https://doi.org/10.53656/math2024-2-6-inv

Overview
Обзор

ИЗСЛЕДВАНЕ НА ВРЪЗКАТА МЕЖДУ ФОНЕМНИЯ СЪСТАВ И ЕМОЦИОНАЛНАТА ВАЛЕНТНОСТ НА ТЕКСТ

Велина Славова, Филип Андонов
Нов български университет – София (България)

Резюме. Статията се основава на резултатите от публикувани авторови изследвания на връзката между звуковия състав на езика и емоцията, довели до направените тук изводи и обобщения. За да се разкрият по-дълбоки зависимости между езика и емоциите беше разработена методика и съответна автоматизирана система, позволила прилагане на статистически методи над фонемния състав на големи обеми от текстови данни. Данните от английски език и резултатите от приложените статистически методи потвърждават наличието на зависимости между звуковия състав, изразен като двойки от гласна и съгласна, и емоционалната валентност (позитивна-негативна емоция) за текстове на новини от интернет.

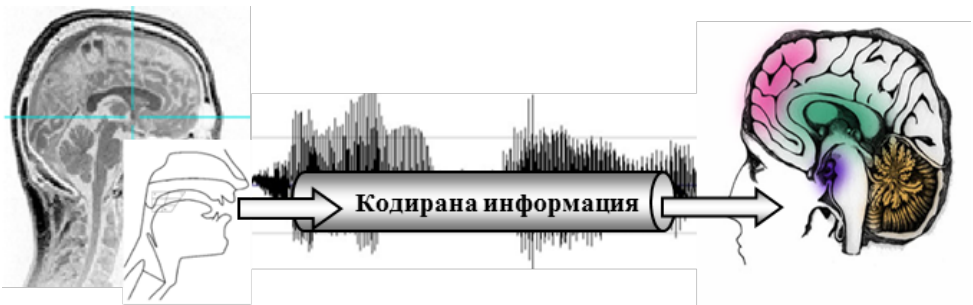
Представените изследвания наложиха разработване на комплексна автоматизирана система, която беше реализирана с участие на студенти по информатика, в рамките на редовни учебни дисциплини. Опитът ни говори, че включването на научноизследователски тематики в учебния процес на бакалавърско равнище спомага както за разбирането на ролята на информационните технологии за научни изследвания, така и на ограниченията на системите, използвани като „изкуствена интелигентност“.

Ключови думи: фоносемантика; емоционален звуков символизъм; разпознаване на емоции в текст; обучение по информатика

1. Увод

Предлаганата тук статия обобщава и анализира резултатите от публикуваните досега изследвания на авторите по темата. Методиката, обобщена тук, се основава на проследяване на резултатите в много области, например когнитивна наука, изследване на емоции, лингвистика и т.н., довели до направените догадки за връзката между емоция и фонемно съдържание на изказваното. На фиг. 1. е показана обща схема на обмен посредством говор, залегнала в направените изследвания. Човешкият слухов апарат е генетично устроен и допълнително трениран при контактите в социума така, че да разграничава спектри на произнасяните от друг човек честоти и амплитуди, да декодира фонетичния състав на изказаното, да разграничава думи и фрази и да извлича смисъл. Всички изследвания

в специализирани области показват, че освен смисъл слушателят декодира и емоция. Нататък в разсъжденията е прието, че в сигнала от говорна реч са закодирани както смисъл, така и емоция.



Фигура 1. Обща схема на комуникация чрез говор, залегнала в основата на изследванията

Прозодичните характеристики на речта – интонация, ритъм, паузи, скорост на промените в основната честота и т.н., които са от значение за предаване на емоция, са сравнително добре изучени и се прилагат в системите за разпознаване на емоция в говор. Тези системи обичайно не разпознават фонетичния състав и думите в аудиосигнала.

При разпознаване на емоции в текст проблемът е значително по-сложен и в много малка степен решен, защото текстът няма прозодични характеристики, например интонация. Системите за разпознаване на изразено отношение (opinion mining) работят с вход от текст и обичайно използват списъци от ключови думи, препинателни знаци, емотикони и т.н. Те са приложими предимно за класификация на потребителски отзиви за даден продукт или към кратки текстови обмени в социалните мрежи. Когато се търси каква е емоцията, кодирана в по-дълъг текст, например на новини, се подразбира, че читателят я „декодира“ от смисъла на написаното. „Сигнал“ за емоция обаче би могъл да бъде закодиран и във фонетичния състав на изложението. Такава хипотеза не следва да се проверява, преди да бъде направен анализ на наличните до момента резултати в различни научни области.

2. Резултати от други научни области

В обобщените тук публикации (Slavova 2019; Andonov & Slavova 2019; Slavova 2020; Slavova & Andonov 2021; Slavova 2021; Slavova &

Andonov 2022) е направен обстоен анализ на публикуваните резултати от други области, които имат пряко отношение към поредицата представени изследвания. Тук е обобщено най-съществено от направения обзор и са цитирани само източниците с особено голямо значение за разработената методика.

