TCA2 – nástroj pro zarovnávání paralelních textů

Translation Corpus Aligner 2 (TCA2) je nástroj k interaktivnímu zarovnávání paralelních textů, vyvinutý Øysteinem Reigemem[1] na institutu *Unifob Aksis* (dnes *Uni Digital*) při Univerzitě v Bergenu, jako reimplementace staršího, neinteraktivního programu TCA, vytvořeného počátkem 90. let Knutem Hoflandem a Stigem Johanssonem pro projekt norsko-anglického paralelního korpusu. Verze TCA2 byla implementována v jazyce Java[2] a je tedy snadno použitelná na všech platformách, má přívětivé grafické uživatelské rozhraní a podporuje standardy TEI/XML a Unicode pro snadnou práci s texty v různých jazycích. Nástroj je schopen zarovnávat zvolené elementy dvou textů (již předem segmentovaných), a to jak ručně, poloautomaticky tak i zcela automaticky.

Uživatelské rozhraní a práce s TCA2

Grafické rozhraní využívá v horizontálním plánu dvou panelů s paralelními texty, rozdělených vertikálně na třetiny. V horní třetině stojí na obou stranách seznam barevně odlišených párů již zarovnaných textových elementů, uprostřed je prostor pro právě zarovnávaný blok elementů, kam lze volně přidávat či ubírat elementy podle potřeby ze seznamu následujících, ještě nezarovnaných elementů v dolní třetině. Po odsouhlasení skupiny paralelních elementů ve střední části se tento pár odsune do vrchní třetiny a je možné sem přidávat elementy pro další paralelní skupinu.

Do každého sloupce je možné načíst jeden dokument ze souboru ve formátu XML. Kdykoliv v průběhu práce je možné aktuální výsledky uložit, a to ve třech různých podobách najednou: 1) jako pár původních XML souborů doplněných o atributy „corresp“, které u všech již zarovnaných elementů udávají identifikátory odpovídajících paralelních elementů v druhém textu[3] – v této podobě lze texty později načíst zpět do TCA2 a pomocí zvláštního tlačítka nechat programem přeskočit všechny již spárované elementy a vrátit se tak zpět na místo, kde párování končí, a pokračovat v nedokončené práci; 2) jako pár textových souborů ve formátu „newline“, kde každý řádek odpovídá právě jedné pozici (jednomu zarovnanému páru),[4] a který lze použít ve starších aplikacích typu ParaConc nebo Corpus WorkBench (po úpravě) – tyto soubory ovšem již neobsahují původní XML strukturu;[5] 3) jako samostatný XML soubor s pouhým seznamem spárovaných identifikátorů paralelních elementů obou textů (neobsahuje žádný text). TCA2 nepodporuje automatické časované ukládání v průběhu práce.

Program nabízí tři různé módy zarovnávání: 1) manuální mód, kdy je uživatel nucen seskládat ručně a odsouhlasit každý jednotlivý pár paralelních elementů; 2) plně automatický mód, kdy program použije vlastních algoritmů k vytvoření párů paralelních elementů; 3) poloautomatický interaktivní mód, kdy program automaticky zarovná jen ty páry elementů, které si podle jeho odhadu odpovídají v poměru 1:1, a zastaví se k ruční kontrole navrhovaného zarovnání, pokud vyhodnotí, že poměr je jiný. Po odsouhlasení uživatelem (či ruční korekci) pak lze automaticky pokračovat k dalšímu takovému místu. Pro poloautomatický mód je také možné nastavit maximální limit počtu elementů, po jejichž automatickém zarovnání se program zastaví ke kontrole bez ohledu na to, zda i nadále považuje elementy za spárovatelné v poměru 1:1.

Hlavní výhodou druhé verze TCA je (vedle interaktivity) především značně rozšířená a zjednodušená možnost nastavení parametrů použitých při automatickém vyhodnocování kandidátů

