

О. М. Зубань

ЗАДАЧИ И МЕТОДЫ АВТОМАТИЧЕСКОГО МОРФЕМНОГО АНАЛИЗА В КОРПУСЕ УКРАИНСКОГО ЯЗЫКА

Корпус украинского языка [1] (в оригинале Корпус української мови – КУМ) имеет исследовательский характер и по своему назначению ориентирован на решение широкого круга лингвистических задач, в частности

в области морфемики и словообразования, хотя создатели большинства корпусов славянских языков не ставят задачу параметризации текстов на морфемном уровне: по нашим сведениям, кроме Корпуса украинского языка, только Национальный корпус русского языка [2] имеет поиск по словообразовательным параметрам, но без аннотирования посткорневой зоны. Возникает ряд вопросов: нужна ли в корпусе морфемная (словообразовательная) аннотация текста? Какие исследовательские перспективы открывает перед лингвистом корпусноориентированный автоматический морфемный и словообразовательный анализ? Как методически правильно создать такой анализ?

Работая над проблемами конструирования параметризованных лингвистических баз, коллектив лаборатории компьютерной лингвистики Киевского национального университета имени Тараса Шевченко разработал методику компьютерного моделирования морфемной структуры украинского слова и в 2001 г. создал автоматизированную систему морфемно-словообразовательного анализа (АСМСА) украинского языка [3; 4]. АСМСА – лексикографическая система (база данных), выполняющая различные задачи автоматической систематизации лексики (исходных форм) в области морфемики и словообразования: группирует лексику в общекорневые и общеаффиксальные классы; классифицирует слова общекорневой выборки по словообразовательным тактам; классифицирует лексику по структурно-морфным моделям слов; создает алфавитно-частотные корневые и аффиксальные словари.

Система АСМСА создавалась на базе ≈ 170 тыс. слов, разбитых на морфемы по принципам морфемного членения словаря И.Т. Яценко [5]. Морфемный анализ проводился автоматизировано с помощью удобного компьютерного инструмента, в котором каждое слово анализировалось в отдельной электронной карточке (рис. 1). По методике моделирования структурно-функциональных связей морфов в слове каждому слову приписывалась модель, определяющая границы и функциональный тип морфа в морфемной структуре слова.

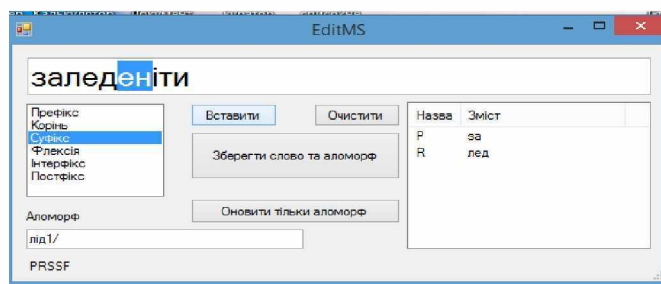


Рис. 1. Электронная карточка морфемного анализа слова *заледеніти*

Маркировка морфем в электронной карточке слова автоматически приписывает каждому слову формулу программной процедуры морфемной сегментации: *заледеніти* P2R5S7S8F10, где буквы обозначают функциональный тип морфа (P – приставка, R – корень, S – суффикс, F – флексия, I – интерфикс, X – постфикс), а цифры – границы морфа по порядковому номеру (с начала слова) конечной графемы каждого морфа (за – 2; лед – 5; ен – 7; і – 8; ти – 10). Таким образом, формула программной процедуры морфемной сегментации отображает функциональную структуру слова на морфемном

уровне – PRSSF, а субстанциональное выражение этой структуры представляется графемно-цифровой моделью, автоматически переводимой в базе данных на латиницу по номеру буквы в алфавите. Информация о слове с программной процедурой раньше записывалась в res.ini-файлы (рис. 2), объединяемые в единую базу данных. Сегодня работа с базой данных происходит в on-line режиме.

