

СТАТИИ / СТАТЪИ / ARTICLES

*Еудженио Пики (Пица, Италия), Джузепе Камуля (Пица, Италия),
Мария Йовчева (София, България), Моня Камуля (Пица, Италия)*

**ПРОГРАМАТА DBT – ПЕРСПЕКТИВИ ПРИ КОМПЮТЪРНАТА
ОБРАБОТКА НА СРЕДНОВЕКОВНИ ТЕКСТОВЕ**

DBT (Data Base Testuali) е софтуерен продукт за текстов анализ и за пълно-текстово търсене, изработен и развиван от д-р Еудженио Пики в Института за компютърна лингвистика към Националния съвет за научни изследвания в Пица. Още от самото си създаване DBT играе авангардна роля в Италия и Европа в областта на компютърната лингвистика и в приложението на информатиката в хуманитарните науки.

Програмата DBT успя да се наложи и вече е на път да се превърне в стандарт за текстологичните изследвания в Италия. Тенденциите на развитието на системата отговарят на актуалните изисквания за текстово проучване. Модулната ѝ природа дава възможност за интеграция с процеси, които не са включени в стандартната версия на DBT. Освен това продуктът може да се използва за големи текстови бази от данни, както например показва LIZ (Letteratura Italiana Zanichelli)¹ – компактдиск на италианската литература, чието второ издание бе през 1995 г. Така че ако прилагането му върху литературни текстове вече е добре познато, то на DBT съществуват и серия програми за по-тесни лингвистични изследвания: изработване на различни видове индекси, търсене в текста на символни низове, търсене чрез ключови думи и др.

Бифункционалната природа на DBT (от една страна – инструмент за проучване на литературни творби, и от друга – за анализ на документи) прави системата един от най-добрите продукти за текстово изследване между съществуващите досега. Това е резултат от постоянното съобразяване на програмата с потребностите на научната общност. Пряката връзка между програмирането и потребителите, която характеризира DBT във всички фази на нейното развитие, е една от причините за успеха ѝ.

Текстовете, обработвани с DBT, са структурирани според системата за кодиране на данни на DBT и могат да бъдат използвани като входни данни за други приложни процедури в системата.

Основните компоненти на DBT са:

I. DBT Query System – Система за търсене (Sistema di interrogazione)

II. DBT-Corpus – Система за търсене върху еднородни текстови корпуси (Sistema di interrogazione di corpora testuali omogenei)

¹ Компактдискът съдържа произведения на италианската литература. Включени са всички творби на най-известните писатели и поети, а от останалите са подбрани само някои.

- III. DBT-INDICI – Отпечатване на различни индекси и конкорданс (Stampa concordanze ed indici vari)
 - IV. DBT-Coocc – Търсене на статистическа съчетаемост (Ricerca Cooccorrenze statistiche)
 - V. DBT-Cerca – Последователно търсене (Ricerca Sequenziale)
 - VI. DBT и изображенията (DBT e la iconografia)
 - VII. PiTAGGER – Автоматична лематизация (Lemmatizzazione/Tagging automatica)
 - VIII. DBT-Tree – Концептуални структури (Strutture concettuali)
 - IX. DBT CercaL – Последователно търсене върху лематизирани текстове (Ricerca Sequenziale di testi lemmatizzati)
 - X. PiMORFO – Морфология на италианския език (Morfologia della lingua italiana)
 - XI. DBT-Lemmat – Автоматизирана лематизация (Lemmatizzazione assistita)
 - XII. WinDBTL – Система за търсене върху анотирани текстове (Query system per testi annotati)
 - XIII. DBT-Synchro – Система за синхронизация и търсене на паралелни текстове (sistema di „sincronizzazione“ e di interrogazione di testi paralleli)
 - XIV. Лексикографско работно място (Stazione di lavoro Lessicografica)
 - XV. DBT SGML/TEI – DBT и Стандартизираният универсален маркиращ език/Инициатива за кодиране на текстове
 - XVI. DBT Internet – Система DBT, вградена в Интернет (Sistema DBT disponibile in Internet: versione client-server e versione WEB)
 - XVII. WinDBT – Нелатински азбуки (Alfabeti non latini)
- Първият пример за приложение, при който се използват структурирани текстове в средата DBT, е системата за търсене on line – Query System.

I. DBT QUERY SYSTEM. СИСТЕМА ЗА ТЪРСЕНЕ

- Системата за управление и търсене на текстови бази от данни е в непрекъснато развитие от десет години насам. Най-важните ѝ характеристики са:
- управление на текстове с латински и нелатински азбуки и кодови таблици с различни символи;
 - изпълнение в реално време и интерактивно на всички функции на системата за пълнотекстово търсене;
 - високо качество, бързина и гъвкавост;
 - пълно зачитане на лингвистичните качества на текстологичния материал и запазване целостта на изходните текстове, от които започваме работата;
 - възможност за оптимизация и на оперативната, и на използваната част от външната памет;
 - възможност за управление на много големи корпуси от текстове.
- DBT работи върху:
- неанотирани текстове;
 - структурирани текстове;
 - лематизирани и/или морфологично анотирани текстове;
 - текстове с произволни азбуки.
- DBT дава възможност за:

– интегриране с електронни речници. Това означава възможност за свързване на модули от лексикални програми, които функционират като морфологичен двигател за развиване и търсене в текста на всички форми от определена лема. Освен това системата позволява свързването с речници с флективни форми, създадени не автоматично, а чрез лематизация;

– интегриране с таксономия – възможно е да се свържат указатели (reference points), които ще бъдат използвани на етапа на търсене. Те могат да бъдат: във вид на таксономия, тезауруси, структурирани концептуално и др.

