Jednání 12. konference o jazykových zdrojích a hodnocení (LREC 2020), s. 3782-3789
Marseille, 11.–16. května 2020

OpusTools a paralelní Corpus diagnostika

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, Jorg Tiedemann
Ústav digitálních humanitních studiíUniversity
of Helsinki, Helsinki/Finsko{
mikko.aulamo, umut.sulubacak, sami.virpioja, jorg.tiedemann}@helsinki.fi

Abstraktní

Tento článek představuje OpusTools, balíček pro stahování a zpracování paralelních korpusů obsažených ve sbírce OPUS corpus.Balíček implementuje nástroje pro přístup ke komprimovaným datům v archivovaném formátu vydání a umožňuje snadnou konvertovat mezi běžnými formáty.OpusTools obsahuje také nástroje pro identifikaci jazyka a filtrování dat a nástroje pro import dat z různých zdrojů do formátu OPUS.Použití těchto nástrojů ukazujeme v paralelní tvorbě korpusu a diagnostice dat.Ta je zvláště užitečná pro identifikaci potenciálních problémů a chyb v rozsáhlém souboru údajů.Pomocí těchto nástrojů nyní můžeme sledovat platnost datových souborů a zlepšit celkovou kvalitu a konzistenci sběru dat.

Klíčová slova: Korpus (tvorba, Anotace atd.);Strojový překlad;Nástroje, systémy, aplikace

1. Úvod

Opus (Tiedemann, 2012) je největší sbírkou otevřeně dostupných paralelních korpusů.Kolekce neustále roste v průběhu let a je široce používán v práci na strojový překlad a cross-lingvistický výzkum.V současné době obsahuje 57 uvolněných sborů pokrývajících více než 700 jazyků a jazykových variant, které vytvářejí více než 70 000 bitextů ve smyslu sladěných jazykových párů napříč všemisbory ve sbírce.Velikost a popularita OPUS vyžaduje vybudování efektivní infrastruktury, která umožňuje různým uživatelům získat data a přístup k nim, a tento dokument zavádí dva balíčky, které poskytují nástroje pro tento účel.Cílem těchto balíčků je usnadnit stahování, převod a zpracování dat obsažených v OPUS z příkazového řádku nebo z aplikací využívajících knihovny, které tyto nástroje implementují.Oba balíčkyse vztahují na knihovnu Python s nástroji příkazové řádky a komplementárnímmodulem Perl, které jsou poskytovány jako open source a s volitelnými licencemi.

V níže uvedených sekcích uvádíme nástroje a jejich základní využití a také diskutujeme o tom, jak jsme tyto nástroje aplikovali k vytvoření nových datových souborů a k systematické diagnostice celé databáze.S dostupností OpusTools je nyní možné provádět pečlivé kontroly zdravého zdraví rozsáhlých datových souborů k ověření platnosti kódování, k nalezení poškozených odkazů a struktur a k identifikaci dalších problémů s daty.

2. Charakteristika OPUS

Opus zahrnuje paralelní corpora z široké škály zdrojů.Každý z nich přichází s jejich vlastní zvláštnostia vlastnosti se mohou podstatně lišit v závislosti na původních dat a jejich distribuce.Filozofií OPUS je zachovat přirážku a anotaci co nejvíce,ale sjednotit základní datový formát tak, aby byl přístup k paralelním údajům co nejtransparentnější.To znamená, že corpus data jsou převedena na samostatný (schema-free) XML, který udržuje původní přirážku, ale důsledně přidává základní přirážku, která je nezbytná pro zarovnání a další jazykovézpracování.Zarovnání je uloženo jako standoff anotace ve formátu XCES Align (pro zarovnání věty) a „Moses formát“ (pro zarovnání slov).Pomocí tohoto principu mohou být data uchovávána odděleně od zarovnání anotace, která umožňuje efektivní implementaci a ukládání masivně paralelních dat a v případě potřeby také umožňuje alternativní zarovnání.Obrázek 1 znázorňuje příklad patové anotace použité v OPUS pro určení spojení mezi větami.Každý soubor sezarovnání věty může obsahovat libovolný počet prvků linkGrp pro sladění dokumentů ze sběru dat.Dokumenty jsou specifikovány pomocí cesty vzhledem k XML kořeni OPUS sub-corpus a linkové prvky poskytují zarovnání věty sadami identifikátorů věty, které jsou odděleny středníkem.Vytvoření alternativního zarovnání se jednoduše provádí vytvořením novéhosouboru zarovnání věty a není třeba provádět další úpravy s původními corpusovými daty.Všimněte si, že zarovnání věty je dvojjazyčné, jak je uvedeno v příkladu.Nicméně, standoff anotaceumožňuje sladit masivně paralelní datové soubory napříč všemi jazykovými páry bez duplikace některé z propojených datových souborů.Kromě toho mohou existovat alternativní soubory corpus s různými úrovněmi anotace, aniž by bylo nutné tyto alternativní soubory přeladit.Obrázek 2 ukazuje příklady takových anotovaných souborů, které jsou zarovnány stejným způsobem se zarovnáním věty od patové věty uložené v externích souborech.Více informací o datových strukturách v OPUS naleznete na Wiki OPUS.[1]

Dalším principem v OPUS je poskytnout data v jiných běžných formátech, aby byly snadno přístupné pro širokou škálu aplikací.Tyto datové formáty jsou však právě generovány ze základního kódování založeného na XML, které slouží jako hlavní kopie každého corpusu.Uživatelé OPUS dat si obvykle nejsou vědomi těchto principů a stahují datový formát, který nejvíce vyhovuje jejich potřebám.

Myšlenka OpusTools je nyní sjednotit přístup k hlavním datům v XML a k dalším generovaným formátům poskytnutímzákladních knihoven a nástrojů příkazové řádky pro získávání a konverzi corpus dat.Poskytují také pohodlné nástroje pro základní filtrování a náhodný přístup v archivovaných datech v jejich komprimované formě, která se používá pro distribuci dat.To druhé je obzvláště důležité, protože velikost některých kor—

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

```
<?xml verze=„1.0“ enkódování=„utf-8“?>
<!DOCTYPE cesAlign PUBLIC
                    „-//CES//DTD XML cesAlign//EN“ ""> <cesAlign
version=„1.0“>
<linkGrp targType=„s“
        fromDoc=„en/0/1089124/4995691.xml.gz“
        toDoc=„fr/0/1089124/4588599.xml.gz“>
Odkaz id=„SL0“ xtargets=„1;1“ překrývá =„0.331“/>
Odkaz id=„SL1“ xtargets=„2 3;2“ překrývá =„0.560“/>link id=„SL2“
xtargets=„4;“/>
Odkaz id=„SL3“ xtargets=„5 6;3“ překrývá=„0.854“/>link id=„SL4“
xtargets=„7 8 9;4“ překrývá=„0.699“/>link id=„SL5“ xtargets=„10 11;5“
překrývání=„0.776“/>
```

Obrázek 1:Příklad vyrovnané věty ve formátu XCES Align.Element linkGrp určuje páry dokumentů, které jsou zarovnány a v linkovýchprvcích jsou uvedeny odkazy mezi jednotlivýmivětami.Volitelnéatributy překrývání v tomto příkladu se vztahují na poměry časového překrývání, které se používají jako funkce v zarovnání titulků.

