú skóre atribútov pre každý pár viet.Napríklad páry viet v korpuse opensubtitles majú prekrývacie skóre, ktoré naznačujú, do akej miery sa časové pečiatky oboch segmentov prekrývajú. opus_read je schopný filtrovať páry viet, ktoré neprekračujú daný prah skóre atribútu.Okrem toho je možné odstrániť páry segmentov nazáklade skóre spoľahlivosti jazykovej identifikácie.Jazykové štítky a skóre spoľahlivosti môžu byť pridané do vetyXML súborov s opus_langid skriptom.

opus_express je skript postavený na opus_read, ktorý môže extrahovaťready-to-use školenia, vývoj, a testovacie sady pre jazykový pár z jedného alebo viacerých OPUS korpusov.Postupnajprv vyplní stanovenú kvótu viet pre testovaciu sadu, potom pokračuje rovnako pre vývojovú sadu, a zvyšok vypustí do tréningového setu.Skript môže voliteľne-pre-shuffle dáta pred rozdelením, alebo naopak, označiť a zachovať hranice dokumentov cez splity pre modely na úrovni dokumentov. opus_express tiež obsahuje možnosť využiť skóre atribútov, ako sú hodnoty prekrývania, ktoré extrahuje opus_read v jeho uvedomenie kvality-toggle, ktorý-uprednostňuje vyššie-confidence páry vety presahujúce konfigurovateľný prah, ktorýsa má triediť do testovacích a vývojových súborov.

opus_cat sapoužíva na čítanie jednojazyčných korpusov z OPUS alebo jednotlivých súborov v rámci týchto korpusov.Súbory môžu byť vytlačené vo formáte XML alebo môžu byť prevedené do jednoduchého textu. opus_cat je užitočný pre manuálnu kontrolu domény alebo kvality jedného korpusu, pretože je schopný čítať súbory priamo zo ZIP archívov v OPUS korpusoch.

opus_get je skript pre sťahovanie paralelných korpusových súborov z OPUS.Pred stiahnutím môžu byť korpusy vyhľadávané a uvedené podľa ich názvu, zdrojového jazyka a cieľového jazyka.Napríklad je možné stiahnuť súbory pre konkrétnu jazykovú dvojicu v jednom korpuse, všetky súbory páru jazykov v

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

Surový formát XML:

```
<?xml verzia=„1.0“ kódovanie=„utf-8“?>
< dokument>
   <CHAPTER ID=„1“>
      <P id=„1“>
         <s id=„1“>Obnovenie relácie,/s>
      </P>
      <SPEAKER ID=„1“ NAME=„Prezident“>
         <P id=„2“>
            <s id=„2“>Vyhlasujem obnovené zasadnutie Európskeho parlamentu prerušené vo štvrtok, 14 jún 2001. schôdzu;/s> </P>
```

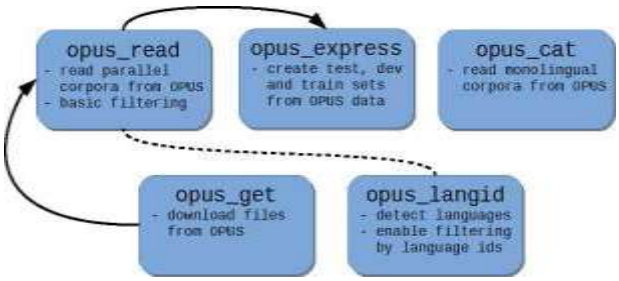
Tokenizovaný (anotovaný) formát XML:

```
<?xml verzia=„1.0“ kódovanie=„utf-8“?>
<dokument> ≥lt;CHAPTER ID=„1“> ≥lt;P id=„1“>
<s id=„1“>
<chunk typ = „NP“ id=„c-1“>
   <w hun=„NN“ strom=„NN“ lem = „opätovný začiatok“ pos=„NN“ id=„w1,1“>Resumption lt;/w>
</chunk>
<chunk typ = „PP“ id=„c-2“>
   <w hun=„IN“ strom=„IN“ lem=„in“ pos=„IN“ id=„w1,2“>of lt;/w>
</chunk>
<chunk typ = „NP“ id=„c-3“>
   <w hun=„DT“ strom=„DT“ lem=„the“ pos=„DT“ id=„w1,3“>the-art;/w>
   <w hun=„NN“ strom=„NN“ lem=„session“ pos=„NN“ id=„w1,4“>session lt;/w>
</chunk>
</s>
```

UD Parsed XML formát:

```
<?xml verzia=„1.0“ kódovanie=„utf-8“?>
< dokument>
<CHAPTER ID=„1“>
   <P id=„1“>
     <s id=„1“>
       <w xpos=„NOUN“ hlava=„0“ výkony=„Číslo=Sing“ UPOS=„NOUN“ lemma=„Resumption“ id=„1,1“ deprel=„koreň“>Resumption lt;w> <w xpos=„ADP“
       head=„1.4“ UPOS=„ADP“ lemma=„of“ id=„1.2“
       <w xpos=„DET“ head=„1,4“ výkony=„Definite=Def|PronType=Art“ UPOS=„DET“ lemma="„id="1,3„deprel=“det„>“det">
     </s>
```

Obrázok 2:Príklady XML kódovaných dát v OPUS.Rôzne druhy anotácií môžu byť pridané bez zničenia zarovnania vety, ktoré je uložené ako standoff anotácia väzieb medzi viet ID je.



Obrázok 3:Päť Pythonových scenárov založených na OpusTools.Každý zo skriptov môže byť použitý samostatne. opus_express je postavený na opus_read a opus_read používa opus_get na stiahnutie OPUS súborov. opus_langid musí byť aplikovanýna súbory vety, aby bolo možné jazyk id filtrovanie pre opus_read.

jeden korpus alebo všetky súbory pre konkrétny jazyk v celom OPUS. opus_read používa opus_get pre automatické sťahovaniepožadovaných korpusových súborov.

opusJangid sapoužíva na pridávanie jazykových identifikačných štítkova skóre spoľahlivosti pre každú vetu v danom súbore XML vety.Identifikácia jazyka sa vykonáva pomocou dvoch nástrojov „off-the-shelf“:Pycld2,[3]Python väzby pre Compact Language Detector 2 a langid.py[4] (Lui a Baldwin, 2012). opus_langid musí byť aplikovaný na vety XML súbory predtým, než opus_read môže filtrovať páry vietpodľa ich jazykových štítkov.Obrázok 4 znázorňuje príklad súboru vety, ktorý bol spracovaný s opus_langidom.

