ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 53, КН. 1, СБ. А, 2015 – ФИЛОЛОГИЯ,
PAISII HILENDARSKI UNIVERSITY OF PLOVDIV – BULGARIA
RESEARCH PAPERS, VOL. 53, BOOK 1, PART A, 2015 – LANGUAGES AND LITERATURE

КУРС ПО ИНФОРМАЦИОННИ ТЕХНОЛОГИИ ЗА ФИЛОЛОЗИ – КОНЦЕПЦИЯ И ИЛЮСТРАЦИИ

Руси Николов
Пловдивски университет „Паисий Хилендарски“

A COURSE IN NATURAL LANGUAGE PROCESSING FOR LINGUISTS – CONCEPTION AND ILLUSTRATIONS

Roussi Nikolov
Paisii Hilendarski University of Plovdiv

Beyond its usual applications, information technologies offer a powerful and complementary approach to the study – both scientifically and practically – of human languages. The goal of the course is two-fold: to familiarize students in linguistics with the main types of language resources and tools, but also to teach them the basics of the design of such tools. As an example, a computational method is presented here for evaluating the phonological distance between words, fairly consistent with the perceptual data, as well as its application in the computer program TREFL for automatic generation of didactically useful associations between words.

Key words: Computational Linguistics, Phonological Similarity, Foreign Language Learning

Изучаването на информационните технологии от филолози е целесъобразно дотолкова, доколкото компютърните приложения и количествените методи повишават производителността и разширяват възможностите при работа със или върху естествените езици. Да работим със или върху естествените езици, означава съответно да ги използваме или да ги изучаваме, а да ги изучаваме, може да означава да ги усвояваме, за да можем да ги използваме пълноценно или да ги изследваме с научни методи.

Методите и инструментите за автоматична обработка и анализ на естествените езици са важно допълнение, а не алтернатива на традиционните методи чрез нашите сетива, преценки и когнитивно мис-

лене. Един език е естествен, когато възниква и се развива спонтанно и функционира благодарение на естествените, вродени способности на човека (сетивни, моторни, интелектуални) за удовлетворяване на неговите комуникативни и експресивни потребности. Следователно релевантните (т.е. съществените) елементи и отношения в системата на един език се обуславят изцяло от функционалното съчетаване на тези способности в комуникативния процес. Показателен пример за това е фактът, че акустичната вълна като продукт на речевия апарат на говорещия и като материален носител на речево съобщение се „разчита“ с технически средства чрез извличане на несъизмеримо малка част от масата на акустичните данни, но със съществена информационна стойност от перцептивна гледна точка. Разпознаването на устната реч предполага и наличие на голям обем от референтни езикови данни. Като цяло развитието на езиковите технологии следва това на лингвистичните модели и данни, от една страна, и на изчислителните средства и методи, от друга страна; лингвистичните и компютърните науки са тясно преплетени и взаимно обусловени в интердисциплинарната област на автоматичната обработка на естествените езици.

Най-общите разграничения при работа със или върху естествените езици са следните:

1. Форма на езиковия материал – устна или писмена.

2. Материален носител на езиковия материал. За устната реч това са акустичната вълна (подлежаща единствено на еднократен и субективен слухов анализ), магнетофонният запис (подлежащ на неограничен слухов анализ, с възможност за приложение на статистически методи) и дигиталният запис (подлежащ допълнително на произволни контролирани манипулации, на обективен и автоматичен инструментален анализ, с количествени и визуални, абсолютни и сравнителни резултати). Езиковият материал се представя в писмена форма с помощта на символен код (фонетична, правописна или друга азбука); символният низ може да бъде изписан върху материален носител (хартия) или дигитализиран (с помощта на кодова таблица, дефинираща съответствие между символ и число).