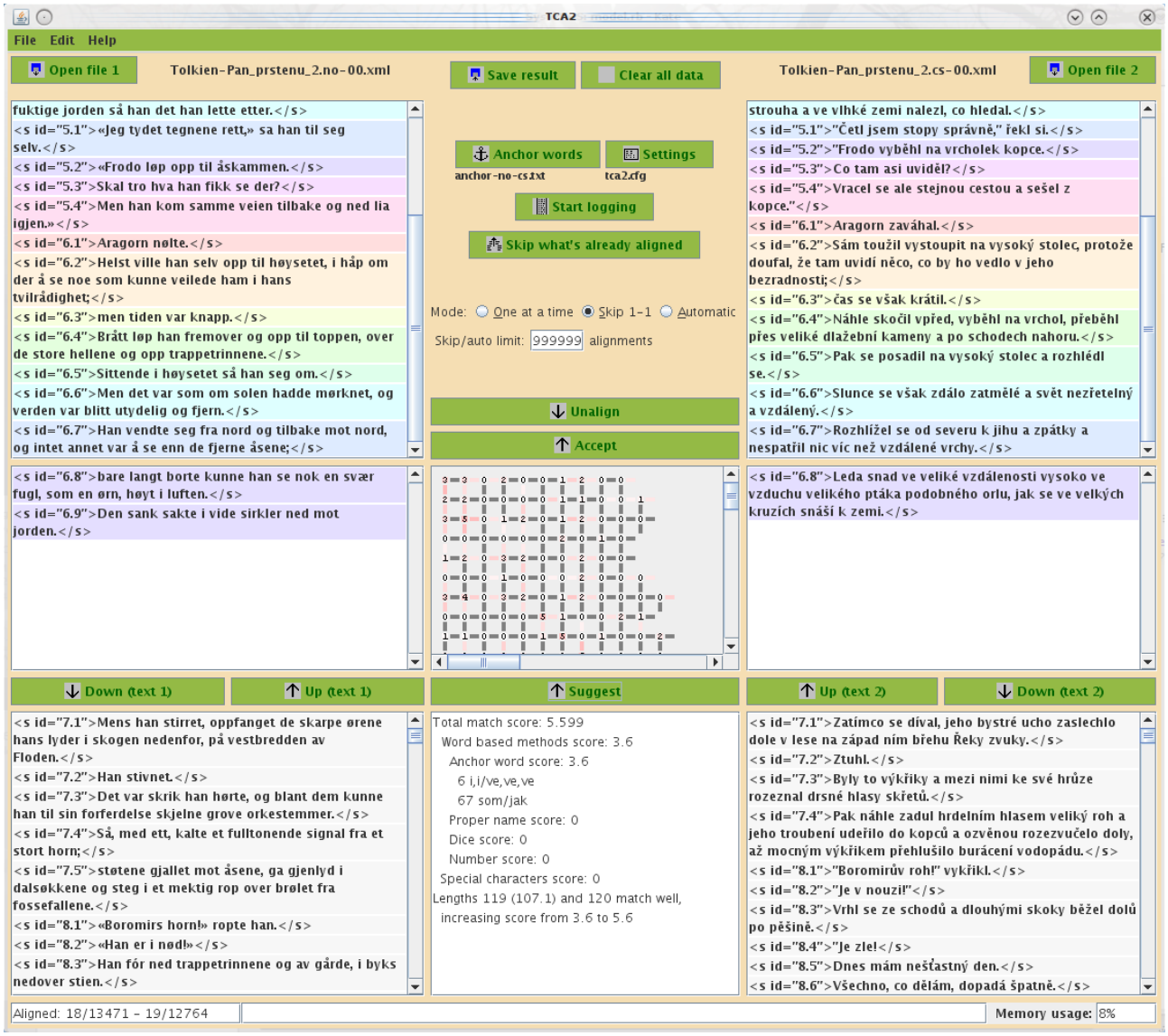
1 K vývoji přispěli v různé míře např. i uživatelé programu z Univerzity v Tromsø aj.

2 Aktuálně existují verze pro jazyky Java verze 1.5 a verze 1.6.

3 U elementů bez paralely v druhém dokumentu je do atributu zapsán alespoň identifikátor nejbližšího nadřazeného elementu v protějším textu – tedy při zarovnávání vět obvykle identifikátor aktuálního odstavce, kde věta v protějším textu chybí.

4 Shodné číslo řádku tedy identifikuje pár paralelních elementů mezi oběma soubory. Elementy seskupené do jedné skupiny stojí tak vždy na jednom řádku, zatímco na pozici, kde daný text neposkytuje žádný element paralelní k elementu na odpovídající řádce v druhém dokumentu, se nachází jen prázdný řádek.

