**МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ**

**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЛІНГВІСТИЧНИЙ УНІВЕРСИТЕТ**

# «КОМП'ЮТЕРНА ЛІНГВІСТИКА: СУЧАСНЕ ТА МАЙБУТНЄ»

## Матеріали міжнародної науково-практичної конференції

**КИЇВ 2012**

**ОРГАНІЗАТОРИ:** Київський національний лінгвістичний університет
Лабораторія комп'ютерної лінгвістики

**РЕКОМЕНДОВАНО ДО ВИДАННЯ** рішенням Вченої ради КНЛУ
*(Протокол № 6 від 30 січня 2012 р.)*

**Технічний редактор, комп'ютерна верстка** Льон Олександр Васильович

**Комп'ютерна лінгвістика: сучасне та майбутнє.** Матеріали Міжнародної науково-практичної конференції – К.: КНЛУ, 2012.– 52 с.

Збірник сформовано за матеріалами Міжнародної науково-практичної конференції: «Комп'ютерна лінгвістика: сучасне та майбутнє» (23-24 лютого 2012 р., Київський національний лінгвістичний університет м. Київ).

**Валентина Перебийніс, Тетяна Бобкова**
Київський національний лінгвістичний університет

# Історія лабораторії комп'ютерної лінгвістики КНЛУ

Лабораторія комп'ютерної лінгвістики КНЛУ була створена у 2002 році з ініціативи Сергія Микитовича Назарова, тодішнього проректора з наукових питань КДЛУ. Напрямок досліджень лабораторії – комп'ютерна навчальна лексикографія, укладання перекладних навчальних словників. За 10 років у лабораторії укладено серію англо-українських й україно-англійських навчальних словників для початкового та середнього рівня викладання англійської мови – всього 4 словники, при цьому словники першого рівня витримали три видання у паперовому форматі, а в комп'ютерному форматі англо-український словник має звуковий супровід.

Крім перекладних словників у лабораторії укладено словники-довідники «Труднощі англійського слововживання для українців» та «Морфологія англійського дієслова: система та функціонування», де наводяться дані про словозмінну парадигму англійського дієслова та про вживаність кожної словозмінної форми біля 300 найуживаніших дієслів у сучасній художній прозі, драмі, наукових та суспільно-політичних текстах. Всі ці словники можна побачити на книжковій виставці, організованій бібліотекою КНЛУ. Ще однією фундаментальною працею лабораторії є навчально-методичний комплекс, який складається з англо-українського та україно-англійського навчальних словників з методичними коментарями для другого рівня навчання англійської мови. Він існує в комп'ютерному форматі, планується його видання у паперовому форматі. В лабораторії укладаються не лише навчальні словники. Так, на замовлення московського видавництва «Астрель» укладено «Російсько-український розмовник» і «Російсько-український словник» на 210 тисяч статей. Т.В. Бобкова уклала англо-українсько-російський словник термінів з комп'ютерної лінгвістики, що містить переклад і тлумачення близько 1000 термінів.

Значну увагу приділяють в лабораторії використанню та створенню корпусів текстів. Так, на матеріалі мільйонного корпусу документів НАТО укладено ЧС на замовлення Міністерства освіти і науки України (2007 р.). Розроблено кілька корпусів: Багатомовний корпус субтитрів до кінофільмів – розробник К.М. Лебедєв; В.О. Коломієць працює над створенням навчального корпусу англійських текстів, написаних носіями української мови (UCLE); В. Орел створює корпус анотацій до англійських статей з комп'ютерної лінгвістики.

За останні п'ять років співробітники лабораторії виступили з доповідями на 15 конференціях, зокрема на дистанційному занятті-семінарі з дистанційного навчання іноземним мовам, організованим Львівським університетом безпеки життєдіяльності.

Існує ще одна сфера діяльності лабораторії комп'ютерної лінгвістики – забезпечення навчального процесу на відділенні прикладної (комп'ютерної) лінгвістики, яке відкрилося, знову з ініціативи Сергія Микитовича Назарова майже одночасно з лабораторією. Ця робота була пов'язана з великими труднощами, оскільки в Україні склалася дуже несприятлива ситуація для структурно-математичної лінгвістики. В середині минулого століття ця галузь мовознавчих наук набула значного розвитку в колишньому СРСР, у тому числі й в Українській РСР. Працювали

відділення в кількох університетах, а також в АН УРСР, відкрили кілька рад по захисту дисертацій, друкувалася значна кількість наукової літератури. Але на початку 80-х років ситуація різко змінилася: було закрито аспірантуру, припинено підготовку кадрів з цього фаху, припинилося друкування наукової літератури, ледве животіли відділення структурної лінгвістики в університетах.

В інших же республіках СРСР розвиток математичної лінгвістики продовжувався, тому Україна значно відстала і від цих республік, і від світового рівня. І коли на початку XXI ст. в Україні зіткнулися з гострою проблемою комп'ютерного опрацювання інформації, знову почали відкривати відділення (але не аспірантуру!), 20 років відставання далися взнаки: не було кадрів, наукової літератури, були втрачені традиції, отже доводилося починати з нуля. У такій самій ситуації був і наш університет. Крім лабораторії комп'ютерної лінгвістики нікому було очолити цю роботу. Почалися пошуки фахівців, які могли б викладати, літератури для забезпечення навчального процесу, концепції відділення. Оскільки відкрилося воно під назвою «Прикладна лінгвістика» то його не співвідносили з колишньою структурно-математичною лінгвістикою і спочатку визначали напрямок підготовки як будь-яке опрацювання інформації: документалістика, бібліотечна справа тощо. І тільки після аналізу навчальних планів відповідних відділень російських навчальних закладів та вимог до комп'ютерного опрацювання інформації стало зрозуміло: треба готувати спеціалістів з математичної лінгвістики, які вміли б програмувати лінгвістичні задачі. Цим шляхом і пішли. Треба відзначити, що таке спрямування є унікальним в Україні: ніякий інший заклад не готує лінгвістів-програмістів.

З самого початку розвивали у студентів творчий підхід до навчання, яке б поєднувалося з дослідницькою роботою. Починаючи з другого курсу студенти пишуть курсові роботи, які є не рефератами з опрацьованої наукової літератури, а самостійними дослідженнями. Життя показало, що ми на правильному шляху. Наші студенти з успіхом виступають на наукових студентських конференціях, одержують дипломи на конкурсах молодих науковців, а Т.В. Бобкова нагороджена почесною грамотою міністерства за керівництво науковою роботою студентів (2008 р.). Студенти беруть активну участь у науково-дослідній роботі лабораторії, проходять в ній практику, за результатами своїх досліджень друкують наукові статті, їх можна побачити на книжковій виставці в бібліотеці.

Важливе місце в роботі лабораторії становить забезпечення студентів навчальною літературою, враховуючи те, що відповідні наукові праці, посібники та підручники друкувалися в середині минулого століття й зараз є раритетами, причому це переважно російськомовні видання. Тому в лабораторії було створено кілька посібників: «Статистичні методи для лінгвістів» (В.І. Перебийніс), «Традиційна та комп'ютерна лексикографія» (В.І. Перебийніс, В.М. Сорокін), «Збірник лінгвістичних задач» (Т.В. Бобкова), підручник англійської мови для студентів-прикладників (В.О. Коломієць). Укладено три хрестоматії в електронному форматі: з формально-граматичного аналізу, математичної лінгвістики, лінгвістичного моделювання (В.І. Перебийніс, О.В. Льон). На сучасному етапі укладається посібник з математичної лінгвістики.

**Людмила Алексієнко**
Київський національний університет імені Тараса Шевченка

# Концепція структурної морфології української мови

Сучасна граматика потребує, з одного боку, поглиблення теоретичних узагальнень у сфері формальної будови мови на різних рівнях системи; іншого – диференціації, різновекторності, адекватній граматичним об'єктам і, відповідно, дисциплінам з їх предметом та понятійною базою.

Слово як центральна одиниця мовної системи та об'єкт морфології пов'язане практично з усіма підсистемами і має складну граматичну структуру – морфемну, словотвірну (у дериватах) та словозмінну.

Виробленню концепції цієї дисципліни сприяла багаторічна практика викладання морфології української мови на філологічному факультеті Київського національного університету імені Тараса Шевченка.

В основу триєдиного курсу морфології (морфеміка, словотвір, парадигматика) покладено теоретичні положення і дихотомії структурної лінгвістики, що забезпечує необхідну точність і формалізм як у теорії, так і в процедурній частині курсу. Це відомі положення про систему і структуру; мову і мовлення; інваріанти і варіанти мовних одиниць; синхронію та діахронію стосовно морфології та ін.

Структурний вектор опису морфології мови, зокрема української, потребує несуперечливої аксіоматики і точності в термінології та процедурах аналізу. Виконання таких завдань можливе за умов розумних й адекватних цим завданням компромісів. Так, у сучасній українській парадигматиці необхідними є компроміси щодо частиномовної класифікації (відносно займенників і дієприкметників); у питаннях грамемної структури деяких категорій (відмінка в іменних частинах мови; способу в дієслові); щодо кількості основ у дієслові; щодо типології парадигм, їх повноти / неповноти / обнуління тощо.

Важливим завданням парадигматики є реалізація аксіоми про парадигматичний і синтагматичний аспекти опису слова, що пов'язано з необхідністю розмежування категорійних і функціональних форм і забезпечує належний рівень точності й несуперечливості у морфології в цілому та парадигматиці зокрема.

Для процедурної частини парадигматики важливими є правила виведення основоформ (словозмінних основ), що виключає можливість графемного представлення лексем і словоформ і потребує їх фонемного запису. Інакше порушується формально-структурний принцип опису парадигматичних одиниць як інваріантно-варіантних структур та втрачається можливість визначення морфонем і модифікованих основ як засобів конструювання словоформ.

**Никанор Бабырэ**
Молдавский государственный університет

# О новом методе анализа динамики артикуляторного процесса и его применение к молдавской (бессарабской) речи румынского языка (по данным кинорентгенографирования)

Современные научные достижения должны внедрять, в том числе и словесники в целом, и фонетисты-экспериментalисты, в частности, при решении проблем общего, частного и сопоставительного языкознания.

Особое внимание в этом плане следует уделять вопросам исследования артикуляторной динамики молдавской (бессарабской) речи румынского языка, а также речи других языков: русского, гагаузского, болгарского, украинского и др.

Как и ряд других наук, языкознание, особенно при изучении фонетических проблем, оперирует двумя методами: субъективным и объективным. Среди экспериментальных приемов особое место занимает рентгенографирование (статический рентген) и кинорентгенографирование (динамический рентген). Первым русским ученым, получившим отчетливые снимки ротовых артикуляций звуков речи при помощи рентгена, был Р.Д. Енько, который опубликовал результаты своей работы в 1912 г. После него, особенно в послевоенное время, рентгенографирование получило широкое распространение и применение. Изучению звуков молдавской речи при помощи рентгенографирования посвящены работы Д.В. Бэдэрэу (молдавские монофтонги в сопоставлении с английским) и исследования Г.М. Гожина в области вокализма и консонантизма.

Благодаря развитию методик на основе приема кинорентгенографирования, стало возможным изучение соотношения между физиологическим и акустическим аспектами речи в их динамике. Для этой цели пользуются синхронной записью артикуляции на кинорентгеновском аппарате и акустического речевого сигнала, что дает возможность установить, как данная артикуляция коррелирует с соответствующим акустическим эффектом. С помощью кинорентгена можно достаточно объективно наблюдать как динамику гласных, дифтонгов, согласных, так и явления коартикуляции, т.е. совмещения артикуляции соседних звуков в словах и предложениях. Термин ***динамика*** употребляется в методике, как отмечает проф. Скалозуб Л.Г., в значении развития артикуляторных движений, необходимых для произнесения слогов и слов. Артикуляторный динамический анализ позволяет выбрать из всех кадров (с учетом длительности кадра 10 мс и паузой между кадрами 10 мс), приходящихся на данный звук (от 13-15 для кратких монофтонгов и до 18-19 для долгих монофтонгов и дифтонгов), те из них, которые относятся к основной фазе т.е. к основному типу монофтонга, дифтонга или согласного. Если учесть, что основная фаза, определяющая звук для гласных длится в среднем только одну треть артикулирования звука, а 2/3 занимают переходные моменты, то станет ясной ценность кинорентгенографирования и артикуляторного динамического анализа для определения эталонных значений фонем и их оттенков. Нами эта новая методика артикуляторной динамики впервые использована при исследовании монофтонгов и дифтонгов молдавской (бессарабской) речи. Учитывались движения следующих

органов речи: языка, нижней челюсти, увулы, подъязычной кости, задней стенки фаринкса; наблюдались также движения по вертикали и горизонтали нижних очертаний преларингальной зоны фаринкса.

Проанализированный способ артикуляторной динамики выявил определенную целостность, имеющую свои границы и вершины, что даёт полное основание считать артикуляцию монофтонгов /a/, /e/, /i/, /u/, /i/ (краткого) и дифтонгов /'ai/, /'au/, /'ei/ в потоке молдавской речи происходящей в соответствии с моделью слога.

Дифтонги образуют слог особого типа и строения, динамическая организация и артикуляторные признаки компонентов слога взаимообусловлены. Анализ дифтонгов также показал, что дифтонги /'ai/, /'au/ молдавской (бессарабской) речи являются скаленными, т.е. неравноценными по интенсивности, частоте основного тона и длительности его компонентов и падающими (нисходящими). Примером таких скаленных дифтонгов является /'au/, /'ai/ молдавской (бессарабской) речи. В процентном отношении длительность первого компонента дифтонгов (53,7 %) больше, чем второго (46,3 % ).

Данные исследования призваны способствовать реализации в Республике Молдова инженерно-лингвистических задач коммуникации и обработки информации, созданию диалоговых систем, автоматического перевода на иностранный язык с озвучиванием синтезатором устной речи. Такие промышленные линии будут способствовать созданию многоязычной автоматической телефонной связи. Согласно мнению академика Р.Г. Пиотровского, научно-технический потенциал Республики Молдовы и наши экспериментальные исследования, а также исследования академика А.Н. Попеску, позволяют надеяться, что задача создания румынского и румыно-язычного человеко-машинного диалога вполне решаема.

**Ася Бобкова**
Киевский национальный лингвистический університет

## Лексическое ядро языка избранной поэзии Иосифа Бродского

Традиция статистических исследований русского художественного текста насчитывает несколько десятилетий. Начало этим исследованиям положили составленные в 50-60 гг. XX в. частотные словари языка А.С. Пушкина, Ф.М. Достоевского, А.С. Грибоедова и М. Цветаевой. Сегодня в России традиции составления частотных словарей продолжают коллективы исследователей Института русского языка им. В.В. Виноградова, Курской лингвофольклористической школы, Томского государственного педагогического университета, Смоленского государственного педагогического университета. В частности, на кафедре истории и теории литературы Смоленского государственного педагогического университета создано 42 частотных словаря поэтов XIX-XX вв., «Войны и мира» Л. Толстого, «Доктора Живаго» Б. Пастернака [2]. На материале этих словарей выделена общая лексика, которая с наибольшей частотой встречается в текстах поэтов XIX-XX вв.

По произведениям И. Бродского составлено два частотных словаря: словарь сборника «Часть речи» [4] и собрания сочинений, размещенных в сети Интернет [3],

включающего более 570 стихотворений и поэмы. Однако отсутствует частотный словарь на материале сборника, подборкой стихов для которого занимался бы сам поэт. Известно, что И. Бродский скептически относился к идее посмертного избранного и неоднократно высказывался против подобных изданий, но в 1988 г. поэт составил проект собственного идеального сборника, положенного в основу издания «Избранное. Перемена империи».

Для определения лексического ядра словаря Бродского тексты 83 стихотворений указанного сборника были собраны в один массив и автоматически поделены на слова от пробела до пробела. В массив были включены названия стихотворений, посвящения, эпиграфы, иноязычные слова, арабские и римские числа. При этом под словом понималась единица текста от пробела до пробела, используемая поэтом в качестве слова: буквы, части слова, сокращения, сочетание морфемы и слова, варианты слов, сочетания слов и предложения. Для автоматического выведения словарных форм использовался пакет программ (РУТА-ПЛАЙ), разработанный в лаборатории компьютерной лингвистики Киевского национального университета им. Тараса Шевченко.

Известно, что определение частоты отдельных форм слов в тексте, их разграничение и сведение в лемму усложняется грамматической и лексико-грамматической омонимией, «связанной с сосуществованием в одной звуковой форме разных семантико-грамматических категориальных значений, воплощенных в разных лексических единицах – омонимических словах разных частей речи» [1, с. 84]. Снятие лексико-грамматической омонимии проводилось вручную с помощью контекстного анализа. В результате лемматизации и снятия омонимии установлено, что список словоформ, используемых И. Бродским в текстах избранного, составляет 11909 единиц. Словарь языка избранного включает 8078 слов.