ДВТ позволява:

– да се следи позиционното местонахождение с указания за евентуална страница, за да покаже локализацията на всички словоупотреби на определена лексема в текста (разбира се, като използваме едно основно издание). Освен това индикацията на реда в страницата се управлява автоматично. Отбелязването на страницата може да се променя според номера на страницата в ръкописа или според неговата пагинация (recto или verso);

– да се показва логическото местонахождение с указания за текстовата единица. Номерата на реда или на стиховете, или (ако специално е отбелязано) на параграфите се управляват автоматично, за да се даде логическата локализация на различните словоформи на думите в текста. Схемата за логическо местонахождение може да бъде структурирана на йерархически степени (позволени са максимално 9 степени) и програмата може да управлява оперативно тези структури;

– да се управляват различни конотатори. Предоставена е възможност да се поставят конотатори, за да бъдат окачествявани отделни текстови единици или част от тях. Конотаторите са данни, които включват индикация от конотаторен тип и реално съдържание. Например валидни са следните конотатори: R – Метафора (Similitudine), за да отбележи конотатор от типа „риторична фигура“, който се отнася към метафора, или A – Икономика (Economiā) за конотатор от типа „аргумент“ и стойност, свързана с икономика, за да позволи отбелязването на топика с такова значение в определена текстова единица. Конотаторът може да има две полезни употреби: от една страна, да се търсят директно всички конотатори от определен тип и с определена стойност, позволявайки автоматичното локализиране на всички появи на дадено явление в определен текст; от друга страна, конотаторите се използват за ограничаване на валидността на полето при търсене на думи или на лингвистични явления в текста (ограничение върху конотаторите);

– да се извършва класификация на думите. Възможно е да се класифицират думи в текста след като дадем някаква характерна особеност. Например да бъдат отбелязани всички собствени имена, разделени по типологията им: лични, местни, имена на богове. Тези класификации могат да бъдат използвани на етапа на търсене и/или за получаване на особени частични специализирани индекси към текста;

– да се работи с многоезична среда. Дава възможност за разграничаване на части в текст, написан на различни езици, и като резултат да се създават отделни индекси (един за всеки език) и освен това бързо да се индивидуализира речникът, който е използван за всеки от езиците в анализирания текст. Даден език (стил) може да служи за идентифициране на части от текст, написан на различни езици или приписван на различни извори като например цитати от други автори или библейски цитати в средновековните

произведения;

- да се следят текстове с критически апарат. Може да се въведат и търсят текстове от критически апарати с варианти или да се управляват бележки и анотации, намиращи се извън текстовете;

- да се работи с текстове с изображения или ръкописни извори. Позволява да се прибавят в текста изображения, част от самия текст или репродукция на извор (ръкопис), от който е взет текстът. Изображенията могат да бъдат разглеждани на екрана, като се сравнява текст и картина, а освен това – да бъдат анализирани с функциите за обработката на образи, вградени в DBT. Това е особено полезно при изследване на средновековни ръкописи, защото може да се визуализира и сравнява оригиналният текст с електронния и с евентуалните илюстрации в ръкописа, за да се провери кодификацията, да се поправят грешките и изобщо да се приложи комплексен подход към конкретния кодекс, като се съчетаят лингвистичното, палеографското и изкуствоведското му проучване;

- да се управляват текстове със звук (в момента тази процедура все още се разработва). Възможно е да се сравняват звук, текст и изворът на звука. Например – италианско-английски речник, при който са съединени графичният вид на една дума с нейния звуков облик;

- да се изследват контрастивни текстове. Потребителят може да сравнява или съпоставя текстове, които са свързани по определен начин: например да се проследи отношението между оригинал и превод или между различни преводи от един оригинал и др.

Първата стъпка при използване на системата е да прехвърлим текстовете във вътрешния формат на DBT. Тази процедура е проста, бърза и след изпълнението на определени конкретни инструкции е напълно автоматична. След като прехвърлим текста във формат DBT, могат да се използват процедурите на системата за търсене, които ни предоставят множество възможности: да влизаме в отделен текст или в даден корпус от текстове и да получаваме като резултат елементи или комбинация от елементи. С помощта на DBT потребителят може:

- да разглежда на екрана целия текст или част от текста (текстовете), върху които работи;

- да търси форми на думи, които съдържат определена комбинация от символи (букви и други знаци), като използва подбрани от самия него кодове: например сродни думи, лексеми с еднаква представка или наставка и др.;

- да изчислява честоти;

- да дефинира функции за търсене, в които думите са свързани по различен начин и да получава всеки контекст, отговарящ на специфични условия на търсене: устойчиви словосъчетания, съчетания от две думи, дори ако те не са в контактна позиция;

- да създава конкорданси и да определя особени условия за създаването им;

- да избира интересувания го език, когато много езици или типове езици се употребяват в текста или текстовете;

- да отбелязва анализиран текст, използвайки функцията block-note, вградена в системата.

Друга важна характеристика на системата е, че съдържа процедури за управление на изображения, представени в текста. База от данни с изобра-

жения е свързана с текста и потребителят динамично може да работи с тях. Условието, при което резултатите от търсенето са визуализирани на екрана или са отпечатани, може да се определи от потребителя в зависимост от неговите лични изисквания или вкусове. Системата е много гъвкава и лесна за използване.

II. DBT-CORPUS. СИСТЕМА ЗА ТЪРСЕНЕ ВЪРХУ ЕДНОРОДНИ ТЕКСТОВИ КОРПУСИ

На базата на отделните DBT текстове е възможно да се създадат много големи корпуси, които да бъдат обработвани със специална DBT програма, притежаваща същите функции, каквито има версията за отделен текст. Тя дава възможност за бързо формиране на обемни корпуси, изхождайки от вече структурирани и индексирани архиви.

На етапа на търсене може:

а) да се селектира корпусът, който ще бъде изследван, като се подбере библиографската информация или определен вид информации, свързани с всеки текст;

б) да се използват функции за DBT търсене, които са на разположение в реално време за целия работен корпус.

III. DBT-INDICI. КОНКОРДАНСИ И ИНДЕКСИ

Освен функции на система за търсене с директно и интерактивно влизане DBT съдържа серия от процедури за създаване на конкорданси и различни индекси. Като резултат от процедурите могат да бъдат създавани ASCII файлове или файлове на формат RTF (Rich Text Format), които позволяват следващо обработване с текстообработваща програма във вече форматирани и номерирани страници в зависимост от персоналните изисквания на потребителя.

