Мср Јована Миловановић*

811.163.41'322.2
811.133.1'322.2
https://doi.org/10.18485/melissa.2018.17.1.ch14
Прегледни рад
Примљен: 04.07.2018
Прихваћен: 11.09.2018.

АУТОМАТСКО УПАРИВАЊЕ НА ПРИМЕРУ ДВОЈЕЗИЧНОГ ФРАНЦУСКО-СРПСКОГ КОРПУСА

Сажетак

Предмет овог рада су анализа и евалуација аутоматског упаривања француско-српских паралелних текстова. У раду су најпре представљени различити приступи и процеси у аутоматском упаривању (статистички метод базиран на дужини реченица, употреба речника, итд). Посебна пажња посвећена је двама програмима за аутоматско упаривање текстова, *Alinéa* и *Hunalign*. Наведени програми испитани су на корпусу ЛМД, састављеном за потребе овог истраживања, који чине текстови на француском и српском језику објављени у листу *Le Monde Diplomatique*. Након упаривања, урађена је евалуација упарених реченица, а парови реченица распоређени су у три категорије: тачно упарени, делимично нетачно упарени и нетачно упарени. Такође, у раду су приказане грешке у упаривању, те су предложени кораци у припремној фази корпуса који би могли такве грешке предупредити или умањити. Циљ рада је да се испита да ли су програми за аутоматско упаривање двојезичних текстова учинковити код француско-српских текстова, да се утврди који је програм најподобнији за наведени пар језика, као и да се укаже на значај и примене ових програма у развоју двојезичних корпуса.

Кључне речи: корпусна лингвистика, аутоматско упаривање, двојезични корпуси, француски језик, српски језик, alinéa, hunalign.

1. Увод

Аутоматско упаривање заузима значајно место у рачунарској лингвистици и изради великих вишејезичних корпуса. Овај вид упаривања се може спровести на различитим нивоима, попут нивоа речи, реченице, пасуса, итд. Неки програми упаривање врше помоћу статистичких модела који не изискују посебне податке за анализи-

* jovana-m@hotmail.com

ране језике, док други користе посебне лексиконе за сваки језик из корпуса. У овом раду анализирамо оба приступа у упаривању на нивоу реченица и разматрамо који приступ даје боље резултате на француско-српском корпусу. Циљ овог рада је да прикаже учинковитост анализираних програма и објасни честе грешке у аутоматском упаривању француско-српског текста.

2. Корпус

За потребе истраживања, сачинили смо корпус ЛМД који се састоји из текстова објављених у листу *Le Monde Diplomatique.* Корпус садржи 280 025 појавница, односно 49 текстова изворно објављених на француском језику и њихових превода на српском језику. Критеријум за одабир текстова техничке је природе: издвојени текстови на српском језику бесплатно су доступни на вебсајту *Недељника*[1]. Текстови на француском преузети су са званичне странице *Le Monde Diplomatique*[2]. Корпус је у текстуалном формату .txt, а енкодирање је UTF-8.

При стварању корпуса ЛМД, текстови нису измењени, те садрже имена аутора, фусноте и додатна објашњења, који у неким случајевима постоје само у изворном тексту или преводу, што може представљати потешкоћу за упаривање. За разлику од корпуса коришћеном у неким истраживањима, попут Brown, Lai et Mercer (1991 : 172), наведене информације у корпусу ЛМД нису посебно обележене или издвојене. У оквиру корпуса постоје и текстови са непреведеним или скраћеним пасусима, што према Véronis et Langlais (2000 : 373), такође отежава упаривање. Важно је напоменути и да датуми објављивања текстова и називи фотогорафија и илустрација објављени у електронској верзији текстова нису укључени у корпус јер не представљају интегрални део текста.

Када је реч о потешкоћама у упаривању, неки истраживачи сматрају да врста текста утиче на квалитет упаривања, односно да су неки жанрови тежи за упаривање. Према томе, Simard и Plamondon

1 http://www.nedeljnik.rs/lmd
2 https://www.monde-diplomatique.fr

(1996 : 79) тврде да се текстови из сфере права лако упарују јер су преводи правничких текстова и документације дословни, док су књижевни текстови знатно захтевнији за упаривање услед слободнијих превода. Новински текстови налазе се на средини овог спектра пошто преводи нису дословни као што је то случај у сфери права, али садрже бројне термине и цифре који олакшавају аутоматско упаривање.

