UDK: 81’322.4

ДОБИЈАЊЕ СИНТАКСИЧКИХ ПРАВИЛА ПРЕВОЂЕЊА ИЗ ПАРАЛЕЛНЕ БАНКЕ СТАБАЛА

Михаела Колхон, mcolhon@inf.ucv.ro, Универзитет у Крајови Катедра за рачунарство A.J. Куза 13, 200585 Крајова, Румунија

С енглеског превела Јелена Бајић

Апстракт

Студија коју представљамо се бави развојем система машинског превођења у коме се користи паралелни алгоритам за екстракцију синтаксичких фраза да би се добио богати и робусни скуп синтаксичких образаца за превођење. Да би се овај приступ учинио изводљивим, разматрају се само поравнања фраза са фразама из двојезичне банке стабала са синтаксичким конституентима.

Знање садржано у овој студији се дефинише путем синтаксички мотивисаних преводних секвенција записаних у виду морфосинтаксичких спецификација, пошто су настале у оквиру пројекта MULTEXT-East (Erjavec, 2010).

Кључне речи

Паралелни синтаксички шаблони, Превођење засновано на фразама.

1. Увод

Основни приступи у машинском превођењу су *директно превођење, превођење засновано на трансферу* и *статистичко превођење*. У директном машинском превођењу, превод се врши реч по реч или фраза по фраза, док се у приступу који се заснива на трансферу, не обраћа пажња на лексички ниво, пошто се превод генерише уз помоћ механизма за трансфер дефинисаног на структуралном нивоу да би синтаксичка репрезентација конструкција изворног језика била пренета на одговарајуће конструкције у циљном језику. Систем за статистичко машинско превођење се састоји од модела преводног односа, а правила превођења се добијају аутоматски из извора који служи за обуку (у овом случају, то је двојезични корпус).

У обради природних језика, корпус се користи да би се систему за машинско превођење обезбедили емпиријски и статистички подаци. Паралелни корпуси се могу користити да би се генерисало изузетно вредно језичко знање које им омогућава да подржавају аутоматску идентификацију сегмената текста који представљају реципрочне преводе (Tufiş & Ion, 2007). Два сегмента текстова из *битекста* (паралелног текста) који представљају реципрочне преводе чине *јединицу превода* (Tufiş & Ion, 2007). Јединице превода које одговарају синтаксичким фразама се могу користити за генерисање других реченица у циљном језику система за машинско превођење: уместо генерисања превода појединачних речи у изворном језику, генеришу се преводи фраза и коначни превод се склапа њиховим пермутовањем (Yamada & Knight, 2001).

Паралелни корпуси могу да подрже све врсте система за машинско превођење: у директним приступима, корпуси се користе ради екстракције информација о лексичким јединицама (како се одређена реч преводи у извесном окружењу), у приступима који се заснивају на трансферу, корпуси се користе за екстракцију правила трансфера, док се у статистичким приступима корпуси користе за екстракцију правила превођења и додељивање вероватноће могућим преводима.

Паралелна банка стабала је посебна врста паралелног корпуса у коме су реченице синтаксички анотиране. Синтаксичка стабла битекста из паралелне банке стабала су поравната на подреченичном нивоу (на нивоу речи, фразе или клаузе), што се у литератури назива *фразно поравнавање*, Паралелне банке стабала могу да служе као корпус за аутоматско извођење правила трансфера или за екстракцију двојезичних речника или уопште за проучавање превођења (Samuelsson & Volk. 1993).

Методи машинског превођења све више користе не само формални апарат синтаксе, него и лингвистичке структуре типа стабла из изворног језика, циљног језика или оба. Технике за екстракцију фраза у оквиру статистичког машинског превођења које се базира на фразама, иако нису синтаксички мотивисане пружају веома велику покривеност (Ambati et al., 2009). Основни овакви системи раде са паровима фраза који су у складу са поравнавањем речи: речи унутар фразе су непрекинуте секвенције које се састоје од речи поравнатих једних према другима, а не према речима изван секвенције (Wenniger et al., 2010).

Машинско превођење које се заснива на синтаксичким стаблима је веома много проучавано последњих година, пошто постоји општа потреба да се побољша учинак најновије генерације статистичког машинског превођења које се базира на фразама (Araùjo & Caseli, 2010).

Поравнавање чворова два паралелна стабла парсирања може да пружи информације о структуралном поравнању представљеног пара реченица, прецизније, може да помогне у експерименту којим се тестира изводљивост

аутоматског међујезичког трансфера синтаксичких конституената. Уопштено говорећи, компонента трансфера је систем правила који повезује речи и структуре у једном језику са речима и структурама у другом језику (циљном језику).

Фразама се, традиционално, сматрају синтаксички конституенти реченице. Фразе се могу веома добро поравнавати, чак и ако све речи двеју фраза нису поравнате. Поравнавањем унутрашњих чворова двају паралелних стабала парсирања, остварује се веза између фраза представљених тим чворовима, а директан резултат тога је поравнавање подстабала чији корени се налазе у тим чворовима.

Савремена пракса у хијерархијском превођењу заснованом на фразама подразумева екстракцију исправних фраза и хијерархијских правила из битекстова поравнатих на нивоу речи (Gispert et al., 2010). Чак и ако модели засновани на поравнавању нису примерени за директну примену у превођењу, и даље могу да дају обиље корисних информација, као што је статистика потребна за прављење шаблона за превођење.

Ове технике су показале да почињање од великих табела синтаксичких фраза и давање предности синтаксичким када се преклапају са несинтаксичким фразама дозвољава стицање „преводилачког знања“ Оне показују повећање брзине декодирања и побољшање квалитета превода које произилази из прецизности ових фраза мотивисаних синтаксом (Ambati et al., 2009).

Очекује се да ће већина фраза идентификованих у стаблима парсирања бити преведена без мешања са другим фразама или речима. Уопштено гледано, именичке фразе се много више придржавају наведеног правила. Супротно од њих, глаголске фразе обично трпе структурне модификације током превођења, што је изазвано померањем адјункта (Colhon, 2011).

Применом поравнавања чворова између токена паралелних стабала парсирања, могу бити екстраховани модели превођења за синтаксичко машинско превођење. Прикупљањем паралелних синтаксичких образаца из двојезичне банке стабала, можемо генерисати скупове образаца за превођење доброг квалитета које треба да научи статистички систем за машинско превођење које се базира на синтакси. Главни циљ овде представљене студије је стварање образаца за превођење мотивисаних синтаксом.

1.1 Енглеско-румунски корпус. Шта садржи?