3.2. Knižnica OpusTools Python

Okrem toho, že príkazový riadok skripty, opus_read,opus_cat,opus_get a opus_langid sú spojené s modulmi Python, ktoré môžu byť importované a použité v niečí vlastné skripty.Moduly poskytujú rovnakú funkcionalituako nástroje príkazového riadku a tiež podrobnejšie dáta spravujúce ovládanie pomocou podmodulov a funkcií.Všetky Python kód je napísané v Python 3.

Modul OpusRead možno inicializovať s parametrami, ktoré zodpovedajú vlajkám daným na opus_read a používa sa na sťahovanie a konverziu korpusových súborov z OPUS.Vnútorne OpusRead používa moduly na analýzu XML-zo subknižky parse, ktorá je súčasťou balíka Opus- Tools Python.Podknižnica obsahuje moduly pre analýzu súborov nastavenia XCES a vety.Ou

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

```
<?xml verzia=„1.0“ kódovanie=„utf-8“?>
<text>
<p id=„1“>
    <s cld2=„en“ cld2conf=„0,99“ id=„s1.1“ LangID=„en“ langidconf=„1.0“>
      Vyhlásenie predsedu vlády pána Ingvara Carlssona o politike vlády na otvorení švédskeho parlamentu v utorok 4. októbra 1988.
    </s>
</p>

<p id=„2“>
    <s cld2=„en“ cld2conf=„0,98“ id=„s2.1“ LangID=„en“ langidconf=„1.0“>
      Vaše Veličenstvo, Vaše kráľovské výsosti, pán predseda, poslanci švédskeho parlamentu.
    </s>
</p>
```

Obrázok 4:Príklad súboru s vetou, kde boli do štítkov vety pridané jazykové štítky a skóre spoľahlivosti.

Modul AlignmentParser analyzuje daný súbor odkazu XCES a inicializuje moduly SentenceParser pre analýzu súborov-vety.AlignmentParser výstupy jednejvety pár segmentov, zatiaľ čo SentenceParservýstupy jednotlivévety z oboch strán zarovnania.LinksAlignmentParser môže byť použitý v prípade, že sú potrebné iba odkazy XCES a parsing vety môže byť vynechaný.Na analýzu vety existuje aj alternatívnymodul ExhaustiveSentenceParser,ktorý je robustnejší ako SentenceParser, ale o niečo pomalší pri analýze len malej časti veľkého korpusu.Každý z modulov v parse podknižke môže byť individuálne importovanýdo Python skriptu a použitý na extrakciu jednotlivýchvety, párov viet alebo XCES odkazov.

OpusCat je modul Python používanýskriptom opus_cat a oba majú rovnakú funkcionalitu čítania jednojazyčných súborov vety z OPUS.OpusCat využíva upravenú verziu modulu SentenceParser:Pri čítaní súborov s jednou vetou proces analýzy viet nemusí sledovať poradie špecifikované v súbore zarovnania a SentenceParser v OpusCat jednoducho vypustí každú vetu v súbore.Oba OpusCat a SentenceParser môžu byť importované ako moduly Python, aby mali podrobnú kontrolu nad čítaním jednojazyčných súborov.**OpusGet** modul napája opus_get skript s korpusmi-sťahovanie schopností.Importom modulu do kódu Python je možné získať podrobné informácie o korpusoch OPUS v rámci dátových štruktúr Python.Tieto informácie zahŕňajú počet párov zarovnania, počet dokumentov, počet žetónov a veľkosť v kilobajtoch okrem iných položiek.

OpusLangid modul má rovnakú funkcionalitu ako opus_langid skript:pridanie jazykových štítkov a skóre spoľahlivosti jazykovej identifikácie do súborov XML vety.Okrem toho OpusLangid obsahuje triedu LanguageIdAdder, ktorá môže byť použitá pre získanie-jazykových štítkov a identifikačné skóre spoľahlivosti z pycld2 a langid.py pre jednoduché textové vety s jedinou funkciou volanie.

3.3. Modul OpusTools Perl

Doplnkový balík nástrojov OPUS je poskytovaný ako Perl modul dostupný s povoľnou MIT licenciou.5 It obsahuje nástroje príkazového riadku, ktoré sú užitočné najmä pre vytváranie nových súborov údajov, ale aj všeobecne pre rýchly prístup k dátam v rôznych formátoch.Niektoré funkcie súteraz nahradené implementáciou v Python knižnici popísané vyššie, a tu sa zameriame na nástroje, ktoré podporujú ďalšie prípady použitia.Tieto nástroje patria najmä do týchto troch kategórií:

Nástroje na konverziu:Nástroje, ktoré môžu byť použité na import a export dát v rôznych formátoch súborov a dátových značiek.Hlavným účelom je importovať nové súbory dát do OPUS a vytvárať dátové súbory, ktoré sú uvoľnené v rôznych formátoch.

Nástroje na zosúlaďovanie: Zarovnanie viet a slov možno použiť rôznymi spôsobmi a tieto nástroje poskytujú niektoré pohodlné operácie na vrchole zladených bitextov.

Ostatné nástroje na spracovanie:Táto kategória zahŕňa nástroje na anotáciu a indexovanie.