3. Одабир програма

Бројни програми за аутоматско упаривање испитани су на корпусима који садрже француски језик. У оквиру овог истраживања, одабрали смо два програма за упаривање корпуса : *Alinéa*[3] и *LF Aligner*[4] (базиран на алгоритму *Hunalign*). Пресудан утицај у одабиру програма имали су резултати истраживања Kraif (2001 : 30) и Varga et al. (2005 : 595) који указују на то да наведени програми постижу висок степен успеха при упаривању различитих језичких парова и могу с прецизношћу упарити више од 90 % садржаја у анализираним корпусима. Оба програма ослањају се на алгоритам истаживача Gale et Church (1998) заснован на статистичком приступу применљивом на различите језике. Резултат оваквог аутоматског упаривања је паралелни текст аутоматски подељен на мање јединице, које не одговарају увек реченицама, а у којем сваки део изворног текста има одговарајући део на страном језику. Наведене појединачне делове ћемо у даљем раду означавати појмом *сегмент*.

Alinéa је заснована на алгоритму који користи когнате, односно формално сличне речи, потпуне кореспонденте, попут бројева, и дужину речи. На основу њих, програм пореди сегменте и упарује их. Реч је о приступу који не изискује посебне ресурсе за сваки језик понаособ, већ функционише на основу статистичких модела који користе ограничене количине информација (такозвани *knowledge-poor* приступ (Kraif 2006 : 19)). *Alinéa* нуди могућност аутоматске евалуације упареног корпуса, на основу референтног корпуса, које програм пореди како би утврдио степен подударности, односно одступања.

3 http://turing3.u-grenoble3.fr/olivier.kraif/index.php
4 https://sourceforge.net/projects/aligner/

LF Aligner, као и *Alinéa*, врши упаривање у неколико етапа. Упаривање почиње припремном фазом у којој се корпус сегментира, а потом се речима аутоматски уклањају наставци како би се уклонио њихов утицај на процес упаривања. Потом се прелази на упаривање у коме се, за разлику од приступа у програму *Alinéa*, користи двојезични лексикон уз статистички приступ (Varga et al. 2005 : 592). *LF Aligner* не нуди могућност аутоматске евалуације, али је помоћу овог програма могуће уносити измене директно у упареном тексту. Оба су програма доступна бесплатно путем интернета.

4. Евалуација

Упаривање је извршено најпре помоћу програма *Alinéa*, а потом и *LF Aligner*. У оба случаја коришћен је корпус ЛМД, али је резултат упаривања унеколико различит. Наиме, *LF Aligner* се показао двоструко бржим у упаривању корпуса (2,5 минута наспрам 5 минута за *Alinéa*). Такође, два анализирана програма нису понудила исти број парова сегмената. *LF Aligner* је корпус ЛМД сегментирао на 4 686 парова, док је са програмом *Alinéa* исти документ сегментиран на 5 104 пара.

	Alinéa	***LF Aligner***
време	5 минута	2,5 минута
број парова	5104	4686

Табела 1. Упаривање корпуса ЛМД у програмима Alinéa и LF Aligner.

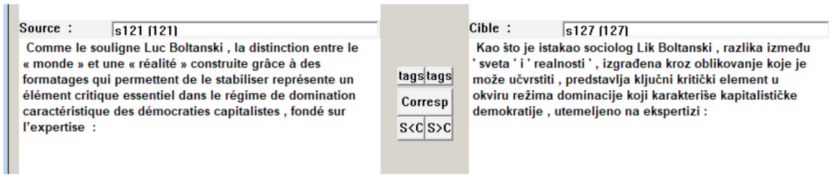
Након упаривања, уношењем измена у аутоматски упарен корпус уз помоћ програма *Alinéa* направили смо референтни корпус. Затим смо документе добијене помоћу два програма упоредили са референтним корпусом. У евалуацији су заступљене две различите методе. У првој методи су резултати евалуације исказани ф-мером, јединицом која је коришћена у бројним истраживањима о аутоматском упаривању, попут Langlais, Simard et Véronis (1998), Chiao, Kraif et al. (2006), Kraif (2001) и Varga et al. (2005). Ф-мера се заснива на прецизности и одзиву. Прецизност представља однос тачних парова

сегмената и укупног броја сегмената, док одзив представља однос тачних парова сегмената и укупног броја сегмената у референтном корпусу (Véronis et Langlais 2000 : 376).