Pora je rozsáhlá takovým způsobem, že vyžaduje, aby běžné souborové systémy zpracovávaly data v surové, nekomprimovanéformě.Například nejnovější opensubtitles corpus obsahuje zhruba 3,7 milionu individuálních dokumentů ve 67 jazycích se zarovnáním ve více než 3 600 bitextech.Jedním z nejnovějších dodatků, JW300 pokrývá 380 jazyků ve více než 46 000 bitextů.Celkem existuje více než 9,2 milionu jednotlivýchdokumentů pouze v nejnovějších vydáních všech sborů a toto číslo je zdvojnásobeno různými typy předzpracování, které jsou poskytovány, surový text a tokenizovaný sbor, které jsou částečně anotovány dodatečnými jazykovými informacemi.Kromě toho se bitexty uvolňují v nativním formátu XML(viz obr. 2), formátu prostého textu a překladové paměti(TMX).Vydání v současné době zabírají celkem 5,9 TB prostoru v komprimovaném formátu.

Výše uvedenáčísla ilustrují potřebu vhodných infrastruktura účinných nástrojů pro správu různých datových souborů.To je motivace k implementaci volně dostupných OPUS nástrojů popsaných níže.Vytvoří pohodlnou knihovnua sadu nástrojů pro stahování, extrahování a konverzi datze sbírky OPUS.Kromě toho pomáhají provádět systematickou diagnostiku sběru pro identifikaci chyb a problémů v souborech dat.Níže budeme nejprve prezentovat oba balíčky a jejich funkčnost.Poté poskytujeme informace o jejich využití při vytváření nových datových souborů a nakonec podáváme zprávu o aplikaci OPUS nástrojů pro diagnostické studie a kontrolu zdraví.

3. Balíček OpusTools

Balíček OpusTools je sada nástrojů pro stahování a správu paralelních corpora dat z OPUS.Balíčekse skládá z knihovny Python a souvisejících skriptů příkazových řádků.Navíc existuje balíček Perl pro vytvoření nových datových souborů a přístup k paralelním datům.

3.1. Nástroje pro příkazovou linku

Balíček OpusTools obsahuje pět skriptů založených na příkazovém řádku Python 3:Opus_read, opus_express, opus_cat, opus_get a opus_langid.2[Skripty] umožňují stahování OPUS dat, výstup dat v konkrétních formátech,extrahování školení, vývoje a testovacích souborů z dat, a další.Obrázek 3 zobrazuje přehled skriptů.

opus_read je skript pro stahování paralelních korpusů a jejich převod do požadovaných formátů.Opus corpora obsahuje soubory pro zarovnání formátu XCES, které ukazují na dva soubory XML vět v různých jazycích.Formát zarovnání XCES spojuje věty ve zdrojových souborech s větami v cílových souborech pomocí věty ID. Soubory věty v OPUS corpora jsou komprimovány do archivů ZIP a opus_read umožňuje číst data přímo z komprimovaných souborů. opus_read parsuje daný zarovnání soubor a produkuje výstup v jednom ze čtyř formátů:normální, můstkové, TMX nebo XCES odkazy. opus_readse nejprve pokusí číst OPUS soubory z lokálních adresářů.Pokud požadované soubory nejsou nalezeny, nástroj nabízí možnost jejich stažení.Větové soubory lze stáhnout v surovém, tokenizovaném nebo parsovaném formátu.

opus_read obsahuje základní filtry pro odstranění nežádoucích párů vět před vytvořením výstupního souboru.Nonalignments, kde je zdroj nebo cílový segment prázdný, lze vynechat.Alternativně lze specifikovat určitý počet zdrojových a cílových segmentů, např. je možné do výstupu zahrnout pouze jedno zarovnání.Některé korpora obsahují atribut skóre pro každý pár vět.Například, páry vět v opensubtitles corpus mají překrývající skóre, které ukazují, do jaké míry se časové razítka obou segmentů překrývají.opus_read je schopen odfiltrovat páry vět, které nepřekračují daný práh skóre atributu.Kromě toho lze páry segmentů odstranit nazákladě skóre důvěry v jazykovou identifikaci.Jazykové štítky a skóre důvěry lze přidat do větovýchXML souborů pomocí skriptu opus_langid.

opus_express je skript postavený na opus_read, který dokáže extrahovatpřipravené k použití tréninku, vývoje a testování pro jazykový pár z jednoho nebo více OPUS corpora.Postup-nejprve vyplňuje specifikovanou kvótu vět pro zkušební sadu, poté pokračuje stejným způsobem pro vývojovou sadu a zbytek se vloží do tréninkové soupravy.Skript může volitelně předmíchatdata před rozdělením, nebo naopak označit a uchovat hranice dokumentů napříčrozdělením pro modely na úrovni dokumentů. opus_express také obsahuje možnost využít skóre atributů, jako jsou hodnoty překrývající se hodnoty, jak jsou extrahovány opus_read ve svém přepínání uvědomění kvality, cožupřednostňuje dvojice vět s vyšší spolehlivostí, které převyšují konfigurovatelný práh, který má-být seřazen do testovacích a vývojových sad.

opus_cat sepoužívá pro čtení monojazyčného sboru z OPUS nebo jednotlivých souborů v rámci těchto sborů.Soubory mohou být vytištěny ve formátu XML nebo mohou být převedeny na prostý text. opus_cat je užitečný pro manuální kontrolu domény nebo kvality jednoho korpusu, protože je schopen číst soubory přímo z archivů ZIP v OPUS corpora.

opus_get je skript pro stahování paralelních corpus souborů z OPUS.Před stahováním může být corpora vyhledána a uvedena podle jejich jména, zdrojového jazyka a cílového jazyka.Například, jeden může stahovat soubory pro konkrétní jazykový pár v jednom corpus, všechny jazykové dvojice souborů v

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

Surový formát XML:
```
<?xml verze=„1.0“ enkódování=„utf-8“?>
< dokument>
   <CHAPTER ID=„1“>
      <P id=„1“>
         <s id=„1“>Použití relace </s>
      </P>
      ID=1" NAME="předseda">
         <P id=„2“>
            „2“>Prohlašuji zasedání Evropského parlamentu přerušené ve čtvrtek 14. června 2001. Prohlašuji za obnovené zasedání Evropského parlamentu přerušené ve
         čtvrtek 14. června 2001.
```

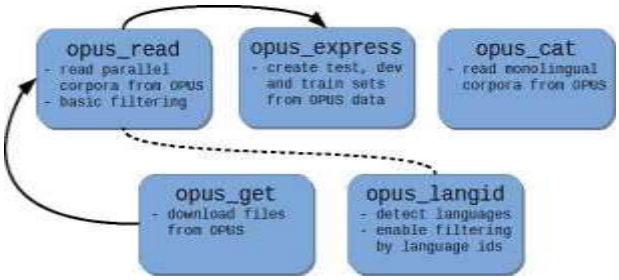
Tokenizovaný (neoznámený) formát XML:

```
<?xml verze=„1.0“ enkódování=„utf-8“?>
Dokument> CHAPTER ID = „1“> P id=„1“>
<s id=„1“>
Typ <chunk=„NP“ id=„c-1“>
   <w hun=„NN“ tree=„N“ lem=„resumption“ Pos=„NN“ id=„w1.1“>Resumption</w>
</chunk>
Typ <chunk=„PP“ id=„c-2“>
   „IN" strom = "IN" lem="Pos="IN" id="w1.2"> of</w>
</chunk>
Typ <chunk=„NP“ id=„c-3“>
   <w hun="DT" tree="DT" lem="Pos="DT" id="w1.3"> the</w>
   <w hun=„NN“ tree=„NN“ lem= „session“ pos=„NN“ id=„w1.4“>session</w>
</chunk>
</s>
```