3. Инструменти за обработка и анализ на езиковия материал. Нашите сетивни системи, и по-специално слуховата, зрителната и тактилната, са необходими и в много случаи дори достатъчни „инструменти“ при изследването на естествените езици; в други случаи използването на информационните технологии открива нови перспективи не само на приложно, но и на чисто научно, фундаментално равнище.

Динамичното развитие на информационните технологии и на компютърните устройства (лаптопи, таблети, смартфони…) към по-голяма мобилност, капацитет и изчислителна мощност превърнаха тези устройства в основен информационен носител, организатор и канал за данни на естествен език, изместващи съответно книгата на хартия, библиотеката и класическия телефон например. Необходимостта от бърз и евтин достъп до информацията предопределя позитивния характер и необратимостта на тези промени. Бърз и евтин достъп до информацията в дигитална форма не е синоним, за съжаление, на лесен достъп. Известен е парадоксът, че компютрите не правят живота ни по-лесен; и наистина, ако на тях гледаме като на интелектуална протеза, проблемът е, че тя е жива и се развива по собствени закони (при това невинаги в правилната посока), а за нас като потребители остава принудата непрекъснато да се адаптираме към нея. В този смисъл е наистина важно усилията ни за „адаптация“, т.е. за актуализиране и повишаване на нашите компетенции като потребители (и евентуално като създатели) на такива софтуерни инструменти, да са оправдани с оглед на приноса им за повишаване на нашата производителност, в случая при работа върху и/или чрез естествените езици.

Първата цел е постигане на практически умения за използване на компютърни системи за решаване на конкретни задачи; постигането на тези умения предполага усвояването по естествен начин, т.е. като осезаема необходимост, на задълбочени знания в някои области на информационните технологии (напр. бази от данни), на лингвистиката и на други дисциплини (напр. превод, фонетична акустика). В този смисъл това не е традиционен курс по компютърна лингвистика: във връзка с представянето на конкретни приложения на компютърната лингвистика се разглеждат теми, които са разнородни от теоретична гледна точка, но са свързани с това, че са елементи на концептуалната основа на представените софтуерни приложения за обработка на естествените езици.

Втората цел е да се представят на базата на личния опит на автора конкретни възможности за създаване със сравнително прости средства на полезни инструменти за събиране, организиране, консултация, анализ и интерпретация на езикова информация в устна или писмена форма. Програмирането се изучава традиционно в инженерните специалности; в действителност филолозите с оглед на развитието на информационните технологии, за което стана въпрос в самото

начало, имат не по-малко нужда от инженерите от такива умения. Тази необходимост има две измерения.

От една страна, решаването на всяка достатъчно сложна задача се свежда в резултат на нейния анализ до решаване на система от по-прости задачи. При съвременното състояние на дигитализация на информацията съществена е вероятността да се намери ефективно софтуерно решение за някои от елементарните задачи за обработка на символни низове. Проектите, които се разчленяват на елементарни задачи, не са задължително интердисциплинарни, а елементарните филологически задачи не се решават задължително от интердисциплинарни екипи. Съображенията за икономическа ефективност изискват в такива случаи филологът да притежава елементарни умения да програмира, които да му позволяват да създава т.нар. ad hoc програми, или с други думи – компютърни програми, които решават толкова специфични задачи, че се превръщат на практика в програми за еднократно ползване.

От друга страна, нови гледни точки към даден лингвистичен проблем, които се основават на математическия формализъм и на експерименталния характер на компютърната лингвистика, могат да се окажат продуктивни не само на приложно, но и на чисто научно, фундаментално равнище.

Тези съображения ще бъдат илюстрирани с една комплексна, интердисциплинарна задача с конкретна приложна цел: фонологично разстояние между словоформите в психолингвистичното пространство с приложение в чуждоезиковото обучение. Предложеното решение е цялостно и оригинално както на концептуално, така и на практическо равнище под формата на ергономично софтуерно приложение с отворен код.