Резултатите, получени с такива процедури, които могат да работят с цялата серия от параметри в зависимост от личните нужди на потребителя, са следните:

а) конкорданси (concorda). Могат да се изискват автоматични конкорданси на всичките форми в даден текст, представени в азбучен ред, в сравнение със същите форми и в реда на текстовете (текста) вътре във всяка група с повече от един контекст, отнасящ се за съответната форма. Могат да се изискват и конкорданси само за някои лексеми, които са избрани от потребителя според условията за търсене, свойствени на DBT. Големината и условията на контекста могат да се определят от потребителя;

б) азбучни фреквенции (frequenze alfabetiche) – за честотите на всички думи в един текст, разположени в азбучен ред. При изброяването на честотите, например в един текст на италиански, може да се изисква автоматично сравнение с основния речник на италианския език, създаден от Тулио де Мауро²: отбелязват се думите, които се намират в различни полета на ос-

² Мауро, Т. Vocabolario di base della lingua Italiana. 1991. Основен речник на италианския език, формиран според статистически критерии. Отражава най-използуваната част от речниковия фонд. В него влизат първите 5000 леми според честотния речник на италианския

новния речник, и накрая се дава статистика на тези стойности;

в) низходящи фреквенции (*frequenze decrescente*) – за честотната употреба на всички думи в даден текст, разположени в низходящ ред в зависимост от стойността на фреквенциите. Така в началото ще стоят думите с най-висока честота и в края – думите с най-ниска честота;

г) индекс на мястото (*index locogum*) – за създаването на индекс с местата на всички думи в един текст. Подредбата е по азбучен ред, като се отбелязва мястото на думата в текста. Индикирането може да бъде според логически или топографски признак. За избягване на прекалено обемни изходни данни потребителят може да сложи граница, над която системата няма да даде списък на появите, а само абсолютната честота;

д) индекс на читаемост (*indice di leggibilita*) – DBT съдържа програма, която позволява да се получи процентът на читаемост на един текст според формулата GULPEASE на Лучизано и Пиемонтезе³ (адаптация на оригиналната формула на Flesh⁴). Индексът на читаемост може да бъде пресметнат или за целия текст, или за отделна текстова единица;

е) статистики (*statistiche*) – позволява броенето на някакви стойности в целия текст и/или в отделни части от него, броенето на различни пунктуационни знаци, честотата на всички диграми и триграми (серия от два или три съседни знака).

IV. DBT-СООСС. ТЪРСЕНЕ НА СТАТИСТИЧЕСКА СЪЧЕТАЕМОСТ

Чрез използването на формула, основана на статистически оценки, се търсят думи, които в един текст се появяват като по-тясно свързани с друга дума или група от думи, посочени от потребителя. Особеностите на системата са:

а) използва се специална формула за търсене на двойки от думи, за всяка от които се дава оценка, основана на стойностите на честотата както на самите думи, така и на тези думи в част от текста с определени размери. Тази стойност показва в правопрпорционална зависимост силата и статистическата стойност на вероятността тези две думи да се срещнат заедно, давайки възможната стойност на корелация между тях в изследваната част от текста. Например *New York*;

б) формулата може да се приложи върху цял текст и всички думи в него;

в) възможно е да се избират от потребителя думите, за които се търси статистическа съчетаемост.

език (Bartolini, U., C. Tagliavini, A. Zampoli. *Lessico Italiano di Frequenza*. Milano, 1972), към които е интегрирана друга част от лемн, подбрани по различни начини.

³ Индексът на читаемост представя математическа формула, която чрез статистическо изчисление може да предскаже трудността на текста на основата на предварително определена скала от стойности на неговите параметри като например дължината на думите или на изреченията. Формулата GULPEASE е създадена на базата на италианския език и изчислява дължината на думите чрез броя на буквите в тях.

⁴ Индексът на читаемост Flesh се ползува с най-голям успех от всички формули на читаемост. Той взема под внимание средната дължина на думите, изразена чрез броя на сричките в дадена лексема, и средната дължина на изреченията, изразена чрез броя на думите в изречението. Създаден е на основата на английския език.

Параметрите, вградени в системата, са:

1. Windows Size – показва се броят на думите отляво и отдясно в зависимост от представителната дума за използването им при броенето. Тези стойности ограничават частта от текста, в която трябва да се търсят статистически свързани думи.

2. Използува се списък от празни думи на предварително приготвен ASCII файл, в който са изброени лексеми с висока честота, оставащи извън броенето. Например в него могат да се включат служебни думи като съюзи, предлози или някои повтарящи се наречия.

3. Може да се въведе и минимална граница на честотата на две думи в един и същ прозорец, под която резултатите няма да се визуализират. Тази стойност е удобна за двойки с висока честота с по-малка дисперсия във фазата на представянето на резултатите.

4. Начинът на подреждане може да бъде азбучен или в индекс, получен чрез прилагането на специалната формула.

5. Могат да се търсят предлози в част от текста, като се изхожда от избраните активни думи: например започвайки от всички форми на даден глагол в текста, автоматично да се търсят и изведат всички предлози, които този глагол управлява.

6. Съществува възможност за визуализиране на контекст. В един прозорец могат да се получат едновременно всички контексти на една специфична двойка от думи между онези двойки, които са намерени автоматично чрез статистическата формула.

7. Контекстът може да се разширява. Всеки създаван контекст е възможно да бъде разширен до целия текст, така че потребителят да анализира по-добре стойността и смисъла на специфичния контекст.

V. DWT-CERCA. ПОСЛЕДОВАТЕЛНО ТЪРСЕНЕ

В този случай е възможно да се търсят граматични категории върху лематизирани и/или аотирани текстове:

- а) търсене на елементи на текста като за нелематизирани текстове;
- б) използване на лема като елемент за търсене;
- в) употреба на граматически код (TAG) като елемент за търсене;
- г) употреба на морфолого-синтактична информация като елемент за търсене.

VI. DWT И ИЗОБРАЖЕНИЯТА

Системата DWT съдържа възможността да управлява изображения както ако те са оригинални графични образи от един и същ текст, така и ако са илюстрации, включени в самия текст.

Присъствието на илюстрации, включени в текста, е обозначено и е възможно да бъдат извикани по желание на потребителя.

Всички образи от даден текст могат да бъдат събрани в една „библиотека“ за изображения, която потребителят да преглежда със средствата на DWT. Освен това е предвидена възможност от изображенията да се връща отново към текста, който ги съдържа.

Показаните на екрана изображения могат да бъдат графично преработени, като се използва сложна „библиотека“ за графични функции (Lead Tools for Windows).