Изследването на емоции в последните десетилетия е интензивно, особено с оглед машинното им разпознаване. Научните изследвания показват, че влиянието на емоциите е на всички нива – физиологично, психическо и когнитивно. Два основни модела на емоции се използват понастоящем. Първият е дискретен, в който на емоциите са дадени имена (радост, тъга, гняв и т.н.), чийто брой и състав е променлив според авторите на изследванията. Вторият (известен със съкращението VAD – Valence-Arousal-Dominance) е непрекъснат и представя емоционалните състояния с разположението им в пространство, най-често по две оси. Едната, наричана Валентност (Valence) показва дали и в каква степен емоцията е положителна, т.е. приятна, или е отрицателна, т.е. неприятна, а втората е Възбуда (Arousal) и показва дали и в каква степен емоцията е свързана с възбудено психическо състояние. Представените тук изследвания на авторите изучават емоционална валентност на текстове в термините на втория модел поради неговата универсалност и наличието на резултати за неговите предимства (по-подробен анализ в Slavova 2019; Slavova 2020).

Изследванията на мозъчна дейност показват, че емоционалните състояния пораждат ясна активност върху мозъчната кора. Например има неезикови области на кората, реагиращи само на отрицателни думи. От особено значение за представения подход е, че при четене на текст мозъкът „автоматично включва“ в обработката и аудиторната част от кората (свързана със слуховата модалност) и прочетеното се обработва подобно на вход от говорна реч.

Множество изследвания на езици (руски, английски, немски и др., поезия и проза) достигат до извода, че съществува връзка между емоция и език. Езиковеди и психолингвисти откриват връзки на отделни емоции с определени фонеми, намират зависимости между позициите на фонемите в думата и нейната емоционална натовареност и т. н. В последните десетилетия се появиха резултати в областта на „звуковия символизъм“, който, най-общо, изследва връзката между фонетичния състав и смисъла на думите (по-подробен литературен анализ на тази област е направен в изброените авторски изследвания). Трябва да се подчертае, че разкритите връзки не се наблюдават пряко, а за откриването им са прилагани статистичес-

ки методи над корпуси от езикови данни. Особен интерес от гледна точка на методиката, представена тук, е изследването (Kawahara & Shinohara 2012), разкриващо, че двойките гласна – съгласна, представени изолирано като акустичен стимул, пораждат у обследваните лица усещане както за големина, така и за емоция. Този резултат доведе до предположението, че съществуват сублексикални части от говорната реч (съставни части на думите), които пренасят емоционална натовареност посредством фонетичното си съдържание. Това предопредели приложения нататък гещалт-подход към текста, т.е. текстът се разглежда като единно цяло, съставено от обвързани нива – изречения, думи и части от думи, като всяко ниво има значение за кодирането на емоционална валентност. Прилагането на гещалт-подход към текста в последното десетилетие доведе до съществени разкрития за връзката текст – емоция. Например в поредица корпусни изследвания на немския език се установява, че всички изброени по-горе нива на езика, включително сричките, са преносители на емоционална натовареност, виж например (Aryani et al. 2016; Ullrich et al. 2021).

Анализът на резултатите от други области, изложен по-горе накратко, доведе до първото изследване (Slavova 2019) на връзката между сублексикални единици и емоционална валентност на текст. В сравнение с цитираните резултати, в изследването се разглежда текст на естествен език, като думите не се разделят на срички, а в тях се изолират съзвучията от съгласна и гласна, т.е. в „обърнат“ ред в сравнение с тези, изследвани в (Kawahara & Shinohara 2012). Нататък в представените тук изследвания сублексикалните единици съставените от две фонеми – една съгласна и една гласна в прав и обратен ред, се наричат *бифони*.

3. Начално изследване

На фиг. 2. е показана общата схема на началното изследване (Slavova 2019). Двадесет текста от по една страница, подбрани от книги в оригинал на английски език, бяха обособени като материал за експеримент, данните за който се съхраняват локално в база. Експерименталните текстове бяха оценени за емоционална валентност от независими оценители. Текстовете бяха разградени до словоформи, като числителните и имената на хора, градове и др, умишлено са изключени от нататъшна обработка, защото тяхното използване в езика е неизбежно независимо от изразяваната емоция. На всяка словоформа е намерена фонетичната ѝ транскрипция и са изолирани бифоните от тип съгласна – гласна.

Например, *to discover* се произнася и съответно транскрибира като *ˈtuː dɪskʌvər* и се представя с посочения тип бифони като *tuː*, *dɪ*, *kʌ*, *və*.