V prvej kategórii máme nástroje na dovoz, ako napríklad Moses2opus, tmx2opus a xml2opus.Ex
Port skripty zahŕňajú opus2moses, tmx2moses, opus2text a opus2multi.

xml2opus je jednoduchý skript, ktorý pridáva hranice vetydo ľubovoľných XML dát.Detekcia hranice viet sa vykonáva pomocou nástrojov uvoľnených s Europarl paralelným-korpusom (Koehn, 2005) a zabalených v module Perl Lingua-::Sentence.V budúcnosti budú integrované ďalšie nástroje založené na klasifikátoroch UD Treebank.Inline značky, ktoré pridávajú značku vo vete, nie sú momentálne podporované.

Moses2opus číta jednoduché textové súbory, ktoré sa bežne používajú v strojovom preklade so zarovnanými vetami na rovnakom riadku.[5]Nástroj konvertuje dáta do jednoduchého-samostatného XML pre korpusové dáta a formát XCES Align pre zarovnanie standoff vety, pretože sa používa v OPUS.V súčasnosti je podporovaný iba dvojjazyčný vstup.Jednoduché textové súbory neobsahujú hranice vety, ale stále môžu obsahovať zarovnanie viet, ktoré nie sú individuálne.Preto moses2opus pridáva vetu značenie pomocouLingua::Vedenie a podľa tohoupravuje zarovnanie vety.Skript podporuje aj rozdeleniebitextov na menšie časti.Prázdne riadky v zdroji a

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

[5]Názov pochádza z Mojžišovho balíka, ktorý spopularizoval tento formát.

cieľový jazyk môže byť použitý na označenie hraníc dokumentu.Okrem toho korpus možno rozdeliť na rovnako veľké časti pomocou prahu dĺžky pre maximálny počet prekladových jednotiek zahrnutých v jednej časti.

tmx2opus prevádza prekladové pamäte vo formáte TMX do OPUS XML.Nástroj pridáva hranice vety rovnakým spôsobom ako moses2opus.To tiež umožňuje pripojiť niekoľko TMX súborov cez nástroj konverzie a je schopný zlúčiť informácie v prípade prekrývajúcich sa viet, ktoré sú zahrnuté vo viacerých prekladových jednotkách.To je užitočné pri spracovaní dát, ktoré prichádzajú ako rôzne bitexty, ale pokrývajú rovnaký obsah.Preto sú vo výslednom OPUS XML uložené iba jedinečné vety, aj keď sa objavujú v rôznych prekladateľských jednotkách s zarovnaním do rôznych jazykov. tmx2opus môže tiež spracovávať prekladové spomienkys viac ako dvoma jazykmi v prekladateľskej jednotkea vytvorí dvojjazyčné súbory zarovnania vety pre všetky jazykové páry, ako sú potrebné v OPUS.Okrem toho je tiež možné rozdeliť dáta na menšie časti podobné tomu, čo robí moses2opus.Vlastnosti z TMX súborov môžu byť tiež skopírované do prevedených dát, aby sa zachovali ďalšie meta dáta.Uplatnenie tmx2opus na vytvorenie dovážaného korpusu ParaCrawlv OPUS je opísané v oddiele 4.
Export skriptov prevažne vykonáva konverziu dát v opačnom-smere. opus2moses a opus2textkonvertujú dátaOPUS XML na obyčajný text a sú väčšinou zastarané a nahradené implementáciou balíka Python, ktorý bol zavedený skôr. tmx2mosesje pohodlný skript na extrahovanie zosúladených viet z ľubovoľných súborov TMX a neobmedzuje sa na OPUS dáta.

opus2multi je nástroj, ktorý dokáže vytvárať multiparallelné súbory dát z korpusov OPUS.V OPUS, všetky súbory dát sú zladené bilinguálne, ale v niektorých prípadoch jeden by chcel mať zarovnanie, ktoré zahŕňa viac ako dva jazyky.Za týmto účelom opus2multi môže pomôcť spojiť dvojjazyčné zarovnanie vetya extrahovať odkazy na väčší počet jazykov.-Nástroj pracuje na zarovnanie súborov standoff vety a využíva pivotný jazyk na vytvorenie prekladovýchjednotiek vo všetkých daných jazykoch.Na tento účel sa rozširuječiastočne prekrývajúce sa zarovnanie viet, až kým všetky jazyky nebudú pokryté bez ďalších konfliktov vo výslednej prekladateľskejjednotke (t. j. žiadne zostávajúce prekrývanie s ostatnými jednotkami).Výsledkom tohto procesu je nastavenie vety-súbory, ktoré sú (pre pohodlie) tlačené bilinguálne pomocou formátu XCES Align, ktoré potom môžu byť ďalej spracované pomocou OpusTools extrahovať skutočné páry zarovnania.Existuje tiež možnosť kontrolovať maximálnu veľkosť prekladovej jednotky (v počte viet v jednom jazyku), pretože veľkosť môže rásť bez obmedzení v procese expanzie.Súčasťou je aj experimentálna črta zahrnutia vnútrojazyčných spojení pre ďalšie tranzitné mapovanie.Je to užitočné pre súbory dát, ako sú otvorenétitulky, v ktorých môžu byť použité alternatívne súbory titulkov na prepojenie medzi rôznymi jazykmi.