5 V nastaveních však lze zvolit možnost přidání údajů o nadřazeném elementu.



na paralelní páry. V nastaveních je možné regulovat význam jednotlivých poměřovaných veličin (viz níže), zvolit, které XML elementy jsou vlastně určeny k zarovnávání (může jich být více) a jaké jsou jejich nadřazené elementy, zvolit sadu znaků, které nemají být považovány za součásti slov (interpunkční a jiná znaménka) a naopak zvolit znaky, které zvyšují skóre, pokud se vyskytují v elementech obou dokumentů. Nastavení je možné ukládat do externího souboru a opět načítat. Ukládat a načítat lze také seznam frekventovaných kotevních slovíček (pomocný slovník), který program též využívá při automatickém zarovnání.

Princip automatického zarovnávání

TCA není nástroj používající stochastických, jazykově zcela nezávislých metod porovnávání textů, ale využívá několika jednoduchých statistických veličin k poměřování paralelnosti textových řetězců mezi oběma dokumenty na základě podobnosti mezi jazyky. Poměřuje se jednak poměr délek jednotlivých řetězců (v nastaveních lze zvolit, jaký poměr je pro daný pár jazyků nejobvyklejší), poměr nalezených paralelních kotevních slov[6] ze zadaného pomocného slovníčku,

6 Může se jednat i o víceslovné výrazy, jimž pak lze přisoudit větší váhu.

poměr výskytu podobných slov či sousloví,[7] poměr shody nalezených vlastních jmen,[8] poměr výskytu nalezených shodných číslic a dalších relevantních znaků zadaných v nastaveních[9] (viz výše). Tyto veličiny jsou vyhodnocovány a – na základě váhy, kterou jim lze individuálně přisoudit v nastaveních – dopředu poměřovány vždy pro omezenou skupinu následujících elementů z obou textů,[10] mezi nimiž se vyhledávají nejvhodnější kandidáti pro spárování. Program počítá jen se třemi typy situací,[11] které jsou ale zdaleka nejběžnější: přímá shoda elementů 1:1, vypuštěný element v překladu (1:0) a překlad jednoho elementu pomocí dvou (1:2). Žádný „směr překladu“ není však brán v úvahu a všechny tři možnosti jsou tudíž přípustné v obou směrech současně, takže možných kombinací je celkem pět: 1:0, 0:1, 1:1, 1:2 a 2:1.

Program ve spodní části prostředního sloupečku uživatelského rozhraní průběžně ukazuje výsledné hodnoty poměřovaných veličin pro aktuálně nastavený pár elementů (ať už navržený automaticky nebo sestavený ručně uživatelem), a ve střední části pak i matici předběžných propočtů vzájemných skóre mezi jednotlivými elementy pro celou nejbližší skupinu. Tak je možné snáze určit, proč v některých situacích váhá či proč se v dané situaci rozhoduje tak, jak se rozhoduje, a jak hodlá dále postupovat u nejbližších následujících elementů.

Praktické zkušenosti a úspěšnost automatického zarovnání

Z výše uvedeného vyplývá, že zvolená metoda je tím účinnější, čím příbuznější si jsou oba jazyky a čím přesnější je překlad. Autoři původní metody uvádí chybovost v průměru kolem 2%, s výjimkou textů přeložených velmi volně. Metoda byla celkem úspěšně zkoušena i na zarovnávání textů mezi germánskými a slovanskými či románskými jazyky.

Při zarovnávání textů mezi norštinou a češtinou se program osvědčil přinejmenším v interaktivním poloautomatickém módu, tedy jako poměrně spolehlivý nástroj k identifikací míst, kde překlad není lineární a věty si neodpovídají v poměru 1:1. Úseky textu, které si odpovídají v tomto poměru jsou programem identifikovány celkem spolehlivě. Jen výjimečně je mezi takto spárovanými větami dvojice, kde věta v jednom jazyce zahrnuje i část kontextu věty v paralelním textu, a správné zarovnání by tedy mělo být spíše v poměru 2:2. Poměrně často však program volí kombinaci 1:0 a následně ji kompenzuje opačným poměrem 0:1 i v situacích, kdy si věty skutečně odpovídají přímo 1:1. Většinou však takové chování indikuje právě nějaký jiný poměr, který je třeba upravit ručně (např. zmíněný 2:2 nebo jiné složitější kombinace, se kterými program už nepočítá). Pro interaktivní zarovnávání s ruční kontrolou je toto chování rozhodně vhodnější, než aby došlo omylem k automatickému spárování vět, které k sobě nepatří. Při práci je tak možné se spolehnout na automatické zarovnání 1:1 a věnovat se pouze podezřelým místům s jiným poměrem a ta ručně korigovat.