Фигура 2. Обща схема на началното изследване

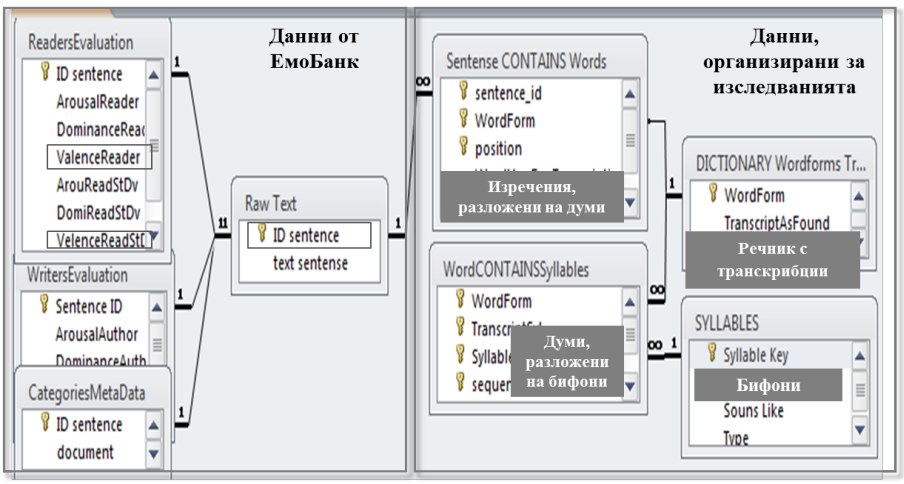
Възникнаха множество технически проблеми: при разработване на бифонен парсер, при запис на фонетични символи за обработка с SQL, при подаване на данни в продукт за статистическа обработка и т.н., дължащи се на използваните фонетични означения, разлики във фонетичните стандарти, типографски особености в изписването, различен брой знаци в един бифон и т.н., които са по-подробно коментирани и публикувани в (Andonov & Slavova 2019).

Резултатът показва наличие на корелация между честотите на определени бифони и оценките на читателите. Подробности са дадени в (Slavova 2019). Полученият обнадеждаващ резултат доведе до задълбочаване на търсенето на връзка между фонемния състав и емоционалната валентност на текст. Анализът показа, че е необходимо текстовете да се покрият по-плътно със сублексикални елементи. От особено значение за статистическата достоверност беше да се намерят обеми текстов материал, надлежно оценен за емоционалната валентност от читатели. Това доведе до търсене на достъпен и оценен за емоционална валентност езиков корпус.

4. Корпусен анализ

Корпусът ЕмоБанк[1] (Buechel & Hahn 2022) съдържа 10 000 надлежно оценени за валентност изречения на английски език от 7 различни жанра (вкл. романи, писма, заглавия от вестници и туристически проспекти). Корпусът беше локално разтоварен в релационна

база, в която бяха допълнително организирани необходимите за изследването данни (фиг. 3).



Фигура 3. Част от базата данни за корпусния анализ

Съставеният речник за фонетични транскрипции беше хомогенизиран по британския фонетичен стандарт и нататък попълван при автоматизирано запитване към Оксфордския онлайн речник посредством реализиран за тази цел програмен модул (виж също фиг. 8). Към момента таблицата за фонетични транскрипции съдържа стандартизирано фонетично представяне на около 20 000 словоформи.

Транскрибираните думи са разложени на бифони от разработения бифонен парсер и съхранени в таблица. За работата на парсера са съставени модели на бифоните от два типа – съгласна – гласна и гласна – съгласна. Тези модели, общо 1056 на брой, са получени като елементи на декартово произведение на множеството на всички гласни и това на всички съгласни (и в обратен ред), както съставът на двете множества е дефиниран в английската фонетика (Slavova 2020).

Сублексикалното равнище, което се изследва нататък, е от бифони (от двата типа). Изборът на такъв сублексикален „атом“ е в резултат на разсъждението, че за да се образуват думи в говорна реч е необходимо да се съчетаят гласна и съгласна. Всички езици имат и гласни, и съгласни, но сричкообразуването, както то се дефинира в езикознанието, се различава от език до език. При предложената

методика, представянето на изследваното сублексикалното равнище на фразата *Many stores will pack...*, която се транскрибира като *'meni 'stɔ:z wɪl pæk*, се получава при разлагане на съставни бифони така: *me, en, ni, tɔ:, ɔ:z, wɪ, ɪl, pæ, æk*.

Изреченията на корпуса бяха разложени на словоформи, а съответните им фонетични транскрипции – разложени на бифони. В десетте хиляди изречения на ЕмоБанк се откриват 805 различни бифона.

За статистическия анализ изреченията бяха филтрирани, като оценените като неутрални и тези с голямо разминаване в оценките от читатели бяха изключени, а останалите – групирани по валентност. Групите изречения бяха наречени „синтетични документи“, като целта на това групиране беше полученият в една порция текстов материал да бъде достатъчно дълъг, за да има вероятност всеки от наличните в корпуса 805 бифона да се срещне в него поне веднъж. Получиха се 22 „синтетични документа“ (фиг. 4А), всеки от които е еквивалентен на около три стандартни страници текст. Тези „синтетични документи“ не представляват смислено езиково съобщение, а са съвкупност от отделни несвързани изречения, които са близки по валентност.

