Български език, Приложение / Bulgarian Language, Supplement, 66 (2019), 93–99
Print ISSN: 0005-4283, Online ISSN: 2603-3372

doi: 10.7546/BL.LXVI.19.PR.06

ЯВНИ И НЕЯВНИ ВРЪЗКИ ПРИ МОДЕЛИРАНЕ НА СИНТАКТИЧНИТЕ ОТНОШЕНИЯ

ПЕТЯ ОСЕНОВА
СОФИЙСКИ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
petyaosenova@slav.uni-sofia.bg

Резюме: Статията дискутира преноса на лингвистична информация от конституентен към дependентен тип представяне. Разсъжденията се опират на синтактичния ресурс Бултрибанк, който е анотиран спрямо конституентната теория Опорна фразова граматика, и на същия ресурс, но анализиран в рамките на инициативата за Универсалните зависимости. Разгледани са случаите на експлициране на синтактични връзки в дependентното представяне, които са били неявни в конституентното, като се е разчитало на лесната им изводимост. Преносът на информация е представен чрез явленията, включени в т.нар. разширени зависимости. Направен е извод, че преминаването от един вид представяне към друг помага за откриване на проблемните моменти в лингвистичните теории и в описанието на определени езикови явления.

***Ключови думи*:** конституентност, дependентност, моделиране, разширени зависимости, български език

EXPLICIT AND IMPLICIT LINKS IN THE MODELING OF SYNTACTIC RELATIONS

PETYA OSENOVA
SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
petyaosenova@slav.uni-sofia.bg

Abstract: The paper discusses the transfer of linguistic information from a constituent type to a dependency type of representation. The survey explores the syntactic resource Bultreebank which has been annotated in accordance with the HPSG constituent theory and the same resource analyzed within the Universal Dependencies framework. The author considers cases where the syntactic relations have been made explicit in the dependency representation but have remained implicit in the constituent representation since they could be inferred easily from the annotation. The transfer has been introduced through the so-called enhanced dependencies. A conclusion is made that the transition from one representation to another supports the discovery of problem areas across linguistic theories and in the description of certain linguistic phenomena.

***Keywords*:** constituency, dependency, modeling, enhanced dependencies, Bulgarian

Както е известно, синтактичните отношения могат да бъдат моделирани, най-общо казано, по два основни начина – чрез конституентен анализ, чиято основна единица на изразяване е фразата, и чрез дependентен анализ, чиято основна единица на изразяване е думата. Тук е мястото да отбележа, че има редица лингвистични изследвания, които въвеждат междинни равнища на анализ. Такова предложение е например понятието *катена* (Osborne et al. 2012). Смята се, че катената не е конституент, а е набор от думи, който описва равнище, по-разширено от думата, но невинаги съвпадащо с конституент. Например съчетанието *идвам с* е катена, но не е конституент. Катената отразява по-скоро близост с дependентните теории, отколкото с конституентните. По този начин може да се опишат редица явления, проблемни за всяка една лингвистична теория, като например съчинителното свързване, идиоматиката, елипсите и т.н. Разбира се, трябва да спомена и изследванията, посветени на т.нар. многокомпонентни думи, които също размиват границите между речник и граматика, показвайки сложната природа на естествения език.

От друга страна, интересен е въпросът дали информацията, представена чрез конституентен анализ, се запазва, или частично се изгубва, когато се пренесе в дependентна рамка, и обратното. В областта на езиковите технологии ресурсите, които кодират синтактична информация, са известни под името *трибанки* (treebanks), т.е. банки от дървовидни структури.

Целта на този текст е да дискутира два основни проблема: а) пренасянето на информация от конституентен вид представяне към дependентен вид и б) равнищата на кодиране на информацията в дependентния вид представяне. Конкретните параметри са следните: преносът на информация се осъществява от първоначалния синтактичен ресурс Бултрибанк[1], анотиран спрямо идеите на Опорната фразова граматика, към дependентното му представяне спрямо идеите на Универсалните зависимости[2]. Трансферът на информацията от конституентното представяне беше осъществен на два етапа: към базисни дependетни отношения, а по-късно – и към т. нар. „разширени“ зависимости.