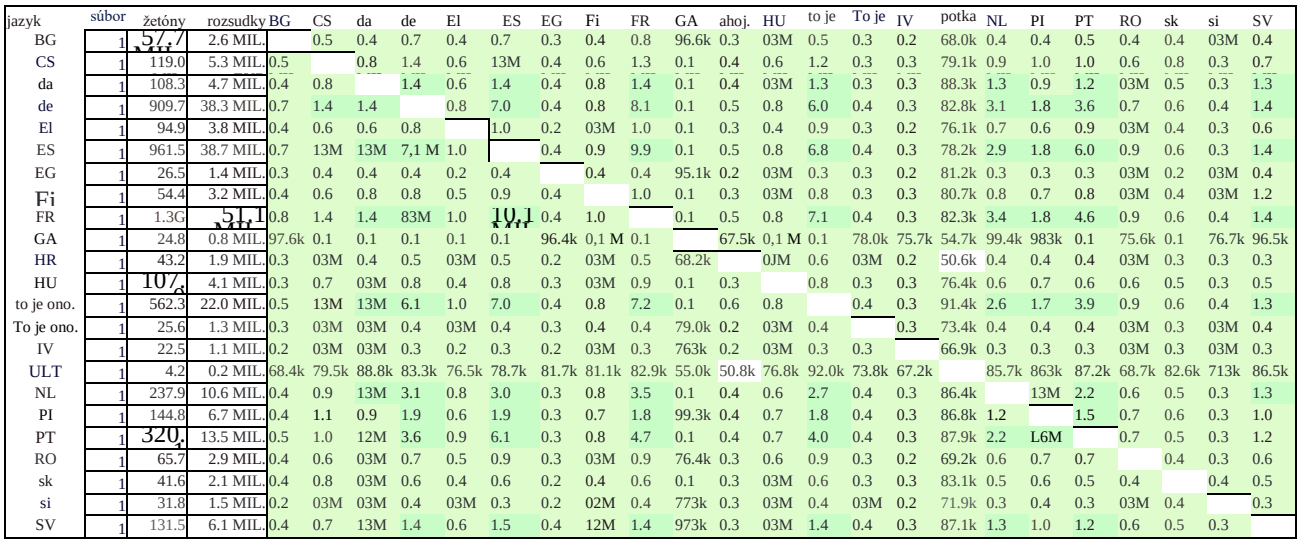
Nástroje na nastaveniev balíku OpusTools pomáhajú spracovávať zarovnanieviet v ich anotačnom formáte standoff. opus—swap —align jednoducho vymení smer zarovnania.Opus poskytuje iba zarovnanie v jednom smere (keďže sú symetrické), ale niekedy je výhodné mať prístup k odkazom aj v opačnom smere.opus-merge —align kombinujesúbory nastavenia vety avymaže duplikáty, ak existujú. opus-split-align rozdeľuje súbory zarovnania vety dosamostatných súborov s jednou na každú skupinu zarovnania, t. j. zladený dokument.Nakoniec, opus— pivoting umožňuje vytvoriť prechodnú vetu zosúlaďovanie medzi dvoma jazykmipomocou pivotného jazyka a odkazmi na pivotný jazyk.To je vhodné pre korpusy, ktoré sú dodávané sbitexty, ktoré nepokrývajú všetky jazykové páry, ale len zladiť so špecifickým jazykom, ako je angličtina.Za predpokladu, že medzi bitextmi dochádza k podstatnému prekrývaniu, povedzme A ^ P a B ^ P, opus-pivotingvýpisy z vetyv A a B, vytvorenie nového bitextu A ^ B. Oddiel 4.ilustruje použitie príkladom vytvorenia MultiParaCrawl.Nakoniec, ďalší nástroj na zarovnanie, opus—pt2dice, extrahuje hrubé pravdepodobnostné dvojjazyčné slovníky z tabuliek fráz-prekladov vytvorených zo zarovnania slov a pomocou SMT nástrojov vychádzajúcich z Moses toolboxu.Tieto slovníky používajú nejakú heuristiku na filtrovanie dát a nástroj tiež vytvára dodatočné skóre Dice ako symetrizovaná hodnota nastavenia z podmieneného prekladu pravdepodobnosti obsiahnutej v pôvodných tabuľkách fráz, čo je užitočné pre dvojjazyčnú lexikón extrakciu (Smadjaa kol., 1996).

Iné nástroje: Posledná kategória nástrojov obsahuje ďalšie nástroje na spracovanieúdajov, ako je opus —udpipe a opus —index.Prvý implementuje obal okolo UDPipe (Straka a Strakova, 2017) pre anotáciu OPUS dát a uložiť výsledok v OPUS-conforming XML.OpusTools môže používať predtrénované modely pochádzajúce z LIN- DAT.[6]V neposlednom radeje opus-index nástrojom na indexovanie korpusov OPUS pomocou stola Corpus Work Bench (CWB) (Evert and Hardie, 2011).Vytvára všetky importné súbory a spúšťa encoder, ak je k dispozícii na vytvorenie-multiparalelných korpusov, dotazovaných pomocou CWB vyhľadávača.

4. ParaCrawl a MultiParaCrawl

V tejto časti by sme radi predstavili import dát ParaCrawl na preukázanie použitia OpusTools.Korpus ParaCrawl[7] bolextrahovaný prehľadaním webu a použitím komplexného potrubia na zosúladenie dokumentov a viet nazáklade balíka Bitextor (Espla-Gomis, 2009).Aktuálna verzia v5.0 pokrýva 24 európskych jazykova projekt poskytuje automaticky vyčistené bitexty prejazyky zosúladené s angličtinou.Veľkosť sa pohybuje od 100 000 prekladových jednotiek (Maltese-English) až po viac ako 50 miliónovjednotiek (francúzsko-anglické) a dátové súbory sú distribuované v jednoduchom texte alebo formáte TMX.Zatiaľ čo existuje niekoľko bonusové jazykové páry, ktoré tiež obsahujú dva bitexty nezahŕňa angličtinu, väčšina z kolekcie je bilingválne zladená s anglickým obsahom.
Cieľom integrácie ParaCrawl do OPUS je sprístupniť dáta prostredníctvom natívneho OPUS formátu a tiež vyčerpávať všetky jazykové páry zahrnuté v kolekcii.Na tieto účely, predtým zavedené