На всеки синтетичен документ беше присвоена мярка за валентност – *валентност на документа*, изчислена като средна аритметична стойност на валентностите на изреченията, които го съставят. Регресионният анализ показа, че валентността на документите се апроксимира много добре от линейна функция (Slavova 2020), което позволи след това да бъдат прилагани линейна регресия и корелационен анализ. Беше търсена корелация между валентността на синтетичните документи и честотите на бифоните в текстовете на тези документи. Честотите, изразени като *коефициент на участие* $ScoBiph_{ij}$ на всеки бифон във всеки документ, беше изчислена като:

$$ScoBiph_{ij} = \frac{NBiph_{ij}}{NWord_j}, \tag{1}$$

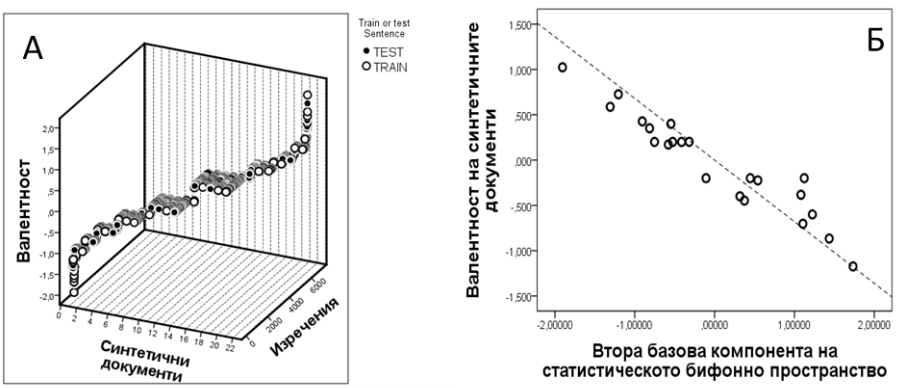
където $NBiph_{ij}$ е броят на появявания на i-тия бифон в j-тия документ, а $NWord_j$ е броят на транскрибираните и разложени думи в този документ.

Теглото $WBiph$ на i-тия бифон беше изчислено посредством наличната корелация между коефициентите му на участие, изчислени съгласно (1), в поредицата от синтетични документи и валентността на тези документи:

$$WBiph_i = r(ValenceDoc_{j=1,k}, ScoBipi_{i,j=1,k}), \qquad (2)$$

където r е корелационният коефициент на Пирсън между валентностите $ValenceDoc_{j=1,k}$ на синтетичните документи (k на брой) и коефициентите на участие $ScoBiph_{i,j=1,k}$ на i-тия бифон в тези k синтетични документа.

Честотите на много бифони в положителни изречения и в отрицателни изречения се различават съществено. Многократното разделяне на двадесетте и две порции изречения на случайни по състав половини показа (фиг. 4А), че изчислената според (1) честота на всеки бифон в половината изречения (train) предсказва много добре каква ще е честотата му в другата половина изречения на синтетичния документ (Slavova 2020). По-нататък в цялостното изследване, обобщено тук, теглата на бифоните, т.е. корелационните коефициенти, се използват като коефициент на влияние на всеки бифон за тестване на връзката между емоционалния тон на текстове извън корпуса и техния сублексикален състав.



Фигура 4. А. Групиране на изреченията от корпуса по валентност в „синтетични документи“ с разделяне по случаен начин на „тренировъчна“ и „тестова“ половина. **Б.** Корелация на втората базова компонента на пространството от бифонни честоти с емоционалната валентност

Анализът на базовите компоненти на статистическото пространство от (относителни) честоти на бифоните показа, че втората базова компонента корелира (Pearson) с оценките за валентност на читателите на 0.96, т.е. тя е практически еквивалентна на емоционална

валентност, изразена с бифони (фиг. 4.Б; Slavova 2021). Резултатите, изведени от анализа на корпуса подкрепят изненадващо добре началната хипотеза.

Този резултат от корпуса потвърди, че и на сублексикално ниво се кодира информация за емоционална валентност. Откритите зависимости по никакъв начин не зависят от смисъла на изписаното, още повече че те са изведени на база на набори от необвързани изречения. Въпросът дали тези зависимости важат и за смислово свързан текст извън корпуса, е особено важен за изясняване.