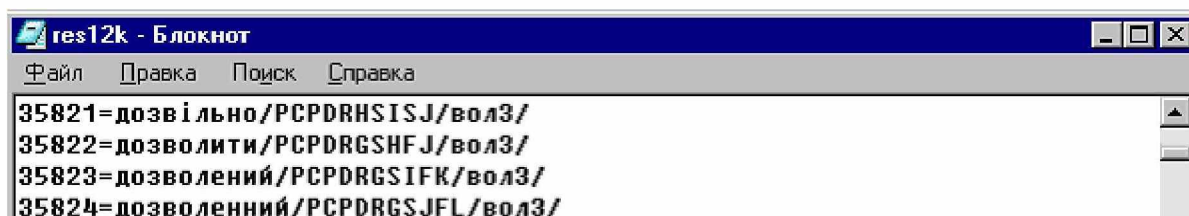


Рис. 2. Фрагмент res.ini-файла

С целью формирования общекорневых выборок в процессе создания морфемной базы данных (МБД) учитывалась омонимия и алломорфия корневых морфов: 1) омонимия снималась приписыванием каждому корню-омониму цифрового кода: в существительном *лід-ер-Ø* приписывается корень *лід2*, а в глаголе *за-лед-ен-і-ти* – корень *лід1*; 2) алломорфным корням приписывался условно первичный алломорф – маркер инварианта корневой морфемы: *за-лед-еніти* – *лід1*, *льод-овий* – *лід1*.

Сегодня система АСМСА объединяет три базы данных: 1) морфемных структур слов (исходных форм) ≈ 200 тыс.; 2) базу корней-алломорфов ≈ 2500; 3) базу корней-омонимов ≈ 3100 корней.

Система АСМСА с начала своего создания ориентировалась на проведение автоматической морфемной сегментации текстовых словоупотреблений. Использование АСМСА в процедуре автоматической морфемной сегментации в Параметризированной базе данных поэтической речи, созданной коллективом лаборатории компьютерной лингвистики в 2004 г., демонстрирует опыт автоматической морфемной аннотации текста.

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	.	

Рис. 3. Фрагмент морфемной сегментации словоупотреблений поэтических текстов Л. Костенко

На рис. 3 показано морфемную сегментацию словоупотреблений текстового фрагмента *...за щитом смарагдових лісів*. Каждое словоупотребление сегментируется на морфы с автоматическим приписыванием двух программных процедур: первая – морфемная модель начальной словоформы лексемы; вторая – морфемная модель словоформы текстового фрагмента.

В 2010 г. началось создание Корпуса украинского языка, составляющего на сегодня корпус текстов, параметризованных на уровне морфологической аннотации, объемом ~ 44 млн. словоупотреблений. В процессе создания автоматического морфемного анализа в Корпусе с целью оптимизации поиска на больших текстовых массивах мы отказались от метода морфемной аннотации текста. Морфемная разметка текстовых словоупотреблений не производится, тексты Корпуса используются только как источник для составления разного рода алфавитно-частотных морфемных словарей по текстовым выборкам с помощью системы АСМСА, исполняющей функцию морфемного модуля-анализатора в Корпусе украинского языка.

На этапе работы морфологического модуля все словоупотребления текста с морфологической разметкой лемматизируются в исходные формы и попадают в морфемный модуль в виде алфавитно-частотного словаря исходных форм с информацией о части речи, частотных характеристиках, контекстах употребления и морфемной модели слова, автоматически приписываемой исходной форме способом сопоставления с этим же словом в АСМСА (рис. 4).