Для выделения лексического ядра языка избранной поэзии И. Бродского использовалась методика, разработанная в Смоленском педагогическом университете. На материале словарей поэтов XIX – XX вв. исследуются 30 наиболее частотных слов каждого автора. В результате сопоставления списков этих слов установлено, что «у большинства поэтов XIX в. (Грибоедов, Рылеев, Лермонтов, Баратынский, Тютчев) среди наиболее частотных мало слов, обозначающих природу. У них втрое больше слов, обозначающих человека, части его тела, элементы его духовного мира» [2]. В соотвествии с этой методикой из словаря И. Бродского были выбраны 30 самых частотных слов, обозначающих понятия.

| | | |
|---|---|---|
| *вещь – 70* | *черный – 38* | *вода – 32* |
| *время – 62* | *пространство – 36* | *воздух – 32* |
| *глаз – 61* | *ночь – 36* | *новый – 31* |
| *жизнь – 60* | *конец – 35* | *город – 30* |
| *лицо – 60* | *место – 35* | *стена – 30* |
| *тело – 48* | *мысль – 35* | *окно – 30* |
| *день – 41* | *вид – 34* | *голова – 30* |
| *рука – 41* | *человек – 33* | *вечер – 29* |
| *свет – 40* | *тень – 33* | *море – 28* |
| *мир – 39* | *земля – 33* | *друг – 28* |

Показатели частоты этих слов в тексте говорят об их значимости для поэта и находятся в пределах от 70 (*вещь*) до 28 (*море* и *друг*). Для выделения групп слов, объединяемых общим значением, были использованы данные толкового словаря С.И. Ожегова. Всего выделено три группы, внутри каждой из них значение слова объясняется с помощью общего понятия или другого слова этой же группы.

Половина из 30 самых частотных слов входит в группу с общим значением «пространство» (*вещь, мир, пространство, место, свет, черный, тень, земля, вода, воздух, город, вид, стена, окно, море*). Почти треть списка (30%) составляют слова, «обозначающие человека, части его тела, элементы его духовного мира» (*глаз, лицо, тело, рука, человек, голова, жизнь, мысль, друг*). Пятую часть списка (20% общего количества) составляют слова с общим значением «время» (*время, день, ночь, конец, новый, вечер*).

Таким образом, лексическое ядро языка избранной поэзии И. Бродского составляют наиболее часто употребляемые поэтом слова, объединяемые в группы с общими значениями «пространство», «человек» и «время». При этом наиболее значимой, включающей половину из 30 самых частотных слов, является группа «пространство» в отличие от словарей поэтов XIX-XX вв., в которых преобладают слова, обозначающих человека.

Литература

1. Дарчук Н.П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник. – К.: Видавничо-поліграфічний центр «Київський університет», 2008. – 351 с.
2. Баевский В.С., Романова И.В., Самойлова Т.А., Смагина О.А. О мере близости частотных словарей русских поэтов XIX- XX веков // Математическая морфология. Электронный математический и медико-биологический журнал. – Т. 3, вып. 2. – 1999. – С. 119-131.
http://www.smolensk.ru/user/sgma/MMORPH/TITL.HTM
3. Орлова О.В. Компьютерный анализ поэтического текста и моделирование ассоциативно-смыслового поля ключевого концепта творчества автора// Открытое дистанционное образование. – Выпуск 1(9), 2003. – С. 57-60.
4. Частотный словарь языка сборника «Часть речи» И. Бродского http://slovari.ru/default.aspx?s=0&p=5316&0a0=1894

**Татьяна Бобкова**
Киевский национальный лингвистический університет

## Составление частотного словаря избранной поэзии Иосифа Бродского

Статистическое изучение художественного текста предполагает составление частотных словарей авторов с целью определения их предпочтений и редких словоупотреблений. Сегодня создание таких словарей, полностью описывающих язык автора, является отдельной отраслью статистической лексикографии. В России за последнее время для различных исследований составлено более 40 словарей русских

авторов: А.И. Куприна, А. Блока, Ин. Анненского, Кирши Данилова, И.С. Шмелёва, А.П. Чехова, А. Введенского, М.Ю. Лермонтова, Э.Я. Мандельштама, А. Ахматовой, Б. Пастернака и других авторов [1]. В том числе составлено два частотных словаря по произведениям И. Бродского: словарь сборника «Часть речи» [4] и собрания сочинений, размещенных в сети Интернет [2]. В этом смысле представляется целесообразным для статистического исследования словаря поэта учитывать его предпочтения в выборе текстов. Поэтому в качестве материала для частотного словаря был выбран сборник «Избранное. Перемена империи», в основу которого положен проект самого И. Бродского.

Для составления частотного словаря И. Бродского тексты стихотворений указанного сборника были собраны в один массив и автоматически поделены на слова от пробела до пробела. В качестве слова, как и в большинстве работ по статистической лексикографии, была принята последовательность знаков между двумя пробелами или пунктуационными знаками [3]. В отдельных случаях, связанных со спецификой поэтического текста, таким разделителем между двумя единицами текста служил конец строки:

“Внешность их не по мне.
Лицами их привит
к жизни какой-то не-
покидаемый вид.”

В массив текста были включены названия стихотворений, посвящения, эпиграфы, иноязычные слова, записанные латиницей или кириллицей, арабские, римские числа и следующие виды авторских слов:

1) отдельные буквы: «что всякая точка в пространстве есть точка “а” и нормальный экспресс, игнорируя “b” и “c”…»,

2) сокращения слов: «и населенье в субботу выстраивалось гуськом, как караван в пустыне, за сах. песком», «см. светило, вставшее из вод»;

3) части слова: «И лобзают образа с плачем жертвы обреза...»

4) морфемы + слово (через дефис): «в мозгу вовсю разгорается лампочка анти-света»;

5) варианты слова: «мостовую пересекаешь с риском быть за{п/к}леванным насмерть»;

6) сочетания слов через дефис (по модели слово-приложение, сложное слово): «бродят в осоке лошади-пржевали», «прими от меня эту рифмо-лепту», «Что есть главный закон тюрьмо-динамики»;

7) предложения, записанные через дефис: «Добрый вечер, проконсул или только-что-принял-душ».

Для лемматизации использовался пакет программ (РУТА-ПЛАЙ) с последующим разрешением грамматической и лексико-грамматической омонимии вручную с использованием конкорданса. В исследуемом массиве текстов в отношения омонимии вступают:

а) часть парадигмы и полная парадигма изменяемых слов («Настоящее пламя пожирало внутренности игрушечного аэроплана» /«Но здесь, где обычно с прошлым смешано настоящее»);

б) части парадигмы изменяемых слов («Так, дохнув на стекло, выводят инициалы

тех, с чьим отсутствием не смириться»/ «Безумье дня по мозжечку стекло в затылок»);

в) грамматические формы изменяемого слова и неизменяемое слово («Чаще всего блестящие где-то в чаще пруды или озера»);

г) словоформы изменяемых слов и неизменяемое слово («Ах, чем меньше поверхность, тем надежда скромней»/ «И уже ничего не сниться, чтоб меньше быть»);

д) неизменяемые слова («Годится только, чтоб выйти вон»/ «Чтоб крикнуть: “вон!”»).

Характерные для текстов избранной поэзиа И. Бродского комплексы омонимичных лексико-грамматических классов представлены в Табл.1.

| № | Класс | Пример |
|---|---|---|
| 1. | существительное-прилагательное | настоящее |
| 2. | существительное-глагол | дали, постой |
| 3. | существительное-наречие | вечером, бегу |
| 4. | существительное-местоимение | ком |
| 5. | существительное-числительное | сорока |
| 6. | существительное-частица | чай |
| 7. | существительное-вводное слово | правда, право |
| 8. | существительное-союз-частица | раз |
| 9. | прилагательное-местоимение | другой |
| 10. | прилагательное-наречие | приятно, выше |
| 11. | прилагательное-частица | похоже, кто |
| 12. | прилагательное-наречие-вводное слово | верней |
| 13. | прилагательное-числительное-местоимение | первый |
| 14. | глагол-вводное слово | значит |
| 15. | глагол-междометие | прощай |
| 16. | местоимение-союз | кто, откуда |
| 17. | местоимение-числительное | один |
| 18. | местоимение-наречие | почему |
| 19. | местоимение-частица | тем, это |
| 20. | местоимение-частица-наречие | сам |
| 21. | местоимение-союз-наречие-частица | так |
| 22. | наречие-числительное | много, мало |
| 23. | наречие-предлог | рядом, около |
| 24. | наречие-частица | что, так, просто |
| 25. | наречие-союз-междометие | пока |
| 26. | наречие-союз | сколько |
| 27. | союз-частица | ни, только |
| 28. | предлог-частица | ради |

Табл.1. Комплексы омонимичных лексико-грамматических классов

В результате лемматизации и снятия омонимии установлено, что список словоформ, используемых И. Бродским в текстах избранного, составляет 11909 единиц. Оригинальные авторские формы слов составляют 0,8% списка: *поблизосте, гортанней* (сравнительная степень), *пенат* (именительный падеж, единственное

число), *музык* (родительный падеж, множественное число), *гудбая* (родительный падеж), *дребезг* (именительный падеж, единственное число) и др. В список словоформ входит также незначительная часть (0,33%) иноязычных слов – немецких, латинских, итальянских, английских, литовских, записанных латиницей (*architekten, dominikanaj*) или кириллицей (*кляйне, фиш*), числа и формулы.

Словарь избранной поэзии И. Бродского включает 8078 слов. При этом больше половины слов встречается в текстах стихотворений только один раз, индекс исключительности лексики составляет 62,4%. К редким словоупотреблениям можно отнести следующие группы слов:

1) архаизмы, составляющие 1,2% всего словаря (*скирос, дабы, ланит*);

2) авторские неологизмы – 0,5% (*сребролюбивый, бюстовать, лошади-пржевали, тюрьмо-динамика*);

3) заимствования, охватывающие 1,3% словаря (*бранзулетка*, *парубки, кривда, пановать*);

4) термины 0,9% всего словаря (*плазма, лейкоциты*), из них третью часть составляют лингвистические термины (*часть речи, глагол, падеж, суффикс, сказуемое*);

5) просторечья и жаргонизмы, составляющие 1% словаря (*мандраж, базлать, алконавт, братан, харя, атас, мусор, кирза*);

6) звукоподражательные слова – 0,46% (*курлы, хруск, муу-танки*).

Более 3% словаря избранного составляют географические наименования и имена собственные (*Флоренция, Арно, Ялта, Чучмекистан, Третьеримск*).

Самым частотным словом является предлог *в*, который встретился в текстах избранной лирики более 1200 раз. В первый десяток вошли служебные слова – предлоги, союзы, частицы и местоимения *я, он*, *тот,* чаще всего встречаемые в текстах (см. Табл. 2).

| | | |
|---|---|---|
| 1 | в | 1233 |
| 2 | и | 989 |
| 3 | не | 589 |
| 4 | на | 432 |
| 5 | что | 404 |
| 6 | как | 398 |
| 7 | с | 336 |
| 8 | я | 307 |
| 9 | тот | 293 |
| 10 | он | 214 |

Табл. 2. 10 самых частотных слов

3037 слов встретились в текстах избранного больше одного раза. Из них два слова употребляются поэтом больше 900 раз (*в, и*), три слова – больше 500 раз (*в, и, не*), 12 слов встретились в текстах больше 200 раз (*в, и, не, на, что, как, с, я, тот, он, от, этот*) и 25 слов – более 100 раз (*в, и, не, на, что, как, с, я, тот, он, от, этот, быть, ты, но, к, из, весь, мы, по, они, за, а, где, она*). Показатели самых частотных слов, обозначающих понятия, находятся в пределах от 70 до 28. Максимальной частотой в этом списке отличается слово *вещь*, минимальной – слова *море* и *друг.*

Литература
1. О мере близости частотных словарей русских поэтов XIX-XX веков // Математическая морфология. Электронный математический и медико-биологический журнал. – Т. 3, вып. 2. – 1999. – С. 119.
http://www.smolensk.ru/user/sgma/MMORPH/TITL.HTM
2. Орлова О.В. Компьютерный анализ поэтического текста и моделирование ассоциативно-смыслового поля ключевого концепта творчества автора// Открытое дистанционное образование. – Выпуск 1(9), 2003. – С. 57-60.
3. Перебийніс В.С., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання. – К.: Наукова думка,1985. – 204 с.
4. Частотный словарь языка сборника “Часть речи” И. Бродского http://slovari.ru/default.aspx?s=0&p=5316&0a0=1894

**Наталія Вовчаста**
Львівський державний університет безпеки життєдіяльності

## Використання програми «Словник пожежно-рятувальних термінів» на практичних заняттях з іноземної мови

Педагогічні програмні засоби є невід’ємним компонентом навчального процесу з іноземної мови у вищих навчальних закладах МНС України.

Серед ППЗ, які пропонуються останнім часом, важливе місце займають електронні навчальні курси, посібники, словники.

Іноземна мова по різному викладається в різних аудиторіях, що залежить від застосування її в тій чи іншій професії. Тому, необхідною умовою навчання є глибокі змістовні програми, ЕП, відповідний допоміжний методичний матеріал з фахових дисциплін.

Вибір ППЗ навчання у ВНЗ МНС України зумовлений: цілями навчання; змістом навчального матеріалу та специфікою предметної області; темпом та терміном процесу навчання; стилем навчання та рівнем педагогічної майстерності педагога; дидактичним та матеріально-технічним забезпеченням процесу навчання; рівнем підготовки курсантів. На практичних заняттях з іноземної мови за професійним спрямуванням на перших, других та п’ятих курсах напрямку підготовки «Пожежна безпека» та «Цивільний захист» широко використовуються Англо-український пожежно-технічний словник -мінімум [5] та Короткий українсько-англійський словник зі сфери надзвичайних ситуацій [4].

Вищезгадані словники є додатковим засобом до навчальних посібників «Англійська мова для рятувальників» (Частина І) [1], «Англійська мова для рятувальників» (Частина ІІ) [2] та «Професійна англійська мова: тексти і вправи»[3]. Електронні та паперові версії словників широко використовуються при вивченні іноземних мов майбутніми фахівцями цивільного захисту та у їх професійній діяльності. Невеликі за об’ємом, ретельно підібрані, вони дають змогу курсантові знайти відповідний переклад слова, словосполучення та необхідні також при роботі з іноземною науково-технічною літературою, документацією на спеціалізоване

обладнання закордонного виробництва тощо, для розширення особистісного тезауруса курсантів.

Щодо Програми «Словник пожежно-рятувальних термінів», то вона створена на основі трьох пожежно-технічних словників (англійсько-український, французько-український та німецько-український). Ця програма призначена для викладачів, курсантів та студентів навчальних закладів МНС України, рятувальників а також студентів і фахівців інших професій, які використовують пожежно-рятувальну термінологію у своїй діяльності .

Застосування програми «Словник пожежно-рятувальних термінів» є доцільним у комп'ютерних класах, електронних читальних залах бібліотеки, під час роботи на переносних комп'ютерах та у домашніх умовах при наявності комп'ютера. При перекладі електронних текстів передбачено введення слова для пошуку шляхом копіювання перших літер з тексту у поле для введення. Позитивним є те, що програма не вимагає інсталяції, що робить можливим її використання з флеш-накопичувачів та інших переносних носіїв інформації. Словники, які використовуються програмою, можна редагувати за допомогою програми Microsoft Access. Основним недоліком електронного словника для персонального комп'ютера є те, що ним не завжди можна скористатися. Зокрема, під час занять, які проводяться за межами комп'ютерного класу, зовні приміщень, на навчаннях, під час виконання професійних обов'язків (за винятком можливості використання переносних комп'ютерів) курсант, студент, викладач чи фахівець не може працювати з таким електронним словником [6].

Таким чином, використання ППЗ на практичних заняттях з іноземної мови за професійним спрямуванням є ефективним способом важливим компонентом у системному та послідовному навчанні.

Література

1. Вовчаста Н.Я. Англійська мова для рятувальників. Частина І. Навчальний посібник з англійської мови для курсантів та студентів напряму підготовки «Пожежна безпека» та «Цивільний захист» Навчальний посібник посібник / Вовчаста Н.Я., Дідух Л.І., Іванів О.В., Ткаченко Т.В. – Львів: Вид-во ЛДУ БЖД, 2011. – 128 с.

2. Вовчаста Н.Я. Англійська мова для рятувальників. Частина ІІ. Навчальний посібник з англійської мови для курсантів та студентів напряму підготовки «Пожежна безпека» та «Цивільний захист» Навчальний посібник / Вовчаста Н.Я., Дідух Л.І. – Львів: Вид-во ЛДУ БЖД, 2011. – 128 с.

3. Вовчаста Н.Я. Професійна англійська мова: тексти і вправи. Навчальний посібник / Бадюк О.О., Вовчаста Н.Я., Дідух Л.І., Іванів О.В., Ткаченко Т.В., Шванова О.В.– Львів: Вид-во ЛДУ БЖД, 2011. –72 с.

4. Вовчаста Н.Я. Короткий українсько-англійський словник зі сфери надзвичайних ситуацій–Понад 5000 термінів і термінологічних сполучень / Гульчевська М.Б., Бугайська О.В., Брига Т.Р., Монастирська Д.М., Лабач М.М.; За ред. Коваль М.С., Шуневича Б.І. – Львів: Вид-во ЛДУ БЖД, 2010. – 184 с.

5. Гульчевська М. Б. Англійсько-український пожежно-технічний словник-мінімум / Гульчевська М. Б., Вовчаста Н. Я., Бугайська О. В. — Львів : ЛІПБ, 2005. — 181 с.

Сольський Р. П. Німецько-український пожежно-технічний словник-мінімум / Сольський Р. П. — Львів : ЛІПБ, 2005. — 59 с.