UD Parsed XML formát:

```
<?xml verze=„1.0“ enkódování=„utf-8“?>
< dokument>
<CHAPTER ID=„1“>
   <P id=„1“>
     <s id=„1“>
      Xpos="NUN" head="0" feats="Number=Sing" UPOS="NOUN" lemma="Resumption" id="1.1" deprel=""ADP="> Resumption< of" id="1.2;>>“
      Xpos="DET" head="1.4" Feats="Definite=Def|PronType=Art" UPOS="DET" lemma="the" id="1.3" deprel=">the</w>"Nopos="NOUN" head="1.1"
      feats="Number=Sing" UPOS="NO"="N“
     </s>
```

Obrázek 2:Příklady XML kódovaných dat v OPUS.Různé druhy anotací lze přidat bez zničení zarovnání věty, která je uložena jako standoff anotace odkazů mezi průkazy věty.



Obrázek 3:Pět Pythonových scénářů OpusTools.Každý ze skriptů může být použit samostatně. opus_express je postaven na opus_reada opus_read používá opus_get ke stahování souborů OPUS. opus_langid musí být aplikovándo větovýchsouborů, aby bylo možné filtrovat jazyk id pro opus_read.

jediné corpus nebo všechny soubory pro konkrétní jazyk v celém OPUS. opus_read používá opus_get pro automatické stahovánípožadovaných souborů corpus.

opusJangid sepoužívá pro přidávání jazykových identifikačních štítkůa skóre spolehlivosti pro každou větu v daném souboru XML věty.Identifikace jazyka se provádí pomocí dvou nástrojů mimo polici:Pycld2,[3]Python vazby pro Compact Language Detector 2 a[4] langid.py (Lui a Baldwin, 2012). opus_langid musí být aplikován na věty XML soubory, nežopus_read může filtrovat větydvojice jejich jazykových štítků.Obrázek 4 ukazuje příklad souboru věty, který byl zpracován s opus_langid.

3.2. Knihovna OpusTools Python

Kromě skriptů příkazových řádkůjsou opus_read, opus_cat, opus_geta opus_langid spojeny s moduly Python, které lze importovat a používat v rámci vlastních skriptů.Moduly poskytují stejnou funkcijako nástroje příkazové řádky a také podrobnější řízení dat pomocí podmodulů a funkcí.Kód Python je napsán v Pythonu 3.

OpusRead modul lze inicializovat s parametry, které odpovídají příznakům zadaným opus_reada slouží ke stahování a konverzi corpus souborů z OPUS.OpusRead interně používá XML parsingové modulyz parse subknihovny obsažené v balíčku Opus- Tools Python.Subknihovna obsahuje moduly pro analýzu souborů se zarovnáním XCES a vět.The

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

```
<?xml verze=„1.0“ enkódování=„utf-8“?>
<text>
<p id=„1“>
    <s cld2=„en“ cld2conf=„0.99“ id=„s1.1“ LangID=„en“ langidconf=„1.0“>
      Prohlášení předsedy vlády Ingvara Carlssona o vládní politice při zahájení švédského parlamentu v úterý 4. října 1988.
    </s>
</p>

<p id=„2“>
    <s cld2=„en“ cld2conf=„0.98“ id=„s2.1“ LangID=„en“ langidconf=„1.0“>
      Vaše Veličenstva, Vaše královská Výsosti, pane předsedo, poslanci švédského parlamentu.
    </s>
</p>
```

Obrázek 4:Příklad větového souboru, kde byly přidány jazykové štítky a skóre spolehlivosti do větových značek.

Modul AlignmentParser analyzuje daný XCES link soubor a inicializuje SentenceParser moduly pro analýzu souborů-vět.AlignmentParser výstupy jednotlivých segmentůpárů vět, zatímco SentenceParser výstupy jednotlivévěty z obou stran zarovnání.LinksAlignmentParser lze použít v případě, že je potřeba pouze XCES odkazy a přeskočení větového souboru.Pro analýzu věty existuje také alternativnímodul ExhaustiveSentenceParser,který je robustnější než SentenceParser, ale o něco pomalejší při analýze jen malé části velkého korpusu.Každý z modulů v parse subknibrary lze individuálně importovatdo Python skriptu a použít k extrahování jednotlivých vět, párůvět nebo XCES odkazů.

OpusCat je modul Python používanýskriptem opus_cat a oba mají stejnou funkci čtení monojazyčných větových souborů z OPUS.OpusCat využívá modifikovanou verzi modulu SentenceParser:Při čtení souborů s jedinou větou se proces analýzy věty nemusí řídit příkazem zadaným v zarovnání souboru a SentenceParser v OpusCat jednoduše vypustí každou větu v souboru.OpusCat a SentenceParser mohou být importovány jako moduly Python a mít podrobnou kontrolu nad čtením monojazyčných souborů.**OpusGet** modul pohání opus_get skript s možnostístahování corpora.Importem modulu v Python kódu je možné získat podrobné informace o OPUS corpora v rámci Python datových struktur.Tato informace zahrnuje počet párů zarovnání, počet dokumentů, počet žetonů a velikost v kilobajtech mimo jiné.

OpusLangid modul má stejnou funkci jako skript opus_langid:přidání jazykových štítků a hodnocení důvěry jazyka do XML větových souborů.Navíc OpusLangid obsahuje třídu LanguageIdAdder, kterou lze použít pro získánízjazykových štítků a identifikačních bodů z pycld2 a langid.py pro prostou textovou větu s jedním voláním funkce.

3.3. Perl modul OpusTools Perl

Komplementární balíček nástrojů OPUS je k dispozici jako modul Perl s povolením MIT licence.5

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

zahrnuje nástroje příkazové řádky, které jsou užitečné zejména pro vytváření nových datových souborů, ale také obecně pro rychlý přístup k datům v různých formátech.Některé funkce jsounyní nahrazeny implementacemi v knihovně Python popsané výše, a my se zde zaměříme na nástroje, které podporují dodatečné případy použití.Tyto nástroje spadají především do těchto tří kategorií:

Nástroje pro konverzi:Nástroje, které lze použít k importu a exportu dat v různých formátech souborů a značkování dat.Hlavním účelem je importovat nové datové soubory do OPUS a vytvořit datové soubory, které se uvolňují s různými formáty.

Nástroje pro zarovnání:Věta a zarovnání slov mohou být použity různými způsoby a tyto nástroje poskytují některé pohodlné operace na vrcholu zarovnaných bitextů.

Ostatní zpracovatelské nástroje: Tato kategorie zahrnuje nástroje pro anotaci a indexaci.