[6] https://lindat.mff.cuni.cz
[7] https://paracrawl.eu



jazyk	súbor	žetóny	rozsudky	BG	CS	da	de	El	ES	EG	Fi	FR	GA	ahoj.	HU	to je	To je	IV	potka	NL	PI	PT	RO	sk	si	SV
BG	1	57.7	2.6 MIL.		0.5	0.4	0.7	0.4	0.7	0.3	0.4	0.8	96.6k	0.3	03M	0.5	0.3	0.2	68.0k	0.4	0.4	0.5	0.4	0.4	03M	0.4
CS	1	119.0	5.3 MIL.	0.5		0.8	1.4	0.6	13M	0.4	0.6	1.3	0.1	0.4	0.6	1.2	0.3	0.3	79.1k	0.9	1.0	1.0	0.6	0.8	0.3	0.7
da	1	108.3	4.7 MIL.	0.4	0.8		1.4	0.6	1.4	0.4	0.8	1.4	0.1	0.4	03M	1.3	0.3	0.3	88.3k	1.3	0.9	1.2	03M	0.5	0.3	1.3
de	1	909.7	38.3 MIL.	0.7	1.4	1.4		0.8	7.0	0.4	0.8	8.1	0.1	0.5	0.8	6.0	0.4	0.3	82.8k	3.1	1.8	3.6	0.7	0.6	0.4	1.4
El	1	94.9	3.8 MIL.	0.4	0.6	0.6	0.8		1.0	0.2	03M	1.0	0.1	0.3	0.4	0.9	0.3	0.2	76.1k	0.7	0.6	0.9	03M	0.4	0.3	0.6
ES	1	961.5	38.7 MIL.	0.7	13M	13M	7,1 M	1.0		0.4	0.9	9.9	0.1	0.5	0.8	6.8	0.4	0.3	78.2k	2.9	1.8	6.0	0.9	0.6	0.3	1.4
EG	1	26.5	1.4 MIL.	0.3	0.4	0.4	0.4	0.2	0.4		0.4	0.4	95.1k	0.2	03M	0.3	0.3	0.2	81.2k	0.3	0.3	0.3	03M	0.2	03M	0.4
Fi	1	54.4	3.2 MIL.	0.4	0.6	0.8	0.8	0.5	0.9	0.4		1.0	0.1	0.3	03M	0.8	0.3	0.3	80.7k	0.8	0.7	0.8	03M	0.4	03M	1.2
FR	1	1.3G	51.1	0.8	1.4	1.4	83M	1.0	10.1	0.4	1.0		0.1	0.5	0.8	7.1	0.4	0.3	82.3k	3.4	1.8	4.6	0.9	0.6	0.4	1.4
GA	1	24.8	0.8 MIL.	97.6k	0.1	0.1	0.1	0.1	0.1	96.4k	0,1 M	0.1		67.5k	0,1 M	0.1	78.0k	75.7k	54.7k	99.4k	983k	0.1	75.6k	0.1	76.7k	96.5k
HR	1	43.2	1.9 MIL.	0.3	03M	0.4	0.5	03M	0.5	0.2	03M	0.5	68.2k		0JM	0.6	03M	0.2	50.6k	0.4	0.4	0.4	03M	0.3	0.3	0.3
HU	1	107.8	4.1 MIL.	0.3	0.7	03M	0.8	0.4	0.8	0.3	03M	0.9	0.1	0.3		0.8	0.3	0.3	76.4k	0.6	0.7	0.6	0.6	0.5	0.3	0.5
to je ono.	1	562.3	22.0 MIL.	0.5	13M	13M	6.1	1.0	7.0	0.4	0.8	7.2	0.1	0.6	0.8		0.4	0.3	91.4k	2.6	1.7	3.9	0.9	0.6	0.4	1.3
To je ono.	1	25.6	1.3 MIL.	0.3	03M	03M	0.4	03M	0.4	0.3	0.4	0.4	79.0k	0.2	03M	0.4		0.3	73.4k	0.4	0.4	0.4	03M	0.3	03M	0.4
IV	1	22.5	1.1 MIL.	0.2	03M	03M	0.3	0.2	0.3	0.2	03M	0.3	763k	0.2	03M	0.3	0.3		66.9k	0.3	0.3	0.3	03M	0.3	03M	0.3
ULT	1	4.2	0.2 MIL.	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	237.9	10.6 MIL.	0.4	0.9	13M	3.1	0.8	3.0	0.3	0.8	3.5	0.1	0.4	0.6	2.7	0.4	0.3	86.4k		13M	2.2	0.6	0.5	0.3	1.3
PI	1	144.8	6.7 MIL.	0.4	1.1	0.9	1.9	0.6	1.9	0.3	0.7	1.8	99.3k	0.4	0.7	1.8	0.4	0.3	86.8k	1.2		1.5	0.7	0.6	0.3	1.0
PT	1	320.	13.5 MIL.	0.5	1.0	12M	3.6	0.9	6.1	0.3	0.8	4.7	0.1	0.4	0.7	4.0	0.4	0.3	87.9k	2.2	L6M		0.7	0.5	0.3	1.2
RO	1	65.7	2.9 MIL.	0.4	0.6	03M	0.7	0.5	0.9	0.3	03M	0.9	76.4k	0.3	0.6	0.9	0.3	0.2	69.2k	0.6	0.7	0.7		0.4	0.3	0.6
sk	1	41.6	2.1 MIL.	0.4	0.8	03M	0.6	0.4	0.6	0.2	0.4	0.6	0.1	0.3	03M	0.6	0.3	0.3	83.1k	0.5	0.6	0.5	0.4		0.4	0.5
si	1	31.8	1.5 MIL.	0.2	03M	03M	0.4	03M	0.3	0.2	02M	0.4	773k	0.3	03M	0.4	03M	0.2	71.9k	0.3	0.4	0.3	03M	0.4		0.3
SV	1	131.5	6.1 MIL.	0.4	0.7	13M	1.4	0.6	1.5	0.4	12M	1.4	973k	0.3	03M	1.4	0.4	0.3	87.1k	1.3	1.0	1.2	0.6	0.5	0.3	

Obrázok 5:Štatistiky z multiParaCrawl korpusu – viacjazyčné rozšírenie ParaCrawl cez zarovnanie v angličtine.Horný-pravý trojuholník dáva veľkosť, pokiaľ ide o zarovnanie vety v jednoduchom textovom formáte, a spodný-ľavý trojuholník ukazuje veľkosť extrahovaných súborov TMX z hľadiska jedinečných prekladových jednotiek na jazykový pár.

