КОМПЮТЪРНА ОБРАБОТКА НА ЕСТЕСТВЕНИ ЕЗИЦИ – АСПЕКТИ НА АВТОМАТИЧЕН СИНТАКТИЧЕН АНАЛИЗАТОР

Румен Рикевски

1. Съвременни тенденции в автоматичния синтактичен анализ

Обработката на естествен език е подобласт на изкуствения интелект и лингвистиката. Тя изучава задачите, свързани с автоматичното генериране и разбиране на естествени човешки езици. Системите за генериране на естествен език преобразуват информация от компютърни бази от данни в обикновен човешки език, а системите за разбиране на естествен език преобразуват текстове на човешки език във формално представяне, по-лесно за обработка от компютърни програми.

Синтактичният анализ е един от най-значимите стадии при обработката на естествен език и има водеща роля при разработването на множество съвременни приложения. В наши дни изследванията в тази област се прехвърлят от простото изречение към сложното, често пъти дори и към параграфа. Въпреки значителните постижения, които се отбелязват през последните години, изследователите продължават да срещат редица трудности, основните от които са:

– Как да бъде постигната синтактична обработка на изреченията по възможно най-гъвкавия, благонадежден и резултатен начин в езици с почти свободен словоред,

– До каква степен познанието за стилистичните категории на текстовете може да подпомогне при синтактичната им обработка,

– На какво се дължат синтактичните особености, които се наблюдават при езиците с почти свободен словоред и по какъв начин може да се интерпретира компютърно, разнообразието от подредби на думите в такива езици,

– Кой са най-елегантните и достъпни компютърни алгоритми и методи, чрез които да се разрешат успешно по-горните проблеми.

2. Синтактична обработка на езици с почти свободен словоред.

Сложните изречения винаги създават проблеми, когато се определя вида на простите изречения, които ги съставят, както и когато се определят главните и второстепенни части на речта, които съставляват простите изречения. Нещата стават още по-трудни, когато човек се занимава с обработка на текстове, където сложните изречения се свързват помежду си чрез съюзи и препинателни знаци, образувайки периоди.

От друга страна, почти свободният словоред на едно изречение подсказва, че думите нямат твърдо установена позиция и по този начин се затруднява до голяма степен, създаването на една завършена съвкупност от правила, които са способни да опишат такива конструкции. Езиците с почти свободен словоред се характеризират със относително стабилна позиция на думите вътре в главните и във второстепенните части на речта и със свободна подредба на частите на речта една спрямо друга. Тези езици имат много сходства с езиците с напълно свободен словоред.

През последните години в областта на обработката на естествен език се появиха доста граматически формализми, които имат претенциите, че могат да бъдат резултатно използвани при синтактичен анализ на изречения в един голям спектър от езици. Най-известните от тях са: Core Language Engine, Functional Unification Grammar, Generalized Phrase Structure Grammar, Lexical Functional Grammar, Categorical Unification Grammar, Head-driven Phrase Structure Grammar и Constraint Logic Grammar. Ако направим един задълбочен анализ на лингвистичните им свойства по отношение на обвързаност към конкретна лингвистична теория, възможност за анализиране на периоди и универсалност при работа с различни езици, стигаме до извода, че никой от тях не може в пълна степен да отговори и на трите изисквания.

Традиционно повечето от тези формализми се използват с успех при езици, които имат строго определена позиция на частите на речта. Когато обаче човек се занимава с езици с почти свободен словоред и трябва да се анализират периоди, формализмите често пъти се явяват безрезултатни, понеже броят на правилата за анализ на периоди става твърде голям. Конкретно за новогръцкия език, който принадлежи към тези езици е преценено, че създаването на една граматика, която би могла да анализира новогръцкия синтаксис и базирана на някой от изброените

формализми би изисквала огромни човеко-ресурси за определянето на правилата за синтактичен анализ, които биха възлезли на хиляди.

3. Подредба на частите на речта в новогръцкия език.

Множество съвременни трудове, които се отнасят до синтактичната типология приемат, че принципно езиците имат някаква основна, синтактично определена подредба на частите на речта. Факт е, че понякога тази подредба може да бъде променена поради причини от прагматичен характер, но основната подредба се счита за главна характеристика на езика, чрез която другите му аспекти могат да бъдат анализирани.

От изследвания, свързани с голям брой гръцки текстове, е констатирано, че подредбата на частите на речта в новогръцкия език е подчинена на прагматични критерии. Това означава, че в новогръцкия език не съществува някаква стабилна, основна подредба на частите на речта, на която всички останали подредби да представляват варианти. Това, което може да се каже е, че най-общо съществува една обичайна подредба на частите на речта, която зависи от вида на изреченията, които се използват, но тя обаче може във всеки един момент да бъде променена заради причини от експресивен характер или заради особения функционален стил на един текст. Също така, новогръцкият език оправдава характеристиката си на език с почти свободен словоред на думите, понеже има относително стабилна подредба на думите в главните и второстепенни части на речта и свободна подредба на тези части помежду им.