Синтактичният ресурс Бултрибанк[3] кодира основните фразови отношения, които са отразени в следните типове без отчитане на словоредния вариант: опора подлог (*Иван идва*), опора комплемент (*Чета вестник*) и опора адюнкт (*Вървя бързо*; *хубава жена*). Отбелязана е опората във всяка една фразова проекция. Допълнително са отбелязани субординационните и координационните отношения; елипсите (структурни и дискурсни); субстантивациите и номинализациите; собствените имена (на хора, локации и организации); кореферентните отношения (включващи явления като контрол и свързване). Трябва да спомена също така факта, че не всички проекции са разглеждани на фразово равнище. Така например *да*-фразите (*да дойда*) и клитичните комплекси (*казах му*, *да му кажа*) са анализирани като лексикални елементи, а не като фрази. При пренос от конституентна

към дependентна рамка се появяват различни проблеми: първо, как отношенията във фразите да се проектират в подходящи дependентни отношения. Това включва не само набора от релации, но и задължителното определяне на опора между думите, което налага „синтактизиране“ на отношенията и следователно изключва лексикалните проекции; второ, как да се пренесат отношения, които надграждат над синтактичните, като например кореференциите, или такива, които дават допълнителна информация за ролите на комплементите и адюнктите.

Трябва да се има предвид също така, че инициативата за Универсалните зависимости представлява особена рамка сред теориите на зависимостите. Тя се опитва да отрази типологическото разнообразие на езиците, като в същото време нейната цел е да осигури ресурси за трениране на съпоставими автоматични системи за синтактичен анализ на текстове на различни езици. Основните дependентни релации[4] включват: основни зависими части (аргументи), неосновни зависими части и зависими части при имената. Основните зависими части включват: именни групи с ролите на подлог, пряк и непряк обект; изречения с ролите на изреченски подлог, изреченски комплемент (със или без явлението контрол). Неосновните дependентни релации включват: именни групи с ролите на косвен dependент, вокатив, експлетив и дислоцирана релация; изречения с ролята на обстоятелствено подчинено изречение; модифициращи думи с ролите на адвербиал и дискурсна дума; функционални думи с ролите на спомагателна дума, копула и маркираща дума. Групата на зависимите части при имената включва: именни групи с ролята на номинален модификатор, апозиция, нумерален модификатор; изречения с ролята на изреченски модификатори; модификаторни думи с ролята на адективни модификатори и функционални думи с ролите на детерминатори, класификатори (само за определени езици) и падеж (в тази концепция името управлява предлога, който го въвежда). От групата на останалите релации (координация, многокомпонентни думи, специални релации и т.н.) ще бъде спомената само координацията.

Извън дефинирането на основните дependентни релации обаче бяха въведени и т.нар. разширени зависимости[5] (enhanced dependencies). Те имат за цел „да направят неявните връзки между пълнозначните думи явни чрез допълнително поставяне на релации и чрез увеличаване на наименованията на релациите“ (Schuster, Manning 2016). Тази стъпка се налага поради поне две причини: първо, защото Универсалните зависимости определят само пълнозначните думи като опори, т.е. изключват възможността функционална дума да бъде опора, а често разстоянието от една пълнозначна дума до друга е дълго и следователно е потенциално проблемно; второ, част от релациите, които се използват, имат по-широка функция и назовават повече на брой отношения. По този начин многозначността нараства, а степента на яснота намалява.

В момента списъкът с разширените зависимости е следният, като идеята е той да се разширява още: нулеви възли за елипсирани предикати; разпространяване на конюнктите; допълнителни релации при подлога с оглед на конструкциите с контрол; аргументи при пасивни конструкции (и други конструкции, при които се променя валентността); кореференция при релативните конструкции; етикети на модификатори, които съдържат предлог или друга информация, променяща падежа.

Ще дискутирам всяко от представените по-горе явления, като указвам интерпретацията му в първоначалния синтактичен ресурс Булгрибанк и кодирането му спрямо идеите на Универсалните зависимости.