nástroje tmx2opus a opus— pivoting sa stávajú šikovnými. tmx2opusje užitočnýnielen pre extrakciu zarovnania z pôvodného zdroja TMX, ale tiež poskytuje funkcionalitu pre-pridanie hranice viet a zníženie nadbytočnosti medzi rôznymi bitextmi.Použitie jedinečnej možnosti tmx2opus zmenšuje veľkosť anglickejčasti korpusu (t. j. 252 miliónov oddelene zosúladených anglických vietv 23 bitextoch) na menej ako 60 % pôvodných dát.Zároveň jedinečnosť umožňuje vytvoriť multiparalelálny korpus tým, žesa v novo vytvorenej jedinečnej sade viet presmeruje na odkazy na angličtinu.Na to jemožné použiť opus— pivoting, ako bolo vysvetlené skôr.Pomocou tohto postupu by sa mohlo vytvoriť ďalších 253bitextov s veľkosťou až 10 miliónov vety zosúladených prekladových jednotiek.Obrázok 5 sumarizuje neanglické bitexty v Mul- tiParaCrawl.

5. Paralelná korpusová diagnostika

Naša diagnostická rutina pre kolekciu OPUS využíva nástroj príkazového riadku opus_read (popísaný v časti3.1.) na získanie zladených textových údajov pre konkrétny jazykový-pár v danom korpuse.Aby to bolo možné,opus_read analyzujenatívny XML dáta na generovanie požadovanej-podmnožiny dát a potom vykoná konverziu na obyčajný textovýformát.Počas tohto procesu diagnostickárutina počúva akékoľvek chyby, ktoré by mohli vzniknúť, a zaznamenáva ich na zostavenie diagnostickej správy pre neskoršie analýzy.Tento postup vykonávame systematicky pre každý pár jazykov dostupných v rámci každého z korpusov OPUS.[8]Pre našu diagnostiku využívame úplnú granularitu OPUS tým, že zhromažďujeme samostatné čítania pre rôzne korpusy,ktoré tvoria bitexty, a tiež udržiavame regionálne variantyjazykov oddelene namiesto ich spájania.Ak chcete vykonať tento druh vyčerpávajúcej analýzy, spustili sme celkom 87.948 CPU úlohy paralelne, s beží od 1 sekundy do 5,2 hodín, a každá úloha používa medzi 4a 128 GB pamäte.Celkove celá diagnostickáanalýza trvala približne 1000 hodín výpočtu, vpriemere 18,2 hodiny na korpus.Zatiaľ čo granularitanašej analýzy bude vnútorne užitočná pre určenie anomálií v OPUS, aby sa uľahčila oprava, zhromažďujeme aj naše dáta, aby sme generovali celoplošné korpusové čísla, ktoré podávame a diskutujeme v tejto časti.

5.1. Analýza chýb

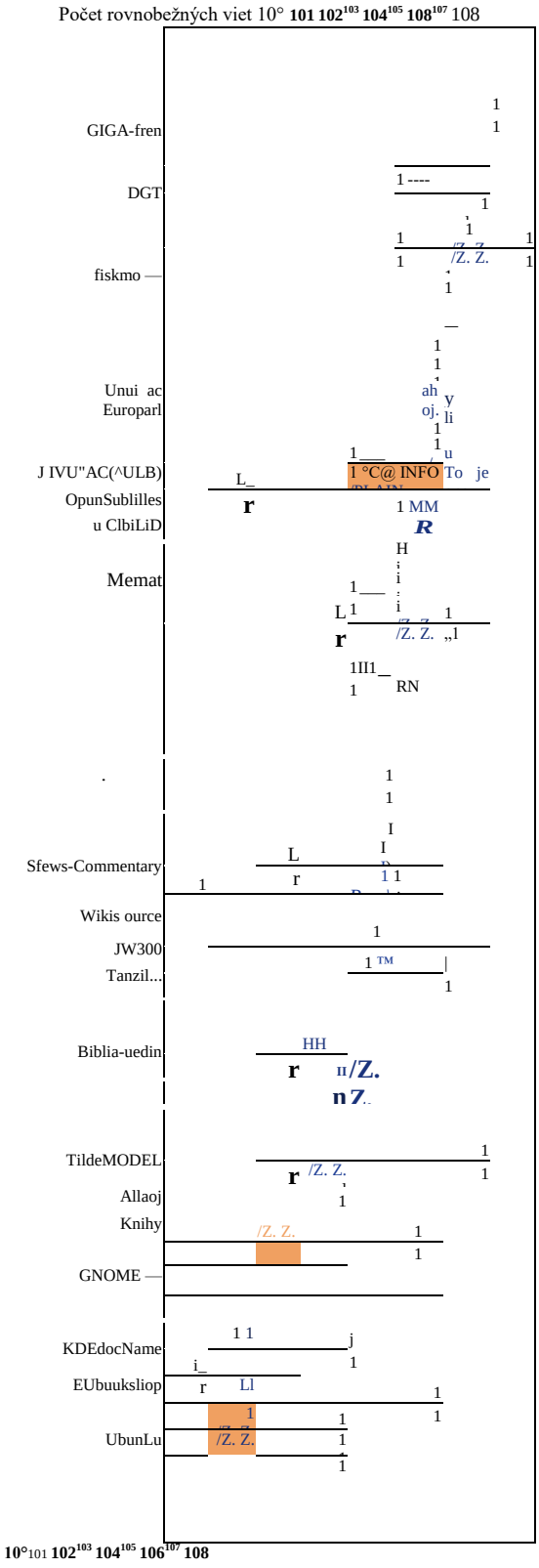
V „diagnózach“ prihlásených v našej správe sú uvedené príčiny každej chyby vyhľadávania, ktorá nám poskytuje prostriedky na spoľahlivú lokalizáciu a opravu.Výsledky zhromažďujú všetky diagnózy a výsledky ukazujú, že zatiaľ čo 37 korpusov je úplne bezchybných, vyhľadávanie údajov sa zastavilo aspoň pre jednu jazykovú dvojicu pre zvyšných 18 korpusov.Množstvochýb pri získavaní v týchto korpusoch sa líši od malého zlomku po celý korpus (obrázok 8).Prevažná väčšinatýchto chýb pochádza z zle formovaných XML dát s neplatnými tokenmi (96,2 %) alebo nesúladnými štítkami (3,5 %).Naše doterajšie čiastočné kontroly naznačujú, že tieto môžu byť pripísané menším chybám konverzie, ako sú neuniknuté špeciálne znakya XML subjekty vyskytujúce sa v pôvodných dátach pred importom do OPUS.Ďalšia veľmi malá časť chýb (0,3 %) indikuje chýbajúce dátové súbory v hlavnom súborovom systéme, kde je OPUS hostovaný, čo pravdepodobne indikuje chyby v kopírovaní a zostáva, aby boli ďalej preskúmané.