1. Разстояние между символни низове (езикови единици)

Всяка конкретна езикова единица (като морфема, словоформа, синтагма, изречение, текст) може да бъде сравнявана с друга единица на същото езиково равнище по количествени критерии, определящи формалното разстояние между тях. Терминът „разстояние“ се използва в смисъл на количествена мярка за отдалечеността, т.е. за степента на различие (или обратно, за близостта, за степента на подобие) между езиковите единици в тяхната правописна или звукова форма. Такива сравнения имат важни приложения, например за обективна оценка на различията между:

- диалекти (диалектологията),
- текстове (плагиатство),
- изречения (превод с помощта на системи за преводаческа памет, при търсене на приблизителни съвпадения),
- думи в правописна форма (при интерактивна корекция на правописа чрез подсказване на близки по правопис думи спрямо грешно написаната),
- думи, изрази и т.н. в звукова форма (автоматично разпознаване на речта).

Интерсен е и случаят, когато езиковите единици са кодирани в писмена форма, но с помощта на фонетична абука. Такива кодиращи системи от фонетични символи и правила, каквато е например международната фонетична азбука, позволяват да правим значително полесно сравненията между произношението на езиковите единици на базата на тяхната фонологична форма, например при контрастивни изследвания на устни диалекти. В приложно отношение, и по-специално в областта на чуждоезиковото обучение, могат по-ефективно да се експлоатират формалните отношения между думите, като се използват не само правописни, но и произносителни асоциации между тях, особено при езиците с т.нар. нефонетичен правопис.

2. Алгоритми за определяне на разстоянието между символни низове

Предложеният тук нов алгоритъм за определяне на разстоянието между символни низове се основава в определена степен на психолингвистични критерии, за разлика от алгоритъма на Левенщайн (Левенщайн 1965: 845) или този на Хеминг (Хеминг 1950: 147). Изискването за определяне на разстоянието между словоформите в психолингвистичното пространство е продиктувано от поставената приложна цел. Възможността да асоцииираме една нова дума при изучаване на чужд език с друга, достатъчно близка по звучене и вече позната дума, ускорява създаването на траен акустичен образ на думата в системата на менталния лексикон.

2.1. Разстояние на Хеминг

Разстоянието на Хеминг между два низа с еднаква дължина е равно на броя на символните позиции, по които се различават двата низа. Тази величина има сама по себе си проста дефиниция и се изчислява лесно, но същевременно поражда не толкова прости въпроси в теорията на комуникацията и намира важно приложение в областта

на кодирането и надеждния пренос на информацията. Доколкото в езикознанието, както и в теорията и приложенията на автоматичната обработка на естествените езици се сравняват еднотипни езикови единици, а не произволни сегменти от речта с равна дължина, разстоянието на Хеминг не представлява интерес за компютърната лингвистика и нейните приложения. В предложения тук нов вид разстояние обаче разстоянието на Хеминг участва като елемент в един по-сложен алгоритъм, който има определени предимства спрямо популярния и широко прилаган алгоритъм на Левенщайн.

2.2. Разстояние на Левенщайн

Разстоянието на Левенщайн между два символни низа се определя от минималния брой на елементарните промени – вмъкване, изтриване и заменяне, – които позволяват да се редактира единият низ така, че да съвпадне с другия.

Кратки програмни кодове за изчисляване на разстоянието по Левенщайн между два произволни низа, написани на различни програмни езици, са свободно достъпни в интернет. Самият факт на тяхното публикуване показва, че те представляват определено интелектуално предизвикателство, особено за непрограмисти, които имат обаче амбицията да придобият умения сами да решават средно сложни задачи в тяхната професионална област с помощта на компютърното програмиране. Затова и тук е представен такъв код, написан на скриптовия език AHK и достъпен на придружаващия сайт <http://web.uni-plovdiv.bg/rousni>.