От друга страна, в новогръцкия език падежът на съществителното се декларира от едно променливо окончание и по този начин изразява функционалността си в изречението. Както е известно, в новогръцкия език има четири падежа, които се различават помежду си по своето окончание. Правилата и окончанията на падежите правят смисъла на едно изречение в новогръцкия език почти независим от словореда: за пример функционалността на едно съществително зависи от склонението му, а не от позицията му в изречението.

Съгласно едно изследване на голям брой гръцки текстове, почти 80% от изреченията са сложни. Сложните изречения, от своя страна се състоят от две или повече прости изречения, които се свързват помежду си със съчинителна (паратаксис) или подчинителна (хипотаксис) връзка. Простите изречения се състоят от главни и второстепенни части на речта.

Към главните части спадат групата на подлога и групата на сказуемото, а към второстепенните – определението, приложението, сказуемното определение, допълнението и обстоятелственото пояснение.

По правило, определянето на вида на изречението се постига след като най-напред се направи подробен синтактичен и понякога семантичен анализ. Иначе казано, видът на изречението се определя веднага след като всички части на речта са синтактично разпознати и често пъти семантично установени. В някои случаи, както при текстове от резюмета, се постига определянето на главните изречения от един предварително обработен текст. Тези случаи са базирани на установяването на семантични, фразеологични и контекстуални показатели, което от своя страна предполага някакъв морфологичен и семантичен анализ на тези текстове. По този начин можем да заключим, че в досегашните подходи определянето на вида на изреченията не е бил опитван да бъде постигнат без взаимното използване на синтактични и семантични познания.

Друг проблем, който се среща при определянето на вида на изречението в един текст е липсата на установен писмен стил (written style). Това означава, че писменият стил е доста субективен и не винаги се подчинява на някакви общоприети правила (т.е. кога точно да се използват запетаи, горни точки, тирета и т.н.). Лошото прилагане, желано или не, на тези правила има като резултат един текст, в който препинателните знаци не са правилно поставени и това затруднява неговия анализ.

От друга страна, както споменахме и по-рано, анализът на изречения в езиците с почти свободен словоред е една доста трудна дейност, понеже подредбата на частите на речта обикновено варира според специалната тежест, която авторът добавя на думите в изречението. Следва един пример, за да бъде показано това:

Ο Πέτρος έδωσε το παγωτό στην Άννα το πρωί.

По-горното изречение се състои от девет думи. В този случай съществуват 362880 (9!) различни варианти за подредба на тези думи, но само 120 (5!) от тях са граматически правилни. Ключовата цифра тук изглежда е 5, което показва броя на лексикалните съвкупности от които е съставено изречението. Така само изразите (*Ο Πέτρος*), (*έδωσε*), (*το παγωτό*), (*στην Άννα*), (*το πρωί*) могат да променят своята позиция в изречението. Всички останали подредби на думите биха довели до

граматически грешно съставено изречение. Нещо повече, ако някой иска да промени смисъла на това изречение, трябва да замени един или повече от по-горните лексикални изрази с други от същата, обаче синтактична категория.

Повечето съвременни граматически формализми обикновено се прилагат към езици със строго определена или полу-свободна подредба на частите на речта. В тези случаи граматическите правила, които се базират на ограничения (constraint grammar rules) се използват с успех при анализа на изречения в такива езици. Въпреки, че изреченията, които се анализират в болшинството си са съставни, не се вземат в предвид случаите на грешна употреба на препинателните знаци в един текст.

Нещо повече, прилагането на тези формализми към езиците с почти свободен словоред, доста често довежда до увеличаване броя на граматическите правила, които се изискват, за да бъдат покрити всички възможни подредби на частите на речта, особено когато обработката засяга текстове с голям обем. По-специално за новогръцкия език погорното прилагане води до един доста сложен процес на трансформиране от една „дълбока" синтактична структура (deep structure) към една „повърхностна" (surface structure). Даже и за случаите, когато се използват видоизменени, специално за новогръцкия език, граматики от йерархията на Чомски, то резултатът пак е едно вътрешно преобразуване от естествения език към един абстрактен език на синтактични категории, който обаче показва доста слабости (напр. изключват се елиптичните изречения, не се разглеждат отклонения от обичайния словоред, и.т.н.).

ЛИТЕРАТУРА

1. **Antworth, E. L.** PC-KIMMO: A Two-Level Processor for Morphological Analysis [Occasional Publications in Academic Computing 16], Summer Institute of Linguistics, Dallas TX, 1990.

2. **Allen, J. F.** Natural Language Understanding, The Benjamin/Cummings Publishing Company, 1987.

3. **Ralli and E. Galiotou, A.** A prototype for a computational analysis of Modern Greek compounds, Asymmetry Conference, Universitй de Quйbec, a Montrйal, May 2001.

4. **Mackridge, P.** The Modern Greek Language, Oxford University Press, 1985.

5. **Τριανταφυλλίδης, Μ.** Νεοελληνική Γραμματική, ΟΕΔΒ, 1986.

6. **Τριανταφυλλίδης, Μ.** Συντακτικό της Νέας Ελληνικής, ΟΕΔΒ, 1992.