Код	wrd	cls	xabs	xsred	sigma	v	R	context	d	morphem
396	вгасати	Г	2	0,0327868852	0,1780783687	5,4313902456	2	Тліє, не вгасає	2,988105	PRSF
397	вголос	Н	1	0,0163934426	0,1269830605	7,7459666924	1	А то думають	1,110223Е-15	PR
398	вгородити	Г	1	0,0163934426	0,1269830605	7,7459666924	1	Виквятила ту с	1,110223Е-15	PRSF
399	вгороду	Н	4	0,0655737705	0,5079322421	7,7459666924	1	Все вгороду та ви	1,110223Е-15	PRS
400	вдавити	Г	1	0,0163934426	0,1269830605	7,7459666924	1	Чому мене не	1,110223Е-15	PRSF
401	вдарити	Г	2	0,0327868852	0,1780783687	5,4313902456	2	Налигачем скр	2,988105	RSF

Рис. 4. Фрагмент морфемной базы данных исходных словоформ
вгасати, вголос, вгородити, вдавити, вдарити

На выходе морфемного модуля – те же исходные формы слов, сегментированные на морфемы, с информацией о функциональном типе морфемы (рис. 5).

morphemid	wrdFK	mname	mseg
799	396	P	в
800	396	R	гас
801	396	S	а
802	396	F	ти
803	397	P	в
804	397	R	голос
805	398	P	в
806	398	R	город
807	398	S	и
808	398	F	ти
809	399	P	в
810	399	R	гору
811	399	S	у
812	400	P	в
813	400	R	дав
814	400	S	и
815	400	F	ти
816	401	R	вдар
817	401	S	и
818	401	F	ти

Рис. 5. Фрагмент базы данных морфемной сегментации
исходных словоформ *вгасати, вголос, вгородити, вдавити, вдарити*

На основании этих баз по заказу пользователя создаются разного рода алфавитно-частотные словари морфем и морфемных структур слов по текстовым выборкам отдельных авторов или подкорпусов стилей. Эти словари выложены в рубрике "Частотні словники" Корпуса как автономные электронные лексикографические системы [6]. Рассмотрим автоматическое составление морфемных словарей на базе ≈ 80 тыс. словоупотреблений поэтических текстов Т. Шевченко [7].

Инвентарь единиц частотного словаря морфем составляется по двум параметрам – тип морфемы и часть речи, выбираемые пользователем перед построением словаря, в данном случае – корни имен существительных. Частотный словарь корней существительных по текстам Т. Шевченко (рис. 6) состоит из 1509 морфем, каждой из которых приписана информация об абсолютной и средней частоте. Словник строится по алфавиту или частотности: спаду / росту абсолютной частоты.

Морфемно-частотний словник Коренів всього записів: 1509					Частотний словник по морфемі: R:люд, частина мови: Іменник Всього записів: 7		
Морфема	Структури	Слова	Абсолютна частота	Середня частота	Слово	Частина мови	Абсолютна частота
бу	Структури	Слова	531	8,70	люд	ім. ч. р.	12
пі	Структури	Слова	380	6,23	люди	ім. множ.	270
ста	Структури	Слова	340	5,57	людина	ім. ж. р.	1
він	Структури	Слова	335	5,49	людоїд	ім. ч. р.	3
бог	Структури	Слова	306	5,02	людомор	ім. ч. р.	1
люд	Структури	Слова	306	5,02	недолюд	ім. ч. р.	3
світ	Структури	Слова	287	4,70	недолюдек	ім. ч. р.	2
да	Структури	Слова	280	4,59	1		
каз	Структури	Слова	280	4,59			
див	Структури	Слова	273	4,48			
1 2 3 4 5 6 7 8 9 10 ... >>							
Всього записів: 5					Частотний словник по морфструктурі: RIRSF з морфемою -люд- Всього записів: 2		
Структура	Абсолютна частота				Слово	Частина мови	Абсолютна частота
PRF	3				людоїд	ім. ч. р.	3
PRSF	2				людомор	ім. ч. р.	1
RF	282				1		
RIRSF	4						
RSF	1						

Рис. 6. Фрагмент частотного словаря корней (инвентарь и лексическая реализация)

Активация опции «Слова» конкретного корня открывает окно лексической реализации этой морфемы (Частотный словник по морфемі): фрагмент иллюстрирует реализацию частотного корня *-люд-* (абсолютная частота – 306 словоупотреблений) в семи словах разной продуктивности, самым частотным из которых является слово *люди* (270 текстовых словоупотреблений). Активация опции "Структура" этого корня открывает список моделей морфемных структур (нижняя первая таблица), в которых реализуется корень *-люд-*, а активация конкретной модели морфструктуры открывает окно лексической реализации выбранной структуры (Частотный словник по морфструктурі): фрагмент демонстрирует реализацию модели RIRSF с корнем *-люд-* в двух словах: *люд/о/їд/Ø/Ø*, *люд/о/мор/Ø/Ø*.