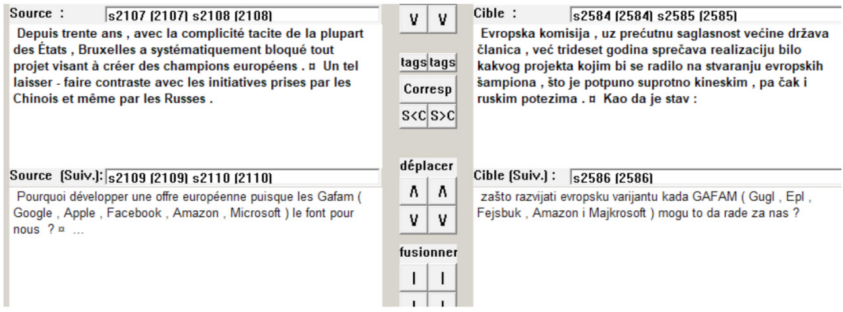
прецизност = (тачни парови сегмената) : (укупан број парова сегмената)
одзив = (тачни парови сегмената) : (укупан број сегмената у референтом корпусу)
ф–мера = 2 ((прецизност*опозив) : (прецизност+опозив))

Илустрација 1. Формуле за израчунавање прецизности, одзива и ф-мере.

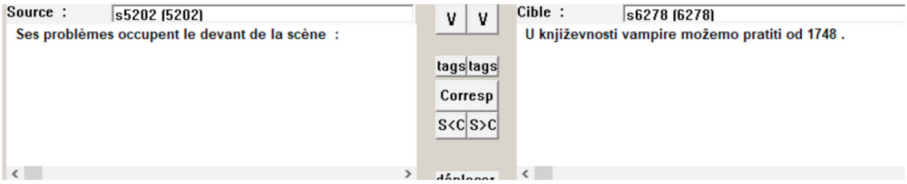
Како је реч о јединицама које парове сегмената анализирају као тачне или нетачне, сматрамо да је сврсисходно увести и други, блажи метод евалуације. Наиме, у другој фази евалуације, анализирали смо добијене парове сегмената и поделили их у три категорије : потпуно тачне, делимично нетачне и потпуно нетачне.



Илустрација 2. Потпуно тачно упарени сегмент.



Илустрација 3. Делимично нетачно упарени сегмент.



Илустрација 4. Потпуно нетачно упарени сегмент.

Потпуно тачни парови сегмената су они парови код којих текст на српском језику у потпуности одговара изворном француском тексту. Делимично нетачни парови су они код којих део текста на српском одговара француском тексту, али други део сегмента није адекватно упарен са изворним текстом. У овом случају, део текста недостаје, или је пак вишак. Најзад, потпуно нетачни парови сегмената су они код којих се ниједан део српског сегмента не подудара са француским сегментом. Верујемо да подела на три категорије омогућава детаљнији увид у учинак анализираних програма пошто се узимају у обзир и делимично успешно упарени сегменти. С друге стране, ф-мера је значајна јединица коришћена у бројним истраживањима, што олакшава поређење резултата различитих истраживања и евалуација.

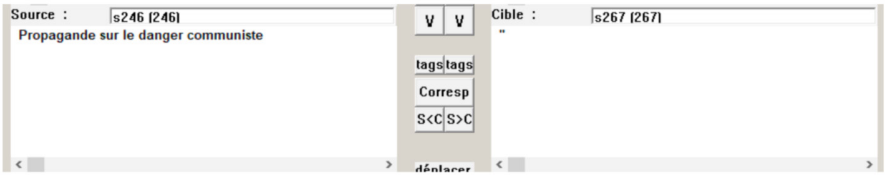
5. Резултати

На основу евалуације упаривања корпуса ЛМД уз помоћ програма *Alinéa* и *LF Aligner*, изведена је ф-мера од 0,66 за *Alinéa* и 0,78 за *LF Aligner*. Када је реч о другој методи евалуације, анализа је показала да је *LF Aligner* потпуно тачно упарио 81,97 % корпуса, док је *Alinéa* упарила само 66,3 % потпуно тачних парова сегмената. С друге стране, *LF Aligner* је делимично нетачно упарио само 4,8 %, док је код *Alinéa* реч о чак 14,97 %. У категорији потпуно нетачних парова сегмената, *LF Aligner* се поново показао учинковитијим од *Alinéa*, те је нетачно упарио 13,23 %, за разлику од *Alinéa* где је реч о 18,73 %.