Разстоянието по Левенщайн е напълно подходящо, например при програмите за корекция на правописа, и по-специално за подсказване на най-близките по правопис думи до несъществуваща в речника дума. От гледна точка обаче на психолингвистичната оценка на формалната близост между езиковите единици (т.е. на менталната оценка на разликите във формата им) разстоянието по Левенщайн има два недостатъка, които се доказват със следните мислени експерименти с лесно предсказуеми резултати.

1. Разстоянието по Левенщайн между две различни по форма еднобуквени думи е единица (напр. предлога „в“ и съюза „и“). Същото е разстоянието и между два текста с дължината на роман, които се различават по една-единствена буква. С пълно основание обаче може да се каже, че между първите два символни низа няма нищо общо, докато вторите два са на практика идентични. За да преодолеем това противоречие, може да се въведе понятието относително разстояние по Левенщайн, като се раздели абсолютното разстояние *LD* (което не

може да бъде по-голямо от броя на символите за по-дългия низ) на броя на символите за по-дългия низ *Lmax*. По този начин ще получим величина, която варира от нула до едно, или от нула до сто процента. Ако трансформираме разстоянието на Левенщайн, като вместо израза *(LD / Lmax) x 100* използваме *((Lmax – LD) / Lmax) x 100*, ще получим в проценти стойност, която отразява степента на съвпадение. Аналогична степен на съвпадение се използва в системите за преводаческа памет за параметриране на функцията *търсене на приблизителни съвпадения* (fuzzy matching).

2. Ако добавим елемент към един обект на възприятие (зрителен или слухов), промяната ще е по-осезаема (при това до такава степен, че оригиналният обект да загуби евентуално своята идентичност), когато новият елемент се вмъкне, вместо да се прикачи към съществуващите елементи. Същото е вярно и при изтриване или замяна на един елемент с друг – като по-радикални се възприемат промените във вътрешността, отколкото по периферията. Вероятно с това е свързан фактът, че афиксите (наставки и представки) са основните елементи на деривативно словообразуване. Разстоянието на Левенщайн не отчита тази зависимост, която е съществена с оглед на поставената цел.

2.3. Обобщено разстояние на Хеминг

Този нов вид разстояние наричаме така, защото се основава на разстоянието на Хеминг, но е приложимо и за символни низове с различна дължина, което е необходимо условие, за да можем да го приложим като оценка на формалната близост между еднотипни речеви единици[1]. Освен това за разлика от разстоянието на Левенщайн то е стъпка в посока да се отчете чрез един формален алгоритъм относителната тежест на всяко елементарно различие по начина, по който тя се отчита на психолингвистично равнище.

Факторите на психолингвистично равнище, които влияят на оценката на близостта между две думи, са многобройни, а тяхното експериментално, обективно установяване и определянето на относителната им тежест е трудно. Рационалният научен подход изисква в подобен случай да се направят хипотези, които да се вложат в информационния модел, чрез който да се експериментира чрез методите на психолингвистиката и така да се направи глобална оценка на психолинг-

[1] За простота можем да използваме и думата „дума“, без да забравяме, че алгоритъмът е приложим за произволен символен низ и следователно за произволен речеви сегмент.

вистичната адекватност на модела. Разумна хипотеза е например следната: разстоянието между „кола“ и „кора“ е по-малко, отколкото между всяка една от тези две думи и думата „кога“. И двете съгласни „л“ и „р“, по които се различават първите две думи, са близки по място на учленяване и се характеризират с помощта на фонетико-акустичните етикети „сонорни“ и „плавни“, докато „г“ е задна съгласна, „шумова“ и „преградна“. Напълно логично е да приемем, че разстоянието между две думи е толкова по-голямо, колкото по-голямо е разстоянието между различаващите се елементи. Но може и да се възрази, че ако експериментираме с тези три думи и съответните три субективни разстояния, резултатът би могъл да опровергае очакванията ни. Един мъж бързо би асоциирал (и в резултат би оценил като близки) думите „кога“ и „кола“, мислейки си „*Кога* ще мога да си купя нова *кола*?“. Друг би асоциирал по-скоро „кола“ и „кора“, но не защото „л“ и „р“ са близки, а защото „Вчера той смени спуканата кора на своята кола“. Психолингвистиката разполага обаче с необходимите методи, за да елиминира факторите от семантично и прагматично естество.