V praxi je úspěšnost a hladkost zarovnání pomocí TCA2 velmi závislá na typu textu nebo konkrétní pasáži v textu. Epické pasáže, sestávající většinou z jednoduchých krátkých vět nebo nekomplikovaných souvětí s přímočarým překladem, program zpracuje velmi snadno, rychle a poměrně spolehlivě. Bloky zarovnatelné v poměru 1:1 mají obvykle v takových pasážích délku 15-50 vět a i v ostatních případech program obvykle správně určí kandidáty pro pár v jiném poměru.

---

7 Podobnost slov se hodnotí podle tzv. Diceova koeficientu podobnosti. V nastaveních lze zvolit, jaké hodnoty lze považovat za relevantní. Tato metoda dokáže identifikovat např. internacionalismy a jiná podobná slova, lišící se mezi příbuznými jazyky pouze pravopisem či koncovkou. Její možný negativní vliv u náhodně si podobných slov je podle autorů TCA zanedbatelný a v TCA2 lze ostatně nastavit i minimální délku slov, která budou takto vůbec poměřována (výchozích nastavení vyžadují alespoň 5 znaků). Program se pokouší také jednoduchým způsobem hodnotit podobnost sousloví, nebo slova a sousloví v protějším textu, a bere tak v úvahu i možnost víceslovných ekvivalentů nebo překladu pomocí kompozita.

8 Vlastní jména se identifikují na základě použití velkých písmen. Jejich podobnost je opět vyhodnocována a program je schopen ignorovat odlišné koncovky (např. při deklinaci).

9 Ve výchozím nastavení například vykřičníky, otazníky nebo znak procenta.

10 V původním TCA to bylo vždy 15 elementů. V TCA2 lze počet elementů, které program dopředu poměřuje, nastavit na libovolnou hodnotu (menší než 20); výchozí nastavení je 10.

11 Zjevně z důvodu co největší efektivity.

Naproti tomu u lyrických nebo filozofických pasáží, obsahujících dlouhá a komplikovaná souvětí,[12] která navíc překladatelé segmentují v cílovém jazyce zcela jinak, je uživatel odkázán především na ruční práci, neboť se tu běžně objevují páry v neobvyklých poměrech jako 2:3, 1:3, 2:4 nebo i 1:4, se kterými TCA2 nepočítá. Naopak vět odpovídajících si v poměru 1:1 je v takových textech menšina. Vyhodnocování velmi dlouhých vět je navíc mnohem náročnější na výpočet a program pak postupuje mnohonásobně pomaleji.

Role slovníčku kotevních slov je pro úspěšnost automatického zarovnávání klíčová. K dobrým výsledkům však stačí už několik málo desítek nejfrekventovanějších slov, která uživatel může postupně přidávat během práce – typicky se jedná zejména o zájmena, spojky, příslovce, běžná slovesa (*řekl, myslí, musí* apod.), a případně částice či předložky. Užitečná jsou především slova s přímými a jednoznačnými ekvivalenty, ale je možné zadat i ekvivalentů několik (na obou stranách současně) a použít hvězdičkové konvence k zadání pouze částí slov (slovních kmenů) a tím k eliminaci flektivních koncovek. Přidat lze též průběžně vlastní jména, která se v překladu a originále nepodobají, ale prostupují celým textem – např. jména hlavních postav románů, zeměpisných názvů, apod. TCA2 umožňuje navíc uvádět ve slovníčku i víceslovné ekvivalenty a celé fráze a navíc jim lze přisoudit vyšší váhu v celkovém skóre.

Při zarovnávání norských a českých textů bylo třeba oproti výchozím nastavením upravit především správný poměr délky ekvivalentních segmentů.[13] Výsledky též zlepšilo výrazné zvýšení váhy shody kotevních slov, která hrají značně významnější roli než samotný poměr délky celých segmentů:[14] zatímco český text je díky syntetičnosti jazyka většinou o něco kratší, u velmi krátkých vět – typicky idiomatických konverzačních frází – si naopak norština vystačí s kratšími slovy a tudíž kratšími obraty, poměr délky se tak obrací a program zbytečně odmítá spárovat věty 1:1. Naproti tomu přidání takových idiomatických obratů nebo jejich částí do slovníčku je mnohem snazší. Na zbytečných problémech se zarovnáváním triviálních jednoslovných replik jako „ano“ či „ne“ se také zpočátku značně projevilo i zcela banální opomenutí specificky českých či (kupodivu) norských typografických uvozovek ve výchozím seznamu znaků, které nemají být považovány za součásti slov. Program tak nebyl schopen identifikovat tato slova, ačkoliv byla explicitně uvedena ve slovníčku jako ekvivalenty. U takových jednoslovných vět pak hraje rozdíl jednoho navíc započítaného znaku podstatnou roli i ve výsledném poměru délky vět.