5.2. Korpusovo-široké štatistiky

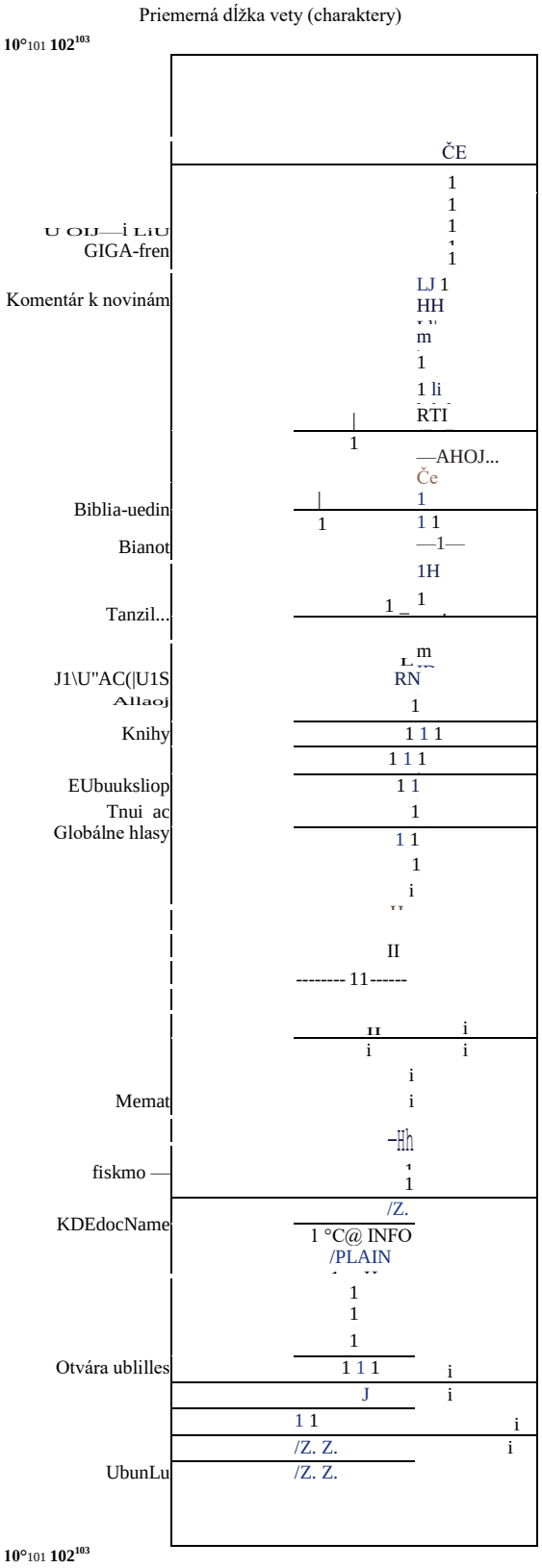
Okrem katalogizácie otázok získavania údajov náš diagnostický postup počíta aj niektoré základné kvantitatívne-štatistiky, ako sú vykázané náklady na výpočet dát a rôzne merania údajov získaných podľakorpusu, jazyka a jazykového páru.Naše zodpovedajúceštatistické analýzy neodhalili významné trendy ani extrémne hodnoty, okrem niektorých meraní, ktorénaznačovali relatívne odchýlky a hladiny hluku v údajochv korpusoch.Na obrázkoch 6 a 7 podávame správy

Na dvoch posledných dodatkochOPUS[8]sme nevykonali diagnostiku:Infopankki a MultiParaCrawl.

rozdelenie dvoch meraní na súbor dostupných lan- priemerná dĺžka vety v znakoch, v uvedenom poradí.Oba páry guage pre každý korpus:priemerný počet meraní sen- bol vizualizovaný pomocou box-and-whisker parcely na desať páry (alebo, presnejšie, prekladové jednotky), a zdôrazniť distribučné rozdiely, kde koncové ukazovatele

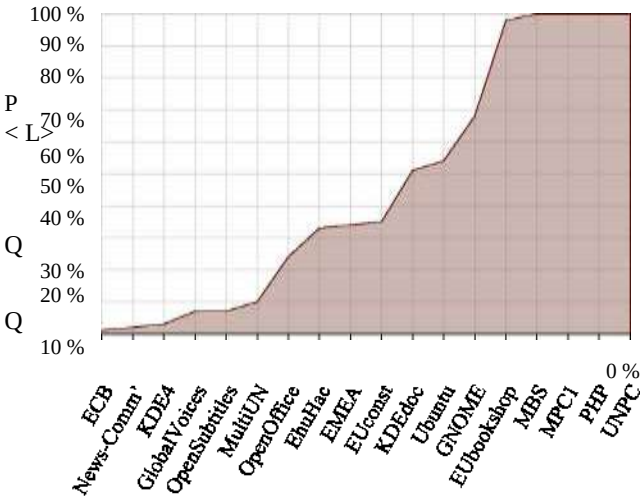


Obrázok 6:Rozdelenie počtu načítateľných paralelných vety párov v sade dostupných jazykových párov pre každý korpus.