	Alinéa	*LF Aligner*
потпуно тачни	66,3 %	81,97 %
делимично нетачни	14,97 %	4,8 %
потпуно нетачни	18,73 %	13,23 %
ф-мера	**0,66**	**0,78**

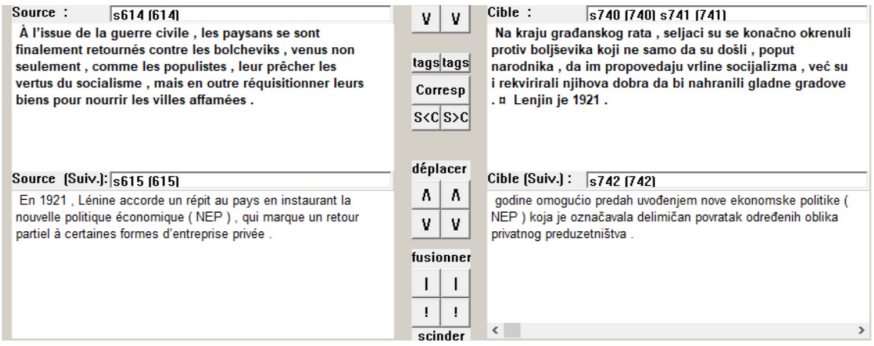
Табела 2. Резултати упаривања корпуса ЛМД уз помоћ програма Alinéa и LF Aligner.

При анализи грешака у упаривању, уочљиво је да програми имају потешкоће са неким типовима сегмената. Први случај односи се на наводнике у француском делу корпуса, који су упарени са неодговарајућим сегментима на српском. Наиме, пошто су у француском језику наводници одвојени размаком од текста на који се односе, програми их одвајају у засебан сегмент. Тај је сегмент у неким случајевима упарен са читавим реченицама у другом сегменту.



Илустрација 5. Потпуно нетачан сегмент, проблем с наводницима.

Други се случај тиче датума у српском језику и њиховог упаривања с француским текстом. Пошто се датуми у српском исказују редним бројевима и пишу с тачком, програми за упаривање тачку препознају као крај реченице и део датума преносе у наредни пар сегмената.



Илустрација 6. Делимично нетачан сегмент, проблем с датумом.

Трећи случај односи се на сегменте који немају кореспондента. У питању су реченице на француском језику које нису преведене на српски или реченице на српском језику које дају додатне информације читаоцима српског издања, те нису присутне у француском тексту. Како програм настоји да упари сегменте једне с другима, сегменти без кореспондената често бивају упарени са неким неодговарајућим сегментом.

3	[s3] L'actualité récente démontre l'urgence d'une relecture des textes saints islamiques par le biais d'une analyse du contexte de la révélation .	[s3] Nalazimo se u trenutku u kom je teško baviti se islamom bez osvrtanja na vesti i sveprisutni strah od " džihada " .

Илустрација 7. Потпуно нетачан сегмент, сегмент без кореспондента.

На основу добијених резулзтата, могуће је закључити да је *LF Aligner* постигао знатно боље резултате. Овај програм тачно је упарио 15,67 % више сегмената од *Alinéa.* Треба истаћи и да је *LF Aligner* имао 5,5 % мање потпуно нетачних парова. Резултати указују на то да је при упаривању у програму *LF Aligner*, у корпусу присутан мањи број делимично нетачних парова сегмента. Из ове чињенице се може закључити да је *LF Aligner* имао мање потешкоћа са поменутим проблемима у упаривању сегмената са наводницима или датумима.

Ипак, ако се упореде резултати упаривања корпуса ЛМД и резултати евалуација спроведених у оквиру истраживања која су спро-

вели Kraif (2001 : 30) и Varga et al. (2005 : 594), уочљиво је да су се програми показали мање учинковитим на корпусу ЛМД. Ова разлика може се, између осталог, објаснити чињеницом да је Kraif (2001 : 30) анализираио енглески и француски језик, који имају значајан број речи које се идентично пишу и које стога олакшавају упаривање. С друге стране, будући да је *LF Aligner* створен најпре за мађарско-енглеске корпусе, овај аргумент се у том случају не може применити. Могуће је ипак објаснити наведену разлику чињеницом да *LF Aligner* користи веома богат лексикон за упаривање мађарско-енглеских корпуса, што је могло допринети успеху на корпусу из поменутих истраживања.