VII. RTAGGER. АВТОМАТИЧНА ЛЕМАТИЗАЦИЯ

Представя процедура на граматическо кодиране (tagging) и автоматична лематизация. Засега се прилага за изследвания върху италиански език чрез използване на:

а) морфологията на италианския език, която позволява морфологичният компонент на тази процедура да препраща всяка отделна форма към всички лемми, към които тя може да принадлежи;

б) статистическия подход – на основата на корпус от текстове (Training Corpus), вече автоматично или неавтоматично анализирани като база за лингвистични знания и чрез данни от корпуса, обобщени статистически, се проучват новите текстове;

в) интерактивен редактор, който дава възможност за контрол на резултатите от автоматичната фаза на тагирането и за евентуално поправяне на грешките.

Последната процедура може да се прилага и към езици, за които съществува автоматична морфологична система, и към езици, за които съществува един обемен корпус за обучение. В резултат:

а) лематизираните текстове ще бъдат съвместими с текстове, лематизирани чрез полуавтоматична процедура DBT Lemmat;

б) текстовете, автоматизирани и чрез помощна процедура от компютър, и чрез автоматична процедура, ще бъдат разположени и съвместими с процедурата на DBT, която управлява и търси лематизирани текстове (WINDBTL).

VIII. DBT-TREE. КОНЦЕПТУАЛНИ СТРУКТУРИ

Процедурата позволява да се използва външен инструмент: йерархична структура за анализ на даден текст. Йерархичната структура е запаметена като файл и представя някакво явление.

Структурата се представя като съвкупност от възли на низходящи йерархични равнища, като се започне от началния възел и се стигне до последния.

Структурата трябва да бъде определена според подходящ синтаксис и впоследствие може да се приложи към различни текстове за измерване и разпознаване на елементите като например възлите и подразделенията, които се откриват в отделните текстове. Могат да се анализират едни и същи резултати от различни гледни точки, за да се установи какви текстове (или части от тях) по-добре се идентифицират с дадена концептуална структура или част от такава структура.

Целта на процедурата е да измери колко и каква част от една йерархична структура е отразена в текста. Така резултатът при прилагането на определена структура към даден текст ще бъде качествена и количествена оценка на съответствието между текста и елементите, записани в структурата.

Възможното приложение на програмата е да се оцени колко различни текстове са отразени в една структура или чрез изследване на един и същ текст с различни структури да се определи коя е най-добре представена в него.

Резултатът от процедурата е визуализирането на избраното дърво, където се отбелязва всеки възел на структурата и честотата, с която той е намерен в определен текст. Визуализирането може да бъде синтетично или аналитично – това означава, че могат да се индикират и отделните думи и/или използваните устойчиви словосъчетания.

За всеки възел или за всяко разклонение на структурата може да бъде изискван контекст, което ще позволи по-обективна оценка.

IX. DBT-CERCAL. ПОСЛЕДОВАТЕЛНО ТЪРСЕНЕ ВЪРХУ ЛЕМАТИЗИРАНИ ТЕКСТОВЕ

В този случай може да се търси върху граматични категории. Възможно е да се извършват същите процедури както DBT Cerga върху лематизирани и/или аотирани текстове.

X. PiMORFO

Съдържание:

1. Индекс (речник на лемите)
2. Морфологично създаване
 - 2.1. Правила за флексиране
 - 2.2. Флексиране на глаголни и неглаголни лемите
3. Морфологичен анализ

Засега като пример за функционирането е системата за морфологичен анализ на италианския език, развита в Института за компютърна лингвистика в Пиза, която е замислена, за да се получи двойна употреба: етап на създаване на всички форми от дадена лема на италианския език и етап на морфологичен анализ, позволяващ да се съотнася всяка форма към лемата (лемите), на която (на които) тя принадлежи или може да принадлежи.

Морфологичната система се основава на група от средства, създадена от три главни компонента (вж. следващите параграфи): индекс на лемите на италианския език, правила за флексиране на лемите и правила за анализ на словоформите.

1. Индекс на лемите на италианския език.

Състои се от архив от около 120 000 лемите. За всяка лема е отбелязана следната информация: лема, граматическа категория, кодът на флексиране и кодът за използване.

Освен експлицитно представените в архива данни има възможност за имплицитно получаване също и на кода на групата.

Флексионният код заедно с кода на групата определя правилото за флексиране, което трябва да бъде приложено към дадена лема, за да получим всичките нейни словоформи. Всички лемите, които развиват флексия на формите си според един и същ механизъм, ще имат един и същ код. Той се отнася само към алгоритъма за създаване на формите без никаква линг-

вистична автоматична импликация. В перспектива, за да използваме правилата за флексиране, свързани с всяка лема, се опитахме да отстраним понятието за неправилност, съществуващо в правилата за флексиране, като флексионният код бе свързан не само към морфологичните особености на неправилните форми, но и към правилните, възвръщайки цялата система към еднородно управление.

2. Алгоритъм на морфологичното създаване (Generazione Morfologica)

На етапа на морфологичното създаване има като входни данни лема (разбира се, намираща се в индекса на лемите), а като изходен резултат се получава редът на всички форми, които принадлежат към парадигмата на тази лема със съответната граматична класификация.

2.1. Правила за флексиране

Съдържат информация за флексирането на лемите. Те показват създаването (като се започне от основната лема) на всички форми на парадигмата, свързвайки всяка от тях със съответната граматична класификация.

2.2. Флексиране на глаголните и неглаголните лемите

Всеки файл на правилата за флексиране, отнасящи се към глаголни или неглаголни лемите, е изграден от два различни типа информация: от една страна – група от таблици, показващи всички възможни окончания, и от друга страна – реалните правила за флексиране, които ще служат като ориентир за тези групи окончания.

3. Морфологичен анализ

Вторият етап на обработка, извършвана от морфологичната система PiMorf, представя морфологичен анализ на форми от италиански език, като позволява да се свързва всяка анализирана форма с лемата, към чиято парадигма тя принадлежи. Разбира се, една дума може да бъде отнесена към повече от една лема: в този случай всички лемите се приемат като еднакво възможни. Фазата на избор на една от няколко възможни форми, която се опитва да намери правилната лема на дадена форма в определен контекст, е следващата фаза на лингвистичен анализ. Тя трябва да има като отправна точка описания по-горе морфологичен анализ. Фазата на автоматичен избор на омографите при анализиране на текста може да се достигне чрез системата PiTagger, в която се прилага статистически подход. Системата може да бъде особено ценна при изследване на синтетични езици като например старобългарския или някои от съвременните славянски езици, в които немалка част от словоформите на дадено име (съществително, прилагателно, местоимение) са получени посредством промяна на падежната флексия.