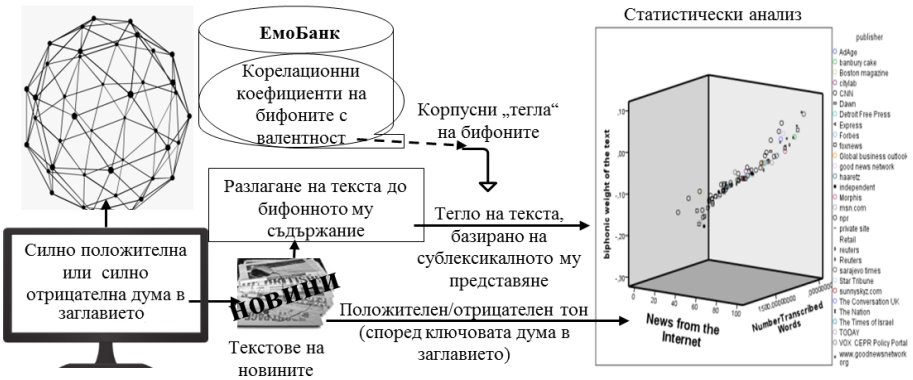
5. Прилагане на получените зависимости извън корпуса

От критично значение е да се провери дали получените зависимости между бифонен състав и емоционална валентност са валидни и извън ЕмоБанк. От текстовия материал в корпуса бяха изчислени корелационните коефициенти (Pearson) на бифоните с валентността и приложени върху текстове извън корпуса. Коефициентите варират по стойност от −0.8 до +0.8, като знакът показва дали са свързани с положителна или отрицателна валентност. Примери за силно отрицателно корелирани с валентността бифони са: *gʌ*, *nəʊ*, *nɒ*, *əʊn*, *feɪ*, *uːm*, *ɔːtʃ*, *ʌd*, *ɔɪd*, *vəʊ*, а за силно положително корелирани с валентността бифони са: *uːtʃ*, *veɪ*, *əʊg*, *eʃ*, *dʒɔɪ*, *ve*, *ʊər*, *jɔː*. Пълен списък на бифоните, които имат статистически значима корелация с емоционалната валентност в ЕмоБанк е преставен в (Slavova 2020).

Необходимо беше да се намерят достатъчно на брой сравнително кратки, смислово завършени и надлежно оценени за валентност текстове. Наличните интернет ресурси със свободен достъп не предлагат подобен оценен за валентност текстов материал. Това наложи търсенето на друг подход за определяне на емоционалния тон на текста.

Достатъчно кратки и смислово завършени са новините в интернет. В направените изследвания (Slavova & Andonov 2021, 2022) са използвани текстове на интернет новини, публикувани от световно известни англоезични агенции (CNN, Fox News, The Independent). Известно правило в журналистиката е заглавието да говори достатъчно ясно за съдържанието. Може да се допусне, че новините с негативен емоционален тон имат в заглавието си негативна дума и – симетрично. Наличните речници за емоционално оценени думи в английския позволиха съставяне на списък със 100 най-силно отрицателни (horrible, kidnapped, terror и т.н.) и 100 най-силно положителни (love, peace, happiness и т.н.) думи, които бяха използвани като ключови думи при търсенето на заглавия. Бяха направени две извадки – извадка А от 40 случайно разтоварени по ключова дума

новини и извадка B – от 100 (фиг. 5).



Фигура 5. Обща схема на изследването за прилагане на зависимостите извън корпуса

С извадка A се целеше да се провери в дали дума с определена силно изразена валентност в заглавието говори за емоционалния тон на съдържанието. Извадката беше подложена на оценяване от читатели. Резултатът показа близо 97% съвпадение между оценките на читателите за емоционален тон на съдържанието и емоционалната валентност на дума в заглавието (подробности в Slavova & Andonov 2021, 2022). Това позволи нататък да се извърши теглене от интернет на новини автоматично, по съставения списък с ключови думи, като се приеме с достатъчно висока вероятност, че ключовата дума в заглавието съответства на емоционалния тон на текста.

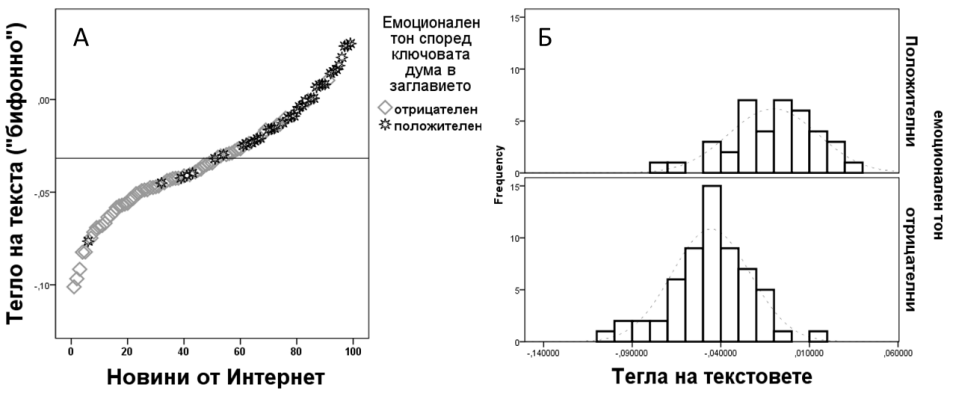
Както е показано на фиг. 5, текстове от новините бяха сублексикално представени като бифонен състав. На всеки бифон беше присъдено *тегло*, равно на корелационния коефициент на бифона с валентността в ЕмоБанк, изчислено по формула (2). Мярката, която бе използвана, за да се оцени фонемната индикация за валентност на текста, е *фонемното тегло* $WNews$ на всеки текст, изчислено с помощта на теглата $WBiph$ на бифоните, от които е съставен текстът:

$$WNews_j = \frac{\sum_i n_i.WBiph_i}{NBiph_i} \tag{3}$$

където $WNews_j$ е фонемното тегло на текста, $WBiph_i$ е теглото на i-тия бифон, изчислено по формула (2), n_i е броят появявания на

i-тия бифон в j-тата новина, а сумата от всички бифонни тегла е нормализирана спрямо броя на бифоните $NBiph_i$ в декомпозирания текст на j-тата новина.

Статистическият анализ показа, че теглата на текстовете, изчислени на база на техния сублексикален бифонен състав, са свързани с емоционалния тон на новината, както е показано на фиг. 6 (Slavova & Andonov 2022).

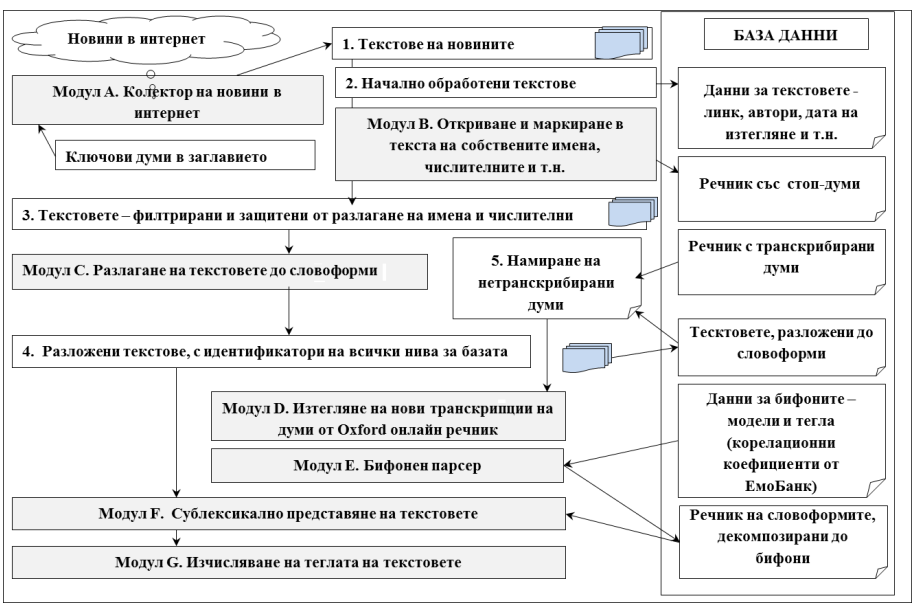


Фигура 6. А. Тегла на текстовете от новините в извадка Б и емоционалният им тон. **Б.** Хистограми на теглата на текстовете с отрицателен и с положителен емоционален тон

За смислово завършени текстове, дълги около страница, откритите на база корпусния анализ зависимости между сублексикалния състав и валентността оказват ясно изразено влияние. Трябва да се подчертае, че текстът се разглежда единствено като фонемен състав и написаното не се разглежда като смисъл. Това показва, че фонемното съдържание на думите не е напълно случайно и поставя въпроси за генезиса на думите като фонетичен състав. Направените анализи на същите текстове на базата на речници с емоционално оценени думи показаха, че използваните думи, като речников състав, не предсказват добре емоционалната валентност на текста. Това не е изненадващо за разглежданите текстове, защото използването на емоционално натоварени думи в новините по правило се избягва. Може да се предположи, че в едно завършено цялостно изложение авторът, вероятно несъзнателно, мислено генерира такива средства и начини на изказ, че неговото отношение личи в цялостния фонемен състав. Обяснение за подобен феномен трябва да се търси в когнитивна наука.

6. Реализация на модулна система като част от учебния процес

Събирането и обработването на данните, необходими за изследванията, описани тук накратко, изисква разработване на достатъчно сложна система. Обичайно подобни трудоемки системи се реализират от много участници, често с включване на докторанти в разработките. Системата, както тя е използвана в последното изследване (Slavova & Andonov 2022), беше реализирана поетапно в течение на няколко последователни години, като в разработката бяха включени студенти от бакалавърската степен по информатика в Нов български университет (НБУ). Както е илюстрирано на фиг. 7, системата се състои от две основни компоненти – база данни (създадена с SQLite) и текстообработващи модули (написани на Python). Системата работи при обмен между двете компоненти. Входът и изходът на всеки от модулите са организирани със съхраняване във файлове, за да могат междинните резултати да се подават към следващ модул за запис в базата, а и процесът на обработка на текстовия материал да се проследи с лекота.