Omezení TCA2

TCA2 je nástroj určený pouze k zarovnávání textů, neumožňuje tedy žádné vyhledávání či vytváření konkordancí z výsledných textů. Aktuální verze neumožňuje provádět ani žádné jiné zásahy v textech, jako např. korekci objevených chyb nebo chybu v segmentaci elementů. Zarovnávání probíhá lineárně od začátku do konce na principu zipu: všechen text před místem aktuální práce je zarovnán a všechen následující text je nezarovnán. Není tedy možné se vrátit doprostřed zarovnaného textu a provádět zde lokální změny bez rozpojení následujícího zarovnání. Uživatel má samozřejmě možnost od místa aktuální práce zpětně po jednom rozpojit každý již zarovnaný pár a vracet se tak k místu chybného zarovnání, ale rozpojené elementy se tak dostanou zpět mezi nezarovnané a je třeba je následně zarovnat znovu.

Další nevýhodou současné verze je, že neobsahuje žádné statistické nástroje, které by umožnily vyhodnotit už zarovnaný text a doporučit tak uživateli vhodné hodnoty pro nastavení, natožpak aby se program sám učil z už hotové práce a zlepšoval tak svou přesnost. Uživatel je tudíž odkázán na to, aby vhodná nastavení sám odhadoval a (podle průběhu práce a úspěšnosti výsledků programu u jednotlivých elementů) případně korigoval. Vzhledem k tomu, jak významný vliv mají špatná

12 Typicky například texty Milana Kundery.

13 Pro české texty se osvědčila hodnota mezi 0.8-0.9 délky norského textu. Samozřejmě je nutné, aby uživatel otevíral texty v jednom jazyce vždy na stejné straně, neboť program není schopen jazyky aktivně rozlišit.

14 Dle dokumentace programu se ovšem zdá, že přiměřený poměr délek segmentů hraje jinou roli než ostatní veličiny, neboť ve skutečnosti sám všechna ostatní skóre přímo modifikuje (zvyšuje nebo snižuje). Není ostatně možné mu uživatelsky žádnou váhu nastavit, neboť je sám vahou všech ostatních.

nastavení na úspěšnost automatického vyhodnocování kandidátů paralelních párů, není TCA2 ideálním nástrojem pro zcela nezkušeného uživatele bez zájmu o funkčnost programu či pro nahodilé zarovnávání jednotlivých textů pokaždé v různých jazycích – nebo alespoň ne v případě, že se uživatel chce vyhnout zbytečně častým manuálním zásahům.

Na dalším vývoji TCA2 se momentálně nepracuje. Od výrazných optimalizací v roce 2007 však velmi dobře a (na moderních počítačích)[15] rychle zvládá i práci s rozsáhlými paralelními texty o více než deseti tisících zarovnávaných elementů (počet vět v mnohasetstránkovém románu). Díky tomu je bez problémů použitelný i pro práci s rozsáhlými beletristickými texty používanými v projektu InterCorp, aniž by bylo nutné je dělit na menší části.

Literatura

Hofland K., 1996, A Program for Aligning English and Norwegian Sentences. In *Research in Humanities Computing 5. Selected Papers from the ACH/ALLC Conference, University of California, Santa Barbara, August 1995,* eds. G. Perissinotto, Clarendon Press, Oxford, 165-178.

Hofland K., S. Johansson, 1998, The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies,* eds. S. Johansson, S. Oksefjell, Rodopi, Amsterdam, 87-100. Dostupné z WWW: http://khnt.hd.uib.no/files/align.htm

Translation Corpus Aligner (TCA) 2. Om programmet. Unifob Aksis, Bergen. Dostupné z WWW: http://gandalf.aksis.uib.no/tca2/

TCA2: Et brukervennlig program for sammenstilling av setninger fra en originaltekst og dens oversettelse(r). Unifob Aksis, Bergen. Dostupné z WWW: http://gandalf.aksis.uib.no/knut/tca2.html

15 Rychlost odezvy při práci na počítačích starších než 3-4 roky může být u dlouhých textů znatelně nižší.