Разстоянието между различаващите се елементи е следователно съществен фактор за оценката на глобалното разстояние между системите, съставени от тези елементи. Това е така, защото елементите от своя страна също могат да са съизмерими. В най-чист вид това се илюстрира при символни низове от цифри, например 66666. Ако това са семестриалните оценки на един студент, не е без значение дали ще заменим една шестица с петица или с двойка (така както не е без значение по колко фонологични признака се различават различаващите се фонеми при сравнение на две думи). Също така не е без значение (вероятно) дали с двойка ще бъде заменена третата или първата шестица например, защото двете оценки се отнасят за различни учебни дисциплини. Аналогично, за да оценим доколко се отдалечаваме от „оригиналната“ форма на една дума, не е без значение дали ще заменим буква/звук в средата, или в края на думата дори когато думите са с еднакви дължина и акцентна структура. Потвърждение на това предположение е статистическият факт, че разлика в средата на думите има по-голяма тежест в лексикалната система на езика, отколкото разлика в началото или края, която често служи за разграничаване на словоформите в една тясна граматическа или семантична парадигма.

Особено показателен за необходимостта да се отчита не само броят, но и тежестта на елементарните различия, е и следният пример. Лесно и бързо могат да се вмъкнат автоматично в един електронен

текст между всеки два съседни символа един или повече специални символи с широчина нула, т.е. невидими знаци, чиито позиции също са скрити. На екрана на компютъра или разпечатани на хартия, оригиналният и манипулираният текст ще бъдат напълно идентични. Въпреки това една автоматизирана система за проверка за плагиатство би могла да пропусне да установи дори стопроцентово преписване, защото различието между дигиталните форми на оригиналния и манипулирания текст може да се контролира по посочения начин в неограничени на практика рамки.

Да се отчита не само броят на елементарните различия, но и тяхната тежест с оглед на конкретното приложение, е целта на този алгоритъм. В голяма степен решението е приблизително и частично, но резултатите показват, че то вече е полезно от приложна гледна точка. На този етап се отчита единствено тежестта на елементарните различия, свързана с позицията на елементите, но не и с вътрешноприсъщите характеристики на тези елементи (правописен или фонологичен символ: фонема, ударение, сричкова граница). Именно затова беше подчертано, че става въпрос само за една стъпка към тази цел.

Обобщеното разстояние на Хеминг се дефинира чрез следната формула:

$$D = \frac{D_{\min} + (L_{\max} - L_{\min})}{L_{\max}},$$

където:

- $L_{\max}$ и $L_{\min}$ са съответно дължините на по-дългата и на по-късата от двете сравнявани думи;
- $D_{\min}$ е минималното разстояние по Хеминг за всички възможни сравнения между по-кратката дума и сегментите от по-дългата дума с дължина, равна на по-късата. Броят на тези сегменти е равен на $L_{\max} - L_{\min} + 1$.

Лесно се установява, че в частния случай, когато двете сравнявани думи са с еднаква дължина, получаваме точно разстоянието по Хеминг, но отнесено към дължината на по-дългата дума[2]. Така величината има граница на вариране от 0 до 1. Разстоянието е нула за думи, които са идентични по графичната си форма, и единица, когато

[2] В този смисъл избраният термин „обобщено разстояние на Хеминг“ не е напълно коректен.

D_{min} (минималното разстояние по Хеминг) е максимално възможното, т.е. равно на дължината на по-късата дума.