XI. DBT-ЛЕММАТ. АВТОМАТИЗИРАНА ЛЕМАТИЗАЦИЯ

Служи за лематизация с помощта на компютър (посредством специализирано работно място), която може да се извърши:

- а) без речникови бележки – т.е. без да се основаваме на речник;
- б) с употреба на речник;
- в) с предварително използване на речници и морфологични справочници;
- г) с речници, които се създават и обогатяват автоматично, докато лематизираме нови текстове.

Методът на работа при взаимодействащата процедура за лематизация

ция на текстове (Interattiva) предвижда лематизация на текста по азбучен ред на формите. За всяка форма, представена в азбучна последователност, се дават всички съотнесени контексти и освен това е възможно да се свърже всяка словоупотреба със съответната лема и да се класифицира граматически. Лемите, към които принадлежат тези форми, могат да са изготвени от потребителя или пък да са предложени като помощ от автоматични морфологични системи: автоматични речници или речници от типа 'форма:лема', предварително създадени автоматично при лематизацията на други еднородни текстове.

В процедурата на лематизация участват програми за създаване на честотни индекси и конкорданси в ред форма:лема и (или) лема:форма. В резултат:

а) лематизираните текстове ще бъдат съвместими с текстове, лематизирани и чрез автоматична процедура на тагиране (виж PiTagger);

б) текстовете, лематизирани посредством процедура, само подпомагана от компютъра, след това могат да се разположат и съвместят с процедурата на DBT за управление и търсене на лематизирани текстове WinDBTL.

XII. WinDBTL. СИСТЕМА ЗА ТЪРСЕНЕ ВЪРХУ АНОТИРАНИ ТЕКСТОВЕ

Представя система за управление и търсене на тагирани и/или лематизирани текстове.

Анотираните и/или лематизирани текстове чрез процедурата за лематизация, осъществявана само с помощта на компютъра и чрез изцяло автоматична лематизация (PiTagger), могат да се прехвърлят в процедурата на DBT, специализирана за този тип текстов материал. Текстът е запаметен в тази процедура според неговата класификация и цялата приложена информация може да се използва на етапа на търсене.

Процедурата WinDBTL може:

- да регистрира и съдържа получената класификация;
- да търси върху текст и чрез новата въведена информация;
- да създава индекси и конкорданси за лема и/или форма;
- да търси синтактични модели в тагирани текстове;
- да служи за база на нови процедури за анализ;
- да бъде интегрирана към системата „Лексикографско работно място“ (Stazione di lavoro lessicografica).

XIII. DBT-SYNCHRO. СИСТЕМА ЗА СИНХРОНИЗАЦИЯ И ТЪРСЕНЕ ВЪРХУ ПАРАЛЕЛНИ ТЕКСТОВЕ

Включва процедура за автоматично подреждане на паралелни текстове на два различни езика.

Процедурата предвижда автоматично подреждане на два текста, при които единият е превод от другия, и позволява на системата за търсене DBT Synchro да работи върху тях. Като резултат се получава успоредно подреждане на двата текста.

Това е полезно особено при съпоставка на преводна творба с нейния

оригинал или на различни редакции и преработки на едно и също произведение, което улеснява в значителна степен изследователите при проучването на писмената традиция на средновековни текстове.

Процедурата е приспособена за работа и алайниране и на сходни текстове на различни езици и/или латински и италиански текстове.

XIV. ЛЕКСИКОГРАФСКО РАБОТНО МЯСТО

Предоставя допълнителна система за помощ при лексикографско редактиране. Тя е създадена и развита като гъвкав и лесноприложим инструмент, който може да бъде използван в етапите на обработка на материала: редактиране на документи, откриване на цитати и получаване на информация за различни текстови извори, форматиране на материал за отпечатване, а така също и за компилиране на структурирани архиви: речници, глосарии, тезауруси, библиографии и др.

Основните характеристики, които правят Лексикографското работно място особено ценна система, са:

- а) висока степен на гъвкавост;
- б) възможност да се предприемат обработки на материала от архива, които да бъдат мощен инструмент за изследване;
- в) възможност за съвместно действие в реално време с основни текстови източници (корпуси) с цел да се получи информация, полезна за съставяне или компилиране на документи или за автоматично извличане на цитати от текстове, с които работим;
- г) възможност за работа с нелатински азбуки както при предварителната обработка на документи, така и при управлението на основните текстови извори (корпуси).

Основните функции, които системата предлага, са:

- а) управление на структурирани лексикални входни данни;
- б) редактор, специализиран за лексикални входни данни с цялата серия функции *ad hoc*, за да улесни и води редакторската работа, допълнен с различни помощни средства (*help environment*), който за всеки поискан аргумент дава точна и подробна информация: включване, удвояване, премахване, възвръщане, свързване, подразделяне, търсене, копиране на поле вътре в документа; копиране, премахване, възвръщане и отпечатване на документа, вмъкване в ASCII файл, функция за прехвърляне от един документ в друг и др.

в) *context cutter* – функция за оптимизирано отрязване на контекст. Позволява от голям документ да се отреже желана част от текста, като се използват специализирани функции за ускоряване на търсенето;

г) допълнение към DBT за анализи и използване на текстове и корпуси. При компилиране на речници или глосарии може да бъде особено важно потребителят да се допита до конкретния контекст, от който да получи полезна информация за компилацията на входната лема. Освен това е ценна възможността да се извлече особен контекст на различни думи, за да се постави автоматично като цитат в структурирането на лемата, върху която се работи. Средата DBT на практика може да бъде извикана за даден текст, така че потребителят да разполага с нейните функции и освен това да има възможността да извлича контекст. Контекстът, автоматично премества

в лемата, ще бъде по-широк и за него е на разположение функцията *context cutter*;