6. Козяр М.М., Кузик А.Д. Використання мобільних телекомунікаційних пристроїв у системі підготовки фахівців оперативно-рятувальної служби цивільного захисту [Електронний ресурс] / Козяр М.М., Кузик А.Д.– Режим доступу: – http://ubgd.lviv.ua/wap-port/html/noaooij_1.html

**Тетяна Грязнухіна, Тетяна Любченко**
Український мовно-інформаційний фонд НАН України

## Електронні словники паронімів та їх використання в системах автоматичної обробки тексту

Паронімія належить до поширених мовних явищ, які через свою причетність до плану змісту мовних знаків вимагають особливої уваги до себе в системах автоматичної обробки текстової інформації. Зокрема це стосується систем машинного перекладу, автоматичного редагування, систем інформаційного пошуку. Ефективність цих систем залежить від вміння їх розпізнавати і виправляти в тексті, що обробляється, ситуацій із неправильним вживанням (у розумінні адекватності передавання авторської думки) одного з компонентів паронімічної пари.

Створюваний в Українському мовно-інформаційному фонді НАН України електронний словник паронімів (ЕСП) передбачається використовувати як основний інструмент ідентифікації паронімів в українських текстах. Крім того, ЕСП в Автоматичному редакторі є джерелом інформації про лексичні значення розпізнаних паронімів.

В Українському мовно-інформаційному фонді паронімічна параметризація входить до завдання автоматичної семантичної розмітки Українського національного лінгвістичного корпусу.

Побудова ЕСП передбачає виконання таких завдань:

- інтепретація на графемному рівні основної характеристики паронімів – „звукова подібність” (стосовно до письмової форми мови);
- формування ЛБД фонетичних паронімів української мови;
- формування ЛБД квазіпаронімів української мови;
- інтеграція ЛБД паронімів із тлумачним словником української мови у загальній лексикографічній системі УМІФ НАНУ.

Формування ЛБД паронімів здійснювалося на базі електронного граматичного словника української мови (обсягом понад 260 тис. словникових одиниць) у автоматизованому режимі.

Словникова стаття ЕСП включає в себе таку інформацію: компоненти паронімічної пари, лексико-граматичний клас, тип паронімів (фонетичні чи квазіпароніми), розрізнювальні ланцюжки графем у паронімічній парі, лексичні значення компонентів паронімічної пари.

**Наталія Дарчук**
Київський національний університет імені Тараса Шевченка
Інститут філології

# Морфологічне анотування Корпусу української мови

Обов'язковою складовою частиною лінгвістичного забезпечення будь-якої системи автоматичного опрацювання тексту є автоматичний морфологічний аналіз (АМА), тому що морфологічний аналіз присутній на всіх етапах аналізу тексту: ані морфемний, ані синтаксичний, ані семантичний аналіз не можуть обійтися без визначення частин мови. Морфологічні ознаки одиниць тексту є інструментом дослідження зв'язку між лексикою і граматикою, між використанням його у мовленні, між парадигматикою (в аспекті розгляду відмінкових форм відмінюваних слів) і синтагматикою (в аспекті лінійних зв'язків слів, сполучуваності у тексті). Роль саме такого «перекидного містка» виконують частини мови.

Внаслідок роботи АМА кожній словоформі тексту приписуються коди частин мови і значення граматичних категорій (рід, число, відмінок, вид, час, особа тощо). Характер цієї інформації, обсяг її й методи, за допомогою яких встановлюється морфологічна інформація, залежать від мети дослідження, у межах якого здійснюється АМА, від типу мови, орієнтації на вид текстів (усне або писемне мовлення). При анотуванні Корпусу української мови, робота над яким розпочалася у 2009 р., був випробуваний третій варіант АМА (перший розроблено у 1991 р., другий – у 1996 р.). **Метою** даної доповіді є висвітлення питань про принципи, покладені в основу нового лінгвістичного забезпечення морфологічного анотування Корпусу української мови, та деякі результати його застосування. АМА, який застосовується до Корпусу української мови, можна вважати автоматичним формально-морфологічним з елементами синтактико-морфологічного аналізу. У здійсненні пропонованого АМА української мови, на відміну від попередніх варіантів, треба пройти тільки два етапи: 1) формально-морфологічний, або флективний; 2) контекстний.

І-й етап – **флективний**. Формально-морфологічний, або флективний етап базується на поєднанні двох мовних інформацій, вміщених у дві таблиці: таблиці квазіоснов (сталої, незмінної частини слова та змінної частини слова без флексії) і сателітної таблиці квазіфлексій (флексії або змінної частини слова із флексіями) із характеристикою частиномовною і категоріальною (рід, число, відмінок, особа, час). Кожній лексемі приписується буквений код приналежності до певної частини мови, а тим, що мають словозміну – номер парадигматичного класу, для якого у сателітній таблиці наводяться форми словозміни, які разом із кодом класу складають двійковий граматичний код. За цими таблицями відбувається АМА у такій послідовності: кожна текстова словоформа порівнюється з основами словника основ на максимальний графемний збіг у колонці «стала частина основи». При позитивній відповіді аналіз продовжується за таблицею квазіфлексій і в разі збігу зі словозмінною формою приписується інформація про граматичні значення слова, а у разі омонімії (кількох однакових графемно виражених слів) – ланцюжок граматичних значень у послідовності двоелементних кодів. Наприклад, при ідентифікації словоформи ***столі***

буде приписаний код ЙП, а ***столи*** – код ЙА і ЙУ, тобто ланцюжок кодів, оскільки наявна граматична омонімія форм наз. і знах. відмінків множини. Таблиця квазіоснов складає 210 тис. одиниць, а відповідно словник словоформ, які породжуються поєднанням інформацій з таблиці основ і сателітної таблиці, – близько 3,2 млн слововживань, що забезпечує практично на 97% присвоювання морфологічної інформації (3% – це оказіоналізми або форми, не унормовані граматикою української мови, або неукраїнські слова тощо).

Морфологічна інформація, передбачена в АМА, в цілому відповідає граматичній традиції, прийнятій більшістю учених-україністів, що є важливим при побудові запита користувача. З іншого боку, ми розуміємо, що наступність аналізів – синтаксичного, семантичного – вимагає бути послідовними у веденні бази даних. Наприклад, розмежування словозмінних парадигм дієслів без –ся і з –ся є принципово важливим, оскільки одне й те саме дієслово має дефектність у парадигмі, яка корелює зі зміною значення (пор. особове і безособове дієслово: *спати – спало і спатися – спалося; довести – доведе і довестися – доведеться*). Алгоритмічно всі дієслова, які належать до одного парадигматичного класу, мають всі часово-особові форми, але у базі даних проти конкретних дієслів проставлено позначки, які не дозволяють розгортати певні парадигматичні форми. На етапах синтаксичного та семантичного аналізу ця інформація буде необхідною при встановленні предикативних центрів речення тощо. Парновидові дієслова мають різні парадигми, оскільки категорія виду словокласифікуюча, що набуватиме великого значення при здійсненні семантичного аналізу. До парадигми дієслова належать дієприслівники, а форми дієприкметників – до ад'єктивного класу слів. До парадигми дієслова відносяться особово-часові форми дійсного способу і форми синтетичного наказового способу та синтетичного майбутнього часу. Форми аналітичного майбутнього часу будуть встановлені на етапі синтаксичного аналізу (*буду читати* на цьому етапі розглядаються як два різних слова з відповідною граматичною анотацією).

Для іменників парадигма повна складається із семи відмінків (включаючи кличний відмінок для форми однини), завдяки чому породжується 13 форм (однина + множина), неповна (або singularia tantum, або pluralia tantum) і нульова, що фіксується у спеціальному коді.

До класу прикметників відносяться прикметники, дієприкметники і порядкові числівники, які мають 24 граматичні відмінково-родово-числові форми (клична форма не передбачена). Числівники утворюють один клас власне числівників. Складені числівники розглядаються як різні лексеми і тільки на етапі синтаксичного аналізу передбачається їх зібрати в один вузол. Займенники поділяються на займенники-іменники (мають шестирядну парадигму) і займенники-прикметники (мають, як і прикметники, 24 форми), їм надається спеціальний частиномовний код, який відрізняє його від класу прикметників. Серед прислівників виділено клас так званих предикативних слів (*можна, треба, необхідно, потрібно* тощо), оскільки вони виконують в реченні функцію предиката. Ім надається спеціальний код – @0. Для прислівників, як і для прикметників, ступені порівняння не вважаються формами словозміни, тому вони фіксуються як окремі реєстрові лексеми у базі даних.

Серед службових слів виокремлюються прийменники, перша літера коду яких позначає клас прийменників, а друга – код відмінка, у керуванні якого бере участь

прийменник, напр., прийменник *на* має код ПВПП, значить, він бере участь у керуванні знахідним і місцевим відмінком. Сполучники поділяються на сурядні (код СС) і підрядні (СП). Частки мають код Ь0, де нуль (друга позиція у коді), як у класі прислівників, позначає невідмінюваність цієї частини мови.

Всі лексико-семантичні розряди, існуючі для частин мови, не враховуються в АМА, оскільки їхні семантичні особливості будуть деталізовані тезаурусними зв'язками на етапі автоматичного семантичного аналізу. Для користувача пропонується перелік граматичних значень, за якими він може одержати всі контексти вживання або певної лексеми, або певної гарматичної форми. Якщо лексема у певному відмінку має паралельні форми, вони будуть представлені у відповіді на запит. Методично ми намагаємося дати подрібнене представлення граматичної категорії, а складні запити (напр., форми род. + знах. відм.одн.) поки що не передбачені.

ІІ-й етап – **контекстний.** Контекстний аналіз (КА) – це визначення мовленнєвих умов, у яких реалізується актуальне значення досліджуваної мовної одиниці. У системі АМА передбачено можливість робити аналіз і синтез парадигми кожного змінюваного слова і, маючи всі словоформи всіх змінюваних і незмінюваних слів та послідовно порівнюючи їх, можна автоматично укласти ланцюжок омонімічних словоформ і слів. Оскільки омонімія – суттєва перешкода для реальної картини морфологічних характеристик слів української мови, важливим є з'ясування всіх можливих омонімів за матеріалами Корпусу і побудова програм зняття граматичної і лексико-граматичної омонімії. В основу КА лягла ідея контекстуальної зв'язаності словоформ з іншими словоформами у тексті, її позиційні характеристики (напр., наявність пунктуаційних знаків або позиція дієслова чи прийменника у реченні тощо). Реалізація цієї ідеї знайшла відображення у створенні автоматичного конкордансу.

Алгоритм укладався за методом навчальної вибірки. Реалізація цієї ідеї знайшла відображення у створенні автоматичного конкордансу, теоретичною основою якого є:

- наявність таких визначальників (детермінант), за якими кожне граматичне значення детермінується в контексті іншими словами, сполученнями слів або іншими текстовими ознаками;
- текстоцентричний аспект підходу до його створення: він укладається на певному масиві текстів для певної словоформи або лексеми. Такий словник-конкорданс вичерпно ілюструє використання даної лексеми і всі її граматичні значення, що дає можливість виділяти всі детермінанти.

На матеріалі мільйонної вибірки з публіцистичних текстів для кожного омонімічного коду визначалися умови зняття омонімії певного ланцюжка кодів тільки контактних ліво- та правосторонніх (тобто Х -1 та Х +1 ) оточень. Детермінантами могли бути певні частиномовні класи слів (напр., предикативне слово код @0), або граматичний код (напр., прийменник, який керує тільки іменником у знахідному відмінку – код ПВ ), або пунктуаційний знак (тире, кома), або конкретне слово (напр., частка *не* або дієслово *бути*) тощо. Лінгвіст, опрацьовуючи кожний омонімічний ланцюжок, формує алгоритмічну ситуацію і натиснувши кнопку «обробити», пересвідчується в тому, чи спрацьовує це правило на даному та інших подібних контекстах.

Як показали результати тестування, найчастіше реалізується група правил (їх 111) «справа іменник» (ступінь ефективності – 79,0 %). На другому місці блок правил (30) «правосторонній займенник-іменник», ефективність якого значно нижча. На третьому місці «правостороннє дієслово» – за кількістю правил ця група більша, але за ефективністю ще нижча, ніж попередня. У лівобічному контексті найбільш ефективними є блок правил «справа прийменник» – за 78 правилами знімається омонімія у 16,77 % омонімічних ланцюжків.

Аналіз правил з позицій використання препозитивного і постпозитивного контексту показав, що основна орієнтація робиться на правий контекст (у 82,6 % омонімічних ланцюжків знімається омонімія). Це і зрозуміло, оскільки розгортання тексту в силу його лінійності відбувається зліва направо. Лише у 17,4 % випадків лівий контекст допоміг у диференціації омонімічної граматичної інформації. До правил КА входили також пунктуаційні знаки і деякі словоформи, разом вони складали зону «word», але ефективність таких правил порівняно невисока.

Принцип роботи програми КА рекурсивний: кожне речення обробляється за правилами кілька разів, доки не вичерпаються правила. Це дає змогу обробляти ланцюжки однорідних іменників або прикметників, між якими стоять коми або сполучники: якщо присутній омонимічний код хоча б в одного з них, омонімія знімається за граматичними характеристиками неомонімічного члена. Залишилися слова, об'єднані в омонімічні класи, для яких інформації про контактні класи слів недостатньо (таких біля 9 %), тобто ми не завжди можемо дати тверде визначення приналежності деяких слів до тієї або іншої частини мови. Потрібний дистантний аналіз, який буде здійснено на синтаксичному рівні, коли аналізуватимуться словосполучення у кожному конкретному реченні, які будуть об'єднуватися певним синтаксичним зв'язком, організуючись у дерево залежностей. Хоча цілком ймовірно, що певна частина її буде знята і на семантичному рівні, тому що жорсткий детермінований підхід до визначення певних частин мови викликає серйозні труднощі, оскільки передбачити поведінку слова для кожного випадку на базі загальної теорії неможливо. Закони мови, як і закони природи, мають статистичний характер, але знають винятки і знають різний ступінь свого виявлення.

До кола завдань АМА віднесено також «стягнення» в один вузол складених прийменників, сполучників і часток, оскільки вони є вторинними, а їхні члени в силу своєї десемантизованості набули ознак службових частин мови. Перед здійсненням контекстного аналізу зі зняття омонімії, за списком аналітичних прийменників (майже 500 одиниць), сполучників (154), часток (24) їм приписується відповідний код службової частини мови, а прийменникові – участь у керуванні певним відмінком іменника.

Отже, ефективність роботи КА досить велика – майже 93–95 % (залежно від типу тексту: найкращі результати на науково-технічному, найгірші – на поетичному) аналізованих словоформ одержали правильну морфологічну інформацію, що є запорукою хороших результатів на наступних етапах АОТ.

**Зоя Дудник**
Київський національний університет імені Тараса Шевченка,
Інститут філології

# Інструментарій програми Praat в курсі «Аналізу й синтезу усного мовлення»

Відомо, що опанування теоретичного матеріалу з будь-якої фонетичної дисципліни потребує посиленої зовнішньої мотивації студентів. Пошук відповідей на питання «як?» і «чому?» значно прискорює процес засвоєння, особливо, коли ці відповіді студенти отримують самостійно у власному дослідженні актуальних питань фонетики на практичних і лабораторних заняттях.

Розвиток навичок роботи студентів з усним матеріалом (у формі акустичного сигналу) в програмі PRAAT, спеціальне призначення якої – аналіз та синтез мовлення, дозволяє значно поглиблювати їхнє уявлення про живі процеси продукування мовлення як на етапі артикуляції, так і на акустичному етапі.

Знання про одиниці продукування мовлення, які потрібні для роботи з програмою, студенти отримують ще в курсі «Загальна і прикладна фонетика». Курс «Аналіз і синтез усного мовлення» спрямовує студентів на систематизацію параметричних акустичних даних голосних і приголосних англійського й українського мовлення, а також на просодичне моделювання таких явищ усного мовлення, як склад, фонетичне слово, синтагма, фраза. Для цього в програмі PRAAT є необхідний інструментарій. Математичне забезпечення програми дозволяє швидко отримувати акустичні параметри мовленнєвого сигналу, а саме значення частоти основного тону, резонансних формант, інтенсивності, тривалості в ділянках, виділених для аналізу й подальшої класифікації. Самостійно створювані акустичні класифікації студенти порівнюють з даними спеціальної літератури. Саме завдяки такому аналізу вони отримують уявлення про межі й діапазон фонетичного (алофонемного) варіювання та вчаться транскрибувати найважливіші варіанти, узгоджуючи їх з артикуляційними класифікаційними ознаками.

Експерименти з моделювання складів, слів, речень шляхом конкатенації (довільного комбінування сегментів, виділених зі зв'язного усного мовлення) у програмі PRAAT – надзвичайно зацікавлює студентів своєю можливістю перевіряти свої уявлення як про фонетичні властивості мовлення, так і теоретичне висвітлення цих питань у практичній фонетиці англійської та української мов. Це дозволяє уважніше ставитися до опанування артикуляційною базою іноземної мови та розуміти обмеження в описі особливостей артикуляції та різноманітних модифікацій звуків.

Досвід опрацювання фонетичних тем зі студентами з використанням програми PRAAT показав, що такі теми, як наголос, складоподіл, інтонаційний контур фрази, позиційні та комбінаторні модифікації успішно засвоюються в експериментах з конкатенативного синтезу мовлення, оскільки в моделюванні це одразу впливає на якість і розбірливість штучно створюваних одиниць мовлення. Саме в такому моделюванні студенти й знаходять відповіді на питання, як реально виявляє себе явище словесного, фразового, логічного наголосу, або чому потрібно звертати увагу

не лише на комбінаторні асиміляції, а й уважно ставитися до позиційних алофонів і складів.