Машинско превођење је тежак и комплексан задатак. Један од начина да се побољшају резултати аутоматског превођења је да се текстови које треба превести ограниче на одређену област у којој су садржаји међусобно слични. Оваквим приступом се смањује присуство двосмислености у процесу превођења и стога лексикон може бити много мањи и одређенији. Системи за машинско превођење који користе овај приступ се често базирају на корпусу, што значи да користе податке за превођење у облику паралелних корпуса да би изградили своје језичке ресурсе. Паралелни корпуси се, такође, могу користити као подршка контрастивним међујезичким проучавањима.

Да би се могао користити за обуку и тестирање, корпус мора претходно да се обради. Фаза претходне обраде обично обухвата дељење на реченице, токенизацију, етикетирање врстама речи, лематизацију.

Скуп за етикетирање врстама речи се разликује од језика до језика, али већина дефиниција скупа етикета се може превести са једног језика на други уз помоћ једноставних табела за превођење. У пројекту MULTEXT-East (Erjavec et al., 2003) *морфосинтаксички описи* (скраћено, MSD) се дефинишу као исти скуп етикета за све језике у пројекту.

Корпус који подржава овде представљену студију се зове *JRC- Acquis* и чини га компилација дела паралелних текстова из домена правне регулативе *Acquis Communautaire*[1]. *Acquis Communautaire* представља скуп свих правних аката Европске уније применљивих на земље чланице. Тај скуп се стално мења и тренутно обухвата текстове написане између педесетих година прошлог века и 2008. на свим језицима земаља чланица ЕУ. *Acquis Communautaire* је збирка паралелних текстова на 22 званична језика, укључујући и румунски.

Корпус *JRC-Acquis* има неколико предности са аспекта коришћења у сврху машинског превођења:

- Посвећен је једном одређеном домену, правним актима Европске уније, али вокабулар није ограничен, пошто укључени текстови обухватају разна поља која потпадају под поменуту правну регулативу.
- У питању је, макар теоретски, конзистентан паралелни корпус, што произилази из званичне природе докумената, чија структура мора бити веома добро организована (Iftene et al., 2010).

За потребе студије која је овде представљена, користили смо корпус на енглеском и румунском језику који потиче из паралелних делова корпуса *JRC-Acquis* на енглеском и румунском. Садржи 1420 реченица и укупно 12.613 токена на оба језика.

1.2 Стварање паралелне банке стабала

Ова студија је спроведена на основу енглеско-румунске банке стабала (Colhon, 2012) састављене од 1420 реченица из енглеско-румунског корпуса који је на *Универзитету Александру Јоан Куза у Јашију* развила *Група за обраду природних језика*[2] са *Факултета за рачунарство*. За потребе стварања овог двојезичног корпуса, коришћени су паралелни делови корпуса *JRC-Acquis* на енглеском и румунском.

Да би се могли користити у поменуте сврхе, сирови текстови двојезичног корпуса су анотирани на више нивоа и кодирани у формату XML, при чему је примењен упрошћени стандард XCES (Ide et al., 2000). Текстови су тако сегментирани, а додате су им лексичке информације. Резултат тога је да су реченичне границе означене и да је сваком токену додељена одговарајућа врста речи (скраћено POS), лема и морфосинтаксичке информације (род, број, лице, падеж, итд), што је учињено извршавањем аутоматског ланца обраде који укључује: сегментацију реченица, токенизацију, етикетирање врстама речи и лематизацију (Simionescu, 2012).

Скупови обележја коришћени за анотирање речи у двојезичном корпусу потичу из морфосинтаксичких спецификација MULTEXT-East–а (најновија верзија тих спецификација је дата у (Erjavec, 2010)).

Да би се направила банка депенденцијалних стабала, људи – анотатори морају да одлуче за сваку реч од које друге речи она зависи. Одређивање међузависности често подразумева процес дубинске интерпретације и то је разлог зашто се понекад дешава да анотатори сматрају да иста секвенца речи може имати различите структуре зависности. Са друге стране, одлуке које треба да доносе људи – анотатори током прављења банака стабала фразних структура често карактерише значајно нижи степен двосмислености (Colhon & Cristea, 2012).

1 Група за језичку технологију Европске комисије је компилирала значајан део текстова из збирке Acquis Communautaire у паралелни корпус под називом JRC-Acquis корпус (Cristea & Forăscu, 2006).

2 Информације о Групи за обраду природних језика са Универзитета Александру Јоан Куза у Јашију се налазе на адреси: http://instrumente.infoiasi.ro/NLPTest/about.jsp

Да би се направила паралелна банка стабала из двојезичног корпуса, био је потребан статистички парсер широког спектра, а требало је и извршити и поравнавање на нивоу речи. Пошто су информације о поравнавању од кључне важности у процесу развоја банке стабала за део корпуса који се односи на циљни језик, одабрали смо да их прецизирамо ручно.

Синтаксичка стабла румунских текстова се генеришу на основу синтаксичких фраза, у паралелним текстовима на енглеском језику, аутоматски добијених применом синтаксичког парсера развијеног на Станфорду (Klein & Manning, 2003). Механизам за генерисање стабала на румунском језику поново употребљава и прилагођава алате и алгоритме за међујезички трансфер синтаксичких конституената и поравнавање синтаксичких стабала. Ручно смо у произашлој банци стабала исправили грешке које се тичу комплетности[3] (конзистентност[4] је обезбеђена стварањем двојезичног корпуса).

Енглеско-румунска банка стабала која се користи у овој студији је сачињена од синтаксичких стабала аутоматски добијених применом добро познатог парсера на део корпуса на енглеском, као и од њихових структура пројектованих на одговарајуће текстове на румунском. Те пројекције се дефинишу на основу поравнавања на нивоу речи, која су спецификована за паралелне реченице у корпусу.

С обзиром на примењени механизам генерисања банке стабала за текстове на румунском, за сваки паралелни пар синтаксичких стабала из произашле енглеско-румунске банке стабала, синтаксички конституенти на енглеском су имплицитно поравнати са фразама на румунском (које представљају секвенције циљних речи). Из ових информација о хијерархијском поравнавању се могу добити трансформациона синтаксичка правила, као што ће бити показано у одељцима који следе.

2. Паралелни обрасци са синтаксичким конституентима

Следећи Гелијев метод описан у (Galley et al., 2004), процес екстракције фраза је подржан паралелним стаблима парсирања из изграђене енглеско-румунске банке стабала. За свако поравнање између унутрашњих чворова синтаксичких стабала, анализирају се потомци поравнатих чворова. Одређене речи или конструкције извесне структуре могу бити истакнуте у листи потомака према сврси екстракције синтаксичких секвенца.

У овом чланку се синтаксичке секвенције описују да би се пружиле информације о начину на који функцијске речи могу да утичу на превод. Из тог разлога су функцијске речи дате у пуном облику, уз потпуне информације о њиховим морфосинтаксичким својствима.