д) допълнение към *Polmone* за запазване на избрани контексти, класифицирани и подредени, които да се използват като рамка (*framework*). Процедурата *Polmone*, включена в системата за текстов анализ, позволява създаването на отделни архиви, наречени *Polmoni*. В тях може да се премести група от избрани контексти, извлечени от един или повече от един текстов DBT архив. Всеки архив *Polmone* се отнася към една-единствена лема и може да бъде създаден както чрез избор на контексти, извършван от лексикограф посредством използването на специална DBT програма, така и чрез специално разчленяване на последователни фази на преработка на текстовете. В началната фаза контекстите, отнасящи се към една и съща лема, са извлечени автоматично от всички анализирани текстове и са преместени в специален архив *Polmone*. Процедурата за управление на тези архиви позволява етап на анализ на групата от словоформи, в която съответните контексти могат да се изследват и класифицират на по-сложна степен в сравнение само с една лема. Управлението на архивите може да бъде извършвано със специална програма или от програма, направена в оперативна, вътрешна лексикографска станция, наречена *Ambiente Polmone*. В нея са разположени функциите за анализ, класификация и възстановяване на групата от контексти. Освен това могат да се избират особените контексти, за да бъдат автоматично включени във входните данни като цитати. За изготвянето на тези цитати е предвидена фазата на *context cutter*;

е) автоматично вмъкване в DBT – позволява автоматичното включване на лексика, създадена с Лексикографското работно място, в структурата на DBT, за да се организират по този начин лексикални бази от данни;

ж) различни оперативни подходи.

Основните елементи на Лексикографското работно място са:

а) редактор за създаването и постоянното попълване на структурирани лексикални входни данни. Той предлага:

– различни специализирани лесноприложими функции, за да подпомага редактирането на структурираните входни данни;

– възможност за взаимодействие със системните файлове и таблици за ръководство и помощ при компилацията на структурираните полета;

– взаимодействие с процедурата на DBT и евентуалните архиви *Polmoni*, за да се използват основните текстови архиви и да се внесат автоматично цитати;

– функции за отрязване на контекст, за да се получат по-кратки и възможно най-показателни цитати.

б) вход/изход за въвеждане и извеждане на материал от лексикалния архив с голяма избирателна способност.

Много важна характеристика е, че по време на извеждането има възможност да се запаметят данните в ASCII формат. При въвеждане също е възможно да сложим в архива данни в ASCII формат. Фазата на запаметяване на данните в ASCII формат е полезна, за да се правят копия на материалите без специално форматиране. Фазата на извеждане и след това на въвеждане позволява преместване на данните от един архив в друг. Възможно е също така потребителят да избира подполета от архива или да определя за всеки отделен архив полетата, които да бъдат извеждани. Разбира се, всички материали остават неизменени на своето място, което означава, че

само са направени техни копия в ASCII формат според инструкциите, дадени от потребителя. При въвеждане той може в случай на добавяне на документи в архив с вече съществуващ ключ на идентификация да избере дали да наслои новия документ към вече съществуващия, или да го замени. Тези възможности, предоставени от процедурите за вход/изход, позволяват серия от операции за решаване на проблеми, които могат да се появят при изпълнението на дадена задача. Например да се компилират отделно от различни хора подполета в краен архив, след което с функциите вход/изход да се създаде един общ архив;

в) WSIndici – за създаване на прости и сложни индекси за отпечатване. Позволява да бъдат представени като специални индекси архивите на Лексикографското работно място. Резултатите може да са подредени в азбучен ред за съдържанието на едно избрано поле или следвайки азбучния ред на списък на отбелязаните полета. Отпечатаните документи могат да бъдат избирани от потребителя чрез определяне на необходимите качества, които те трябва да задоволяват. Освен това е възможно той да определи за всеки документ кое поле да бъде отпечатано. Файлът за принтиране е създаден на RTF, който е стандарт за размяна на данни и форматираните текстове между различните текстообработващи системи, съществуващи на пазара. Този начин регистрира данните, като използва седембитово ASCII представяне и своя собствена система за маркиране, която включва сведения за форматиране и избиране на шрифтовете, характерни за системата за писане. Така потребителят има възможност, използвайки текстообработваща система, която е съвместима с този начин за маркиране, да променя знаците и размерите, пагинацията и всичко, което програмата за писане позволява да се прави;

г) WSmacro – за създаване и прилагане на макроси върху лексикалните входни данни, които да бъдат аналитично използвани за целия архив (корпус). Това е гъвкав инструмент, позволяващ на потребителя да определя и извършва пълната серия от операции върху документите на целия архив според това дали ще отговорят или не на определени условия, като използва повтарящи се цикли на изпълнение. Операциите са запазени в архив на диска с възможност да бъдат извършвани повече от един път според нуждите на потребителя;

д) WSDBT – за форматиране на лексикалния архив на процедурата DBT като структуриран активен DBT архив, върху който да се работи при търсене със специалната програма Query System.

Основните текстови архиви могат да бъдат използвани от лексикограф във фазата на подготовка и съставяне на дадена лема, за да прецени както нейната значимост, така и значимостта на всички нейни форми и варианти в основните текстове. Тази оценка се приближава към истинността на семантичните и синтактичните разлики, които лемата може да има, към установяване на времето на поява и смисъла на думата, към анализиране на промените в семантиката и употребите в течение на времето и в различни изследвани текстове. Текстовите архиви могат да се използват за лексикографски цели – за създаване на рамки, разделени според семантиката на думите, които са направени чрез уместно избрани цитати. Целта, която трябва да се постигне при създаване на допълнителното лексикографско място, е лексикографът да има възможност за анализ на целия материал, който е на негово разположение, да избира значими примери от корпуса,

да прегрупира, класифицира и преподрежда полето на намираните примери и накрая да ги използва за създаване на лема и за включване на цитати и примери, взети направо от текстовете.

Съобразно с типологията за анализ на материала, вида речник, който се създава, и модалностите, използвани при редактирането, някои от оперативните функции са повече или по-малко важни спрямо останалите. По тези съображения бяха създадени три различни подхода за използването на текстовия материал като основна база и за допълването на Лексикографското място, всеки от които се адаптира според конкретния случай. Тези подходи не си противоречат и могат да съществуват съвместно, позволявайки създаването на оптимална оперативност при всеки конкретен случай.

Първото решение предвижда един DBT модул, допълнен към Лексикографското работно място. Този модул дава на разположение на потребителя всички функции за информационно търсене, които са вградени в системата DBT. Те са много полезни при проверка на съществуването и характеристиките на лема, форми и варианти в наличните текстове. Така потребителят може да избере контекст, който да бъде автоматично включен вътре в съставения в този момент лексикален вход. Преместваният контекст има достатъчно голям размер, за да позволи фаза на context cutter.