Фигура 7. Структура на модулната система, реализирана съвместно със студенти от бакалавърска програма по информатика в НБУ

Модул A „събира“ текстове на новини по ключова дума в заглавието. Текстовете се почистват от линкове и се съхраняват, а данните за тях се записват в таблица. Модул B е текстов парсер, ползващ готов продукт за откриване на имената, числителните, датите, топонимите и т.н. Модулът отбелязва в текста откритите имена и др. със специален знак, за да не бъдат те вземани под внимание при разлагането до бифони, като попълва в базата речник със стоп-думи. Модул C разлага текстовете до изречения и до словоформи, съответно с идентификатори и с отбелязана поредност (фиг. 8), като записва резултата както във файл за контрол, така и в базата. Заявка изтегля думите, които не са намерени в базата, в речника с транскрибирани думи и ги подава на входа на модул D. Модул D прави справка с Oxford онлайн речник и записва намерените транскрипции в базата. Модул E е бифонният парсер, той разлага новонамерените думи на бифони и ги записва в съответната таблица в базата. Модул F, базиран на заявки, генерира цялостното сублексикално представяне на текстовете като бифони. Модул G пресмята теглата на декомпозираните текстове като сума от теглата на участващите в тях бифони, нормализирана според дължината на текста.

0	Volunteer visits s41	We			wi:
1	Volunteer visits s41	have			hæv
2	Volunteer visits s41	so			səʊ
3	Volunteer visits s41	much			mʌtʃ
4	Volunteer visits s41	fun			fʌn
5	Volunteer visits s41	just			dʒʌst
6	Volunteer visits s41	talking			ˈtɔːkɪŋ
7	Volunteer visits s41	about			əˈbaʊt
8	Volunteer visits s41	different			ˈdɪfrənt
9	Volunteer visits s41	things			ˈθɪŋz
10	Volunteer visits s41	different			ˈdɪfrənt
11	Volunteer visits s41	subjects			ˈsʌbdʒɪkts
12	Volunteer visits s41	Toia	Toia	name	
13	Volunteer visits s41	said			sed
0	Volunteer visits s42	Shes			ʃiːz
1	Volunteer visits s42	very			ˈveri

78534 of 79826 | No Filter | Search

We	0	wi:	1	wi:	wi:	biph0959
have	1	hæv	1	hæ	hæ	biph0469
have	1	hæv	2	æv	æv	biph0049
so	2	səʊ	1	səʊ	səʊ	biph0749
much	3	mʌtʃ	1	mʌ	mʌ	biph0647
much	3	mʌtʃ	2	ʌtʃ	ʌtʃ	biph0933
fun	4	fʌn	1	fʌ	fʌ	biph0443
fun	4	fʌn	2	ʌn	ʌn	biph0927
just	5	dʒʌst	1	dʒʌ	dʒʌ	biph0255
just	5	dʒʌst	2	ʌs	ʌs	biph0931
talking	6	ˈtɔːkɪŋ	1	tɔː	tɔː	biph0764
talking	6	ˈtɔːkɪŋ	2	ɔːk	ɔːk	biph0152
talking	6	ˈtɔːkɪŋ	3	kɪ	ky	biph0604
talking	6	ˈtɔːkɪŋ	4	ɪŋ	yŋ	biph0980
about	7	əˈbaʊt	1	əb	əb	biph0287

ecord: 1 of 187171 | No Filter | Search

Фигура 8. Илюстрация на междинни резултати от обработката на едно изречение на интернет новините

Показаните на фиг. 7 модули са разработвани с активно участие на студентски екипи от последователни випуски. В учебната програма са включени курсовете „Практика по програмиране и интернет технологии“ и „Практика по програмиране и реализация на бази данни“, в рамките на които беше реализирано участието на студентите. Методиката в посочените практики изисква студентите, в групи от по 3 – 5 души, да работят в екип, за да реализират разработка

по задание от преподавателя. В голямата си част студентските екипи завършиха работата си с работещи програмни модули и подробна документация за разработката си.

Областта на изследване не е близка до компютърната наука и задачите, както и целта, не са интуитивно разбираеми за неспециалисти в обработка на естествен език. Повечето студенти, проявили интерес към тематиката, се запознаха с препоръчани статии от специализирани области. Работата изискваше студентите да участват в продължителни срещи за проследяване на резултата от изпълнение, откриване и обсъждане на проблемите на всички нива. Както може да се съди по илюстрираните на фиг. 8 междинни резултати, работата налага преодоляване на множество нетипични проблеми, свързани както със спецификата на обработка на фонетични означения, така и с представянето на цялостното сублексикално ниво. При достатъчно усилие и контрол от страна на преподавател студентите в бакалавърска степен завършват успешно сложни разработки, намират оригинални решения и проявяват интерес към резултатите в контекста на научни изследвания.