У свакој синтаксичкој структури, можемо идентификовати две главне категорије речи:

- *пунозначне речи* које описују предмете, ентитете, својства, односе или догађаје и које су синтаксички представљене *именицама*, *придевима*, *глаголима* и *прилозима*.
- *функцијске речи* које помажу да се речи сложе у исправан структурни реченични облик. Такође, функцијске речи нам могу указати како су друге реченичне компоненте међусобно повезане. Функцијске речи могу бити *детерминатори*, *квантификатори*, *предлози* или *везници*.

Опсегом чвора *n* на синтаксичком стаблу се сматра подскуп чворова до којих се може допрети из *n* (Galley et al., 2004). Идући одоздо нагоре, алгоритам за екстракцију паралелних синтаксичких образаца „посећује“ свако

3 Комплетност значи да су сваки токен и сваки чвор део синтаксичког стабла (Samuelsson & Volk, 1993).

4 Конзистентност значи да је иста секвенција токена анотирана на исти начин у читавој банци стабала (Samuelsson & Volk, 1993).

синтаксичко дрво на енглеском и отвара све његове унутрашње чворове који су поравнати са најмање једним чвором из румунског паралелног синтаксичког дрвета. Синтаксичка структура и опсези поравнатих фразних чворова на енглеском и румунском чине паралелне обрасце у нашој студији и стога се чувају у бази података (видети **Одељак 2.2**).

Овај метод је довољно брз и једноставан да се може применити на велике скупове података. Наводимо неке од улога које паралелни синтаксички обрасци имају са аспекта научених правила превођења:

- једноставни лексички обрасци за превођење специјалних речи, као што су *функцијске речи*
- обрасци у које су унети необавезни модификатори
- обрасци у којима смо пронашли „лексичке рупе“ које одређује постојање мапирања „један према нула“ између речи, односно, токена у паралелним секвенцијама. На пример, именичке фразе на енглеском језику које се састоје од две именице повезане функцијском речју „оф“. У одговарајућем преводу на румунски сепаратор нестаје.
- Анализирајући велике скупове паралелних образаца можемо идентификовати „склоности врста речи“; обично је познато да преведене речи често припадају истој врсти речи као њихови еквиваленти у изворном језику, али када то није случај, врста речи која се јавља у преводу није случајна.

У овој студији нас занимају варијације или ограничења одређена функцијским речима у конструкцијама у изворном језику. Направили смо базу података која се састоји од неколико паралелних секвенција на енглеском и румунском представљених у различитим форматима. Сваки формат у коме су секвенције представљене има за циљ да обухвати информације о преводу које се тичу паралелних секвенција:

- синтаксичку структуру секвенција,
- поравнавања на нивоу речи
- информације о MSD речи; када су у питању функцијске речи, разматрамо и облик речи

2.1 Формализам репрезентације

Један од захтева који се поставља пред генерички оквир за индуковање правила из паралелних података је да буде способан да инкорпорира све врсте синтаксичких информација које могу долазити са било које стране језичког пара укљученог у машинско превођење (Ambati et al., 2009). Предложени формализам ради са синтаксом конституената са обе стране обраћајући пажњу на значај који функцијске речи имају у процесу превођења.

Предложени алгоритам за екстракцију ради у две фазе да би идентификовао паралелне секвенције из банке стабала. У првој фази чворови из изворног стабла који се поравнавају са чворовима из циљног стабла су екстраховани заједно са њиховим опсезима дефинисаним у смислу морфосинтаксичких етикета (скраћено MSD етикета). Једине лексичке информације које се узимају у обзир, тичу се функцијских речи. Одредили смо да једини облици речи који се појављују у предложеној репрезентацији одговарају функцијским речима. Такве паралелне синтаксичке фразе називамо *синтаксичким обрасцима.*

У другој фази, вршимо екстракцију подстабала из два паралелна синтаксичка стабла чији су корени поравнати и бележимо њихову структуру, придржавајући се начина употребе заграда који је коришћен у пројекту „Банка стабала Пен“[5]. У наставку, овај опис

5 Банка стабала Пен (*Penn Treebank*) је структурно анотирани корпус који се састоји од секвенције реченица (преко 1 милион) извађених из новина *Wall Street Journal*. Свака реченица у корпусу је анотирана етикетама врста речи и фразним структурама.

ће се звати *хијерархијски обрасци*.

Из енглеско-румунске банке стабала са синтаксичким конституентима, екстраховано је 4389 (синтаксичких и хијерархијских) паралелних образаца. Репрезентација за чување образаца може да пружи довољно добре описе домена локалности за фунцијске речи, али није ограничен само на то.

У било ком приступу заснованом на обрасцима, важно је да се правилно одабере формализам описа образаца, као и структура образаца (Kurc et al., 2010). Знање обухваћено овом студијом је представљено синтаксичким обрасцима дефинисаним путем морфосинтаксичких спецификација, развијених током пројекта „MULTEXT-East“ (Erjavec, 2010), као и путем спецификација етикета врата речи и фразних етикета, на начин на који су уведене у пројекту „Банка стабала Пен“[6].

2.1.1 Синтаксичке секвенције на енглеском

Формализам који овде предлажемо подразумева да сваки образац кодира синтаксичке структуре конструкција природног, енглеског, језика и приказује их на један од два начина: синтаксички и хијерархијски. У обе презентације се посебна пажња посвећује функцијским речима које се појављују у приказаној секвенцији, јер те речи сматрамо веома значајним са аспекта система за машинско превођење.

Следи синтаксичка репрезентација синтаксичких образаца на енглеском:

[Phrasal_Tag tag(c_1) … tag(c_n)]

где је Phrasal_Tag обележје за фразу у формализму који се примењује у банци стабала Пен, док tag(c_1) … tag(c_n) представља нотацију која се користи за директне конституенте c_1 … c_n приказане фразе, при чему је:

$$tag(c_i) = \begin{cases} Penn\ Phrasal\ Tag(c_i),\ c_i - a\ phrase\ constituent \\ Penn\ POS\ Tag(c_i) / c_i, c_i - a\ functional\ word \\ Penn\ POS\ Tag(c_i),\ c_i - a\ content\ word \end{cases}$$

Пошто је део банке стабала на енглеском генерисан уз помоћ Станфордског парсера, генерисана су стабла парсирања банке стабала ПЕН. Директна последица тога је да су енглеске речи у банци стабала енглеског језика анотиране уз помоћ ПЕН етикета врста речи, из разлога што Станфордски парсер користи тај стандард. У овом формализму анотације, функцијске речи из текстова на енглеском се могу сматрати реченичним токенима који у формализму ПЕН који се односи на скуп етикета врста речи имају једну од следећих етикета: CC (напоредни везник), DT (детерминатор), IN (предлог/ зависни везник), MD (модал), PRP (лична заменица), PP$ (присвојна заменица), RP (речца), TO (реч *to*), WDT (*wh*-детерминатор), WP (*wh*-заменица), WP$ (присвојна *wh*-заменица), WRB (*wh*-прилог).