Нулеви възли за елипсирани предикати

В Булгрибанк подобни елипсирани предикати са представени като: V-Elip (262 появи) или VD-Elip (255 появи). V-Elip представя по-директната връзка на липсващ глагол или глаголна форма, докато VD-Elip разглежда случаите на елипсирана вербална фраза (VP-ellipsis) и на елипси на копулата. Освен това маркерът VD-Elip е въведен и за елипси на дискурсно равнище, когато е трудно да се идентифицира обхватът на елипсата, камо ли нейната форма. Случаите на дискурсна елипса се обработват само ръчно. Ето пример за V-Elip: *Дадоха сладко, после **V-Elip** [дадоха] кафе*. В подобни случаи в Универсалните зависимости се въвежда нулев възел, т.е. възстановява се елипсираната част и вече възстановена, тя изпълнява ролята си на предикат. Това означава, че отбелязаните позиции в конституентния ресурс директно се пренасят в дипендентния. Намеса трябва да има при маркера VD-Elip, тъй като, както беше казано вече, той отбелязва повече типове елипсирани елементи.

Разпространяване на конюнктите в конструкции със съчинително свързване

При преноса на информация се разчита на имплицитен, но директно изводим анализ. В оригиналния ресурс съчинителните конструкции се смятат за безопорни и следователно – равни структури. Така например във фразата:

[CoordP [ConjArg *уморена*] [Conj *и*] [ConjArg *равна*] *скръб*]

модификаторната релация **amod** може да се установи чрез наличната морфосинтактична и лексикална информация, която идва от елементите на съчинителното свързване. Но тези елементи не са видими на синтактичното равнище. При дипендентното представяне отношенията при съчинителното свързване се йерархизират и увеличават. По отношение на горния пример отношенията са следните: при основните дипендентни отношения вторият конюнкт (*равна*) зависи от първия (*уморена*); съчинителният съюз (*и*) зависи от втория конюнкт (*равна*); първият конюнкт (*уморена*) показва зависимостта на цялата съчинителна конструкция от опората

съществително (*скръб*). В допълнение на това идва разширената зависимост, при която и вторият конюнкт на координацията (*равна*) е дependент на съществителното опора (*скръб*). По този начин се установяват явни връзки между една опора (*скръб*) и два или повече елемента в съчинителна конструкция (*уморена* и *равна*), вместо зависимостта да се изразява само от първия конюнкт.

Допълнителни релации при подлога с оглед на конструкциите с контрол

В оригиналния синтактичен ресурс между неизразения подлог в подчиненото изречение (отбелязан с елемента *pro-ss*) и изразения подлог на главното изречение, когато те реферират към една същина, винаги има релация на тъждество. Например в изречението *Еньо продължаваше да* [pro-ss] *гледа втрещено*[6] изразеният подлог на глагола *продължаваше* се свързва с елемента *pro-ss*, който е при глагола *гледа*, за да представи ситуацията, при която двата подлога се отнасят към едно и също лице.

В дependентното представяне глаголът в главното изречение е свързан със своя дependент – изразен подлог чрез релацията *nsubj*. В разширения вариант на зависимостите обаче се добавя и друга релация *nsubj* – от подчинения глагол (*гледа*) към същия изразен подлог (*Еньо*). Така прехвърлянето от конституентен към дependентен вид предполага само леки структурни разместявания, без да се губи или добавя информация. Елементът *pro-ss* се замества с релацията *nsubj* и се премества на самия глагол от подчиненото изречение.

Аргументи при пасивни конструкции (и други конструкции, при които се променя валентността)

Тук става дума за маркиране на подлозите при страдателни предикативни конструкции като страдателни подлози. В оригиналния ресурс няма специално маркиране на тези аргументи. Например в изречението *Това място бе специално изолирано от общата гравитация* опорната дума на подложната група *място* трябва да се маркира като страдателен подлог. При някои случаи подобна информация може да се извлече автоматично. Така е при причастностраждателните конструкции, тъй като те имат добре отличима парадигма. Други страдателни конструкции не са тривиални, като например пасивите, които се образуват със *се*-спрежение. Известно е, че те въвеждат многозначност между пасив, интранзитивни и детранзитивирани глаголи.