7. Дискусия

Приложената методика за представяне на сублексикалното ниво и направеният статистически анализ на текстове от новини показват стабилно поведение на връзката между бифонния състав на текста и неговата емоционална валентност. За момента не знаем на какво се дължи този ефект. Множество учени от областта на еволюция на езика и на звуков символизъм проявиха интерес към резултата, но до момента не ни е известно специализирано обяснение на наблюдавания феномен.

Интересно е да се потърси подобна зависимост в други езици, включително в българския език. Не разполагаме обаче с корпуси на оценени за валентност текстове, за да предприемем изследване. Изясняването на въпроса дали подобни зависимости са валидни в други езици, както и откриването на тяхната причина остават научни догадки и хипотези, коментирани по-подробно в изброените публикувани авторови изследвания. Предприехме и онлайн експеримент за изследване и оценяване на текстове, които са извадки от романи на английски, дълги около две страници. До момента нямаме статистически стабилен резултат за наличие на връзка между читателските оценки на текстовете и техния сублексикален състав.

Получените до момента резултати показват, че зависимостите фонемен състав – емоционална валентност се проявяват в дълги около страница смислово завършени текстове.

Благодарности

В публикуваните изследвания, обобщени тук, са изказвани писмено поименна благодарност на много от студентите, участвали в разработването на системата. Не е възможно, за съжаление, да бъдат изброени всички. Тук изказваме благодарност на студентите, работили с отдаденост и успех на последните етапи: Ивон Цонкова, Росен Стефанов, Ерик Балиов и Мартин Константинов.

NOTES

1. EmoBank: https://github.com/JULIELab/EmoBank

REFERENCES

ANDONOV, F., SLAVOVA, V., 2019. Technical aspects of the extraction of phonological features for emotion recognition in texts. *International Journal Information Content & Processing*, vol. 6, no. 2, pp. 121 – 130.

ARYANI, A., et al., 2016. Measuring the basic affective tone of poems via phonological saliency and iconicity. *Psychology of Aesthetics, Creativity, and the Arts*, vol. 10, no. 2, p. 191.

BUECHEL, S., HAHN, U., 2022. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv*, arXiv:2205.01996. https://arxiv.org/abs/2205.01996

KAWAHARA, S., SHINOHARA, K., 2012. A tripartite trans-modal relationship among sounds, shapes and emotions: A case of abrupt modulation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34.

SLAVOVA V., ANDONOV, F., 2021. How deeply are emotions encoded in language communication and is this detectable in text. *International Journal Information Theories and Applications*, vol. 1, no. 3, pp. 271 – 299.

SLAVOVA, V., ANDONOV, F., 2022. Bad news or good news when recognizing emotional valence using phonemic content. In *21st International Symposium INFOTEH-JAHORINA*, IEEE, pp. 1 – 6.

SLAVOVA, V., 2019. Towards emotion recognition in texts–a sound-symbolic experiment. *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE)*, vol. 7, no. 2, pp. 41 — 51.

SLAVOVA, V., 2020. Emotional valence coded in the phonemic content–Statistical evidence based on corpus analysis. *Cybernetics and Information Technologies*, vol. 20, no. 2, pp. 3 — 21.

SLAVOVA, V., 2021. On the revealing the emotional valence in communication by text. *Procedia Computer Science*, vol. 192, pp. 1514 – 1523.

ULRICH, S., et al., 2021. Poetry study – On the relation between the general affective meaning and the basic sublexical, lexical, and inter-lexical features of poetic texts–A case study using 57 Poems of HM Enzensberger. *Sound Matters*, vol. 141.

INVESTIGATION OF THE RELATIONSHIP BETWEEN THE PHONEMIC COMPOSITION AND THE EMOTIONAL VALENCE OF A TEXT

Abstract. This paper is based on the results of published authorial research on the relationship between the sound composition of language and emotion, leading to the conclusions and generalizations drawn here. In order to uncover deeper relationships between language and emotion, a methodology and corresponding automated system was developed that allowed the application of statistical methods over the phonemic composition of large volumes of textual data. So far, the English data and the results of the applied statistical methods confirm the existence of dependencies between phoneme composition expressed as vowel-consonant pairs and emotional valence (positive-negative emotion) for news texts from the Internet. The presented research necessitated the development of an automated system. The system was implemented with the participation of computer science students, within the framework of full-time courses. Our experience suggests that incorporating research topics into the undergraduate curriculum helps to understand both the role of information technology for research and the limitations of systems used as "artificial intelligence".

Keywords: phonosemantics; emotional sound symbolism; emotion recognition in text; education in informatics

✉ **Dr. Velina Slavova, Prof.**
ORCID: 0000-0003-3881-8820
New Bulgarian University
Sofia, Bulgaria E-mail: vslavova@nbu.bg

✉ **Dr. Filip Andonov, Ass. prof.**
ORCID: 0000-0002-7237-5042
New Bulgarian University
Sofia, Bulgaria E-mail: fandonov@nbu.bg