V první kategorii, máme import nástroje, jako jsou Moses2opus, tmx2opus a xml2opus.Ex Port skripty zahrnují opus2moses, tmx2moses, opus2text a opus2multi.

xml2opus je jednoduchý skript, který přidává hranice větydo libovolných XML dat.Detekce hranice věty se provádí pomocí nástrojů uvolněných s paralelním corpusem Europarl(Koehn, 2005) a zabalených v Perl modulu Lingua::Sentence.Další nástroje založené na klasifikátorech UD stromových bank budou v budoucnu integrovány.Inline tagy, které přidávají přirážku do vět, nejsou momentálně podporovány.

Moses2opus čte zarovnané jednoduché textové soubory, které se běžně používají ve strojovém překladu se zarovnanými větami na stejném řádku.[5]Nástroj převádí data do jednoduchéhosamostatného XML pro corpus data a formát XCES Align pro vyrovnanou větu, jak se používá v rámci OPUS.V současné době je podporován pouze dvojjazyčný vstup.Prosté textové soubory neobsahují hranice věty, ale stále mohou obsahovat zarovnání věty, které nejsou jedna ku jedné.Proto Moses2opus přidává přirážku věty pomocí-Lingua::Sentence a upraví patové větyodpovídajícím způsobem.Skript také podporuje rozděleníbitextů do menších částí.Prázdné řádky ve zdroji a

[5]Jméno pochází z Mojžíšova balíčku, který popularizoval formát.

cílový jazyk lze použít k označení hranic dokumentu.Kromě toho lze korpus rozdělit na stejně velké části pomocí prahové délky pro maximální počet překladových jednotek zahrnutých v jedné části.

tmx2opus převádí překladové paměti ve formátu TMX do OPUS XML.Nástroj přidává hranice věty stejným způsobem jako moses2opus.To také umožňuje potrubí několik TMX souborů přes nástroj konverze a je schopen sloučit informace v případě překrývající se věty, které jsou zahrnuty v několika překladových jednotek.To se hodí při zpracování dat, které přicházejí jako různé bitexty, ale pokrývají stejný obsah.Proto, pouze jedinečné věty jsou uloženy ve výsledném OPUS XML pro každý jazyk, i když se objevují v různých překladových jednotkách se zarovnáním do různých jazyků. tmx2opus může také zpracovat překlad pamětis více než dvěma jazyky v překladové jednotce,a to bude vyrábět dvojjazyčné věty zarovnání souborů pro všechny jazykové páry, jak jsou nezbytné v OPUS.Dále je možné data rozdělit na menší části podobné tomu, co dělá moses2opus.Vlastnosti z TMX souborů lze také zkopírovat do převedených dat, aby byla zachována další meta data.Použití tmx2opus pro vytvoření dovezeného korpusu ParaCrawlv OPUS je popsáno v oddíle 4. Export skripty provádějí především konverzi dat v opačném směru. opus2mosesa opus2textpřevést OPUS XML data na prostý text a jsou většinou zastaralé a nahrazeny implementací balíčku Python zavedeného dříve. tmx2mosesje vhodný skript pro extrahování zarovnaných vět z libovolných TMX souborů a není omezen na OPUS data.

Opus2multi je nástroj, který může vytvářet multiparalelní datové sady z OPUS corpora.V OPUS jsou všechny soubory dat zarovnány dvojjazyčně, ale v některých případech by se chtělo zarovnat více než dva jazyky.K tomu může opus2multi pomoci spojit dvojjazyčné větya extrahovat odkazy z většího počtu jazyků.Nástroj pracuje na stabilizačních souborech zarovnání věty a používá otočný jazyk pro konstrukci překladovýchjednotek ve všech daných jazycích.K tomu se rozšiřuječástečně překrývající se zarovnání věty, dokud nejsou všechny jazyky pokryty bez dalších konfliktů ve výsledné překladovéjednotce (tj. žádné zbývající překrývání s ostatními jednotkami).Výsledkem tohoto procesu je zarovnání věty-soubory, které jsou (pro pohodlí) tištěné dvojjazyčně pomocí formátu XCES Align, který pak lze dále zpracovávat pomocí OpusTools k extrahování skutečných párů zarovnání.K dispozici je také možnost kontrolovat maximální velikost překladové jednotky (v počtu vět v jednom jazyce), protože velikost může růst bez omezení v procesu expanze.Součástí je také experimentální funkce zahrnutí vnitrojazyčných odkazů pro další tranzitivní mapování.To je vhodné pro datové soubory jako opensubtitles, ve kterých mohou být použity alternativní souborytitulků pro propojení mezi různými jazyky.

Nástroje pro zarovnání v balíku OpusTools pomáhají zpracovávat zarovnánívět ve formátu standoff anotace. Opus-—swap —align jednoduše přemění směrzarovnání.Opus poskytuje pouze zarovnání v jednom směru (stejně jako symetrické), ale někdy je vhodné mít přístup k odkazům i v opačném směru.Opus-merge —align kombinujesoubory se zarovnáním věty asmaže duplikáty, pokud existují. opus-split-align rozdělí soubory zarovnání vět do samostatných souborů s jedním zarovnáním skupiny, tj. zarovnání dokumentu.A konečně, opus— pivotingumožňuje vytvořit přechodné zarovnání vět mezi dvěma jazykypomocí otočného jazyka a odkazy na otočný jazyk.To je vhodné pro corpora, která je dodávána sbitexty, které nepokrývají všechny jazykové páry, ale pouze sladit s konkrétním jazykem, jako je angličtina.Za předpokladu, že existuje podstatné překrývání mezi bitexty, řekněme A ^ P a B ^ P, opus-pivoting výtažky vazby mezi větamiv A a B, vytvoření nového bitext A ^ B. Oddíl 4.ilustruje použití příkladu vytvoření MultiParaCrawl.A konečně, další nástroj pro zarovnání, opus-pt2dice,extrahuje hrubé pravděpodobnostní dvojjazyčné slovníky z frází-translation-tabulek vytvořených ze zarovnání slov a pomocí SMT nástrojů vycházejících z Mojžíšovy nástrojové schránky.Tyto slovníky používají některé heuristiky filtrovat data a nástroj také vytváří další Dice skóre jako symetrizované hodnoty zarovnání z podmíněnéhopřekladu pravděpodobnosti zahrnuté v původních frázových tabulkách, což je užitečné pro dvojjazyčné extrakce lexikonů (Smadja et al., 1996).

Ostatní nástroje: Poslední kategorie nástrojů obsahuje další nástroje pro zpracovánídat, jako je opus —udpipe a opus —index.První implementuje obal kolem UDPipe (Straka a Strakova, 2017) k anotaci OPUS dat a uložení výsledku v OPUS-conforming XML.OpusTools mohou používat předškolené modely pocházející z LIN-DAT.[6]V neposlední řaděje opus-index nástrojem pro indexaci OPUS corpora pomocí Corpus Work Bench (CWB) (Evert and Hardie, 2011).Vytvoří všechny importní soubory a spustí enkodér, je-li k dispozici pro vytvoření multiparalelníhosboru, který má být dotazován pomocí CWB vyhledávače.