**Анатолій Загнітко, Ганна Ситар, Ілля Данилюк**
Донецький національний університет

## Структура і модель бази даних «українські частки та їхні еквіваленти»

1. Група лінгвістів Донецького національного університету під керівництвом проф. А. П. Загнітко з 2009 року працює над створенням бази даних українських часток і їхніх еквівалентів, а також паперової версії відповідного словника. Основними завданнями дослідження є визначення максимально повного реєстру часток різного типу з урахуванням напрямків еволюції, динамічних процесів у межах частин мови, опрацювання принципів їхнього лексикографічного опису [3].

2. Ідея створення щонайповнішого й адекватного словника часток із простеженням різного функційного вияву, регулярності / нерегулярності синонімії, антонімії, омонімії, частоти тощо значною мірою зумовлена опрацюванням попереднього словника прийменників та їхніх еквівалентів [1] і необхідністю написання викінченої граматики службовості з адекватним висвітленням усіх площин категорійної семантики та розкриттям функційного навантаження службових елементів у реченні, висловленні, тексті, окресленням ємності функційно-семантичної парадигми кожного зі службових компонентів та його еквівалентів.

3. Створення бази даних українських часток реалізовано за допомогою програмних продуктів Microsoft Word і Microsoft Access 2010. На сьогодні база даних є таблицею з 38 полів, що корелює з відповідною інформацією про частки, і більше 200 рядків (записів) (за кількістю виявлених на цей момент одиниць – як власне-часток, так і одиниць, використаних у їхній функції). Для зручності роботи з базою була сконструйована форма з 9 вкладок, кожна з яких відбиває суттєві параметри часток, і вмонтованих процедур автоматичного опрацювання даних. Кожна вкладка містить відповідні текстові поля та елементи керування.

4. У першій вкладці «Характерологія» наявні такі ознаки часток:

а) структурний тип часток (проста, складна, складена – з розмежуванням з-поміж складених нечленованих і членованих, і диференціацією власне-часток і часток-виразів (*собі майже, лиш сам, тільки майже*); б) походження частки – відзайменникова, відприслівникова тощо [2].

5. Друга вкладка «Функційні вияви» об’єднує такі параметри, як статус часток, функційний тип, дистинктивний тип і семантику [3].

6. Третя вкладка відображає комунікативні параметри часток, з-поміж яких наразі розмежовано актуалізаційно-тематичні й актуалізаційно-рематичні. Та сама частка може в різних ситуаціях реалізовувати різний комунікативний статус.

7. Четверта вкладка подає особливості синтагматики часток з виявом відповідної регулярності останньої.

8. Семантико-парадигматичні ознаки часток висвітлено у п’ятій вкладці, що охоплює дані про їхні омоніми, синоніми, антоніми та варіанти часток. У цьому разі

значущим є врахування функційного навантаження частки, тому що низка часток постає повторюваною в різних синонімічних, антонімічних, омонімічних утвореннях.

9. Шоста вкладка «Квантитативні характеристики» відображає кількісні параметри часток, абсолютну частоту використання, зафіксовану в спеціально створеному дослідницькому корпусі текстів загальною ємністю 10 млн. слововживань. У ньому наявні тексти різних функційних стилів – художнього, наукового, офіційно-ділового та ін.

10. У сьомій вкладці подано динаміко-еволюційні параметри часток, розглянуті крізь призму відповідного часового зрізу для вияву розширення їхнього функційного тла, збільшення кількісного складу часток. Для цього проаналізовані Словарь української мови (Борис Грінченко), Історичний словник українського язика (за ред. Євгена Тимченка), Словник української мови (В 11-ти т.), Великий тлумачний словник української мови та ін.

11. У восьмій вкладці наведено опис етимології частки відповідно до її розгляду в Етимологічному словнику української мови (наявні 5 т.).

12. В останній дев'ятій вкладці репрезентовано матеріали щодо часток, подані у першому томі нового 20-томного словника української мови, тобто часток, що починаються на літери А та Б.

13. Концепція словника, його структура та основні характеристики поставали неодноразово предметом розгляду на симпозіумах різного рівня – міжнародних, всеукраїнських, регіональних, а також у рамах Всеукраїнського Граматичного наукового семінару (Донецьк, 2009; Донецьк, 2010; Донецьк, 2011), що потім знайшли вияв у різних публікаціях матеріалів цих авторитетних наукових форумів [3; 4]. Завдяки цьому було напрацьовано загальну концепцію словника часток, класифікаційні основи часток у різних вимірах – структурному, функційному, етимологічному тощо.

Література

1. Загнітко А.П., Данилюк І.Г., Ситар Г.В., Щукіна І.А. Словник українських прийменників. – Донецьк: ТОВ ВКФ «БАО», 2007. – 416 с.
2. Загнітко Анатолій. Частки в системі службових частин мови: типологійний і лексикографічний вияви // Лінгвістичні студії: зб. наук. праць / Донецький нац. ун-т; наук. ред. А.П. Загнітко. – Донецьк: ДонНУ, 2011. – Вип. 22. – С. 104-115.
3. Загнитко А.А. Функионально-семантическая типология частиц: внутрипредложенческий и контрастивный аспекты // Příspěvky k aktuálním otázkám jazykovědné rusistiky (3): Aktuální otázky současné jazykovědné rusistiky. – Brno: Tribun EU, 2009. – S. 146-154.
4. Загнітко А. Типологія службовості-допоміжності в реченні і тексті // Wyraz i zdanie w językach słowiańskich. – T. 7: Opis, konfrontacja, przekład / Pod red. Michała Sarnowskiego i Wlodzimierza Wysoczańskiego / Acta Wratislaviensia. – CL. – Wrocław Wyd-wo Uniwersytetu Wrocławskiego, 2009. – S. 303-311.

**Оксана Зубань**
Київський національний університет імені Тараса Шевченка
Інститут філології

# Морфемний аналіз у Корпусі української мови

У межах наукового проекту «Корпус української мови» колектив лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка розробив методику комп'ютерного моделювання структурних відношень одиниць різних мовних рівнів, на основі якої було створено комп'ютерні інструменти – пакети програм, що забезпечують отримання лінгвістичної інформації із тексту і створення різноманітних електронних словників та картотек в автоматичному або автоматизованому режимі.

Корпус українських текстів має модульну будову:

1. **Модуль-текст**, який репрезентує корпус параметризованих текстів (17 млн. словоформ);

2. **Модуль-аналізатор** – інструмент лінгвістичних досліджень великих текстових масивів, що включає пакети програм, які можуть виконувати такі функції: забезпечують зв'язок корпусу текстів з лінгвістичними базами даних: морфологічною, морфемно-словотвірною; синтаксичною; забезпечують роботу в автоматичному режимі пошуку і класифікації лексики за різними параметрами, а також морфологічного, морфемного і статистичного аналізів; формують словник-конкорданс мінімальних контекстових слововживань;

3. **Модуль-словник**, в якому систематизується різнотипна лінгвістична інформація в результаті аналізу текстів;

Структура Корпусу українських текстів відображає таку логіку лінгвістичного аналізу: модуль-текст є текстовим матеріалом дослідження, на базі якого за допомогою модуля-аналізатора в автоматичному або автоматизованому режимі укладаються різноманітні лінгвістичні словники (модуль-словник).

Одним із структурних блоків модуля-аналізатора є автоматичний морфний сегментатор, що використовується у лінгвістичних дослідженнях морфемної та словотвірної структури слова у різних типах українського дискурсу, а саме: у процесі укладання алфавітно-частотних словників усіх типів морфів та морфних структур на базі текстів різних стилів та авторів; встановлення та об'єднання афіксальних аломорфів у морфему (аломорфія коренів переноситься з базового словника на текстову словоформу автоматично); встановлення системних і функціональних характеристик морфем; автоматичного конструювання спільнокореневих гнізд (індекс омонімії коренів переноситься автоматично з базового словника на текстову словоформу).

Морфний сегментатор українського тексту – це система, на вході якої знаходяться лексеми (або словоформи) аналізованого тексту, представлені у вигляді алфавітно-частотного словника, а на виході – ті ж самі лексеми (словоформи), індексовані кодами граматичної належності до певної частини мови та розчленовані на морфи – кореневі, афіксальні з відповідними індексами. Словоформи сегментуються на морфи за загальноприйнятими теоретичними принципами

виділення морфів через автоматичне зіставлення із лінгвістичною моделлю морфної структури слова у Морфемно-словотвірній базі даних (МСБД), де просегментовано 190 тис. лексем, включаючи ономастичну й топонімічну лексику. Після завершення процедури морфного сегментування словоформ морфна структура слова запам'ятовується у символах лінгвістичної моделі, яка визначає межі і тип кожного морфа, дозволяє автоматично описати кожну морфну структуру через програмну процедуру.

На сьогодні в Корпусі української мови проведено автоматичне морфне сегментування словоформ поетичних текстів Л.Костенко і укладено різні типи морфемних словників, які будуть представлені у доповіді. Фрагмент результатів автоматичного морфного сегментування словоформ поетичних текстів Л.Костенко:

| id | cls | morfem | morfema | comm |
|---|---|---|---|---|
| 378422 | ПТ | R | за | за \| RC \| RC |
| 378423 | ЙИ | R | щит | щитом \| RDFE \| RDFF |
| 378424 | ЙИ | F | ом | щитом \| RDFE \| RDFF |
| 378425 | АИ | R | смарагд | смарагдових \| RHSJFL \| RHSJFL |
| 378426 | АИ | S | ов | смарагдових \| RHSJFL \| RHSJFL |
| 378427 | АИ | F | их | смарагдових \| RHSJFL \| RHSJFL |
| 378428 | ЙИ | R | ліс | лісів \| RDFE \| RDFF |
| 378429 | ЙИ | F | ів | лісів \| RDFE \| RDFF |
| 378430 | X | N | . | |

У межах проекту «Корпус української мови» розпочато роботу над створенням Словника афіксальних морфем. Ця робота здійснюється в автоматизованому on-line режимі лінгвістами з метою розроблення бази даних словотвірних значень, яку можна вважати базою знань про інваріанти й варіанти афіксальних морфем сучасної української мови, що містить список усіх афіксальних морфем української мови з інформацією про їхню структурну позицію, семантичну структуру та словотвірний потенціал.

Словник знаходиться на стадії формування і накопичення знань про кожний афікс у кожному слові загального базового словника української мови у 190 тис. слів. Цей словник буде використаний як база знань для формування запитів користувачів щодо семантичної структури афіксальних морфем, явищ морфного шва у дериватах, для автоматичної побудови морфемно-словотвірних гнізд аналізованих текстів, представлених у Корпусі української мови.

Таким чином, автоматичний морфний сегментатор, що забезпечує проведення морфемного аналізу у Корпусі української мови – це зручний лінгвістичний інструмент, який допоможе лінгвісту в автоматичному режимі проводити дослідження з морфеміки та словотвору на базі величезного ілюстративного текстового матеріалу Корпусу української мови, що дозволить отримати нові знання про семантичну та формальну структуру українського слова, проводити різноманітні класифікаційні аналізи лексики за кількісно-морфними моделями; створювати кореневі, афіксальні та словотвірні словники різних стилів та дискурсів; проводити морфний аналіз нових словоформ.

**Євгенія Карпіловська**
Інститут української мови НАН України

# Комп'ютерне моделювання мовних змін: система мови і текст

1. Поява комп'ютера відкрила нові можливості для вивчення мови, що й сприяло формуванню в межах мовознавства нової дисципліни – комп'ютерної лінгвістики (далі – КЛ). Проте статус КЛ і сьогодні спонукає до дискусій. Спектр гасел дискутантів, філологів і нефілологів, досить широкий: від намагань звести всю проблематику КЛ до технологій опрацювання мовної інформації і створення відповідного забезпечення комп'ютерних систем різного призначення, а отже, залучити КЛ до комплексу технічних і фізико-математичних наук, до тверджень, що КЛ є передусім лінгвістичною фундаментальною дисципліною, яка за допомогою комп'ютера покликана розв'язувати власне мовознавчі, теоретичні та практичні, проблеми, а отже, належить до циклу наук філологічних. Ми обстоюємо друге гасло, адже базовим словом у самому визначенні обговорюваної дисципліни є *лінгвістика*. Однак такі дискусії – аж ніяк не вправи для допитливого розуму і відсторонюватися від них, не обстоювати, не роз'яснювати свою позицію нам, філологам, щонайменше, непередбачливо. Наша мовчанка може призвести до розмивання основ КЛ як самостійної спеціальності, вихолощування її власне мовознавчого змісту. Крім того, дивлячися вперед, ми повинні бути свідомими й того, що окреслення змісту спеціальності закладає основу для відповідної підготовки для неї кадрів. І якщо ми переконані в тому, що це дисципліна філологічна, то нам слід передусім дбати про надійну базову підготовку філологів, здатних за допомогою комп'ютера розв'язувати мовознавчі завдання, актуальні для сучасних науки і суспільства. КЛ ще не входить до номенклатури спеціальностей у системі підготовки кадрів вищої кваліфікації: кандидатів і докторів наук. Проте реальну загрозу вихолощування філологічних галузей інтердисциплінарного статусу ми вже бачимо на прикладі спеціальності 10.02.21 – структурна, прикладна та математична лінгвістика, що до 2010 р. за номенклатурою ВАК України, як і раніше ВАК СРСР, проходила за напрямом «Філологічні науки», а нині об'єднана з напрямами «Технічні» та «Фізико-математичні науки». Доцільність і перспективність такого об'єднання саме з огляду на підготовку кадрів у мене викликає великі і небезпідставні сумніви.

2. Моделювання мовної динаміки, змін у системі та структурі мови є тим фундаментальним актуальним завданням сучасного українського мовознавства, у виконанні якого може КЛ може переконливо продемонструвати свій потенціал. Результати вивчення розвитку української мови, її реального сучасного стану вкрай важливі для підготовки комплексу фундаментальних праць, які становлять теоретичне й практичне підґрунтя для мовного будівництва в незалежній Україні, зокрема, для зміцнення державного статусу української мови. Це: 1) формування корпусів української мови різного типу й призначення; 2) створення нової української академічної граматики і на її базі практичних граматик, підручників для навчання української мови як рідної та іноземної; 3) укладання словників української мови нового покоління, загальних і галузевих, одно- і багатомовних, 4) підготовка нової редакції українського правопису. Усі ці завдання взаємопов'язані і

взаємоналаштовані: без аналізу текстів немає реального бачення мовної діяльності суспільства, як без знання складу системи мови, її норм та критеріїв кодифікації мовного матеріалу не можливе розуміння тенденцій реалізації системи мови у мовленні.

Встановлення змін у формі, змісті та вживанні мовних одиниць, у способах їх упорядкування в складі системи мови передбачає їх структурування, а отже, входить до комплексу завдань структурної лінгвістики – однієї з предтеч КЛ. Комп'ютер дає змогу дослідникам змоделювати змінні ділянки системи, зокрема, сучасної української мови, і завдяки цьому відтворити в комп'ютерному середовищі картину розвитку її лексикону та граматичного ладу за роки, що минули від часу публікації академічної граматики «Сучасна українська літературна мова» у 5-ти томах (К., 1969-1973), тому «Словотвір сучасної української літературної мови» (К., 1979), академічного тлумачного «Словника української мови» в 11-ти томах (К., 1970-1980) та нормативних словників радянської доби інших типів. Створені моделі становлять не лише унаочнення змін у мові, а й підґрунтя та інструмент для її дальшого вивчення. Відділ структурно-математичної лінгвістики Інституту української мови НАН України (упродовж 1968-2011 рр. відділ працював у складі Інституту мовознавства ім.О.О.Потебні НАН України) протягом останнього десятиліття зосередив свої зусилля на створенні саме таких комп'ютерних дослідницьких моделей для розв'язання фундаментальних завдань, зокрема для вивчення динаміки українського лексикону кінця XX – початку XXI ст.

3. Виявлення змін передбачає еталон для порівняння різних станів як системи мови, так і її реалізації в текстах. За такий еталон обрано стан українського лексикону, засвідчений словниками радянської доби і змодельований за їх матеріалами в комп'ютерному морфемно-словотвірному фонді української мови. Фонд сформовано у відділі впродовж 1988-1991 рр. Результати оновлення українського лексикону за матеріалами текстів різних функціональних стилів, новими словниками і працями дослідників подано в комп'ютерному фонді інновацій, формування якого розпочато у відділі 2006 р. Фонд складається з корпусу мікроконтекстів нових номінацій різних типів (новотворів, неосемантизмів і неозапозичень), одно- і кількаслівних (він налічує сьогодні близько 20 тис. одиниць), та параметризованої бази, яка і слугує моделлю динаміки сучасного українського лексикону. У базі подано інформацію про форму й семантику нових номінацій, а також відомості про їх функціональний потенціал у системі мови і в текстах. До уваги взято передусім ті інновації, що беруть активну участь у формуванні нових структурованих множин у складі українського лексикону. Такі активні ресурси сучасної української номінації передбачено представити в складі певних концептуальних полів у спеціальному словнику ідеографічного типу, роботу над яким нині завершує колектив відділу.

**Марина Кауль**
Российский Государственный Гуманитарный Университет

# Принципы составления англо-русских и русско-английских учебных словарей

1. Данное выступление посвящено светлой памяти недавно ушедшей от нас Сары Соломоновны Хидекель (1914-2011), инициатора и автора 12 учебных англо-русских/русско-английских учебных словарей.