За овај тип репрезентације, разматрамо две верзије: једну која обележава фразне чворове њиховим индексима и конституенте фразних чворова индексима румунских еквивалената са којима су поравнати и другу верзију која уклања све те информације о индексима и концентрише се на синтаксичку структуру фразних чворова.

При представљању секвенција на енглеском, у облику хијерархијских образаца, придржавамо се репрезентације синтаксичких стабала са конституентима уз коришћење заграда (Taylor, 1996). С обзиром да се у репрезентацији синтаксичких конституената уз помоћ заграда прецизирају и облици речи фразе, применом таквог формализма се могу добити лексичке информације за представљену фразу уз комплетан приказ њене структуре.

У наставку дајемо два примера презентације синтаксичких секвенција на енглеском.

6 Веб адреса пројекта је http://www.cis.upenn.edu/~treebank/.

Пример 1. Конструкција „the preparation" узета из корпуса ће бити приказана на следећи начин:

- синтаксички образац:
  [NP-512[7] {3}:DT/the {3}:NN][8] (са информацијама о индексу)
  [NP DT/the NN] (без информација о индексу)
  У оваквој презентацији се води рачуна о синтаксичкој структури енглеске фразе, тако што се визуелно истичу функцијске речи које бивају идентификоване, не само преко својих синтаксичких етикета, већ и преко свог облика (видети конструкцију DT/the).
  У презентацији са информацијама о индексу, бројеви у угластим заградама, који претходе синтаксичким етикетама, представљају поравнања у паралелној румунској фрази која одговара овој фрази на енглеском означеној као NP-512. Из приложених информација о индексу, примећује се да се цела конструкција кодирана са DT/the NN поравнава са једним јединим токеном у фрази на румунском који има индекс 3.
- хијерархијски образац:
  [NP-512 [[DT/the] [NN preparation]]]
  У овој презентацији, дата је читава структура, укључујући терминалне чворове - секвенције речи - која одговара синтаксичкој фрази NP-512. Ова фраза се састоји од детерминатора [DT/the] након кога следи именички део [NN preparation].

Пример 2. Конструкција "the proposal from the commission" је приказана на следећи начин:

- синтаксички образац:
  [NP-517 {506}:NP {507}:PP] (са информацијама о индексу)
  [NP NP PP] (без информација о индексу)
  У овој презентацији је кодирана структура именичке фразе на енглеском (означене као NP-517 у формализму са информацијама о индексу) која је састављена од именичке фразе поравнате са конституентом поравнате фразе на румунском који има индекс 506 и предлошке фразе поравнате са конституентом чији индекс износи 507 у паралелној фрази на румунском.
- хијерархијски образац:
  [NP-517
  [[NP-508 [DT/the] [NN proposal]] [PP-516 [IN/from] [NP-515 [DT/the]
  [NN commission]]]]]
  У овој презентацији, дата је именичка фраза NP-517 са целим подстаблом којим управља овај фразни чвор.

Примењујући исте презентације, одговарајуће синтаксичке секвенције на румунском су кодиране у сличном формату, уз само једну разлику: за румунске речи примењујемо морфосинтаксичке анотације MULTEXT- East, пошто сматрамо да су те спецификације одговарајуће с обзиром на богатство флексије румунског језика.

2.1.2 Синтаксичке секвенције на румунском

Румунска синтаксичка стабла из банке стабала су аутоматски образована применом алгоритма који генерише стабла одоздо нагоре, и кога воде поравнања на нивоу речи у корпусу (Colhon, 2012). Из двојезичног корпуса на основу кога је сачињена банка стабала смо сачували анотације MULTEXT-East специфика-

7 Што се тиче унутрашњих чворова стабла парсирања, одабрали смо да их означимо фразним обележјима и вредностима њихових индекса. На тај начин можемо да разликујемо два унутрашња чвора стабла парсирања који имају исто фразно обележје.

8 Видети Апендикс А у коме је дат опис етикета из банке стабала Пен коришћених у овом раду.

ција речи из корпуса, пошто ти подаци укључују довољно морфосинтаксичких детаља потребних у свакој синтаксичкој студији, док се за означавање фразних конституената користе фразне етикете из ПЕН банке стабала.

Директна последица тога је да су румунске функцијске речи они токени или речи који у формализму скупа етикета MULTEXT-East имају етикете морфосинтаксичког описа са следећим префиксима: P_ (заменица) као што су Pd_ (показна заменица), Ps_ (присвојна заменица), Px_ (повратна заменица), D_ (детерминатор), T_ (члан), S_ (апозиција), C_ (везник), Q_ (речца).

На синтаксичке обрасце румунског језика је примењен исти формализам презентације који је коришћен за синтаксичке обрасце енглеског са једном једином разликом: речи су анотиране етикетама морфосинтаксичког описа MULTEXT-East, уместо етикетама врста речи из банке стабала Пен. Код нас је румунски синтаксички образац приказан у облику листе:

[Phrasal_Tag tag(c1) … tag(cn)]

где је Phrasal_Tag је фразна ознака у формализму банке стабала Пен, а $tag(c_1) \ldots tag(c_n)$ нотација коришћена за директне конституенте фразе $c_1 \ldots c_n$ где је:

$$tag(c_i) = \begin{cases} Penn\,Phrasal\,Tag(c_i),\ c_i - a\ phrase\,constituent \\ MSD\,Tag(c_i)/c_i, c_i - a\ functional\,word \\ MSD\,Tag(c_i),\ c_i - a\ content\,word \end{cases}$$

У презентацији хијерархијских секвенца на румунском је примењивана репрезентација са угластим заградама за представљање синтаксичких стабала са конституентима (Samuelsson & Volk. 1993).

У **примерима 3** и **4**, фразе на румунском које одговарају фразама на енглеском, датим у **примерима 1** и **2** су наведене заједно са својим обрасцима презентације.

Пример 3. Румунска конструкција "pregătirea" поравната са енглеском секвенцијом "the preparation" је представљена на следећи начин:

- синтаксички образац:
  [NP-513 {3}:Ncfsry][9] (са информацијама о индексу)
  [NP Ncfsry] (без информација о индексу)
  У овој презентацији је кодирана именичка фраза која се састоји од именице чији је индекс 3 у реченици, а чија морфосинтаксичка спецификација гласи Ncfsry.
- хијерархијски образац:
  [NP-513 [Ncfsry pregătirea]]
  Ова презентација даје синтаксичко подстабло које одговара чвору који је обележен са NP-513.