Кореференция при релативните конструкции

Представянето в оригиналния ресурс не предлага явна информация за отношението между релатива и модифицирания от релатив елемент. Подчиненото релативно изречение се маркира специално като CLR. Проекцията на именната група, която то пояснява, е NPA (опора адюнкт). На-

пример [NPA [NP *Човекът*, [CLR *който закъсня*]]]. В дependентното представяне първоначално са въведени следните релации: релацията *acl*, при която опорното съществително (в случая *човекът*) управлява глагола в релативното изречение (в случая *закъсня*). По-късно в рамките на разширените зависимости са добавени още две релации: релацията *nsubj* между съществителното опора и глагола в подчиненото изречение, както и релацията *ref*, която е насочена от съществителното опора към релатива (*който*). Следователно от конституентния ресурс неявната връзка на референция между съществителното и релатива се експлицира в дependентния вид представяне. Подобно експлициране се осъществява между релацията *nsubj* и предиката в подчиненото изречение.

Етикети на модификатори, които съдържат предлог или друга информация, променяща падежа

Тъй като българският език е аналитичен, предложната връзка е основна. Например [NP *стая* [PP *за игра*]]. В оригиналния ресурс е приложен стандартният анализ, при който предлогът (*за*) е опора и взема именната фраза (*игра*) като свой комплемент, проектирайки предложна фраза (*за игра*). В дependентния анализ обаче именната група, въведена от предлога, е опората. В основните дependентни релации е въведена релацията *case*, която е насочена от опората (*игра*) към предлога (*за*). Като част от разширените зависимости е въведена релацията *nmod*, към която се копира самият предлог като низ. Срв. *nmod:за*. Следователно тази стъпка може да се направи автоматично.

В заключение може да се каже следното. Трансферът на лингвистична информация от един тип представяне към друг (както в случая от конституентен към дependентен) дава ценна информация за проблемните моменти в лингвистичните теории и модели, както и за степента на трудност при описанието на определени езикови явления. В зависимост от типа, особеностите и пълнотата на изходната анотационна схема при прехвърлянето на анализи може да се наложи експлицирането на голяма част от неявните връзки или пък структурното разместване на релациите.

БЕЛЕЖКИ / NOTES

[1] www.bultreebank.org

[2] https://universaldependencies.org

[3] По-подробно за лингвистичния подход виж у Осенова, Симов/Osenova, Simov 2007.

[4] За повече информация виж https://universaldependencies.org/u/dep/all.html

[5] За повече информация виж https://universaldependencies.org/u/overview/enhanced-syntax.html

[6] В трактовката на П. Осенова и К. Симов (Осенова, Симов/Osenova, Simov 2007) *да*-формите се разглеждат като изречения. Това обяснява начина на анализ на примера в текста. В традиционните граматики конструкцията *продължаваше да гледа* се анализира като съставно глаголно сказуемо.

ЛИТЕРАТУРА

Осенова, Симов 2007: *Осенова, П., К. Симов*. Формална граматика на българския език. София, ИПОИ.

Osborne et al. 2012: *Osborne, T., Michael Putnam, T. Groß*. Catenae: Introducing a novel unit of syntactic analysis. – Syntax, 15(4), pp. 354–396.

Schuster, Manning 2016: *Schuster, S., Ch. Manning*. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. – In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris, France, ELRA.

REFERENCES

Osborne et al. 2012: *Osborne, T., Michael Putnam, T. Groß*. Catenae: Introducing a novel unit of syntactic analysis. – Syntax, 15(4), pp. 354–396.

Osenova, Simov 2007: *Osenova, P., K. Simov*. Formalna gramatika na balgarskiya ezik. Sofia, IPOI.

Schuster, Manning 2016: *Schuster, S., Ch. Manning*. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. – In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris, France, ELRA.

✉ *Проф. д-р Петя Осенова*
Катедра по български език, Факултет по славянски филологии
Софийски университет „Св. Климент Охридски“
Бул. „Цар Освободител“ 15, 1504 София, България

✉ *Prof. Petya Osenova, PhD*
Department of Bulgarian Language, Faculty of Slavic Studies
Sofia University “St. Kliment Ohridski”
15 Tzar Osvoboditel Blv., 1504 Sofia, Bulgaria

Публикувано: 3 октомври 2019 Published: 3 October 2019