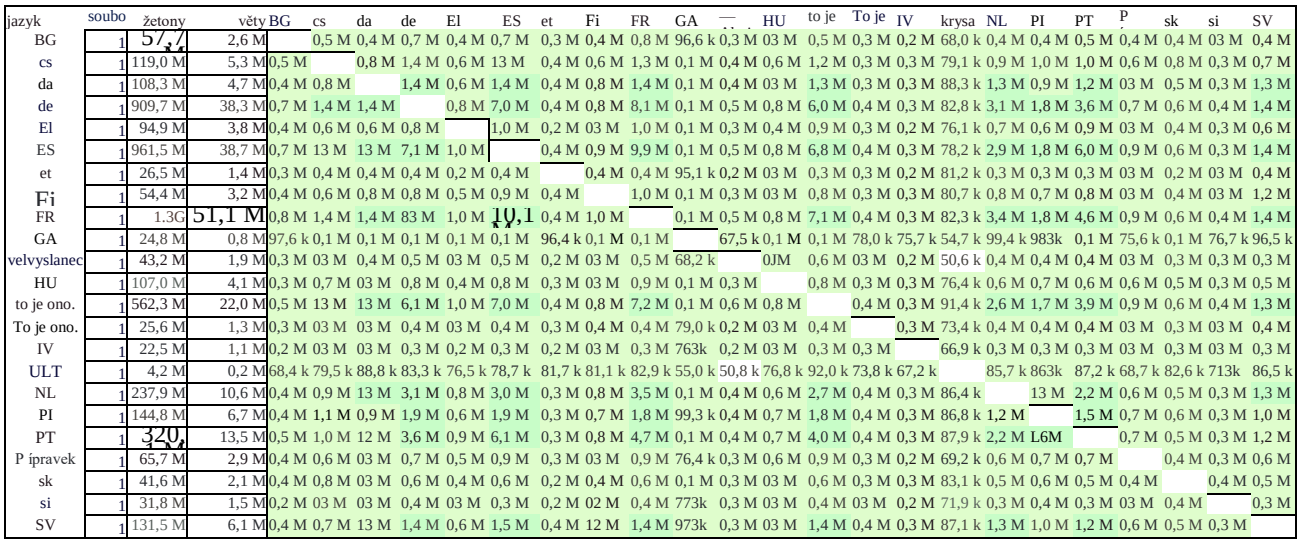
4. ParaCrawl a MultiParaCrawl

V této části bychom rádi ukázali dovoz údajů ParaCrawl, abychom prokázali používání OpusTools.ParaCrawl corpus[7] bylextrahován tak, že se proplížil po webu a použil komplexní potrubí pro úpravu dokumentů a větzaložených na Bitextorovém balíčku (Espla-Gomis, 2009).Aktuální verze v5.0 zahrnuje 24 evropských jazyků a projekt poskytuje automaticky očištěné bitexty projazyky v souladu s angličtinou.Velikost se pohybuje od 100 000 překladových jednotek (Maltese-English) až po více než 50 milionůjednotek (francouzsko-anglicky) a datové soubory jsou distribuovány v jednoduchém textu nebo formátu TMX.Zatímco existuje několik bonusových jazykových párů, které také obsahují dva bitexty, které nezahrnují angličtinu, většina kolekce je dvojjazyčně v souladu s anglickým obsahem.

Cílem integrace ParaCrawl do OPUS je zpřístupnit údaje prostřednictvím nativního formátu OPUS a vyčerpávajícím způsobem pokrýt všechny jazykové páry zahrnuté do sběru.Pro tyto účely byly dříve zavedeny