Пример 4. Конструкција „propunerea Comisiei" је представљена на следећи начин:

- синтаксички образац:
  [NP-508 {506}:NP {507}:NP] (са информацијама о индексу)
  [NP NP NP] (без информација о индексу)
  У овој презентацији је кодирана структура именичке фразе која носи ознаку NP-508 и састоји се од две именичке фразе чији је индекс у румунском стаблу парсирања 506, односно, 507.
- хијерархијски образац:
  [NP-508 [[NP-506 [Ncfsry propunerea]] [NP-507 [Ncfsoy Comisiei]]]]
  У овој презентацији је кодирано синтаксичко подстабло са кореном у чвору који носи обележје NP-508.

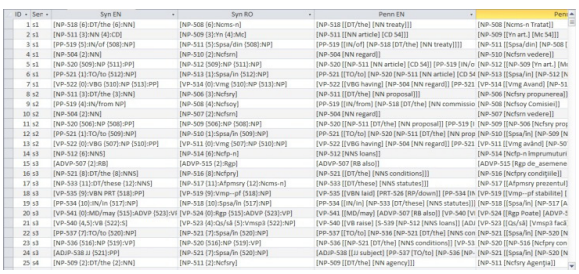
2.2 Језички ресурс са синтаксичким обрасцима

Настале енглеско-румунске паралелне секвенције из двојезичне банке стабала се чувају у бази података (видети **Слику 1**) којој се могу даље постављати упити да би се добиле корисне информације са тачке гледишта син-

9 Видети Апендикс А у коме је дат опис етикета морфосинтаксичког описа коришћених у овом раду.

таксичког трансфера између фраза на енглеском и румунском.

Подаци из базе података кореспондирају са енглеско-румунским паровима поравнатих фраза узетих из банке стабала и кодирају њихову структуру са шест записа. Синтаксички обрасци су представљени у четири поља (енглески и румунски обрасци са и без информација о индексу), а у два поља се налазе хијерархијски обрасци енглеских, односно, румунских фраза. Примери синтаксичких образаца из ове језичке базе података су дати у **Табели 1** и **Табели 2**.



ID	Ser	Syn EN	Syn RO	Penn EN	Penn
1	s1	[NP-518 (6):DT/the (6):NN]	[NP-508 (6):Ncms-n]	[NP-518 [[DT/the] [NN treaty]]]	[NP-508 [Ncms-n Tratat]]
2	s1	[NP-511 (3):NN (4):CD]	[NP-509 (3):Yn (4):Mc]	[NP-511 [[NN article] [CD 54]]]	[NP-509 [[Yn art.] [Mc 54]]]
3	s1	[PP-519 (5):IN/of [NP-508]:NP]	[NP-511 (5):Spsa/din [NP-508]:NP]	[PP-519 [[IN/of] [NP-518 [DT/the] [NN treaty]]]]	[NP-511 [Spsa/din] [NP-508 [
4	s1	[NP-504 (2):NN]	[NP-510 (2):Ncfsrn]	[NP-504 [NN regard]]	[NP-510 [Ncfsrn vedere]]
5	s1	[NP-520 [509]:NP (511):PP]	[NP-512 [509]:NP (511):NP]	[NP-520 [[NP-511 [NN article] [CD 54]] [PP-519 [IN/o	[NP-512 [[NP-509 [Yn art.] [Mc
6	s1	[PP-521 (1):TO/To (512):NP]	[NP-513 (1):Spsa/în (512):NP]	[PP-521 [[TO/to] [NP-520 [NP-511 [NN article] [CD 54	[NP-513 [Spsa/în] [NP-512 [N
7	s1	[VP-522 (0):VBG (510):NP (513):PP]	[VP-514 (0):Vmg (510):NP (513):NP]	[VP-522 [[VBG having] [NP-504 [NN regard]] [PP-521	[VP-514 [Vmg Având] [NP-51
8	s2	[NP-511 (3):DT/the (3):NN]	[NP-506 (3):Ncfsry]	[NP-511 [[DT/the] [NN proposal]]]	[NP-506 [Ncfsry propunerea]]
9	s2	[PP-519 (4):IN/from NP]	[NP-508 (4):Ncfsry]	[PP-519 [[IN/from] [NP-518 [DT/the] [NN commissio	[NP-508 [Ncfsry Comisiei]]
10	s2	[NP-504 (2):NN]	[NP-507 (2):Ncfsrn]	[NP-504 [NN regard]]	[NP-507 [Ncfsrn vedere]]
11	s2	[NP-520 [506]:NP (508):PP]	[NP-509 [506]:NP (508):NP]	[NP-520 [[NP-511 [DT/the] [NN proposal]] [PP-519 [I	[NP-509 [NP-506 [Ncfsry proj
12	s2	[PP-521 (1):TO/To (509):NP]	[NP-510 (1):Spsa/în (509):NP]	[PP-521 [[TO/to] [NP-520 [NP-511 [DT/the] [NN prop	[NP-510 [Spsa/în] [NP-509 [N
13	s2	[VP-522 (0):VBG (507):NP (510):PP]	[VP-511 (0):Vmg (507):NP (510):NP]	[VP-522 [[VBG having] [NP-504 [NN regard]] [PP-521	[VP-511 [Vmg având] [NP-50
14	s3	[NP-512 (6):NNS]	[NP-514 (6):Ncfp-n]	[NP-512 [NNS loans]]	[NP-514 [Ncfp-n împrumuturi
15	s3	[ADVP-507 (2):RB]	[ADVP-515 (2):Rgp]	[ADVP-507 [RB also]]	[ADVP-515 [Rgp de_asemene
16	s3	[NP-521 (8):DT/the (8):NNS]	[NP-516 (8):Ncfpry]	[NP-521 [[DT/the] [NNS conditions]]]	[NP-516 [Ncfpry condiţiile]]
17	s3	[NP-533 [11]:DT/these [12]:NNS]	[NP-517 [11]:Afpmsry (12):Ncms-n]	[NP-533 [[DT/these] [NNS statutes]]]	[NP-517 [Afpmsry prezentul]
18	s3	[VP-535 (9):VBN PRT (518):PP]	[VP-519 (9):Vmp--pf (518):NP]	[VP-535 [[VBN laid] [PRT-526 [RP/down]] [PP-534 [IN	[VP-519 [Vmp--pf stabilite] [
19	s3	[PP-534 [10]:IN/in (517):NP]	[NP-518 [10]:Spsa/în (517):NP]	[PP-534 [[IN/in] [NP-533 [DT/these] [NNS statutes]]]	[NP-518 [Spsa/în] [NP-517 [A
20	s3	[VP-541 (0):MD/may (515):ADVP (523):VP]	[VP-524 (0):Rgp (515):ADVP (523):VP]	[VP-541 [[MD/may] [ADVP-507 [RB also]] [VP-540 [VI	[VP-524 [Rgp Poate] [ADVP-5
21	s3	[VP-540 (4,5):VB (522):S]	[VP-523 (4):Qs/să (5):Vmsp3 (522):NP]	[VP-540 [VB raise] [S-539 [NP-512 [NNS loans]] [ADV	[VP-523 [Qs/să] [Vmsp3 facă
22	s3	[PP-537 (7):TO/To (520):NP]	[NP-521 (7):Spsa/în (520):NP]	[PP-537 [[TO/to] [NP-536 [NP-521 [DT/the] [NNS con	[NP-521 [Spsa/în] [NP-520 [N
23	s3	[NP-536 (516):NP (519):VP]	[NP-520 (516):NP (519):VP]	[NP-536 [[NP-521 [DT/the] [NNS conditions]] [VP-53	[NP-520 [NP-516 [Ncfpry con
24	s3	[ADJP-538 [J] (521):PP]	[NP-521 (7):Spsa/în (520):NP]	[ADJP-538 [[J] subject] [PP-537 [TO/to] [NP-536 [NP-	[NP-521 [Spsa/în] [NP-520 [N
25	s4	[NP-509 (2):DT/the (2):NN]	[NP-511 (2):Ncfsry]	[NP-509 [[DT/the] [NN agency]]]	[NP-511 [Ncfsry Agenţia]]