Obrázok 7:Rozdelenie priemerných vetových dĺžok (v znakoch) cez súbory dostupných jazykových párov pre každý

korpus.



Obrázok 8:Percentuálne podiely jazykových párov v korpusoch OPUS, za ktoré načítanie dát pomocou nástrojov OPUS vracia chyby.Bezchybné korpusy boli vynechané z grafu.

Uveďte najnižšie a najvyššie hodnoty[9]a dve polovice rámčeka predstavujú druhý a tretí kvartil hodnôt oddelených mediánom.

Jedným z najvýraznejších detailov na obrázku 6 je kontrast-medzi rozptylmi.Viac ako tretina korpusov vykazuje veľmi málo až žiadny rozdiel medzi jazykovými pármi, čo znamená úplne multiparallelné dáta, zatiaľ čo iné ako JW300 a otvorené titulky naopak vykazujú veľmi vysokú varianciu, kde rozdiel vo veľkosti dostupných dát môže siahať niekoľko rádov.Keď sa pozrieme konkrétne na prvé kvartily, zdá sa, že niektoré korpusy ako QED a Tatoeba majú značnú časť jazykových párov obsahujúcichveľmi málo prekladových jednotiek, čo pravdepodobne naznačuje vysokú jazykovú detekciualebo zvuk zarovnania vety.Na obrázku 7 sa zdá, že prvý kvartil má podobný relatívny rozsah pre niektoré korpusy, čo znamená, že vety obsahujú v priemere len niekoľko znakov pre niektoré z dostupných jazykových párov.Pravdepodobne nie je náhoda, že tieto prípady väčšinou zodpovedajú korpusom, ktoré boli zostavené z prirodzene hlučných údajov.Okrem toho najmenšie a najväčšie stredné hodnoty na obrázku 7 poukazujú na výnimočne krátke a výnimočnedlhé „typické“ vety v príslušných korpusoch,ktoré môžu naznačovať silný kontrast v segmentácii textu alebo-zreteľne odlišné dátové domény.Napríklad tri korpusy s najnižšími mediánmi zahŕňajú prekladypočítačového softvéru, zatiaľ čo dokumenty z OSN majú najvyššiu strednú dĺžku.

6. Závery a budúca práca

V tomto dokumente predstavujeme OpusTools, open-source balík knižníc a nástrojov príkazového riadku pre efektívny a pohodlný prístup k paralelným korpusom do rozsiahleho zberu dát OPUS.Balík implementuje nástroje pre sťahovanie, konverziu, filtrovanie a spracovanie paralelnýchdátových súborov a uľahčuje prístup ku komprimovaným a archivovaným súborom zo zbierky.Poskytuje tiež pytónovú knižnicu pre programový prístup k dátam, vďaka čomu je možné jednoducho začleniť spracovanie dát do vývoja iných

[9]Neobmedzené koncové ukazovatele naznačujú, že extrema presahuje hranice osi x.

.Okrem toho predstavujeme nástroje na konverziu a zosúlaďovanie údajov, ktoré je možné použiť pri príprave nových súborov údajov z rôznych zdrojov.Ich použitie demonštrujeme pomocou príkladu nedávno pridaného multiParaCrawl korpusu, ktorý rozširuje pôvodný súbor dát o pivotné zosúlaďovanie všetkýchjazykových párov, čo prispieva k rastúcemu pokrytiudatabázy OPUS.

Hoci udržať kolekciu tak veľké, ako OPUS dokonale robustné je docela náročné, riešenie problémov bude jednoduchšie a rýchlejšie s diagnostikou plne zmapované.Celkovo, zatiaľ čo chyby pri získavaní údajov obsahujú jasné akčné body, zdá sa, že štatistické analýzy skôr naznačujú výraznú kvalitatívnua kvantitatívnu rozmanitosť medzi korpusmi OPUS, s trendmi zdanlivo v rámci očakávaní a okrajovými prípadmi, ktoré možno pripísať hluku v pôvodných údajoch.Naším zámerom je vyriešiť všetky problémy týkajúce sa získavania dát tak, aby používanie nástrojov OPUS bolo bezproblémovým zážitkom pre všetkých užívateľov a tiež zefektívniť našu rutinu ako diagnostický nástroj, ktorý by sa stal štandardnou súčasťou procesu rozširovania OPUS o nové korpusy.

Uznanie

ERC Táto práca je súčasťou projektu FoTran financovaného Európskou radou pre výskum (ERC)

H Der Program Európskej únie v oblasti výskumua inovácií Horizont 2020 (dohoda o grante narok[771113),] ako aj projekt MeMAD financovaný z programu Európskej únie pre výskum a inovácie Horizont 2020 (dohoda o grante na rok 780069).

7. Bibliografické odkazy

Evert, S. a Hardie, A. (2011).Pracovný stôl z dvadsiateho prvéhostoročia:Aktualizácia architektúry dotazu pre nové tisícročie.In *Proceedings of the Corpus Linguistics2011, University of* Birmingham, UK.Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996).Translačné kolokácie pre dvojjazyčné lexikóny:Štatistický prístup.*Computational Linguistics*, 22(1):1 – 38.

8. Jazykové odkazy na zdroje

Espla-Gomis, Miquel.(2009).*Bitextor:Free/Open- source Softvér na zber prekladu Spomienky z viacjazyčných webových stránok.*

Koehn, Philipp.(2005).*Europarl:Paralelný korpus pre štatistický preklad stroja*.AAMT. – AAMT.

Lui, Marco a Baldwin, Timothy.(2012). *langid.py:Nástroj na identifikáciu jazyka off-theshelf*.Asociáciapre počítačovú lingvistiku.

Straka, Milan a Strakova, Jana.(2017).*Tokenising, POS Tagging, Lemmatising a Parsing UD 2.0 s UDPipe*.Asociácia pre počítačovú lingvistiku.

Tiedemann, Jorg.(2012).*Paralelné dáta, nástroje a rozhraniav OPUS*.Európska asociácia jazykových zdrojov(ELRA).