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

[7] https://paracrawl.eu



jazyk	soubo	žetony	věty	BG	cs	da	de	El	ES	et	Fi	FR	GA	—	HU	to je	To je	IV	krysa	NL	PI	PT	P	sk	si	SV
BG	1	57,7	2,6 M		0,5 M	0,4 M	0,7 M	0,4 M	0,7 M	0,3 M	0,4 M	0,8 M	96,6 k	0,3 M	03 M	0,5 M	0,3 M	0,2 M	68,0 k	0,4 M	0,4 M	0,5 M	0,4 M	0,4 M	03 M	0,4 M
cs	1	119,0 M	5,3 M	0,5 M		0,8 M	1,4 M	0,6 M	13 M	0,4 M	0,6 M	1,3 M	0,1 M	0,4 M	0,6 M	1,2 M	0,3 M	0,3 M	79,1 k	0,9 M	1,0 M	1,0 M	0,6 M	0,8 M	0,3 M	0,7 M
da	1	108,3 M	4,7 M	0,4 M	0,8 M		1,4 M	0,6 M	1,4 M	0,4 M	0,8 M	1,4 M	0,1 M	0,4 M	03 M	1,3 M	0,3 M	0,3 M	88,3 k	1,3 M	0,9 M	1,2 M	03 M	0,5 M	0,3 M	1,3 M
de	1	909,7 M	38,3 M	0,7 M	1,4 M	1,4 M		0,8 M	7,0 M	0,4 M	0,8 M	8,1 M	0,1 M	0,5 M	0,8 M	6,0 M	0,4 M	0,3 M	82,8 k	3,1 M	1,8 M	3,6 M	0,7 M	0,6 M	0,4 M	1,4 M
El	1	94,9 M	3,8 M	0,4 M	0,6 M	0,6 M	0,8 M		1,0 M	0,2 M	03 M	1,0 M	0,1 M	0,3 M	0,4 M	0,9 M	0,3 M	0,2 M	76,1 k	0,7 M	0,6 M	0,9 M	03 M	0,4 M	0,3 M	0,6 M
ES	1	961,5 M	38,7 M	0,7 M	13 M	13 M	7,1 M	1,0 M		0,4 M	0,9 M	9,9 M	0,1 M	0,5 M	0,8 M	6,8 M	0,4 M	0,3 M	78,2 k	2,9 M	1,8 M	6,0 M	0,9 M	0,6 M	0,3 M	1,4 M
et	1	26,5 M	1,4 M	0,3 M	0,4 M	0,4 M	0,4 M	0,2 M	0,4 M		0,4 M	0,4 M	95,1 k	0,2 M	03 M	0,3 M	0,3 M	0,2 M	81,2 k	0,3 M	0,3 M	0,3 M	03 M	0,2 M	03 M	0,4 M
Fi	1	54,4 M	3,2 M	0,4 M	0,6 M	0,8 M	0,8 M	0,5 M	0,9 M	0,4 M		1,0 M	0,1 M	0,3 M	03 M	0,8 M	0,3 M	0,3 M	80,7 k	0,8 M	0,7 M	0,8 M	03 M	0,4 M	03 M	1,2 M
FR	1	1.3G	51,1 M	0,8 M	1,4 M	1,4 M	83 M	1,0 M	10,1	0,4 M	1,0 M		0,1 M	0,5 M	0,8 M	7,1 M	0,4 M	0,3 M	82,3 k	3,4 M	1,8 M	4,6 M	0,9 M	0,6 M	0,4 M	1,4 M
GA	1	24,8 M	0,8 M	97,6 k	0,1 M	0,1 M	0,1 M	0,1 M	0,1 M	96,4 k	0,1 M	0,1 M		67,5 k	0,1 M	0,1 M	78,0 k	75,7 k	54,7 k	99,4 k	983k	0,1 M	75,6 k	0,1 M	76,7 k	96,5 k
velvyslanec	1	43,2 M	1,9 M	0,3 M	03 M	0,4 M	0,5 M	03 M	0,5 M	0,2 M	03 M	0,5 M	68,2 k		0JM	0,6 M	03 M	0,2 M	50,6 k	0,4 M	0,4 M	0,4 M	03 M	0,3 M	0,3 M	0,3 M
HU	1	107,0 M	4,1 M	0,3 M	0,7 M	03 M	0,8 M	0,4 M	0,8 M	0,3 M	03 M	0,9 M	0,1 M	0,3 M		0,8 M	0,3 M	0,3 M	76,4 k	0,6 M	0,7 M	0,6 M	0,6 M	0,5 M	0,3 M	0,5 M
to je ono.	1	562,3 M	22,0 M	0,5 M	13 M	13 M	6,1 M	1,0 M	7,0 M	0,4 M	0,8 M	7,2 M	0,1 M	0,6 M	0,8 M		0,4 M	0,3 M	91,4 k	2,6 M	1,7 M	3,9 M	0,9 M	0,6 M	0,4 M	1,3 M
To je ono.	1	25,6 M	1,3 M	0,3 M	03 M	03 M	0,4 M	03 M	0,4 M	0,3 M	0,4 M	0,4 M	79,0 k	0,2 M	03 M	0,4 M		0,3 M	73,4 k	0,4 M	0,4 M	0,4 M	03 M	0,3 M	03 M	0,4 M
IV	1	22,5 M	1,1 M	0,2 M	03 M	03 M	0,3 M	0,2 M	0,3 M	0,2 M	03 M	0,3 M	763k	0,2 M	03 M	0,3 M	0,3 M		66,9 k	0,3 M	0,3 M	0,3 M	03 M	0,3 M	03 M	0,3 M
ULT	1	4,2 M	0,2 M	68,4 k	79,5 k	88,8 k	83,3 k	76,5 k	78,7 k	81,7 k	81,1 k	82,9 k	55,0 k	50,8 k	76,8 k	92,0 k	73,8 k	67,2 k		85,7 k	863k	87,2 k	68,7 k	82,6 k	713k	86,5 k
NL	1	237,9 M	10,6 M	0,4 M	0,9 M	13 M	3,1 M	0,8 M	3,0 M	0,3 M	0,8 M	3,5 M	0,1 M	0,4 M	0,6 M	2,7 M	0,4 M	0,3 M	86,4 k		13 M	2,2 M	0,6 M	0,5 M	0,3 M	1,3 M
PI	1	144,8 M	6,7 M	0,4 M	1,1 M	0,9 M	1,9 M	0,6 M	1,9 M	0,3 M	0,7 M	1,8 M	99,3 k	0,4 M	0,7 M	1,8 M	0,4 M	0,3 M	86,8 k	1,2 M		1,5 M	0,7 M	0,6 M	0,3 M	1,0 M
PT	1	320,	13,5 M	0,5 M	1,0 M	12 M	3,6 M	0,9 M	6,1 M	0,3 M	0,8 M	4,7 M	0,1 M	0,4 M	0,7 M	4,0 M	0,4 M	0,3 M	87,9 k	2,2 M	L6M		0,7 M	0,5 M	0,3 M	1,2 M
P ípravek	1	65,7 M	2,9 M	0,4 M	0,6 M	03 M	0,7 M	0,5 M	0,9 M	0,3 M	03 M	0,9 M	76,4 k	0,3 M	0,6 M	0,9 M	0,3 M	0,2 M	69,2 k	0,6 M	0,7 M	0,7 M		0,4 M	0,3 M	0,6 M
sk	1	41,6 M	2,1 M	0,4 M	0,8 M	03 M	0,6 M	0,4 M	0,6 M	0,2 M	0,4 M	0,6 M	0,1 M	0,3 M	03 M	0,6 M	0,3 M	0,3 M	83,1 k	0,5 M	0,6 M	0,5 M	0,4 M		0,4 M	0,5 M
si	1	31,8 M	1,5 M	0,2 M	03 M	03 M	0,4 M	03 M	0,3 M	0,2 M	02 M	0,4 M	773k	0,3 M	03 M	0,4 M	03 M	0,2 M	71,9 k	0,3 M	0,4 M	0,3 M	03 M	0,4 M		0,3 M
SV	1	131,5 M	6,1 M	0,4 M	0,7 M	13 M	1,4 M	0,6 M	1,5 M	0,4 M	12 M	1,4 M	973k	0,3 M	03 M	1,4 M	0,4 M	0,3 M	87,1 k	1,3 M	1,0 M	1,2 M	0,6 M	0,5 M	0,3 M	

Obrázek 5:Statistiky z MultiParaCrawl corpus - vícejazyčné rozšíření ParaCrawl přes otočné zarovnání přes angličtinu.Horní pravoúhlý trojúhelník dává velikost, pokud jde o zarovnání věty v prostém textovém formátu, a levý dolní trojúhelník ukazuje velikost extrahovaných TMX souborů z hlediska unikátních překladových jednotek na jazykový pár.

nástroje tmx2opus a opus— pivoting se stávají praktickými. tmx2opus je užitečnýnejen pro extrahování zarovnání z původního zdroje TMX, ale také poskytuje funkci propřidání příznaku hranice věty a snížení redundance mezi různými bitexty.Použití unikátní možnosti tmx2opus snižuje velikost anglickéčásti korpusu (tj. 252 milionů samostatně sladěných anglických větv 23 bitextech) na méně než 60 % původních dat.Jedinečnost zároveň umožňujevytvořit multiparalelní corpus otočením na odkazech na angličtinu v nově vytvořených unikátních větách.K tomu lze použít opus—pivoting, jak bylo vysvětleno dříve.Pomocí tohoto postupu by mohlo být vytvořeno 253 dalších bitextůs velikostí až 10 milionů překladových jednotek.Obrázek 5 shrnuje neanglické bitexty v Mul- tiParaCrawl.

5. Paralelní diagnostika korpusu

Naše diagnostická rutina pro sbírku OPUS používá nástroj opus_read příkazové řádky (popsaný v bodě3.1.) k načtení zarovnaných textových dat pro konkrétní jazykovýpár v daném corpusu.Za tímto účelem opus_read parsujenativní XML-formátovaná data pro generování požadované-podmnožiny dat a pak provede konverzi na jednoduchý textovýformát.Během tohoto procesu diagnostickárutina naslouchá případným chybám, které by mohly vzniknout, a přihlásí je k sestavení diagnostické zprávy pro pozdější analýzu.Tento postup provádíme systematicky pro každý pár jazyků dostupných pod každým z OPUS corpora.[8]Pro naši diagnostiku využíváme plnou granularitu, kterou poskytuje OPUS tím, že shromažďujeme samostatné údaje pro různé sbory,které tvoří bitexty, a také ponecháváme regionální variantyjazyků oddělené, než je konfulujeme.Abychom provedli tento druh vyčerpávající analýzy, provedli jsme paralelně celkem 87 948 úloh CPU pole, přičemž runtime se pohybuje mezi 1 sekundou a 5,2 hodiny a každou práci s využitím 4 až128 GB paměti.Celková diagnostickáanalýza trvala přibližně 1000 hodin na výpočet,průměrně 18,2 hodiny na korpus.Zatímco granularitanaší analýzy bude vnitřně užitečná pro určení anomálií v OPUS pro usnadnění oprav, shromažďujeme také naše údaje pro generování celokorpusových čísel, které podáváme a diskutujeme v této části.