Слика 1. База података са паралелним енглеско-румунским обрасцима

Овај тип извора се може користити у приступу машинском превођењу између енглеског и румунског заснованом на правилима и обрасцима. Двосмисленост представља један од значајних отежавајућих фактора при извршавању задатка синтаксичке трансформације. Стога трансформациони модел обично мора да анализира лексичке информације како из изворног, тако и из циљног језика.

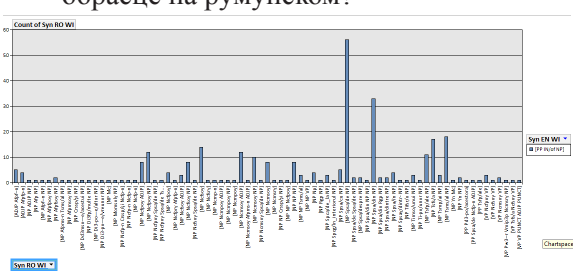
Паралелни синтаксички обрасци могу да се користе за обуку машинског превођења заснованог на правилима, док хијерархијски обрасци могу бити употребљени за обуку статистичког машинског превођења, у оквиру кога се може узимати у обзир, не само хијерархијска структура сваке паралелне енглеско-румунске секвенце, већ и њихови опсези, односно, лексичке информације.

3. Лингвистичка анализа двојезичних образаца

Хијерархијски обрасци су значајни у сваком систему за машинско превођење због информација које могу да пруже о томе како фразе мењају место и како одређене речи (на пример, функцијске речи) из изворног језика утичу на фразне структуре током превођења.

Такође, развијена енглеско-румунска база података са синтаксичким обрасцима се може користити за добијање различитих статистичких података, као што су:

- *који су могући обрасци на румунском за дати синтаксички образац на енглеском?* (видети **Слику 2** и **3**)
- *који обрасци из енглеског имају највећи могући број преводних образаца на румунском?* Прецизније, овде се поставља питање утврђивања које структуре из енглеског могу да представљају велики проблем током превођења на циљни језик
- *који обрасци из енглеског одговарају јединственим обрасцима у румунском?* Другим речима, проблем се може дефинисати на следећи начин: који обрасци из енглеског имају јединствене преводне обрасце на румунском?

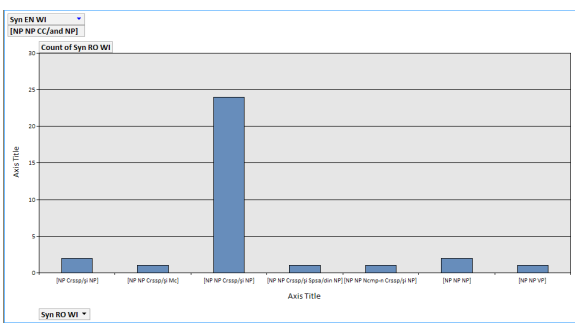


Слика 2. Румунски синтаксички обрасци за образац [PP IN/of NP] у енглеском

Ако је потребна компаративна студија о синтаксичкој структури паралелних изворних и циљних секвенција да би се направила општа схема трансфера за систем за машинско

превођење, онда се могу користити презентације синтаксичких образаца без информација о индексу. У **Табели 1** су дати неки паралелни синтаксички обрасци на енглеском и румунском без информација о индексу.

Из података представљених у **Табели 1**, може се видети, на пример, да превод енглеског прилошког обрасца [ADVP IN/for] има један једини образац у коме је предлог „for" замењен румунским предлогом „pentru", док за енглеску предлошку фразу [PP IN/for NP] постоји више од 10 румунских образаца из банке стабала.



Слика 3. Румунски синтаксички обрасци који одговарају синтаксичком обрасцу [NP NP CC/and NP] у енглеском

Синтаксички обрасци дати у форми представљеној у **Табели 1** су корисни ако су потребни општи обрасци превођења за развој одређене фазе процеса превођења. Али, обично, су потребне детаљније информације, попут података о промени места или поравнавању токена/речи.

Табела 1. Енглеско-румунски паралелни синтаксички обрасци без информација о индексу