Можем да дадем едно рационално тълкуване на формулата, изреждайки я по следния начин. Мислено поставяме по-късата дума върху по-дългата и я плъзгаме, докато открием максимално съвпадение по Хеминг със съответния сегмент от по-дългата дума. (Същото бихме направили, ако търсим приликата например между една дълга и една къса пръчка от дърво.) Придържайки се отново към елементарната, но ефективна логика на Хеминг, добавяме празни позиции от двете страни на по-късата дума, така че двете думи да съвпаднат по дължина. Това ни позволява да направим отново сравнение по Хеминг и така стигаме до израза $D_{min} + (L_{max} - L_{min})$. Полученото разстояние разделяме на дължината L_{max}, за да получим относителната тежест на различията между двете думи. Колкото до позиционно обусловената относителна тежест на елементарните различия (свързана с това дали различието е в периферията, или по средата), анализът на конкретни примери може да покаже, че тя се отчита в процеса на „плъзгането“, т.е. на намирането на D_{min}.

На Фигура 1 по-долу е представена класация на френските думи по разстоянието им до думата *heureux*, изчислено по разгледаната формула и при избрано ограничение $D < 0.25$. В тази класация *heureuse* и *peureux* се намират преди думата *euro*. Подредбата е друга (и поточно невъзможна) при използване на разстоянието на Левенщайн, защото тогава всички тези думи се оказват на едно и също разстояние до *heureux*. Нека обаче направим аналогия между думата *heureux* и менталния образ на едно щастливо човешко лице. Ако към това лице прикачим нов елемент, например допълнително ухо, но без да скриваме нищо от лицето, то това ново лице ще остане лесно разпознаваемо. Не е така обаче, ако в лицето заменим например носа с ухо, така както получаваме думата *euro*, заменяйки /ø/ с /o/ в думата *heureux*.

От друга страна, в случая на сравнението между думите *heureux* и *euro* трябва да се отчетат още два фактора, които биха повлияли на нашата субективна преценка за разстоянието между тях. Гласните /ø/ и /o/ са орални, затворени и закръглени и се различават по един-единствен признак: едната е предна, а другата задна. (Тази разлика дори е проблематична по принцип за носителите на българския език.) Затова не би било изненадващо, ако субективната преценка в случая е различна от тази, която получаваме чрез обобщеното разстояние на Хеминг. Втори-

ят фактор е семантичният, защото между двете понятия някой лесно би могъл да установили пряка причинно-следствена връзка.

heureux /Œ'Rø/

heureux	/Œ'Rø/	0.00
heureuses	/Œ'Røz/	0.20
heureuse	/Œ'Røz/	0.20
peureux	/pŒ'Rø/	0.20
euros	/Œ'Ro/	0.25
euro	/Œ'Ro/	0.25

Фиг. 1. Френската дума heureux и близки до нея по произношение с разстояние по Хеминг, по-малко от 0.25

Обобщеното разстояние на Хеминг е приложено на практика за френски и английски език като специализиран модул в многофункционалната компютърна програма за обработка и анализ на естествени езици *TREFL*[3]. Фигура 1 представлява екранна снимка и дава представа за графичния интерфейс на програмата при извеждането на резултатите на екрана. За произволна дума, въведена в правописната ѝ форма, програмата намира най-близките по произношение думи, включително сред собствените имена, регистрирани в съответния произносителен речник. Описанието на произношението включва фонемната, сричковата и акцентната структура на думите. Генерирането на необходимите за това произносителни речници е отделен и сам по себе си голям и интересен въпрос, включващ например съставянето на система от формални правила за автоматично сричкоделение и приложението им в компютърни програми.

Литература

Левенщайн 1965: Левенштейн, В И. Двоичные коды с исправлением выпадений, вставок и замещений символов. // *Доклады Академий Наук СССР*. Москва, 1965, №163 (4): 845 – 848.

Хеминг 1950: Hamming, R. W. Error detecting and error correcting codes. // *Bell System Technical Journal* 1950, 29 (2): 147 – 160.

[3] http://web.uni-plovdiv.bg/rousni/