5.1. Analýza chyb

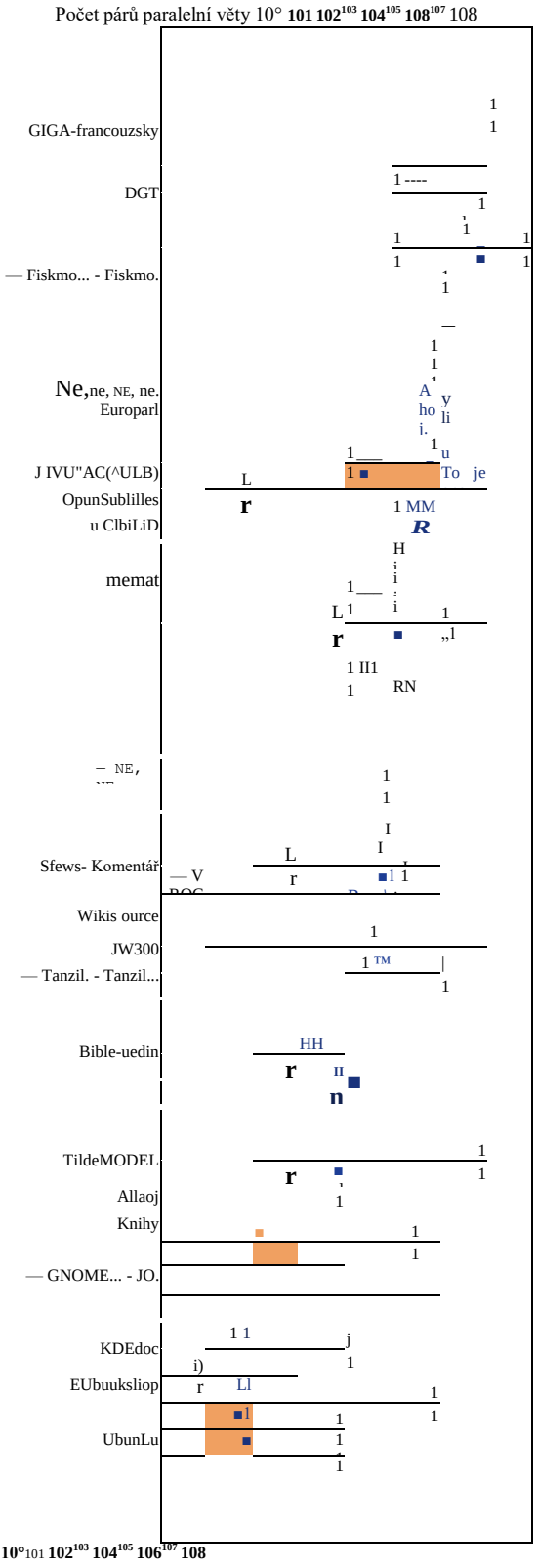
„diagnostiky“ zaznamenané v naší zprávě uvádí příčiny každé chyby vyhledávání, což nám poskytuje prostředky, jak je spolehlivě lokalizovat a opravit.Shromažďování všech diagnóz, výsledky ukazují, že zatímco 37 z sboru je zcela bezchybný, vyhledávání dat zastavil alespoň pro jeden jazykový pár pro zbývající 18 sboru.Hojnostchyb při vyhledávání v těchto tělesech se liší od nepatrného zlomku celého korpusu (viz obrázek 8).Drtivá většinatěchto chyb pochází ze špatně vytvořených XML dat s neplatnými tokeny (96,2 %) nebo nesrovnatelnými tagy (3,5 %).Naše dílčí kontroly zatím naznačují, že tyto chyby lze připsat drobným chybám konverze, jako jsou neupravené zvláštní znakya XML entity, které se v původních datech objevily před importem do OPUS.Další velmi malá část chyb (0,3 %) indikuje chybějící datové soubory v hlavním souborovém systému, kde je OPUS hostován, což pravděpodobně indikuje chyby kopírování, a je třeba je dále prošetřit.

5.2. Statistika corpus-Wide

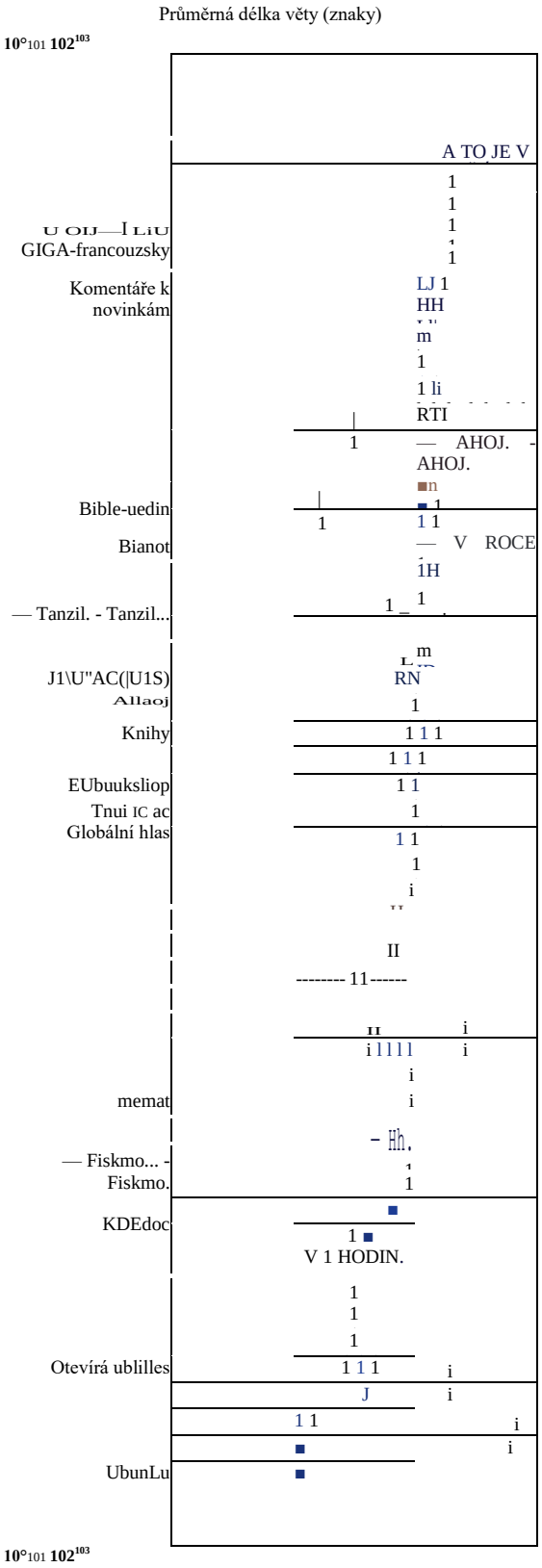
Kromě katalogizace problémů s vyhledáváním dat, náš diagnostický postup také vypočítá některé základní kvantitativnístatistiky, jako jsou vykázané výpočetní náklady na vyhledávání dat, a různá měření načtených dat podle-corpus, jazyk, a jazyk pár.Naše odpovídajícístatistické analýzy většinou neodhalily pozoruhodné trendy ani odlehlé hodnoty, s výjimkou některých opatření, kteránaznačovala relativní odchylky a hladiny hluku v datechnapříč korporou.Na obrázcích 6 a 7 podáváme zprávu

[8]Neprovedli jsme diagnostiku dvou posledních přírůstků-OPUS:Infopankki a MultiParaCrawl.