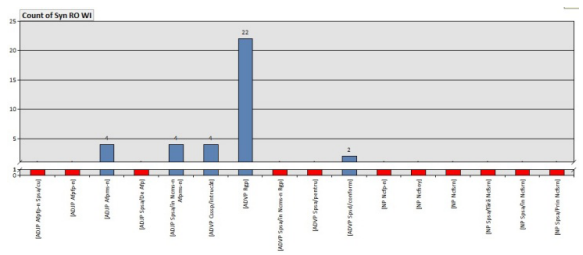
Енглески обрасци	**Румунски обрасци**
[ADJP JJ CC/and JJ]	[ADJP Afpms-n Crssp/şi Afpms-n]
[ADJP JJ PP]	[ADJP Afpfp-n NP]
[ADJP JJ]	[ADJP Afpfp-n], [ADJP Afpfsrn], [ADJP Pd3fsr/CEEA], [ADJP Rgp], [NP Ncmsry], [VP Vmis3s]
[ADVP IN/below]	[ADVP Spsa/de Rp Rgp]
[ADVP IN/by NP]	[NP Spca/în_conformitate_cu NP]
[ADVP IN/for]	[ADVP Spsa/pentru]
[CONJP CC/but RB]	[CONJP Ccssp/ci Crssp/şi], [CONJP Ccssp/dar Spsa/fără]
[NP ADJP NNS]	[NP ADJP Spsa/în Ncms-n], [NP Ncfp-n ADJP], …
[NP CD NN]	[NP Mc Ncmp-n], [NP Mc Yn], [NP Mc], [NP Tifsr/o Ncfsrn]
[NP DT/a JJ NN]	[NP Ncfsrn Afpfsrn], [NP Ncms-n Spsa/de Ncfsrn], [NP Tifso/unei Afpfson Ncfson], [NP Tifsr/o Di3fsr---e/altă Ncfsrn], [NP Tifsr/o Ncfsrn Afpfsrn], [NP Timso/unui Ncms-n Afpms-n], [NP Timsr/un Di3ms/anumit Ncms-n], [NP Timsr/un Ncms-n Afpms-n], [NP Timsr/un Ncms-n]
[NP NN NN]	[NP Ncfp-n Afpfp-n], [NP Ncfp-n Spsa/de Ncfsrn], [NP Ncfp-n], [NP Ncfsoy Afpfson], [NP Ncfsrn Spsa/de Ncfsrn], [NP Ncfsrn Spsa/în Ncmsry], [NP Ncms-n Afpms-n], [NP Ncms-n], [NP Ncmsry Ncms-n]
[PP CC/either PP CC/or PP]	[NP Ccssp/fie NP Ccssp/fie NP], [NP Ccssp/fie NP NP]
[PP IN/as NP]	[PP Cscsp/precum_şi NP], [NP Rc NP], [NP Rgp NP], [NP Spsa/pentru NP], [NP Spsd/conform NP],
[PP IN/for NP]	[ADJP Afpms-n], [NP Ncfp-n], [PP Spcg/în_vederea NP], [PP Spsa/cu NP], [PP Spsa/de NP], [PP Spsa/din NP], [PP Spsa/în NP], [PP Spsa/la NP], [PP Spsa/pe NP], [PP Spsa/pentru NP], [PP Spsa/spre NP], [PP Spsg/contra NP], [PP Spsa/de ADJP]
[VP MD/shall VP]	[VP Px3--a--------w/se VP], [VP Vaip3p VP], [VP Vmip3p NP]

У последњем сценарију, репрезентације са информацијама о индексу су примереније. Додајући те информације, при превођењу се могу искористити лингвистичке везе између речи и фраза у изворном и циљном језику. Примери паралелних синтаксичких образаца са информацијама о индексу су дати у **Табели 2**.

Табела 2. Пример паралелних синтаксичких образаца са информацијама о индексу који одговара структури [NP DT/a JJ NN] у енглеском језику

Енглески образац	Румунски образац
[NP-515 {5}:DT/a {6}:JJ {5}:NN]	[NP-508 {5}:Ncms-n {6}:Ncms-n]
[NP-524 {9}:DT/a {11}:JJ {10}:NN]	[NP-553 {9}:Timso/ unui {10}:Ncms-n {11}:Afpms-n]
[NP-522 {7}:DT/a {8}:JJ {9}:NN]	[NP-513 {7}:Tifso/ unei {8}:Afpfson {9}:Ncfson]
[NP-549 {12}:DT/a JJ {13}:NN]	[NP-517 {12}:Timsr/ un {13}:Ncms-n]
[NP-586 {29}:DT/a {31}:JJ {30}:NN]	[NP-543 {29}:Timsr/ un {30}:Ncms-n {31}:Afpms-n]
[NP-530 {14}:DT/a {16}:JJ {15}:NN]	[NP-525 {14}:Tifsr/o {15}:Ncfsrn {16}:Afpfsrn]
[NP-555 {19}:DT/a {20}:JJ {21}:NN]	[NP-534 {19}:Timsr/ un {20}:Di3ms/ anumit {21}:Ncms-n]
[NP-513 {3}:DT/a {4}:JJ {5}:NN]	[NP-552 {3}:Tifsr/o {4}:Mofsrln {5}:Ncfsrn]
[NP-591 DT/a {43}:JJ {41}:NN]	[NP-553 {41}:Ncms-n Spsa/de {43}:Ncfsrn]
[NP-603 DT/a {31}:JJ {30}:NN]	[NP-555 {30}:Ncms-n {31}:Afpms-n]

Један начин да се издвоје алтернативне румунске структуре за дати образац на енглеском језику за који постоји више могућих превода је да се елиминишу структуре које су идентификоване у корпусу један једини пут и да се у обзир узимају само оне структуре које се у корпусу појављују више пута (видети **Слику 4**). На тај начин, елиминишемо „шум“ у бази података који настаје због неправилног поравнавања на нивоу речи.



Слика 4. “Елиминација шума” у румунским обрасцима који одговарају обрасцу [ADVP RB] у енглеском језику

4. Закључци

У овом раду се описује рад на стварању збирке богато анотираних синтаксичких секвенца поравнатих на више нивоа. У предложеном формализму, циљни језик је флективан и директна последица тога је чињеница да се посебна пажња посвећује морфолошким информацијама које се могу извести из обрађене циљне конструкције из коришћеног корпуса. Намена такве збирке је да буде коришћена у систему за машинско превођење који се развија са циљем да се пређе са речи на фразе као основну јединицу превођења.

Да би се направио систем за машинско превођење, потребно је обезбедити паралелне синтаксичке и лексичке секвенције, како на изворном, тако и на циљном језику ради обуке система. Што има више секвенција и што су прецизније, резултат превода ће бити бољи.

Намеравамо да овај ресурс употребимо за развој система за машинско превођење заснованог на синтакси. Тренутни резултати наговештавају да су, са оваквим подацима, поравнавања са високим нивоом одзива пожељнија од поравнавања са високим нивоом прецизности, јер дају боље резултате у

машинском превођењу. Из тог разлога, обрасци генерисани у овој студији, прецизније, румунски обрасци, морају бити генерализовани да би обухватили више примера варијација морфосинтаксичких обележја.

5. Захвалност

Ово истраживање ауторке, М. Колхон је финансирано путем стратешке донације бесповратних средстава POSDRU/89/1.5/S/61968, Project ID 61986 (2009), коју је суфинансирао Европски социјални фонд у оквиру Секторског оперативног програма развоја људских ресурса за период 2007-2013.

Ауторка жели да изрази захвалност Групи за обраду природних језика Факултета за рачунарство Универзитета Александру Joan Куза у Јашију што је обезбедила енглеско-румунски корпус на основу кога је урађена ова студија. Посебну захвалност ауторка изражава проф. др Дану Кристеи за сталну подршку и савете.

Литература

Ambati, Vamshi, Alon Lavie, and Jaime Carbonell. 2009. Extraction of Syntactic Translation Models from Parallel Data using Syntax from Source and Target Languages. In *MT Summit XII Proceedings of the twelfth Machine Translation*.