В них отразился опыт прекрасного педагога с многолетним стажем, работавшего в ведущих языковых ВУЗах страны, и лингвиста-теоретика, специалиста по лексике и семантике английского языка, которая ещё в 60-х годах 20 века была в центре бурного развития лингвистики, участвовала в разработке всех, тогда новых, понятий, способствовала превращению описательной дисциплины «лексикология» в более или менее точную науку. Именно этот теоретический опыт лёг в основу её лексикографической практики.

2. Одим из теоретических вопросов, которые тогда бурно обсуждались и разрабатывались, был вопрос о валентности, сочетаемости частей речи и словосочетании как единице языка, параллельно с морфемой и словом.
Именно понимание словосочетания как единицы и языка, и речи, со своей семантически значимой структурой и речевой вариативностью, далеко не безграничной, стало основным дидактическим и лексикографическим принципом описываемых словарей.

Насколько знание сочетаемости расширяет словарь студента, можно судить хотя бы по сопоставлению количества входных вокабул словаря и реального количества их употреблений в свободных словосочетаниях: в «Англо-русском учебом словаре тематической лексики» (Астрель, 2008): на 1000 словарных статей приходится более 20000 словосочетаний.

3. Англо-русская/русско-английская сочетаемость является основанием для конрастивно-сопоставительных штудий, является основанием для установления пределов сочетаемости и сохранения её аутентичности в речи.

Поэтому контрастивность является ещё одним из основополагающих принципов описываемых словарей. В словарях делается акцент на расхождении в сочетаемости и структуре эквивалентых словосочетаний. Результатом таких контрастивных исследований стал словарь «Трудности английского словоупотребления» (Астрель, 2002) где выделено около 2500 случаев английского словоупотребления, вызывающих затруднения у русских студентов.

4. Дидактические соображения, которыми руководствуется каждый педагог, ориентированный на развитие навыков речи студентов, акцентируют внимание на повторяемости языковых и речевых моделей. В лексикографическом воплощении этот тезис диктует расширенную сочетаемость, примеры употребления в предложениях, как речевых контекстах. Во всех словарях приводится разветвленная сочетаемость и обильное количество примеров (обычно отделённых от общего текста ***).

5. Как учебно-направленные словари, предназначенные для студентов, думающих в нашем случае по-русски, они учитывают необходимость учительского

комментирования лексики. Все словари снабжены лексико-грамматико-семантическими замечаниями - *Notes* к трактуемой лексике.

Необходимость таких комментариев привела к созданию «объяснительных» словарей. На самом деле в современной терминологии это когнитивные словари (когнитивность понимается как вскрытие внутренней формы слова). В них в обобщенной форме приводятся результаты англо-русских и русско-английских сопоставлений внутренних форм трактуемых в словаре форм. Эта информация приводится в специальных «боксах» сразу после заголовочной части словарной статьи и в дальнейшем интерпретируется в самой структуре статьи.

Организующим принципом самой статьи являются понятия моделей, которые расположены в порядке их структурного усложенения.

6. Для Сары Соломоновны принципиально важным было в каждом новом словаре дать какое-то специфическое описание лексики, её какую-то новую трактовку. Так появились: Словарь дискурсивных слов «Коннекторы и модификаторы» (Русский язык, 2001; Астрель, 2006), не имеющий прецидента как по направленности, так и по степени и специфике комментирования; «Англо-русский учебный словарь тематической лексики с заданиями» (Астрель, 2008); “Living English in Real Situations” (Астрель, 2006) – употребление наиболее частотной лексики в речевых ситуациях.

7. Основным принципом, лежащим в основе всех словарей, была частотность трактуемой лексики – принцип, сегодня вряд ли нуждающийся в комментировании, ставший давно трюизмом, но не всегда последовательно применяемый в лексикографии.

8. Все словари этой серии обильно снабжены индексами и теоретическими обобщениями. Первые делают словари двунаправленными («словарь в словаре»), вторые превращают словари в бесценное учебно-теоретическое пособие.

**Мариола Кобылецка, Татьяна Бобкова**
Киевский национальный лингвистический университет

## Принципы составления иллюстрированного польско-русско-украинского словаря

Настоящий словарь предназначен для школьников-иностранцев и преподавателей русского и украинского языка как неродного. Основная цель данного словаря – предоставить в распоряжение пользователей иллюстрированный лексический материал, организованный по тематическому принципу.

При составлении словаря учтены также показатели употребительности слов и их понятийного веса в пределах методически существенных для начального этапа тем, таких как: Человек, Дом, Питание, Одежда, Школа, Город, Профессии, Больница, Село, Отдых, Праздник, Путешествие, Спорт, Животные, Здоровье, Погода, Время, Противоположные понятия, Действия.

Целевым для составления словника данного словаря является русский язык. Поэтому для составления словника использовалась система градуальных лексических

минимумов современного русского языка, представляющая лексическое ядро русского языка XXI века [1].

В соотвествии с методическими и дидактическими требованиями, изложенными в работах проф. В.И. Перебенос, основу словника составляет лексический минимум, соответствующий начальному этапу обучения и насчитывающий не более 2 тысяч лексических единиц [2]. Поэтому объем словника в основном ограничен V списком указанной системы градуальных лексических минимумов, включающим 2500 самых важных русских слов.

В процессе перевода русских слов на польский – базовый для пользователей и на украинский язык объем словника увеличился. В настоящее время словник включает 3131 слово.

Определение конкретного пользователя словаря – дети школьного возраста, изучающие русский и украинский язык, дает возможность оптимально организовать словарную статью, не перегружая ее дополнительными сведениями. Словарная статья включает слово словника и его переводы на украинский и русский язык.

При этом целесообразным представляется первоначальная подача лексического материала с помощью иллюстративного материала – ситуаций по указанным выше темам, а затем в конце словаря дается алфавитный индекс слов. Использование иллюстративных ситуаций демонстрирует употребление слова в определенном лексическом окружении. Описанная структура словаря обеспечивает доступность пользования и эффективность поиска необходимого лексического материала.

Иллюстрированный польско-русско-украинский словарь предназначен для преподавателей и учащихся, и может быть использован как для аудиторной работы, так и для самостоятельного изучения русского и украинского языка на начальном уровне.

Литература

1. Система лексических минимумов современного русского языка /Под ред д. фил. н., проф. В.В. Морковкина. – М.: Астрель. АСТ, 2003. – 768 с.
2. Перебейнос В.И. Принципы построения ученого словаря /Лексика и лексикография. Сб. научных трудов. – Вып. 7. – М.: Институт языкознания РАН, 1996. – С. 86-91.
3. Перебейнос В.И., Назаров С.Н., Бобкова Т.В. Типология учебных словарей/ Международная конференция «Прикладная лингвистика без границ». Материалы конференции. – Санкт-Петербург: Инфо-да, 2004. – С. 108-117.

**Валентина Коломієць, Сергій Котик**
Київський національний лінгвістичний університет

## Спеціальний навчальний корпус текстів UCLE: сучасний стан і перспективи використання

Одним із актуальних напрямів корпусної лінгвістики, що приваблює увагу дедалі більшої кількості дослідників, є створення спеціальних навчальних корпусів текстів, які містять зразки іншомовного мовлення школярів і студентів.

У лабораторії комп'ютерної лінгвістики КНЛУ укладено український навчальний корпус есе, написаних студентами, які вивчають англійську мову (Ukrainian Corpus of Learner English, скорочено UCLE). На даний момент корпус складається з 325 есе загальним обсягом понад 180 тисяч слововживань. 135 есе були написані під час аудиторних занять, 191 есе – вдома.

У якості інформантів корпусу виступили 102 студенти третього курсу, 142 студенти четвертого курсу і 81 студент п'ятого курсу. Серед інформантів 292 дівчини і 33 хлопці. 218 інформантів навчалися на факультеті англійської мови, 85 – на факультеті перекладачів, 21 – в економіко-правовому інституті КНЛУ. Рівень володіння мовою був визначений викладачами як просунутий.

Для пошуку даних у корпусі й отримання статистичної інформації створено вбудований базовий корпусний менеджер, який дозволяє укладати частотний список словоформ і будувати конкорданси (KWIC і повні конкордансні списки), здійснювати пошук окремих слів і словосполучень, сортувати списки слів, відображати знайдені словоформи у необмеженому контексті, отримувати статистичну інформацію про окремі елементи корпусу.

Передбачається, що в майбутньому корпус буде розширено як за рахунок есе студентів першого і другого курсів та школярів старших класів, так і за рахунок інших жанрів текстів (приватних і офіційно-ділових листів, звітів тощо). Для цього розробляється спеціальна комп'ютерна програма побудови корпусу. Також планується розмітка помилок у текстах есе. Створений корпус може розглядатися як пілотний.

Розробка спеціального навчального корпусу є доцільною із практичної точки зору. Такий корпус дозволяє отримати важливу інформацію про типові помилки в англомовних текстах, створених українськими студентами, особливості словника письмових творів студентів, граматичні аспекти засвоєння іноземної мови тощо. Очевидно, що використання навчального корпусу вплине на розробку та оцінювання методик навчання, створення навчальної літератури, словників.

Прикладом використання спеціального навчального корпусу може бути якісний аналіз актуального англомовного словника студентських есе.

У дослідженні використовувалися 135 есе студентів 3-5 курсів факультетів англійської мови і перекладачів, які були написані в умовах, наближених до екзаменаційних. Аналіз текстів есе здійснювався за допомогою VocabProfile, одного з інструментів Compleat Lexical Tutor (версія 6.2), розробленого Т. Коббом (http://www.lextutor.ca/). Отримані результати були порівняні з результатами аналізу контрольного корпусу з 45 есе із сайту IELTS-Blog [http://www.ielts-blog.com/ielts-writing-samples-essays-letters-reports/], які за оцінкою незалежних експертів відповідають рівням B2, C1 і C2 загальноєвропейської шкали рівнів володіння іноземною мовою (ШРВІМ).

Лексична різноманітність словника есе основного і контрольного корпусів, яка підраховувалась як відношення кількості різних слів до загальної кількості слів у тексті (type token ratio, скор. TTP), відображена в таблиці 1.

Таблиця 1.

Порівняння лексичної різноманітності словників досліджуваних корпусів

| Курс навчання | TTP | ШРВІМ | TTP |
|---|---|---|---|
| 3 | 0,15 | B2 | 0,22 |
| 4 | 0,12 | C1 | 0,27 |
| 5 | 0,15 | C2 | 0,37 |

Аналіз свідчить, що словник есе студентів КНЛУ є менш різноманітним, ніж словник есе контрольного корпусу. Лексична різноманітність словника студентів різних курсів суттєво не відрізняється, в той час як лексична різноманітність есе контрольного корпусу зростає від рівня до рівня.

Оскільки показник різноманітності словника свідчить лише про повторюваність слів у есе і нічого не говорить ні про частоту вживання респондентами високочастотних і низькочастотних слів, ні про те, які саме слова вживаються в есе, словник есе було також проаналізовано з точки зору частотності його вживання у Британському національному корпусі (БНК) та в академічній англійській мові (згідно списку академічної лексики Academic Word List [2]). Отримані результати представлені в таблицях 2 і 3.

Таблиця 2.

Покриття есе студентів КНЛУ словами різної частотності

| Курс навчання | Слова з першої тисячі найчастотніших слів у БНК | Слова з другої тисячі найчастотніших слів у БНК | Слова з Academic Word List |
|---|---|---|---|
| 3 | 84,75% | 6,11% | 3,83% |
| 4 | 84,33% | 5,27% | 4,39% |
| 5 | 85,68% | 5,43% | 3,43% |

Таблиця 3.

Покриття контрольних есе словами різної частотності

| ШРВІМ | Слова з першої тисячі найчастотніших слів у БНК | Слова з другої тисячі найчастотніших слів у БНК | Слова з Academic Word List |
|---|---|---|---|
| B2 | 84,15% | 6,07% | 5,68% |
| C1 | 81,16% | 6,69% | 6,54% |
| C2 | 77,57% | 6,67% | 9,65% |

Хоча значних відмінностей у вживанні слів зі списку перших двох тисяч найчастотніших слів у БНК між двома корпусами не виявлено, з'ясувалося, що в есе студентів КНЛУ значно рідше вживаються слова з Academic Word List. Відсоток цієї лексики є незмінним у есе студентів третього, четвертого і п'ятого курсів КНЛУ, а в контрольному корпусі цей відсоток зростає при переході на вищий рівень за ШРВІМ.

Проведене дослідження свідчить про те, що словник есе студентів старших курсів КНЛУ недостатньо багатий і різноманітний. Лексична «бідність» есе може

бути викликана різними причинами, зокрема нерозумінням особливостей писемного мовлення, незнанням вимог до жанру есе тощо. Для з'ясування цих причин потрібні окремі дослідження. Проте зрозуміло, що виявлена проблема потребує поглибленого аналізу, осмислення і детальнішого вивчення актуального словникового запасу студентів та пошуку шляхів підвищення ефективності його формування.

Література

1. Програма з англійської мови для університетів / інститутів (п'ятирічний курс навчання): Проект / колектив авт.: С.Ю. Ніколаєва, М.І. Соловей (керівники), Ю.В. Головач та ін., Київ. держ. лінгв. ун-т та ін. – К.: Британська Рада, 2001. – 245 с.
2. Coxhead, A. A New Academic Word List / Averil Coxhead // TESOL Quarterly. – 2000. – V. 34. – No. 2. – P. 213 – 238.

**Валентина Коломієць, Вероніка Орел**
Київський національний лінгвістичний університет

## Корпус анотацій наукових статей із комп'ютерної лінгвістики: стан розробки і перспективи використання

Важливим напрямком сучасної корпусної лінгвістики є розробка спеціальних корпусів текстів, які використовуються у процесі навчання професійно-орієнтованої англійської мови. Такий інтерес пояснюється відсутністю на ринку якісних сучасних посібників із англійської мови професійного спрямування для різних освітньо-кваліфікаційних рівнів, спеціальностей і спеціалізацій. Використання спеціальних корпусів уможливлює дослідження лексичних і граматичних особливостей підмов різних предметних сфер, відбір мовного матеріалу і розробку комунікативно-спрямованих навчальних матеріалів для забезпечення підготовки спеціалістів різних галузей у повній відповідності до загальноєвропейських та світових стандартів.

Метою цієї статті є представлення результатів першого етапу роботи над корпусом анотацій наукових статей із комп'ютерної лінгвістики та деяких попередніх результатів аналізу корпусного матеріалу. Корпус складається з анотацій наукових статей, опублікованих у журналі «Computational linguistics», який є офіційним виданням міжнародної асоціації комп'ютерної лінгвістики ACL (Association for Computational Linguistics), у збірнику «Компьютерная лингвистика и интеллектуальные технологии», який базується на доповідях, прийнятих до презентації на щорічній міжнародній конференції з комп'ютерної лінгвістики «Діалог» (Росія) та в збірниках наукових статей, які базуються на доповідях, представлених на щорічній міжнародній конференції з комп'ютерної обробки природної мови і комп'ютерної лінгвістики CICLing (Conference on Intelligent Text Processing and Computational Linguistics). Загальний обсяг створеного корпусу – 1475 текстів анотацій, які містять близько 150 000 слововживань. Співвідношення анотацій із різних джерел відображено в таблиці 1.

Усі анотації мають метатекстову розмітку, яка дозволяє відбирати з усього масиву дослідницький підкорпус, а також аналізувати склад корпусу і коригувати

його у процесі поповнення. Метатекстова розмітка включає інформацію про автора (ім'я, місце роботи, чи є англійська рідною/нерідною мовою) і текст (тема, назва, обсяг, джерело, дата публікації). Метаінформація зберігається у окремій базі даних.

Таблиця 1.

Структура корпусу анотацій наукових статей із комп'ютерної лінгвістики

| Джерело | Період | Кількість текстів | Кількість слововживань |
|---|---|---|---|
| Computational linguistics | 2000-2011 | 227 | 34491 |
| Компьютерная лингвистика и интеллектуальные технологии | 2006-2011 | 555 | 41367 |
| Conference on Intelligent Text Processing and Computational Linguistics | 2000-2011 | 693 | 73938 |

Аналіз текстів анотацій можна здійснювати за допомогою доступних автономних інструментів, які здатні працювати з нерозміченими текстами, таких як Compleat Lexical Tutor (http://www.lextutor.ca/), WordSmith Tools (http://www.lexically.net/wordsmith/).

На наступному етапі роботи над корпусом передбачено морфологічне анотування текстових даних і створення корпусного менеджера, який дозволить здійснювати первинний аналіз текстового матеріалу: укладати частотний список словоформ і будувати конкорданси різних типів.

Після створення корпусу було здійснено попередній аналіз лексики, ужитої у текстах анотацій. З цією метою використовувався VocabProfile, один із інструментів Compleat Lexical Tutor (версія 6.2), розробленого канадським дослідником Т. Коббом в університеті Квебека у Монреалі. VocabProfile дає змогу визначати відсоток високочастотних і низькочастотних слів у письмовому тексті.

З'ясувалося, що для точного розуміння змісту анотацій, яке потребує розуміння принаймні 95% загальної кількості слововживань у текстах [1], потрібно мати словниковий запас обсягом проблизно 13 000 слів (див. табл. 2).

Таблиця 2.

Покриття текстів анотацій словами різної частотності

| Ранг слів у Британському національному корпусі | Міра покриття текстів анотацій | Міра покриття текстів анотацій разом із попередніми словами |
|---|---|---|
| 1000 | 68,68% | 68,68% |
| 2000 | 12,44% | 81,12% |
| 3000 | 3.37% | 84,49% |
| 4000 | 3,12% | 87,61% |
| 5000 | 2,21% | 89,82% |
| 6000 – 8000 | 2,88% | 92,7% |
| 9000 – 13000 | 2,54% | 95,21% |
| 14000 – 20000 | 0,85% | 96,06% |

Понад 70% від загальної кількості слововживань у текстах анотацій складають загальновживані слова, приблизно 15% відсотків – загальнонаукова лексика, ще приблизно 13% – термінологічна лексика, власні імена тощо.