Частотний словник збірки "Твори в п'яти томах". Тарас Шевченко	
Конкорданс до слова: люди (ім. множ.)	
Морфемна структура: RF /люд/и/	
Контекст	Джерело
Пошли ж ти їй долю — вона молоденька, Бо люде чужі їй засміють.	>>
На чужині не ті люде — Тяжко з ними жити!	>>
Кохайтесь, чорнобриві, Та не з москалями, Бо москалі — чужі люде, Роблять лихо з вами.	>>
Серце в'яне співаючи, Коли знає, за що; Люде серця не побачать, А скажуть — ледащо!	>>
Кохайтесь ж, чорнобриві, Та не з москалями, Бо москалі — чужі люде, Згнуцаються вами.	>>
Нехай собі ті люде, Що хотять, говорять: Вона любить, то їй не чує, Що вкралося горе.	>>
Тойді Катерина Буде собі московкою, Забудеться горе; А поки що, нехай люде, Що хотять, говорять.	>>
Батько, мати — чужі люде, Тяжко з ними жити!	>>
ПРИЧИННА ("Рече та стогне Дніпр широкий..."):ТВОРИ:ТАРАС ШЕВЧЕНКО:ПОЕТИЧНА	
МОВА	
Стиль:поетичні тексти	
Приблизна кількість словоживань: 849	
Видано:Київ: Дніпро:	
Жанр: вірш	
Місце: Україна / Росія: XIX ст.	

Рис. 7. Фрагмент конкорданса лексеми *люди*

Активация выбранного слова в "Частотном словнике по морфемі" (верхняя правая таблица) открывает окно контекстов употребления слова (рис. 7). Каждый текстовый фрагмент опцией "Джерело" связан с экстралингвистической информацией о произведении в Корпусе: на рис. 7 демонстрируется информация о произведении по первой цитате конкорданса.

Автоматический частотный словарь морфемных структур слов имеет такую же организацию: частотный словарь морфструктур, реализация конкретной морфструктуры в лексиконе выборки, реализация слова выбранной морфструктуры в текстах. Рис. 8. иллюстрирует реализацию высокочастотной морфструктуры PRSF (3804 текстовых словоупотреблений) в 95 существительных по текстам Т. Шевченко.

Всього записів: 39			Частотний словник по морфструктурі: PRSF, частина мови: Іменник Всього записів: 95 Всього записів: 95						
Структура	Абсолютна частота	Середня частота	Слово	Частина мови	Абсолютна частота	Слово	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності
RF	16323	267,59	пожар	ім. ч.	20	11	0,327868852459016	0,740432117418768	2,25831795812724
RSF	7482	122,66		р.					
R	6605	108,28							
PRSF	3804	62,36	невольник	ім. ч.	15	11	0,245901639344262	0,644160088668475	2,61958436058513
PRF	1631	26,74		р.					
RSSF	898	14,72	пророк	ім. ч.	12	7	0,19672131147541	0,697058741818665	3,54338193757822
RS	604	9,90		р.					

Рис. 8. Инвентарь морфструктур и лексическая реализация морфструктуры PRSF

Использование автоматического морфемного анализа в Корпусе украинского языка раскрывает новые возможности исследования морфемной организации украинского слова в словаре и тексте. Полученные статистические данные могут использоваться в системных стилистических исследованиях [8]: сравнение частотных характеристик морфемных единиц в разных текстовых выборках Корпуса и в системе языка прогнозирует вероятность реализации морфемных единиц в разных текстах и формирует понятие статистической структуры текста на морфемном уровне как его стилеметрической модели. Систематизация морфструктур позволяет анализировать включение морфемных моделей в образование новых слов Корпуса, исследовать морфемную длину и глубину слов в украинских текстах разных стилей, а также морфотактику разных типов морфем. Связь частотных морфемных словарей с конкордансом позволяет проанализировать разноаспектное функционирование морфем и морфструктур в предложениях.