Второто решение предвижда създаването и използването на архив, междинен между текстовите архиви и реалната процедура на лексикографско редактиране. Междинната фаза се състои в създаването на специален архив, наречен технически *Politone*, в който лексикографът може да съхранява всички контексти, които той ще изтъква чрез анализа на текстовете. Архивът ще бъде от рода на кутия, една за всеки лексикален вход, където да бъдат поставени всички части от текста, необходими на потребителя с цел компилиране на входните данни. Различните контексти вътре в архива се управляват от потребителя, използвайки фазата „класификация“, по време на която всеки контекст се асоциира с кода на идентификация на смисъла, като дава възможност за подразделяне с детайлизиране до две степени.

Третото решение прилича на второто по използването на архива и по всички функции на входа и обработката, но е различно за подходите при избора на контекста, който ще бъде включен. Докато при второто потребителят използва и свободно избира отделните контексти за включване в архива един по един, тук е предвидена фаза на анализ на текстовете един по един и фаза на класификация и избор, така че създаването на архива да бъде последователна автоматична процедура.

XV. DBT И СТАНДАРТИЗИРАНИЯТ УНИВЕРСАЛЕН МАРКИРАЩ ЕЗИК/ИНИЦИАТИВАТА ЗА ТЕКСТОВО КОДИРАНЕ

DBT има собствена система за кодификация и вътрешно структуриране на данните, която е резултат от десетгодишен опит в областта на текстовия анализ. Съществува процедура на въвеждане/извеждане между DBT и SGML/TEI. При прехвърлянето от SGML/TEI в DBT цялата информация, съдържаща се в текста, може да бъде прекодирана във формат DBT. Стремелът е да се осигури максимално взаимодействие между двете кодификации без разработките на Института в Пиза да бъдат зависими. В момента е във фаза на реализация версия на DBT, която може да използва материал

на SGML/TEI без да го реструктурира, но чрез създаването на DBT индекси, позволяващи изпълнението на всички функции за търсене, свойствени на системата. Фазата на сегментиране може да се извърши предварително с други процедури или от DBT в момента на индексиранието.

XVI. DBT И INTERNET

През последните десет години системата DBT има голямо разпространение и се приема в италиански университети и в много научни институти не само в Италия за лингвистични и в частност – за лексикографски проекти. За да направим един текст читаем във формат DBT, се изискват твърде прости процедури, които отнемат кратко време. Това означава, че съществуват вече многобройни текстове, структурирани във формат DBT, и тяхното количество бързо нараства. В настоящия момент в научната общност се усеща важността и необходимостта от инструменти за лексикални и лингвистични изследвания. Наистина засега ние се опитваме да направим такива инструменти, които да се използват от всички според разпоредбите, прилагани за защита на авторските права. Затова при изработването на архивите DBT се постаряхме да ги направим достъпни не само за местни потребители, но и за учени и студенти от различни части на света. Именно Интернет предоставя тази възможност. Ние направихме подобна стъпка, следвайки две линии на развитие, неизбежно противоположни една на друга и които се стараят да отговорят на изискванията на потребителя. Едното решение бе да се използва най-известният съществуващ стандарт за разпространението на информацията на Интернет, т.е. WWW и неговите системи за хипертекстова навигация. Най-голямото преимущество от използването на тази технология е, че тя улеснява навигацията в мрежата, защото представя стандарт, независим от платформата и от системата, с която потребителят разполага. Все пак, въпреки нейните положителни характеристики, приемането на философията 'страница след страница', използвана от технологията WWW, води до ограничаване на системата DBT, която бе развита с висока степен на взаимодействие с потребителя и с голяма бързина на времето за отговор.

Затова ние решихме да развием едно приложение към DBT, което работи на Интернет и съдържа всички функции на версията със самостоятелна програма (stand alone). В тази версия, наречена DBT-Net, процедурите са разделени на два отделни компонента: сървър, включващ индекси и текстови архиви, и клиент, който дава на потребителя достъп до архивите на сървъра. Основната цел бе да предоставим един и същ интерфейс и едни и същи функции на самостоятелната програма заедно с възможно най-краткото време за отговор. При това най-голямата част от информацията се запазва при клиента, така че да се намалят средствата, заети при сървъра, успоредно с намаляването на операциите между клиента и сървъра.

DBT-Net позволява на потребителя да влиза, да прави справки и да търси в банката от данни. Тъй като се състои от два компонента (клиент и сървър), чрез протокол TCP/IP се дава възможност на институтите, които имат архиви на DBT сървъра, да общуват с всички, които се нуждаят от преглеждане на тези архиви. Системата клиент бе развита в средата Windows, работеща с обикновени персонални компютри, с цел да задоволи

изискванията на много от потребителите, които обикновено не разполагат с големи компютри. Сървърът, основан предимно при университети или научни институти, освен на версията Windows е в развитие и в средата UNIX.

Първоначалната преценка установи голямата бързина на системата, като твърде приемливо се оказа времето за преместване и натоварване на картините. След кодирането и индексирането на данните с нормалната DBT процедура и разполагането им чрез сървъра, използването на DBT-Net е много лесно. Трябва само персоналният компютър на клиента да е свързан с Интернет. Достатъчно е да стартираме програмата; потребителят се намира точно в една и съща среда със самостоятелната програма на DBT и може да използва всички функции, свойствени на системата. Единствено когато е на Интернет по този начин, той може да работи с големи банки от данни, а не само върху ограничено количество, което всеки потребител има на разположение.

Като предоставяме два различни начина за използване на DBT с Интернет, ние даваме възможност за работа и на тези потребители, които искат да извършват прости търсения без да научават нова система с нейния собствен език, и на онези, които вече използват DBT и искат да извършват много сложни търсения на данни, но ги нямат на разположение при себе си.