Araùjo, Josué G. and Helena M. Caseli. 2010. Alignment of Portuguese-English syntactic trees using part-of-speech filters. In *Workshop in Natural Language Processing and web-based technologies*, 31-40.

Colhon, Mihaela and Dan Cristea. 2012. Automatic extraction of syntactic patterns for dependency parsing in NP chunks. In *International Simpozion "The Syntax and Semantics of Specificity"*, Faculty of Foreign Languages and Literatures, University of Bucharest, Romania. (accepted).

Colhon, Mihaela. 2011. A Contrastive Study of Syntactic Constituents in English and Romanian Texts. In *Proceedings of the Workshop "Language Resources and Tools with Industrial Applications"*, eds. Iftene A., Trandabăţ D.-M., 11-20.

Colhon, Mihaela. 2012. Language Engineering for Syntactic Knowledge Transfer. *Computer Science and Information Systems Journal*, 9(3):1231-1247, ISSN 1820-0214

Cristea, Dan and Corina Forăscu. 2006. Linguistic Resources and Technologies for Romanian Language. *Computer Science Journal of Moldova*, 14(1).

Erjavec, Tomaž, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić and Duško Vitas. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages, In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, 25-32, Budapest, Hungary.

Erjavec, Tomaž. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC'10*), European Language Resources Association (ELRA), ISBN 2-9517408-6-7.

Ide, Nancy, Patrice Bonhomme and Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*. Paris: European Language Resources Association (ELRA), 2000.

Iftene, Adrian, Diana Trandabăţ, Alex-Mihai Moruz and Maria Husarciuc. 2010. Question Answering on Romanian, English and French Languages. In *Notebook Paper for the CLEF 2010 LABs Workshop*, Padua, Italy, ISBN 978-88-904810-0-0, ISSN 2038-496322-23.

Galley, Michel, Mark Hopkins, Kevin Knight and Daniel Marcu. 2004. What's in a Translation Rule? In *Proceedings of HLT-NAACL 2004*, 273-

280, Boston: Association for Computational Linguistics.

Gispert, Adrià de, Juan Pino and William Byrne. 2010. Hierarchical Phrase-based Translation Grammars Extracted from Alignment Posterior Probabilities. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 545-554.

Klein, Dan and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 423-430.

Kurc, Roman, Maciej Piasecki and Stan Szpakowicz. 2010. Corpus-based extraction of morpho-syntactic patterns for the automatic acquisition of hypernymy. *Intelligent Information Systems*, 77-90.

Samuelsson, Yvonne, and Martin Volk. 1993. Alignment Tools for Parallel Treebanks. *ACM Trans. Program. Lang. Syst*. 15(5):795-825.

DOI=http://doi.acm.org/10.1145/161468.16147.

Simionescu, Radu. 2012. Romanian deep noun phrase chunking using graphical grammar studio. In *Proceedings of the 8th International Conference "Linguistic Resources And Tools For Processing Of The Romanian Language"*, eds. M. A. Moruz, D. Cristea, D. Tufiş, A. Iftene, H. N. Teodorescu, 135-143.

Taylor, Ann. 1996. Bracketing Switchboard: An Addendum to the Treebank II Guidelines, http://www.seas.upenn.edu/~jmott/prsguid2.pdf.

Tufiş, Dan and Radu Ion. 2007. Parallel Corpora, Alignment Technologies and Further Prospects in Multilingual Resources and Technology Infrastructure. In *Proceedings of SPED 2007*.

Wenniger, Gideon Maillette de Buy, Maxim Khalilov and Khalil Sima'an. 2010. A Toolkit for Visualizing the Coherence of Tree-based Reordering with Word-Alignments. In *Proceedings of the Open Source Convention at the Fifth Machine Translation Marathon* (*MT-Marathon*), Le Mans (France), 97-104.

Yamada, Kenji and Kevin Knight. 2001. A Syntax-based Statistical Translation Model. In *Proceeding of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, 523-530.

Апендикс А — глосар нотације

У табели која следи је дата нотација коришћена у овом раду:

Табела 3. Значење морфосинтаксичке нотације према лексичким спецификацијама MULTEXT-East

Етикета MSD	**Значење нотације**
Afpfp-n	Придев квалификатор позитив женски множина –одређеност
Afpfson	Придев квалификатор позитив женски једнина коси –одређеност
Afpfsrn	Придев квалификатор позитив женски једнина директан –одређеност
Afpms-n	Придев квалификатор позитив мушки једнина –одређеност
Crssp	Везник *r* прост прост позитив
Ccssp	Везник напоредни прост прост позитив
Di3fsr	Детерминатор, неодређен треће женски једнина директан
Di3ms	Детерминатор, неодређен треће мушки једнина
Mc	Број главни
Mofsrln	Број редни женски једнина директан словни -одређеност
Ncfsoy	Именица заједничка женски једнина коси +одређеност
Ncfson	Именица заједничка женски једнина коси -одређеност
Ncfsrn	Именица заједничка женски једнина прав –одређеност

Ncfsry	Именица заједничка женски једнина прав +одређеност
Ncfp-n	Именица заједничка женски множина -одређеност
Ncmp-n	Именица заједничка мушки множина -одређеност
Ncmsoy	Именица заједничка мушки једнина коси +одређеност
Ncmsry	Именица заједничка мушки једнина прав +одређеност
Ncms-n	Именица заједничка мушки једнина -одређеност
Pd3fsr	Заменица показна треће женски једнина прав
Pp3fso-	Заменица лична треће женски једнина коси
Px3--a---------w	Заменица повратна треће акузатив слаб
Rc	Прилог *c*
Rgp	Прилог општи позитив
Rp	Прилог речца
Spca	Адпозиција предлог сложен акузатив
Spcg	Адпозиција предлог сложен генитив
Spsa	Адпозиција предлог прост акузатив
Spsd	Адпозиција предлог прост датив
Tifso	Члан неодређен женски једнина коси
Tifsr	Члан неодређен женски једнина прав
Timso	Члан неодређен мушки једнина коси
Timsr	Члан неодређен мушки једнина прав
Tsms	Члан присвојни мушки једнина
Vaip3p	Глагол помоћни индикатив презент треће множина
Vmip3p	Глагол главни индикатив презент треће множина
Vmis3s	Глагол главни индикатив прошло треће једнина
Yn	Скраћеница именичка

Табела 4. Значење нотација из банке стабала Пен

Етикета из банке стабала Пен	Значење нотација
ADJP	Придевска фраза
ADVP	Прилошка фраза
CONJP	Везничка фраза
CC	Напоредни везник
CD	Главни број
DT	Детерминатор
IN	Предлог или зависни везник
JJ	Придев
NN/NNS	Именица једнина или градивна / именица множина
MD	Модал
NP	Именичка фраза
PP	Предлошка фраза
RB	Прилог
VP	Глаголска фраза