Опыт создания морфемного автоматического анализа показывает, что с целью извлечения из текста реляционно-функциональных характеристик морфемных единиц не обязательно проводить морфемную или словообразовательную разметку текстов. Методика морфемной аннотации словаря лексем (исходных форм) текстов не уменьшает эффективности и оперативности организации лингвистического исследования, а наоборот, за счет систематизации и разноаспектной классификации морфемной информации повышает разрешающую возможность исследования. Изучение организации текста на морфемном уровне посредством морфемного анализа исходных

форм слов, а не словоупотреблений, оправдан практикой создания частотных морфемных словарей в Корпусе украинского языка. Предложенный подход базируется на онтологическом устройстве морфемной структуры слов флективных языков: морфемная структура основы слова, отображающая эксплицитованную семантику слова, при словоизменении остается относительно стабильной: количество морфем не меняется, может возникать только словоизменительная алломорфия, но она учитывается при лемматизации и может автоматически приписываться морфемам как потенциальный признак. Использование методики морфемной аннотации словаря исходных форм в Корпусе украинского языка демонстрирует эффективность и оптимальность этой методики: ≈ 200 тыс. единиц морфемной базы данных АСМСА позволяют получить информацию о морфемной организации ≈ 44 млн. текстовых словоупотреблений с иллюстрацией контекстов их употреблений. Недостаток этой методики – невозможность автоматически снять омонимию слов одной части речи в случае разной морфемной сегментации омографов: *вида-ти* (напечатать), *ви-да-ти* (дать), *вид-а-ти* (видеть). В таких случаях морфемная сегментация редактируется вручную.

Перед разработчиками АСМСА на данном этапе стоит задача автоматического построения на базе реестра АСМСА и лексики текстов Корпуса общекорневых выборов и словообразовательных гнезд с учетом омонимии и алломорфии корней; общеаффиксальных выборов с учетом омонимии и алломорфии аффиксов. В методологии компьютерного моделирования, использованной в построении АСМСА, обобщены теоретические и прикладные идеи современного языкознания, что делает эту систему эффективным и рациональным инструментом лингвистических исследований.

ЛИТЕРАТУРА

1. Корпус украинского языка [Электронный ресурс]. – Режим доступа : <http://www.mova.info/corpus.aspx>. – Дата доступа : 18.04.2016.
2. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа : <http://www.ruscorpora.ru>. – Дата доступа : 18.04.2016.
3. Дарчук, Н. П. Комп'ютерне анування тексту: результати і перспективи : монографія / Н. П. Дарчук. – Київ, 2013.
4. Алексієнко, Л. А. Методика створення автоматизованої системи морфемно-словотвірної аналізу (АСМСА) слів української мови / Л. А. Алексієнко, Н. П. Дарчук, О. М. Зубань // Наукова спадщина професора С. В. Семчинського : зб. наук. пр. – Київ, 2001, – Ч. 1. – С. 38–49.
5. Яценко, І. Т. Морфемний аналіз: словник-довідник / І. Т. Яценко. – Київ, 1980–1981. – Т. 1–2.
6. Частотні словники Корпусу [Электронный ресурс]. – Режим доступа : <http://www.mova.info/article.aspx?l1=210&DID=5215>. – Дата доступа : 18.04.2016.
7. Електронний словник мови Т. Шевченка [Электронный ресурс]. – Режим доступа : <http://www.mova.info/cfqsh.aspx>. – Дата доступа : 18.04.2016.
8. Зубань, О. Особливості морфемної будови слів у поетичних текстах Т. Шевченка (на матеріалі Корпусу української мови) / О. М. Зубань // Українське мовознавство. – № 44/1. – Київ, 2014. – С. 123–133.