XVII. WinDBT. НЕЛАТИНСКИ АЗБУКИ

В началото DBT бе използвана само за латински азбуки и работеше на MS-DOS, което бе достатъчно за обработването на такива текстове. Предвид на нейните големи възможности по-късно програмата бе доразвита и за работа с нелатински азбуки, като засега тя съдържа южноарабска азбука и кирилица. Именно наличието на славянска версия предопредели, от една страна, избора ѝ за осъществяване на съвместния проект между Кирило-Методиевския научен център и Института за компютърна лингвистика в Пиза на тема „Компютърна лингвистика, класическа филология и филологически изследвания“. От друга страна, DBT създава максимално благоприятни условия за изпълнението на основната цел на проекта – чрез прилагането на системата за електронна обработка върху средновековни текстове да се осъществи подготовката на критическо издание на кирило-методиевски текстове. Не е без значение и фактът, че програмата представя един от най-разпространените продукти от този тип в Италия, а вече я използват и много институти извън Италия.

В началото работата се съсредоточи върху текста на старобългарския превод на Стария завет. Започна се с Книгата на пророк Иезекиил по най-стария запазен славянски препис под сигнатура F.I. 461, среднобългарски ръкопис от втората половина на XIV в. от Руската национална библиотека в Санкт-Петербург, вече въведен в компютър, но според изискванията на дипломатическото издаване. Тъй като за неговото въвеждане бе използвана програма, работеща под Windows, и специален пакет от шрифтове AlphaWin, преди всичко бе необходимо да се допълни DBT с програмата Windows, за да може да обработва и този текст.

Процедурите на DBT за нелатински азбуки се основават върху архив от данни в ASCII файл, структуриран във вид на таблица, която съдържа

цялата информация, необходима за правилното функциониране на програмата. Преимуществото на този архив е, че той може да бъде модифициран много пъти според нуждите на потребителя, без да изменя самата програма. За да използваме DBT за целите на съвместния проект бе необходимо таблицата да се изработи на базата на старобългарския шрифт (Cyrillica Vulgarian) от пакета AlphaWin. За структурирането на таблицата предварително бяха изучени правописните и фонетичните особености на ръкописа, по който работим.

Таблицата е изградена така, че всеки ред съдържа приблизително едни и същи елементи. Първият показва типа на символа, който ще бъде определен. (С=централна буква, Р=пунктуация и т.н.) Вторият е една точка, която показва кой знак ще бъде представителен, когато съществуват различни алографи на една и съща буква. Това е необходимо, за да получим честотите на форми, които имат еднаква лексикална стойност, но различна графика. Третият показва знака, който ще се използва при набиране от клавиатура. Четвъртият е знак, към чиято числена стойност трябва да бъде отнесена съответната старобългарска буква на етапа на отпечатване и визуализиране. Петият показва азбучния ред на буквата, който сме избрали. Шестият е знакът, който ще използваме в DBT за търсене на формите.

DBT работи със специфични архиви, създадени чрез процедура, която се вмъква във входния текст под формата на ASCII файл, кодиран според изискванията на програмата. Той употребява всички стойности на горната част на ASCII таблицата освен 255-а позиция, която се използва за вътрешните функции на програмата. Именно таблицата служи за правилната интерпретация на тези стойности.

Всеки файл на DBT е съставен от кодове на текста (азбука, диакритични знаци) и кодове на структурата (указатели, индикатори за шрифтови начертания, конотатори и др.).

За да прехвърлим текста директно в ASCII формат, положихме огромни усилия преди всичко относно липсващите в клавиатурата символи (знаци), които се наложи да набираме чрез Alt-комбинации. Този проблем засяга най-вече текстовете, които използват сложна азбука като например старобългарската. В това отношение бяхме улеснени значително от използването на WinWord, тъй като позволява създаването на файл, чиято оригинална графична система може да бъде визуализирана. Впоследствие бе достатъчно да запазим файла от Word във формат MS-DOS, за да получим ASCII файл с всички текстови елементи (думи и диакритични знаци).

Накрая използването на друга програма за записване на знаците като например MS-DOS Text with Layout при активен FlexType позволи с помощта на текстов редактор под MS-DOS, като например Kedit, да се редактира ASCII файлът във форма, достъпна за четене, за да включим и специалната кодификация на DBT (например различни указатели за структурата на текста: стихове, глави, отделяне на основния пророчески текст от тълкуванията и др.).

Трябва да се има предвид обаче, че стандартното използване на файл, експортиран от Word в друг формат, е особено трудно, тъй като се работи с различни параметри, зависещи от използвания шрифт и от типа на изисканата информация. При това при прехвърлянето голяма част от знаците, които принадлежат на горната част на ASCII таблицата (Alt+128), не могат да бъдат употребявани.

Тъй като шрифтът, служещ за въвеждане на старобългарски текстове в среда на Windows (т.е. Cyrillica Bulgarian), употребява всички символи на горната част на ASCII таблицата, необходимо е схемата да бъде модифицирана, за да се прехвърля от и на Word. Просто решение за този проблем предоставя AlphaWin и нейният Translation Table Editor.

В перспектива съвместната работа ще продължи в две насоки:

1. Кодиране на текстовете в SGML, тъй като голямо предимство на DBT е, че може да се развива, за да има пряко взаимодействие с текстов материал, кодиран по различни начини.

2. Адаптиране и на други компоненти на DBT за работа с нелатински (в частност – кирилска) азбуки. Като най-близка цел в това отношение е създаването на славянска версия на програмата Лексикографско работно място, което ще предостави изключителни възможности за лексикален и лексикографски анализ на текстовете.

В заключение може да се обобщи, че DBT представя система, която се отличава сред съществуващите текстообработващи продукти с това, че обединява различен тип дейности, необходими за цялостния анализ на текста: както статистически, така и възможност за обработка и използване на изображения. С това основно преимущество тя се доближава във висока степен до представата за интегрирана работна среда на филолога и открива изключителни перспективи за проучванията на филолозите-медиевисти.

Eugenio Picchi (Pisa, Italia), Giuseppe Camuglia (Pisa, Italia), Marija Jovčeva (Sofia, Bulgaria), Monia Camuglia (Pisa, Italia)

THE PROGRAMME DBT – PROSPECTS OF THE COMPUTATIONAL PROCESSING OF MEDIEVAL TEXTS

(Summary)

This article deals with the description of DBT (Data Base Textual), a computational system of textual analysis created by Dr. E. Picchi at the Institute of Computational Linguistics in Pisa. The article contains a general description of the main elements, functions and characteristics of the program, showing how researches may use it to study their texts at different levels.

Moreover, particular attention has been paid to the present and future possibilities that DBT offers to analyse texts written with non-Latin alphabets, in particular to analyse texts belonging to the medieval Slavonic tradition.