Планується створення на базі корпусу частотних списків різних видів (лексемних, афіксальних, сполучуваності, термінів), які будуть покладені в основу словникового мінімуму для магістрантів спеціальності «Прикладна лінгвістика».

Література

1. Nation, P., Waring, R. Vocabulary size, text coverage, and word lists / P. Nation, R. Waring // Vocabulary: Description, acquisition, pedagogy / [ed. by N. Schmitt and M. McCarthy]. – New York: Cambridge University Press, 1997. – P. 6 – 19.

**Валентина Критська**
Інститут української мови НАН України

## Алгоритмічна складність формотворення в українській мові (постановка задачі)

Алгоритмічна складність формотворення розглядається на сукупності описів словозмінних парадигм у «Граматичному словнику української літературної мови: Словозміна» (К., 2011). У процесі аналізу парадигм конкретних лексем виявилося, що парадигму, як іменну, так і дієслівну, можна змоделювати як систему з трикомпонентною структурою відповідно до трьох основних словозмінних характеристик – набору флексій, схеми наголошування та сукупності морфонологічних чергувань в основі. Ці характеристики прив‘язані: 1) до типових схем відмінювання (дієвідмінювання), 2) до буквеної послідовності основ вихідної форми. Опис парадигм у вигляді трипозиційного коду здійснено засобами спеціальної метамови. Код є міткою типу словозміни сукупності лексем або одиничної лексеми. Кількість типів словозміни в Словнику дорівнює 1070.

З іншого боку, код як опис типу словозмінної парадигми є основою алгоритму породження парадигм лексем (упорядкованих множин словоформ). Додаткова інформація до алгоритмів подана в словникових статях Словника, а також у таблицях словозмінних характеристик. Наприклад, у словникових статтях лексем іменникових підкласів слів **(полігон** ч -0 7О, **кільцЕ** с -е 12ФОV, **косА** Я1 (знаряддя)) до реєстрових лексем є інформація підкласу (ч – чоловічий рід, с – середній, ж – жіночий: парадигми мають окреме кодування в кожному підкласі); символи **-а, -0, -е** є показниками флексій вихідних форм, що дає можливість відділити основу словникової форми. На наступній позиції словникової статті – код типу словозміни, в дужках – уточнення семантики лексеми. У таблицях словозмінних характеристик зібрано флексійні набори, схеми наголошування та типи морфонологічних чергувань, які мають однозначні позначки; вони є складниками кодів підкласів.

Алгоритмічною складністю типу формотворення будемо вважати кількість кроків алгоритму породження словозмінної парадигми цього типу (упорядкованої множини словоформ), що буде обчислюватися на першому етапі при порівнянні опису типів

парадигм – кодів у межах одного підкласу. Таке розуміння алгоритмічної складності корелює з ідеєю, що лежить в основі так званої колмогоровської складності: складність об'єкта визначається найкоротшою довжиною його опису. Наше дослідження буде проведено на матеріалі кодів іменних частин мови.

Усі алгоритми породження парадигм мають однакові початкові кроки: 1 – визначення підкласу, 2 – виділення основи, 3 – перевірка на наявність морфонологічних чергувань. За значенням символу підкласу визначаються лексеми, що не можуть мати словозміну (прислівник і под., які далі не розглядаються). На другому кроці відсіються лексеми, які не відмінюються (наприклад, незмінні іменники, у яких відсутні символи флексій). Для залишених для подальшої роботи лексем заповнюється таблиця за схемою відмінювання повторенням вихідної основи. На третьому кроці обирається продовження ходу алгоритму, адже наявність морфонологічних чергувань є факультативною характеристикою.

У першому випадку (якщо чергувань в основі немає) алгоритм простіший: подальші кроки враховують тип наголошування (за позначкою), від чого також залежить кількість кроків (наголос на основі, на флексії чи їхні варіації або комбінації). До прикладу, серед іменників жіночого роду без змін в основі близько 40% лексем – з постійним наголосом на основі (найпростіший тип алгоритму), 1% – з наголосом на флексії, з рухомим наголосом – 1.6% лексем. Серед прикметників більшість (більше 94%) з постійним наголосом на основі, понад 6% – на флексії. За будь-якого флексійного набору і незалежно від підкласу парадигми лексем з незмінними основами, постійним наголосом на основі (відповідно, на флексії або інші), породжуватимуться однаковим алгоритмом.

У другому випадку (якщо є чергування в основі) процес породження парадигми ускладнюється. Покажемо це на прикладі утворення парадигми типу 12ФОV іменників середнього роду з флексією вихідної форми **-е** (лексем на зразок **кільцЕ**, кінцеве буквосполученням основи **-льц**). Флексійний набір 12 (однина: **е, я, ю, е, ем, і/ю, е**, множина: **я, ь, ям, я, ями, ях, я**); схема наголошування ФО (рухомий наголос: на флексії в однині, на основі – у множині); чергування V (**ь/е)** в родовому відмінку множини в другій позиції основи, рахуючи з кінця (**кільц-е – кілец-ь**). Отже, алгоритмом спочатку повинно бути враховано чергування в основі (4 крок). Потім до основ приписуються флексії (5 крок), після чого необхідно замінити ненаголошені літери відповідними наголошеними: в однині – голосні флексій (6 крок), у множині – перші голосні з кінця вихідної основи **кільц-** , зокрема і в словоформі родового множини **кілець** (7 крок). Таким чином, алгоритмічна складність цього типу парадигм дорівнює 7. Пор. з найпростішим алгоритмом породження парадигми лексеми **полігОн** з кодом 7О – 4, до перших трьох кроків додається тільки крок приписування флексій до основ. Найдовший код з наведених трьох прикладів має іменник **косА**, в парадигмі якого є 2 варіанти словоформ зн. одн. (**кОсу – косУ:** позначки 1зз, тобто відхилення від основного флексійного набору 1, Ф(з)кО – флексійний наголос переходить на основу в одній з цих форм, а також у формі кл. одн. – **кОсо**), постійний наголос на основі у множині; в род. мн. чергування **о/і – кіс**). Алгоритмічна складність цього типу словозміни – 9. Отже, дійсно, символьна довжина коду свідчить про більшу чи меншу складність алгоритму, і ми маємо підстави використовувати коди при підрахунках складності.

Сукупність алгоритмів характеризує алгоритмічну складність формотворення в українській мові і може розглядатися як один із параметрів мовної системи.

**Кузьма Лебєдєв**
Київський національний лінгвістичний університет

## Створення Багатомовного корпусу паралельних текстів

При використанні паралельних текстів будь-якого стилю для прикладних задач перед розробником постає проблема вирівнювання речень. Звичайно можна написати програму автоматичного вирівнювання речень, але через те, що в деяких випадках одне речення може бути перекладено кількома реченнями або взагалі не перекладатися, ця програма буде допускати помилки.

Вирівнювання речень можна виконувати вручну, але це досить тривалий процес для великого масиву паралельних текстів. Саме тому для розробки Багатомовного корпусу паралельних текстів було вирішено використати файли субтитрів до сучасних англомовних серіалів з їх перекладами. У разі використання файлів субтитрів (.srt) проблема вирівнювання речень у корпусі вирішується автоматично. Нижче на Рис. 1 подано фрагмент файлу субтитрів англійською мовою (зліва) та німецькою мовою (справа).

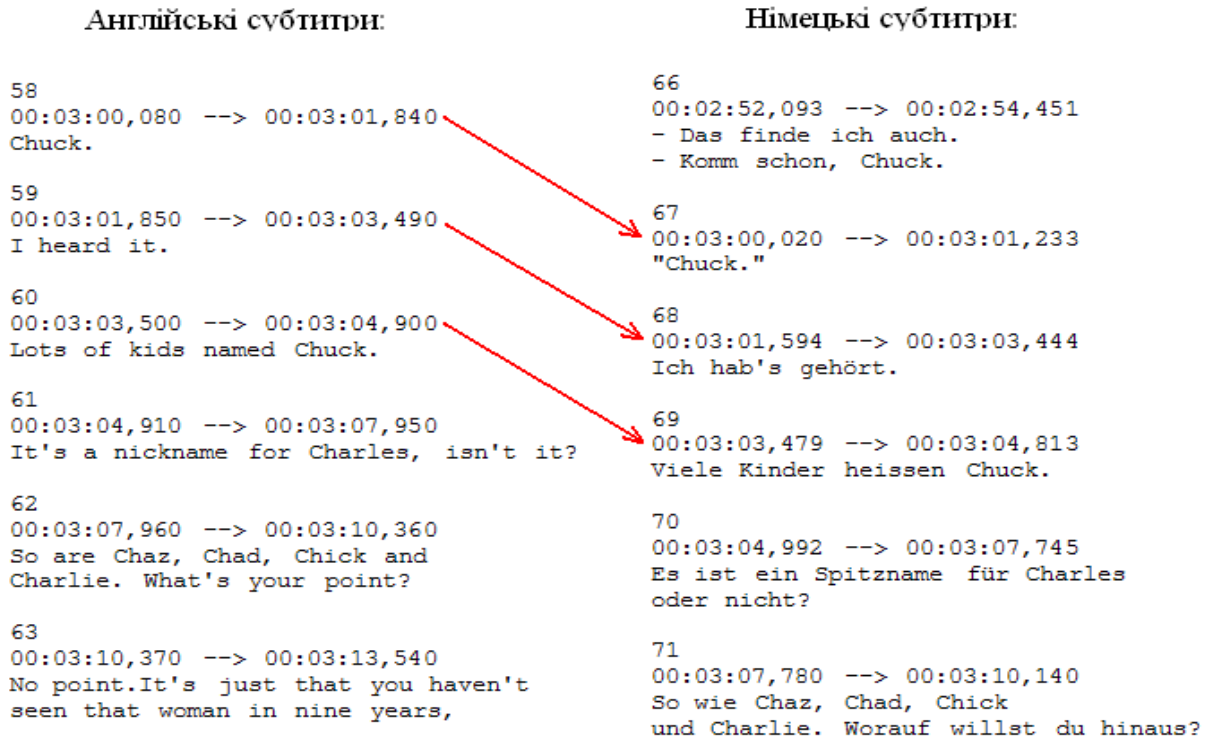



Рис. 1 Фрагмент субтитрів та їхніх перекладів німецькою мовою.

Використовуючи параметр синхронізації часу, можна вирівняти фрагменти паралельних текстів. Описаний підхід має декілька переваг: по-перше, можна розробити програму, яка буде вирівнювати фрази паралельних текстів автоматично і без помилок; по-друге користувач може включати до паралельного корпусу тексти мов, якими він не володіє, і отримувати правильно вирівняні тексти.

Для створення Багатомовного корпусу текстів було обрано субтитри до таких популярних англомовних фільмів та серіалів, як: House M.D. (Доктор Хауз), Heroes (Герої), Smallville (Таємниці Смолвіля), Terminator: The Sarah Connor Chronicles (Термінатор: Хроніки Сари Коннор), Two and half men (Два з половиною чоловіки), Big Bang theory (Теорія великого вибуху), Studio 60 on the Sunset strip (Студія 60 на Сансет Стрип), Desperate Housewives (Відчайдушні домогосподарки). Загальний обсяг корпусу біля 8 млн. слововживань. Оригінальні тексти субтитрів англійською мовою (обсягом біля 2 млн.) перекладено шістьома мовами, зокрема, українською (0,2 млн.) і російською (1,1 млн.), німецькою, грецькою, іспанською, французькою. Отже, створений на базі субтитрів Багатомовний корпус надає відомості щодо сучасної розмовної англійської мови [1, с. 13].

Для збереження даних було розроблено структуру бази даних корпусу, яка включає 8 таблиць. Так, таблиця Subtitles_info містить загальну адресну інформацію: назву серіалу, номер сезону та епізоду, жанр серіалу, дату виходу серії на екрани. Ще 7 таблиць, які містять тексти субтитрів та інформацію про їх вирівнювання (по окремій таблиці для кожної мови корпусу). Дані в базі мають ієрархічну структуру, таблиці поєднані відношенням один до багатьох (до одного запису у таблиці Subtitles_info належить багато записів у таблицях з текстами). Таким чином, розроблено структуру корпуса текстів, яка дозволяє зберігати вирівняні тексти розмовного стилю різними мовами та їхню адресну інформацію.

За допомогою Visual studio 2008 мовою програмування C# було розроблено програму імпорту та вирівнювання текстів. Програма дозволяє користувачеві завантажувати файл субтитрів однією з мов корпусу (за вибором користувача), вводити додаткову інформацію про файл, перевіряти, редагувати правильність вирівнювання текстів та імпортувати файл до бази даних. Крім зазначеної програми імпорту субтитрів було створено програму-конкордансер, яка дозволяє здійснювати пошук за словом, словоформою або словосполученням, отримувати контексти та їх переклади обраними користувачем мовами.

Після розробки локальної версії конкордансеру (Windows Application) було створено веб варіант у двох варіантах інтерфейсу (.aspx та Microsoft Silverlight). Розроблені версії програми-конкордансеру дозволяють здійснювати пошук за словом, словоформою або словосполученням, та отримувати в діалоговому вікні контексти, у яких ця одиниця зустрілась, та переклади контекстів обраними користувачем мовами. На даний момент скористатися створеним конкордансером можна на сайті лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету (http://complinguide.com.ua/Corpus.aspx).

Розроблена структура Багатомовного корпусу й засоби опрацювання англійського тексту можуть бути використані в системах автоматичного морфолого-синтаксичного аналізу, системах машинного перекладу, а також для створення одномовних і паралельних корпусів текстів різних мов.

Література.

Lebedev K. Parallel Multi-Lingual Corpus of Spoken Language / The Sixth International Conference 'Cultural Research: Challenges for the 3rd Millennium'. – Book of Abstracts. – Kyiv, 2010. – P. 12-13.

**Валентина Робейко**
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України

# Особливості автоматичного розпізнавання та синтезу усного спонтанного мовлення

Багато прикладних задач у галузі мовленнєвих технологій пов'язані зі створенням систем розпізнавання та синтезу спонтанного мовлення. Проте, надійність систем розпізнавання спонтанного мовлення суттєво нижча від точності розпізнавання підготованого мовлення. А розбірливість та натуральність озвучених спонтанних текстів ще далека від рівня озвучення літературних і публіцистичних текстів.

Першочерговим завданням під час автоматичної обробки спонтанного мовлення є створення корпусу текстів, що якомога повніше відповідатиме специфіці наукових досліджень та розроблюваних прикладних систем. Акустичний корпус українського ефірного мовлення [1] містить анотовані фрагменти підготованого та спонтанного мовлення (останнє складає більшу частину корпусу), тому великою мірою вирішує проблему збору акустичного та лінгвістичного матеріалу для навчання і тестування систем розпізнавання мовлення.

Українське спонтанне мовлення важко піддається автоматичній обробці в першу чергу через свою варіативність на алофонному та фонемному рівнях. Моделюванню варіативності спонтанного мовлення приділяється багато уваги під час створення словників розпізнавання (генерування кількох варіантів вимови для одного слова [2]) та текстів транскрипцій для синтезу (урахування мовленнєвих особливостей диктора-«донора») [3, 4].

Також спонтанні тексти характеризуються емоційністю, порушенням плавності мовленнєвого потоку, комунікативною спрямованістю та відкритістю словника [5, 6]. На письмі це виражається у неправильній орфографії (навмисній, ненавмисній або зумовленій неграмотністю), специфічній пунктуації (або відсутності розділових знаків), використанні емотиконів та ін. [6]. Для усного спонтанного мовлення властиві паузи хезитації, обмовки, повтори, редукування слів, порушення синтаксичного оформлення речень [7].

Не менш важливою ознакою українських спонтанних текстів є двомовність і суржик.

Відкритий характер словників спонтанних текстів вимагає їх постійного моніторингу та доповнення, а також редагування у вигляді розставляння наголосів. Окрім загального словника для розпізнавання та синтезу створюються додаткові словники, наприклад, словник жаргонізмів, словник суржику, словники власних назв, абревіатур тощо.

Вагомою рисою спонтанності є її індивідуальність для мовлення різних людей. Проблема індивідуалізації частково вирішена для синтезу україномовних текстів. Це використання особливих для диктора-«донора» рис вимови на фонемному рівні [2] та створення його індивідуалізованих інтонаційних контурів у системі озвучення текстів [5].

Проте, статистичні методи, які застосовуються при розпізнаванні мовлення, «усереднюють» мовців. Тому актуальним є створення індивідуальних акустичних моделей і словників транскрипцій, адаптованих до мовлення конкретного диктора з одночасною кластеризацією мовленнєвого сигналу за ознакою індивідуальності мовця.

Розроблений відповідний інструментарій дає змогу досліджувати описані проблеми та будувати більш досконалі моделі акустичного, фонемного, синтаксичного та семантичного рівнів, які покращать якості новітніх мовленнєвих технологій і систем.