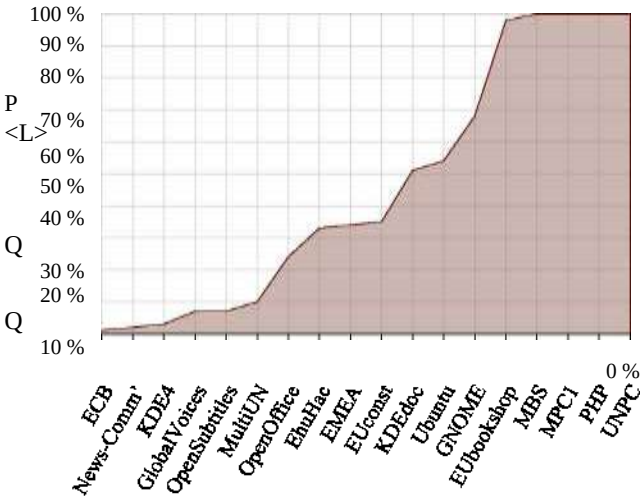
rozdělení dvou opatření po souboru dostupných lan- průměrná délka věty ve znakech.Oba guage páry pro každý corpus:průměrný počet sen- opatření byl vizualizován pomocí box-and-whisker parcely k desetinásobným párům (nebo, přesněji, překladové jednotky) a zdůrazňující distribuční rozdíly, kde koncové parametry



Obrázek 6:Rozdělení počtu retrievable paralelních párů vět nad množinou dostupných jazykových párů pro každý corpus.



Obrázek 7:Rozdělení průměrné délky věty (ve znakech) nad soubory dostupných jazykových párů pro každý corpus.



Obrázek 8:Procento jazykových párů v OPUS corpora, pro které data s nástroji OPUS vrací chyby.Bezchybný sbor byl vynechán z grafu.

Zobrazit nejnižší a nejvyšší hodnoty,[9]a dvě poloviny pole představují druhé a třetí kvartily hodnot, které jsou odděleny mediánem.

Jedním z nejvýraznějších detailů z obrázku 6 je kontrastmezi rozptyly.Více než třetina sboru vykazuje velmi malou až žádnou odchylku v jazykových dvojicích, což znamená plně multiparalelní data, zatímco jiné jako JW300 a otevřené titulky naopak vykazují velmi vysokou rozptyl, kde rozdíl ve velikostech dostupných dat může přesahovat několik řádových řádů.Když se podíváme konkrétně na první kvartily, zdá se, že některé sbory, jako jsou QED a Tatoeba, mají významnou část jazykových párů obsahujícíchvelmi málo překladatelských jednotek, což by mohlo naznačovat vysokou detekci jazykanebo zarovnání vět.Na obrázku 7 se zdá, že první kvartil má podobný relativní rozsah pro některé sbory, což znamená, že věty obsahují pouze několik znaků v průměru pro některé z dostupných jazykových párů.Pravděpodobně není náhoda, že tyto případy většinou odpovídají korpusu, které byly sestaveny z přirozeně hlučných dat.Kromě toho nejmenší a největší mediánové hodnoty na obrázku 7 ukazují na výjimečně krátké a výjimečnědlouhé „typické“ věty v příslušném sboru,které mohou naznačovat silný kontrast v segmentaci textu nebovýrazně odlišné datové domény.Například tři sbory s nejnižšími mediány zahrnují překladypočítačového softwaru, zatímco dokumenty z OSN dosahují nejvyšší mediánové délky.

6. Závěry a budoucí práce

V tomto příspěvku představujeme OpusTools, open-source balíček knihoven a příkazové řádky pro efektivní a pohodlný přístup k paralelnímu sboru v rozsáhlém sběru dat OPUS.Balíček implementuje nástroje pro stahování, konverzi, filtrování a zpracování paralelníchdatových souborů a usnadňuje přístup ke komprimovaným a archivovaným souborům ze sbírky.Poskytuje také Python knihovnu pro programový přístup k datům, takže je snadné začlenit zpracování dat do vývoje dalších nástrojů .Dále představujeme nástroje pro konverzi a zarovnání dat, které lze použít při přípravě nových datových souborů z různých zdrojů.Jejich použití demonstrujeme na příkladu nedávno přidaného corpus MultiParaCrawl, který rozšiřuje původní datovou sadu o otočné zarovnánímezi všemi jazykovými páry, které přispívají krostoucímu pokrytí databáze OPUS.

I když vedení sbírky tak velké jako OPUS dokonale robustní je poměrně náročné, řešení problémů bude jednodušší a rychlejší s diagnostikou plně zmapované.Celkově vzato, i když chyby při získávání údajů zahrnují jasné akční body, statistické analýzy spíše naznačují pozoruhodnou kvalitativnía kvantitativní různorodost mezi OPUS corpora, s trendy zdánlivě v rámci očekávání, a okrajové případy, které lze přičíst hluku v původních údajích.Naším záměrem je vyřešit všechny otázky týkající se vyhledávání dat tak, aby používání nástrojů OPUS bylo hladkým zážitkem pro všechny uživatele a také zefektivnění naší rutiny jako diagnostického nástroje, který by se stal standardní součástí procesu rozšiřování OPUS s novou korporou.

Potvrzování

ERC Tato práce je součástí projektu FoTran, který je financován Evropskou radou pro výzkum (ERC) Program Evropské unie pro výzkuma inovace Horizont 2020 (dohoda o grantu Na771113)' jakož i projekt MeMAD financovaný z programu Evropské unie pro výzkum a inovace Horizont 2020 (dohoda o grantu Na 780069).

7. Bibliografické odkazy

Evert, S. a Hardie, A. (2011).Corpusworkbench jednadvacátého století:Aktualizace architektury dotazů pro nové tisíciletí.V *Proceedings of the Corpus Linguistics2011 konference, University of* Birmingham, Velká Británie.Smadja, F., McKeown, K. R., a Hatzivassiloglou, V. (1996).Překlad kollokací pro dvojjazyčné lexikony:Je to statistický přístup.*Výpočetní lingvistika*, 22(1):1-38.

8. Odkazy na zdroje jazyků

Espla-Gomis, Miquel.2009.*Bitextor: Bitextor:svobodný/otevřený zdroj software pro sklizeň paměti překladu z vícejazyčných webových stránek.*

— Koehn, Philipp.2005.*Europarl:Paralelní Corpus pro statistický strojový překlad*.— AAMT? - ANO.

Lui, Marco a Baldwin, Timothy.(2012). *langid.py:Nástroj pro identifikaci cizího jazyka.*Asociacepro výpočetní lingvistiku.

Straka, Milan a Strakova, Jana.2017.*Tokenizace, POS Tagging, Lemmatising a Parsing UD 2.0 s UDPipe.*Asociace pro výpočetní lingvistiku.

— Tiedemann, Jorg.2012.*Paralelní data, nástroje a rozhranív OPUS.*Evropská asociace jazykových zdrojů(ELRA).

[9]Nezakryté koncové body naznačují extrema za hranicemi osy x.