6. Закључак

Анализа упаривања француско-српског корпуса ЛМД показала је да је од два анализирана програма, *Alinéa* и *LF Aligner*, *LF Aligner* постигао боље резултате. Оваква разлика у учинку, може бити проузрокована разликом у алгоритмима коју анализирани програми користе. Наиме, успех програма *LF Aligner* може се објаснити чињеницом да овај програм не користи само општи, статистички приступ у упаривању, већ се ослања и на двојезични лексикон. Резултати нашег истраживања указују на то да лексичке информације могу позитивно утицати на учинковитост аутоматског упаривања.

Упркос резултатима програма *LF Aligner*, важно је истаћи да ниједан од два програма није постигао довољно висок степен успеха да би се резултати аутоматског упаривања корпуса ЛМД могли употребити без додатне интервенције људског анотатора. Ипак, и уз наведено ограничење, програми за аутоматско упаривање француско-српских двојезичних корпуса могу бити корисна алатка, поготово у стварању и даљем развоју великих двојезичних корпуса. Најбољи резултати постижу се аутоматским упаривањем, уз помоћ програма као што је *LF Aligner*, уз накнадну интервенцијом људског анотатора који би уклонио грешке у упаривању.

Извори и литература

Brown, Peter, Jennifer Lai i Robert Mercer. «Aligning Sentences in Parallel Corpora». *29th Annual Meeting of the Association for Computational Linguistics* (Jun 18 - 21, 1991), Berkli, 1991. 169-176. Štampano.

Chiao, Yun-Chuang, Olivier Kraif i dr. «Evaluation of multilingual text alignment systems: the ARCADE II project». *Proceedings of LREC 2006* (Maj 2006). Đenova, 2006. Štampano.

Gale, William i Kenneth Church. «A Program for Aligning Sentences in Bilingual Corpora». *29th Annual Meeting of the Association for Computational Linguistics* (Jun 18 - 21), Berkli, 1991. 177-184. Štampano.

4. Kraif, Olivier. «Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation» [tekst pripreme za štampu], 2001, 1-44. Veb. 29.9.2018..

Kraif, Olivier. «Qu'attendre de l'alignement de corpus multilingues ?». *Traduire, 4e Journée de la traduction professionnelle*, 2006, 17-37. Štampano.

Langlais, Philippe, Michel Simard i Jean Véronis. «Methods and Practical Issues in Evaluating Alignment Techniques». *98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1,* Montreal, 1998. 711-717. Štampano.

Simard, Michel i Pierre Plamondon. «Bilingual sentence alignment: Balancing robustness and accuracy». *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA).* 1996. 59-80. Štampano.

Véronis, Jean i Philippe Langlais. "Evaluation of parallel text alignment systems". *Parallel Text Processing. Text, Speech and Language Technology : 13.* Springer : Dodrecht, 2000. 369-388. Štampano.

Varga, Daniel, Laszlo Nemeth i dr. «Parallel corpora for medium density languages». *Proceedings of the RANLP 2005*, Budapest, 2005. 590-596. Štampano.

Jovana Milovanovic

AUTOMATIC ALIGNMENT APPLIED TO FRENCH-SERBIAN BILINGUAL CORPORA

Summary

In this paper, we present an analysis and evaluation of automatic alignment of French-Serbian parallel texts. The first point discussed are different approaches and processes in the automatic (statistical lenght-based method, the use of bilingual dictionnaries, etc.). Special attention is given to two alignment tools, Alinéa and Hunalign. The aforementioned tools have been tested on a corpus named LMD, specifically created for the present research, consisting of texts published in the Le Monde Diplomatique in French and Serbian. Following the initial alignment, an evaluation was carried out and the sentence couples were classified into three categories: correct alignment, partially incorrect alignment and incorrect alignment. Furthermore, an analysis of alignment errors is presented, followed by suggestions of modifications in the preparatory phase of alignment that could prevent or minimize errors. The purpose of this research is to examine whether alignment tools for bilingual texts are efficient on texts in French and Serbian, to determine the most adequate alignment tool for this specific language pair and to emphasize the importance and the application of these tools in the development of bilingual corpora.

Keywords: computational linguistics, automatic alignment, bilingual corpora, french, serbian, alinéa, hunalign.