Література

1. Valeriy Pylypenko, Valentyna Robeiko, Mykola Sazhok, Nina Vasylieva, Oleksandr Radoutsky. Ukrainian Broadcast Speech Corpus Development. // Proc. of the 14th International Conference «Speech and Computer: SPECOM'2011». – Kazan, Russia, 2011. – Pp. 435-440.
2. Робейко В.В., Сажок М.М. Багатозначна багаторівнева модель перетворення орфографічного тексту на фонемний // Штучний інтелект. – № 4'2011. – Донецьк, 2011. – С. 117-125.
3. Strik H., Cucchiarini C. Modeling pronunciation variation for ASR: overview and comparison of methods // Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition. – Rolduc, the Netherlands, 1998. – Pp. 137–144.
4. Пилипенко В.В., Робейко В.В. Автоматизированный стенограф украинской речи. // Искусственный интеллект. – № 4. – Донецк, 2008. – С. 768-775.
5. Робейко В.В. Інтонаційна організація спонтанного мовлення. // Мовні і концептуальні картини світу. – Вип. 25. Ч. 3. – Київ, 2009. – С. 225-229.
6. Lyudovyk T., Brozinski S., Noner M., Robeiko V., Sazhok M. Speech Synthesis Applied to SMS reading. // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». – St. Petersburg, Russia, 2009. – Pp. 300-305.
7. Людовик Т.В., Робейко В.В., Пилипенко В.В. Автоматическое распознавание спонтанной украинской речи (на материале акустического корпуса украинской эфирной речи). // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25 - 29 мая 2011 г.). – Вып. 10 (17). – Москва, 2011. – С. 478-488.

**Юлія Романюк**
Інститут української мови НАН України

## Засади алгоритмічного опису української дієслівної словозміни (за матеріалами Граматичного словника української літературної мови)

Граматики української мови, монографії, наукові розвідки подають різноаспектний опис дієслова: і в зв'язку з історією мови, і стосовно структурних та функціональних його особливостей, і стосовно акцентуації. «Граматичний словник

української літературної мови. Словозміна» (К., 2011; далі Словник) побудований так, що подає формальний опис дієслівної парадигми. Традиція формального опису мовних фактів з метою автоматизації лінгвістичних досліджень, вироблена і усталена за чотири десятиліття у Відділі структурно-математичної лінгвістики Інституту мовознавства ім. О.О.Потебні (з 2011 р. – Інституту української мови НАН України), має свою методику, узасаднену практичним досвідом. Згідно з цією методикою аналізується тільки графемний запис словоформи. Графемний запис дієслівних словоформ є підставою для класифікації дієслів: формування парадигматичних класів відбувається за а) графемними варіантами закінчень, б) побуквеним записом чергувань в основі, в) місцем наголосу при словозміні.

Ця класифікація є унаочненням алгоритмічних правил, які укладені та застосовуються для побудови конкретної дієслівної словозмінної парадигми. Цей алгоритм імпліцитно присутній у Словнику: за допомогою таблиць задані правила виведення форм із конкретними закінченнями, наголосом і змінами в основі при дієвідмінюванні. Для прикладу візьмемо найчастотніші парадигматичні класи та спробуємо експлікувати для них правила алгоритму. Отже, найчастотнішим парадигматичним класом дієслів є клас із трипозиційним кодом 7 ОО1 (9421 дієслово з близько 35000). Три позиції коду відповідають класифікації дієслів у Словнику (закінчення, схема наголошення, побуквений запис чергувань в основі). У наведеному для прикладу коді відсутній третій компонент - чергування.

Правила алгоритму можуть бути записані таким чином:

1. Відділити від дієслова суфікс -ти: *проха-ти*.
2. Якщо у першій позиції коду стоїть «7»,то до основи дієслова додається флексійний набір 7: -ю, -єш, -є, -ємо, -єте, -ють (теперішній/майбутній (для дієслів доконаного виду) час), -в, ла, -ло, -ли (минулий час) та –й, -ймо, -йте (наказовий спосіб): *прохаю, прохаєш, прохає, прохаємо, прохаєте, прохають, прохав, прохала, прохало, прохали, прохай, прохаймо, прохайте*.
3. Якщо у другій позиції коду стоїть ОО1, то наголос при дієвідмінюванні залишається незмінним – на тій самій букві основи: *прох=ати – прох=аю...* прох=ав...прох=ай.
4. Якщо у третій позиції коду немає позначки – перейти до наступного дієслова.
5. Другим за частотністю є парадигматичний клас 7 ОО1 F (6916 дієслів). Для цього класу алгоритм дещо ускладниться, на що і вказує додаткова позначка в коді парадигми:
6. Якщо у третій позиції коду стоїть позначка F, то у теперішньому/майбутньому часі та наказовому способі у позиції 2+1 з кінця основи *-ва-* замінюється на 0: *милувати – милую, милуєш... милуй...*

Третім за частотністю є парадигматичний клас із кодом 7 ОО2 F (4702 дієслова). Для цього класу алгоритм буде мати такий вигляд:

Якщо у другій позиції коду стоїть ОО2, то відбувається перехід наголосу у теперішньому/майбутньому часі та наказовому способі на одну букву вліво: *малюв=ати – мал=юю, мал=юєш... малюв=ав,* але *мал=юй, мал=юймо, мал=юйте*.

Наступний за частотністю парадигматичний клас має код 1 ОФ2 Ю (832 дієслова). Для цього парадигматичного класу алгорим буде мати такий вигляд:

1. Відділити від дієслова суфікс -ти: гну-ти.
2. Якщо у першій позиції коду стоїть «1»,то до основи дієслова додається флексійний набір 1: -у, -еш, -е, -емо, -ете, -уть (теперішній/майбутній (для дієслів доконаного виду) час), -в, ла, -ло, -ли (минулий час) та -и, -імо, -іть (наказовий спосіб): *гну, гнеш, гне, гнемо, гнете, гнуть, гнув, гнула, гнуло, гнули, гни, гнімо, гніть*.
3. Якщо у другій позиції коду стоїть ОФ2, то наголос при дієвідмінюванні переходить на флексію у теперішньому/майбутньому часі та наказовому способі: *гн=ути – гн=у... гн=ув...гн=и*.
4. Якщо у третій позиції коду стоїть позначка Ю, то у теперішньому/майбутньому часі і наказовому способі у позиції 1 з кінця основи *-у-* замінюється на 0.

Ми спробували унаявнити правила алгоритму, за якими будується конкретна дієслівна парадигма та виводяться конкретні дієслівні форми у «Граматичному словнику української літературної мови. Словозміна» (К., 2011). Можливість зробити алгоритмічний опис словозміни дієслова, так само як і інших частин мови, виникла з засад подання мовного матеріалу у Словнику – формального опису графемної структури слова.

Література

1. Алексієнко Л.А., Козленко І.В. Граматичний словник українських дієслів. Т. 1 (А-О). – К.: Київський університет, 1998.
2. Зализняк А.А. Грамматический словарь русского языка. Словоизменение. – М.: Русский язык, 1977.
3. Критська В.І., Недозим Т.І., Орлова Л.В., Пуздирєва Т.К., Романюк Ю.В. Граматичний словник української літературної мови. Словозміна / Відп. ред. Н.Ф.Клименко. – К.: Видавничий Дім Дмитра Бураго, 2011.

**Ганна Ситар**
Донецький національний університет

## Особливості структурування та наповнення бази даних «Синтаксичні фразеологізми в українській мові»

Синтаксичний фразеологізм (або фразеологізоване речення) – це особливий тип речень, що складається з постійної та змінної частин, компоненти яких пов'язані ідіоматично, синтаксичні зв'язки і прямі лексичні значення слів послаблені або втрачені на сучасному етапі розвитку мови [1; 2; 4]: *Оце так сміливці! З гарячкою – в… ополонку. «Моржі», які уперше в житті купалися взимку у Дніпрі, лікували застуду морозом та холодною купіллю* (Високий замок. – 07.01.2011); *Що за люба дитина мій Карпо, такий слухняний, такий тихий, хоч у вухо бгай* (Іван Нечуй-Левицький).

Мета цього дослідження – охарактеризувати особливості структури та наповнення створюваної нами бази даних (далі БД) «Синтаксичні фразеологізми в українській мові».

БД становить сукупність структурованих даних про синтаксичні фразеологізми української мови, тобто включає моделі фразеологізованих речень різного типу і їхні різнопланові характеристики. Вона побудована за реляційною моделлю за допомогою програми СКБД Microsoft Access.

Таблиці, що становлять основу БД, складаються з полів і рядків (записів). Кожній моделі синтаксичного фразеологізму в режимі таблиці надано індивідуальний номер.

Для зручності роботи з БД була сконструйована так звана форма з вкладками, що об'єднують кілька полів, елементами керування й навігації. Основним полем у режимі форми обрано поле «Структурна схема речення». Інші поля, що відповідають виділеним у процесі дослідження властивостям синтаксичних фразеологізмів, згруповані для зручної роботи користувача в сім вкладок.

1. Вкладка «*Структура»* відображає такі властивості, як «Тип речення за будовою», «Частиномовний статус незмінного компонента» (див. [5]), «Варіанти структурної схеми», «Наявність поширювачів».

2. Вкладка *«Семантика»* охоплює поля «Типова семантика речення», «Додаткові семантичні відтінки», «Семантичний тип», «Ступінь злитості компонентів» (див. [3]), «Лексичні обмеження», «Образні моделі». Зазначимо, що останні два поля для деяких синтаксичних фразеологізмів залишаються незаповненими, оскільки є синтаксичні фразеологізми, що не мають обмежень на лексичне заповнення змінюваної позиції (наприклад, моделі *Чим не* $N_1$ $Cop_f$, *Який* $N_1$ $Cop_f$), а образні моделі реалізуються в межах не всіх проаналізованих речень.

3. Вкладка *«Синтаксична парадигматика»* включає відомості про компонентний склад синтаксичної парадигми описуваної моделі речення – граматичні модифікації часу, способу, фазові, модальні, заперечні, авторизаційні та інші перетворення базової моделі речення. Парадигма синтаксичних фразеологізмів, звичайно, є неповною, але виявлення обмежень на модифікації для кожної окремої моделі видається значущим для опису її властивостей.

4. Вкладка *«Прагматика*» поєднує три поля – «Прагматична функція», «Прагматичний статус мовця» і «Тип прагматичного зв'язку».

5. Вкладка *«Семантико-парадигматичні властивості»* відображає синонімічні зв'язки моделей синтаксичних фразеологізмів з іншими фразеологізованими моделями й традиційними реченнями, за наявності у ній розмежовано синтаксичні омоніми – нефразеологізовані і фразеологізовані речення, побудовані за однією структурною схемою.

6. Вкладка *«Словникові відомості»* об'єднує дані, які наведено в авторитетних словниках. Оскільки на сьогодні в україністиці моделі речень ще не зазнали лексикографічного опрацювання, мова йде про відомості, почерпнуті із фразеологічних або тлумачних словників, у яких зафіксовані лише окремі синтаксичні фразеологізми.

7. У вкладці *«Приклади»* подано контексти вживання певної моделі синтаксичних фразеологізмів переважно в художньому, публіцистичному й розмовному стилях. Тут подаються приклади основного варіанта моделі, а випадки неелементарної структури, ускладненої семантики й утворення модифікацій ілюструються прикладами в межах відповідних вкладок.

Наступні етапи роботи передбачають не тільки поповнення БД новими моделями, але й розв'язання низки теоретичних і прикладних питань, зокрема, розмежування у межах синтаксичних фразеологізмів членованих і нечленованих одиниць, установлення функційного навантаження варіантності пунктуаційного оформлення та ін.

Література

1. Величко А. В. Синтаксическая фразеология для русских и иностранцев: Учебное пособие. – М.: Изд-во МГУ, 1996. – 96 с.
2. Всеволодова М. В., Лим Су Ён. Принципы лингвистического описания синтаксических фразеологизмов: На материале синтаксических фразеологизмов со значением оценки. – М.: МАКС Пресс, 2002. – 164 с.
3. Личук М. І., Шинкарук В. Д. Ступені фразеологізації речень. – Чернівці: Рута, 2001. – 136с.
4. Русская грамматика: В 2-х т. – Т. 2. Синтаксис / Под ред. Н. Ю. Шведовой. – М.: Наука, 1980. – 709 с.
5. Ситар Г. В. Структурні й семантичні типи синтаксичних фразеологізмів в українській // Мовознавчий вісник: Зб. наук. праць / Відп. ред. Г. І. Мартинова. – Черкаси, 2011. – Вип. 12-13. – С. 178-181.

**Олена Сірук**
Київський національний університет імені Тараса Шевченка

## Підготовка діалектних текстів для корпусного опрацювання

Сьогодні важливість розроблення текстових корпусів, як і в цілому застосування комп'ютерного інструментарію для мовознавчих досліджень вже не викликає заперечень. Але «корпусні студії літературної мови без аналізу діалектного матеріалу, без зіставлення з діалектним матеріалом завжди залишатимуться неповними» [1, с. 162]. Застосування методів та інструментарію корпусної лінгвістики до царини текстової діалектології є тим взаємодоповнюючим поєднанням напрямків, яке забезпечує комплексність, частотну перевіреність і обґрунтованість висновків діалектологічного дослідження.

Ідею створення діалектного корпусу текстів як оптимальної бази для збереження та глибшого опрацювання текстового матеріалу ми реалізуємо у вигляді «Корпусу українських діалектних текстів» (КорУДіТ). Метою діалектного корпусного проекту є відновлення семантичної безперервності у дослідженні говірок шляхом доповнення текстовими даних, отриманих за допомогою питальників, а також забезпечення дослідника базою (структурованим мовним матеріалом та комп'ютерним інструментарієм) для багаторівневого аналізу діалектної мови, зокрема фонетичних, морфологічних, синтаксичних, семантичних, стильових рис на різних етапах її функціонування. Для здійснення вказаної мети потрібно виконати такі комплекси завдань, як зведення сукупності виявлених за різними джерелами українських діалектних текстів, опрацювання цих текстів як елементів єдиної лінгвістичної інформаційної системи, а також забезпечення оперативного доступу користувачів до

цього джерела мовних даних [2, с. 293–300]. КорУДіТ є складовою «Корпусу текстів української мови» (КТУМ), колективної праці фахівців лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка. Наявність діалектних текстів у КТУМ сприяє забезпеченню його репрезентативності, а також розширює коло дослідницьких завдань корпусу, оскільки в територіальних діалектах можна віднайти «як найдавнішу питому мовну специфіку, так і результати інноваційних процесів, потенційних для літературної мови» [1, с. 162].

Зараз ми працюємо над розробленням методики укладання КорУДіТ і створенням його сегмента на текстовому матеріалі західноволинських говірок. Завдання полягає у тому, щоб уможливити корпусне опрацювання зафіксованих в паперовому чи аудіовигляді діалектних текстів шляхом їх переведення в комп'ютерну форму за допомогою програмного забезпечення. Потрібно надати цим текстам відповідне зовнішнє (автор, інформація про видання або приватну текстотеку, про записувачів та інформаторів, анкета текстів тощо) та внутрішнє, або структурне, маркування (номер, початок і кінець тексту, розділу, абзацу, речення, слова тощо), а також власне лінгвістичні розмітки (морфологічну, синтаксичну, семантичну тощо).

Методика роботи з діалектними текстами в КорУДіТ дає можливість упорядкувати й опрацьовувати паралельно три взаємопов'язані підкорпуси (затранскрибованих діалектних текстів, діалектних текстів в орфографічному записі та корпус «перекладених» літературною мовою діалектних текстів). Тексти у фонетичній транскрипції формують підкорпус транскрибованих діалектних текстів; який найбільше зацікавить фахівців-діалектологів. Підкорпус діалектних текстів в орфографічному записі має ширшу аудиторію, зокрема позафілологічну; передбачається можливість досліджувати стилізацію художнього тексту під говірки та розмовний стиль у рамках корпусу усної української мови. Тексти у записі, максимально наближеному до літературної мови, формують підкорпус «перекладених» сучасною літературною мовою діалектних текстів. Цей підкорпус можна застосовувати так само, як і орфографічний підкорпус; він необхідний для автоматичної розмітки тексту, зокрема для коректної роботи автоматичного морфологічного аналізатора, розрахованого перш за все на тексти літературного стандарту української мови. Формування паралельних текстів здійснюється напівавтоматично, для чого передбачена спеціальна комп'ютерна програма, яка на базі фонетичної транскрипції або орфографічного запису створює два інших тексти відповідно до правил базової трансформації текстів, розроблених на основі тестових говірок. Для збереження логіки викладу та подальшого дослідження лінгвістичних змін в комунікації на стику двох різних стильових різновидів мови до корпусу вноситься як текст інформатора, так і текст записувача (окреслення теми розмови або запитання, на які відповідає інформатор). За відсутності слів записувача смислова канва розмови спотворюється, виникає небезпека ототожнення з нормами говірки повторених за інформатором питальних фраз і конструкцій, які в інших комунікативних ситуаціях не вживаються, послідовно замінюючись на інші, не характерні для літературного ідіому, але питомі для говірки.

Сьогодні українська комп'ютерна діалектологія перебуває на етапі формування; корпусів української діалектної мови наразі немає, як немає і публікацій з цієї теми. Тож і теоретичні засади, і практична реалізація нашого корпусного проекту є

новаторськими для української діалектології та корпусної лінгвістики. Розвиток проекту передбачає виконання таких завдань, як шліфування методики опрацювання діалектних текстів на засадах багаторівневого максимально точного їх маркування; розроблення відповідного програмного забезпечення та онлайн-інтерфейсу, «дружніх» як до діалектологів-фахівців, так і до користувачів ширшого профілю; наповнення корпусу текстами українських говірок всіх регіонів України та закордоння.

Література

1. Демська О. Текстовий корпус: ідея іншої форми [Текст] / Орися Демська. – К. : ВПЦ НаУКМА, 2011. – С. 162.
2. Сірук О.Б.: «Корпус українських діалектних текстів» (КорУДіТ) [Текст] / О. Б. Сірук // Мовні і концептуальні картини світу. – К. : ВПЦ «Київський університет». – Вип. 37. – 2011. – С. 293-300.

**Василь Старко**
Волинський національний університет імені Лесі Українки,
**Наталія Чейлитко**
Київський національний університет імені Тараса Шевченка

## Концепція створення Браунського корпусу української мови

Браунський корпус (Standard Corpus of Present-Day Edited American English, або скорочено Brown Corpus) був першим машиночитним загальномовним корпусом, підготовленим до дослідження сучасної англійської мови. Його створили В. Нельсон Френсис та Г. Кучера в Браунському університеті у 1960-х роках. Корпус мав обсяг близько 1 млн слів і складався з 500 уривків. Кожен уривок мав довжину 2000 або більше слів суцільного зредагованого тексту англійською мовою й був опублікований у США 1961 року.

Цей корпус надихнув дослідників на створення цілої родини корпусів – ЛОБ (LOB, Lancaster-Oslo/Bergen Corpus, британський еквівалент Браунського корпусу), Фраун (Frown, Freiberg-Brown Corpus of American English) та Ф-ЛОБ (F-LOB, Freiburg–LOB Corpus of British English). Останні два корпуси – це еквіваленти початкового Браунського корпусу та корпусу ЛОБ, які подають зріз англійської мови початку 1990-х років.

За зразком Браунського корпусу створюють й корпусу інших мов, наприклад, Болгарський Браунський корпус [1]. Побудова таких корпусів має низку переваг. Вони подають збалансовану картину вибраного сегменту мови (зредагована писана проза) й у такий спосіб уможливлюють плідні лінгвістичні дослідження на репрезентативному мовному матеріалі, а також порівняльні дослідження – внутрішньомовні й міжмовні. Браунський корпус може бути зручним інструментом вивчення мови й опанування основ корпусної лінгвістики. До того ж, він дає змогу випробувати й вишліфувати моделі й технології укладання корпусів, а згодом стати основою більшого корпусу із синтаксичним і семантичним маркуванням. Ці

міркування пояснюють доцільність створення Браунського корпусу української мови й слугують дороговказом для його укладачів.

Спираючись на досвід «Браунської родини корпусів» й пристосовуючи їх до особливостей побутування української мови, окреслімо принципи наповнення Українського Браунського корпусу. Попри те, що бажано дотримуватися структури оригінального Браунського корпусу, вважаємо, що варто допустити певні відхилення, як це зробили, наприклад, болгарські лінгвісти. Сформулюймо вимоги до текстів у корпусі.

1. Оригінальні (тобто неперекладні) твори.
2. Твори, створені й опубліковані за відносно короткий проміжок часу (до 10 років).
3. За змоги тексти мають бути зредаговані, окрім тих, щодо яких цю ознаку в принципі неможливо встановити.
4. Підбір текстів за категоріями й підкатегоріями має відповідати первісному Браунському корпусу. Виняток становить категорія F, у якій жанр «вестерн» можна замінити іншою пригодницькою літературою, представленою в Україні.
5. Корпус складається з 500 фрагментів довжиною 2000 слів плюс залишок до кінця речення.
6. Фрагмент повинен бути витяжкою з одного тексту, окрім випадків, коли фрагмент складається з кількох коротких текстів кількох авторів (короткі новини).
7. Тексти мають бути підібрані згідно з класифікацією за типом (інформативні й художні тексти), категорією, підкатегорією. Браунський корпус має 15 категорій, поділених на різну кількість підкатегорій [2].

Порівняння Браунського корпусу й ЛОБ засвідчує, що немає потреби (а іноді й змоги) копіювати структуру взірцевого корпусу до найменших дрібниць. Наприклад, співвідношення між книжками й періодичними виданнями в межах підкатегорії чи щоденними й тижневими періодичними виданнями (в категоріях A–C) може різнитися.

Отже, створення Українського Браунського корпусу за вказаними параметрами дасть змогу створити корисний дослідницький і навчальний ресурс. Інші принципи побудови корпусу варто розглянути окремо.

Література

1. http://dcl.bas.bg/Corpus/home_bg.html – режим доступу: 19.01.2012 р.
2. http://icame.uib.no/brown/bcm.html – режим доступу: 19.01.2012 р

**Богдан Шуневич**
Львівський державний університет безпеки життєдіяльності

## Українсько-англійський комп’ютерний словник пожежно-технічних термінів: лексичні матеріали, програмне забезпечення

Зарубіжний і вітчизняний ринок програмного забезпечення пропонує велику різноманітність комп’ютерних словників. Серед відомих українських комп’ютерних

словників можна назвати, наприклад, інтегровану лексикографічну систему «Словники України» Інституту мовно-інформаційних досліджень НАН України, систему електронних навчальних словників «ГЛОСА» та ін. Лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету, електронний багатотематичний тлумачний словник MultiLock галузевого Нормативно-термінологічного центру нафтогазового комплексу, системи PolyDic v. 1.0, PolyDic ML 3.0 Технічного комітету стандартизації науково-технічної термінології Держспоживстандарту та Міністерства освіти і науки, молоді та спорту України.

Комп'ютерним словником називають «словник, процедури укладання якого здійснює комп'ютер» [1].

Мета доповіді – провести порівняльний аналіз програмного забезпечення PolyDic v. 1.0 [2], за допомогою якого укладається «Англійсько-український комп'ютерний словник з робототехніки» [3], і лінгвістичної бази даних Військового інституту Київського національного університету імені Тараса Шевченка [4], яка запропонована викладачам кафедри іноземних мов та технічного перекладу Львівського державного університету безпеки життєдіяльності для укладання «Українсько-англійського комп'ютерного словника пожежно-технічних термінів» [5].

Для укладання «Англійсько-українського комп'ютерного словника з робототехніки" нами використовується програмне забезпечення з вільним кодом (open source), а саме PolyDic, версія 1.0, яке укладено під керівництвом Романа Мисака (Національний університет «Львівська політехніка») [2].

Система укладання та перегляду електронних словників PolyDic складається з двох програм: PolyDic Editor – для набирання, форматування та редагування словникових баз даних та PolyDic Viewer – для перегляду електронних словників.

Пошук терміна у PolyDic Editor відбувається за першими уведеними літерами в інтерактивному режимі у вікні пошуку. Програма запам'ятовує історію (почерговість) переглянутих статей, в якій зберігається послідовність до десяти переглянутих статей. Особливістю PolyDic та її істотною перевагою над іншими електронним словниками є механізм фільтрів, який можна застосувати до термінів, тексту статей або до них обох одночасно. Наприклад, можна залишити видимими слова, які починаються на "абр", або слова, що мають закінчення "ан". У статтях, у разі потреби, встановлюються зв'язки з іншими статтями, які можна переглянути просто натиснувши курсором миші по відмітці зв'язку. Переклади або тлумачення супроводжуються короткими поясненнями, наприклад, щодо галузі застосування.

У програмі PolyDic Editor передбачено можливість набору словникової бази частинами з подальшим злиттям цих частин в єдину базу. Ця функція корисна під час розподілення праці з уведення інформації у базу між різними операторами-користувачами.

Система PolyDic v. 1.0 розроблена для формування комп'ютерних словників, паперові версії яких уже були укладено або видано, і їх макро- та мікроструктура повністю відповідала паперовим версіям. До недоліків цієї системи можна віднести: обмежену кількість мов перекладних словників (дві); вбудований в систему шрифт не підтримує Unicode і не дає змогу вводити літери з діакритичними знаками; система не підтримує мультимедійні об'єкти.

«Українсько-англійський комп'ютерний словник пожежно-технічних термінів» заплановано створити в рамках лінгвістичної бази даних Військового інституту Київського національного університету імені Тараса Шевченка.

Метою розроблення бази даних є створення загальнодоступного і високоякісного порталу з військової і, в тому числі, пожежно-технічної термінології, а також уніфікація і стандартизація військово-технічних термінів.

На порталі представлена вся необхідна інформація про термін (переклад, пояснення, фото та відео матеріали). Користувачі можуть не тільки користуватися термінологічною базою, а й одночасно приймати активну участь в її розширенні та поліпшенні шляхом додавання відсутньої інформації.

Проект передбачає створення багатомовної термінологічної бази даних військово-технічних термінів, поки що англійською, німецькою, французькою. російською та українською мовами.

Досвід укладання комп'ютерних словників дасть можливість апробувати вітчизняне програмне забезпечення для укладання комп'ютерних словників, а також порівняти різні параметри цих словників, вибрати кращий варіант програмного забезпечення для подальшої словникової роботи в нашому та інших університетах.

Література

1. Карпіловська Є. А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика. – Донецьк: ТОВ «Юго-Восток, Лтд», 2006. – 188 с.
2. Система укладання та перегляду електронних словників PolyDic v.1.0. – Режим доступу до Веб-сторінки: www.lp.lviv.ua. – Заголовок з екрана, 2011.
3. Шуневич Б. Проект англійсько-українського комп'ютерного словника з робототехніки / Б. Шуневич, В. Голтвян, М. Маляр // Лінгвістичні проблеми та інноваційні підходи до викладання чужоземних мов у вищих навчальних закладах, м. Львів, 28-30 жовтня 2010 р. – Львів: ЛДУ БЖД. – С. 83.
4. Лінгвістична база даних Військового інституту Київського національного університету імені Тараса Шевченка. – Режим доступу до Веб-сторінки: http://www.mildic.com/admin. – Заголовок з екрана, 2011.
5. Короткий українсько-англійський словник зі сфери надзвичайних ситуацій / Вовчаста Н.Я., Бугайська О.В. та ін. (За ред. Коваля М., Шуневича Б.). – Львів: Вид-во ЛДУ БЖД, 2010. – 184 с.

**Катерина Яковенко**
Київський національний університет імені Тараса Шевченка

## Створення лінгвістичного корпусу у міжмовних експериментально-фонетичних дослідженнях

Питання створення лінгвістичного корпусу постає перед кожним дослідником на одному з перших етапів його мовознавчих студій. Правильний підбір корпусу даних є запорукою успішності експерименту, адже це матеріал, на основі якого базуватимуться всі подальші висновки і від якого залежатиме істинність отриманих

кінцевих результатів, а також можливість їх використання як достовірного джерела для майбутніх досліджень.

Інформаційна база даних, що слугує основою для експериментів у галузі експериментальної фонетики, суттєво різниться від лінгвістичного корпусу для аналізу граматичних явищ, функціональних аспектів мовних одиниць або вивчення статистичних параметрів мови. Підставою для цього є не лише якісно інший об'єкт і предмет дослідження, але й чинники, що мають бути враховані при його аналізі. Так, якщо зазвичай під поняттям «лінгвістичний корпус» розуміють масиви текстів, то у фонетиці мова йде про масив слів, об'єднаних відповідним до мети фонетичного дослідження принципом.

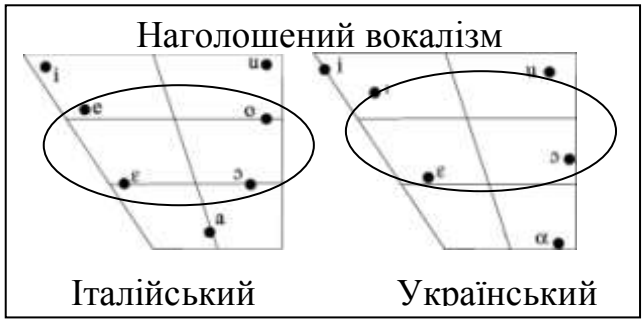
Детальний опис особливостей проведення фонетичних досліджень, практичні поради щодо створення лінгвістичного корпусу, вибору мовців, систем аудіозапису та подальшого аналізу даних подає американський фонетист Пітер Ладефогед у праці «Аналіз фонетичних даних: вступ до експериментально-фонетичних досліджень та інструментальних технологій».

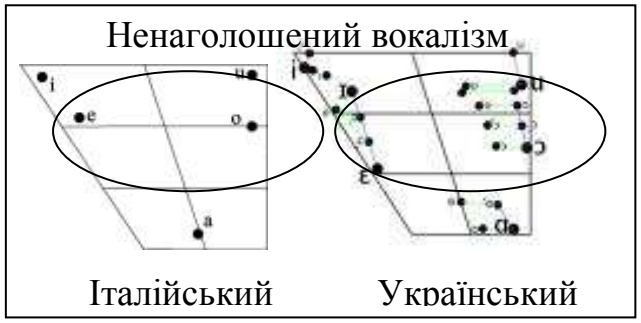
Корпус для міжмовних фонетичних студій і вивчення явища іншомовного акценту має свою специфіку. Так, у ході експериментально-фонетичного дослідження італійського мовлення українців з метою вивчення особливостей засвоєння чужомовного вокалізму, були проаналізовані критерії, які необхідно враховувати при виборі мовного матеріалу для акустичного запису. У їхній основі лежить суть самого явища іноземного акценту, що полягає в накладанні фонетичної системи рідної мови на іноземну за таким алгоритмом:

- звуки та їхні опозиції в іноземній мові (М2), ідентичні або «схожі» на звуки рідної (М1), будуть замінені звуками рідної мови (М1):
$[a_1 \approx a_2 \rightarrow a_1]$, $[a_1{:}b_1 \approx a_2{:}b_2 \rightarrow a_1{:}b_1]$;
- опозиції звуків іноземної мови (М2), відсутні або «нові» для рідної (М1), не дотримуватимуться: $[a_2{:}b_2 \neq M1 \rightarrow ?]$;
- опозиції звуків рідної мови (М1), відсутні в іноземній (М2), все одно зберігатимуться в іноземній мові (М2): $[a_1{:}b_1 \neq M2 \rightarrow a_1{:}b_1]$.

Важливо наголосити на спостереженні за процесом засвоєння звуків, позначених як «нові» і «схожі»: якісно нові звуки іноземної мови засвоюються неносіями мови значно швидше, ніж схожі. Це пояснюється складністю перцепції тонкої межі між акустично подібними чужомовними і вже існуючими у фонетичній базі мовця фонемами, а часом і відсутністю їхнього графічного розрізнення на письмі.

Так, наприклад, досліджуючи засвоєння італійського вокалізму українцями, спочатку варто проаналізувати самі системи вокалізму обох мов у їхніх наголошених та ненаголошених позиціях (використання символів МФА).



З порівняльної таблиці одразу привертає увагу перехід від системи з трьома голосними середнього ступеня підняття в українській мові /ɪ/, /ɛ/, /ɔ/ до системи з чотирма голосними середнього ступеня підняття, що поділяються у свою чергу на високо-середні /e/, /o/ та низько-середні /ɛ/, /ɔ/, творячи мінімальні пари слів за критерієм відкритості-закритості. Іншою є також локалізація голосної /a/ за місцем творення – в італійській мові вона середнього ряду, тобто дещо більш просунута вперед, ніж українська /ɑ/ заднього ряду.

Так, можемо заздалегідь припустити, які труднощі виникають в італійському мовленні українців: недотримання "нових" опозицій голосних середнього ступеня підняття /e/-/ɛ/ та /o/-/ɔ/ і заміна їх українськими /ɛ/ та /ɔ/ відповідно, а також задній, а не середній характер «схожої» голосної /a/.

Найдоцільнішим корпусом для цього дослідження є пари міжмовних омонімів та омофонів: 49 пар слів, у яких всі голосні представлені у різних позиціях і обох мовах мають аналогічне розташування у слові і місце по відношенню до наголосу. Таким чином, близько 6 прикладів припадає на кожну міжмовну опозицію наголошених фонем: /ɔ/-/ɔ/, /o/-/ɔ/, /e/-/ɛ/, /ɛ/-/ɛ/, /e/-/ɪ/, /ɛ/-/ɪ/, /i/-/i/, /u/-/u/, /a/-/ɑ/ – та ненаголошених алофонів: [a]-[ɑ], [o]-[ɔ$^{u}$], [o]-[ɔ], [u]-[u], [e]-[ɛ], [e]-[ɛ$^{ɪ}$], [e]-[ɪ$^{ɛ}$], [i]-[i]. Також важливо, щоб під час запису ключові слова вимовлялися в середині відповідних «штампованих» речень з метою нейтралізувати дію просодичних явищ.

Враховуючи те, що вивчення мови, зокрема засвоєння міжмовних фонетичних відмінностей, великою мірою залежить як від наявності хорошої теоретичної бази, так і іноземного мовного середовища, лінгвістичний корпус буде запропоновано прочитати таким категоріям мовців:

1 група – студенти 3 курсу відділення італійської філології, які мають хорошу теоретичну базу, але орієнтуються переважно на писемне мовлення (приклад аналітичного вивчення мови);

2 група – українці, які живуть в Італії протягом 3 років і вивчали мову з «почутого» в середовищі її природних носіїв, але їм бракує основ граматики;

3 група – українці, що жили в Італії протягом певного часу в дитинстві, сприймали мову на слух і «спонтанно» засвоювали фонетичну базу.

4 група – італійці району Тоскана в Італії (недіалектне мовлення).

Усі зазначені особливості мовної інтерференції були враховані при виборі лінгвістичного корпусу для нашого дослідження, тож результати подальшого зіставлення і аналізу акустичних показників голосних української та італійської мов слугуватимуть підтвердженням або ж запереченням попередньо висунутої гіпотези.

## ЗМІСТ