Т.В. Бобкова, Л.М. Гриднева,
К.М. Лебедев, В.И. Перебейнос

ПРИНЦИПЫ КОДИРОВАНИЯ ЧАСТЕЙ РЕЧИ В АНГЛО-УКРАИНСКОМ КОРПУСЕ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ ДОКУМЕНТОВ НАТО

1. Введение

Данное исследование посвящено одному из важнейших аспектов создания параллельного (двуязычного: английского и украинского) корпуса текстов официально-делового стиля, а именно: принципам кодирования и морфологической разметке текстов документов по определённой тематике. Составляющими корпуса текстов являются тематически однородные, но разножанровые тексты: 1) официальные документы НАТО[1]; 2) дискуссии, касающиеся определённых вопросов и 3) письма[2]. В корпус были включены полные тексты документов, что обеспечило их структурную и лексическую завершенность. Объём английских текстов документов НАТО составляет 306 078, а украинских 310 884 словоупотреблений.

2. Основные принципы кодирования

Лингвистический анализ и переводы текстов с одного языка на другой требуют внимания в первую очередь к морфологической разметке текстов, принадлежащих к языкам разных грамматических систем. Для каждой части речи разработана система буквенно-цифровых кодов. Осуществлена такая задача ранее при

[1] Справочник НАТО. Брюссель, 2001; URL: **www.nato.int/ukraine**
[2] Публікації НАТО. 2004.

создании исследовательских корпусов украинских текстов публицистического стиля и поэтической речи[1].

Создание унифицированной системы кодов для англо-украинского корпуса текстов сопряжено с объективными трудностями, связанными прежде всего с различными типологическими характеристиками исследуемых языков. В украинском языке грамматическое значение в основном выражается синтетически. Для синтетического способа характерным является соединение грамматического средства выражения значения с самим словом[2]. Грамматическое значение выражается с помощью окончания, суффикса, приставки, изменения ударения, внутренней флексии, супплетивного видоизменения.

В отличие от вышесказанного в английском языке преобладает аналитический способ выражения грамматического значения – за пределами слова, с помощью предлогов, союзов, артиклей, вспомогательных глаголов, других служебных слов, порядка слов в предложении[3].

Одна из главных задач исследования двуязычного (параллельного) корпуса текстов на данном этапе состоит в сопоставлении кодов и кодовых цепочек, для того чтобы можно было быстро и качественно соотносить тематически однородные фрагменты, переводить документы с одного языка на другой, создавать словники по одному типу текстов и по группам текстов; сравнивать способы выражения одних и тех же понятий в речевом материале разных языков.

Так, если в украинском языке для существительных фиксируется несколько грамматических категорий: род, число, падеж и

---

[1] *Darchuk N., Sorokin V.* The Text Corpus and Dictionary Hierarchy // Computer Treatment of Slavic and East European Languages. Bratislava, Slovakia, 2007. P. 38–42; URL: **www.mova.info**

[2] *Мечковская Н.Б.* Общее языкознание. Структурная и социальная типология языков. М., 2001. 312 с.

[3] Там же.

pluralia tantum, то в английском эта система представлена только числом и формами possesive case и pluralia tantum (см. табл. 1).

Таблица 1. Коды украинских и английских существительных

Украинский		**Английский**	
Й	Им. м.р.	Й	Им. (Noun)
К	Им. ж.р.		
Л	Им.ср.р.		
И	Pluralia tantum	И	Pluralia tantum
й	Им. м.р., имя собств.	й	Им., имя собств. (Proper Noun)
к	Им. ж.р., имя собств.		
л	Им. ср.р., имя собств.		
и	Pluralia tantum, имя собств.	и	Pluralia tantum, имя собств. (Proper Noun)

Для обозначения падежных и звательных форм украинских существительных в сочетании с приведёнными выше используются буквенные коды для единственного и множественного числа. Для английских существительных актуальными являются коды форм Common Case и Possesive Case для единственного и множественного числа. На данном этапе работы не различаются подклассы английских существительных по способу образования множественного числа.

Система кодирования украинских прилагательных также более разветвлена, чем английских: фиксируется род, число, падеж, степени сравнения; для английских прилагательных – только степени сравнения, так что одному английскому коду соответствует несколько кодов украинских существительных и прилагательных. Подобно этому различаются формы степеней сравнения английских наречий.

При кодировании украинских и английских местоимений различаются местоимения-существительные, местоимения-прилагательные, притяжательные, а в украинских текстах и возвратное местоимения.

Что касается глагола, то английская система гораздо сложнее украинской. Словоизменительная форма английских глаго-

лов, как правило, аналитическая, о чем свидетельствуют объем и статистические характеристики исследуемых текстов[1]. Английский глагол характеризуется сложной и разветвленной системой словоизменительных форм. Эта разветвленность вызвана не столько количеством грамматических категорий, сколько их специфическим соотношением и выражением. В украинском языке видовая пара глагола зачастую образуется с помощью словообразовательных средств: добавлением суффикса (*віддати* совершенный вид – *віддавати* несовершенный вид), или префикса (*робити* несовершенный вид – *зробити* совершенный вид), то в английском языке вид выражается группами форм форм Progressive (Continious) Tenses и Perfect Tenses, то есть такими видо-временными формами, которые объединяют категории вида (aspect) и времени (tense). Так что одному украинскому коду соответствует несколько английских. Например, украинскому глаголу *зробив* соответствуют такие английские формы, как *have done*, *has done*, *had done* и *did*.

Авторы многочисленных грамматик английского языка поразному определяют количество глагольных форм, их значение, особенности функционирования и названия. Это объясняет существование многочисленных синонимических терминов, употребляющихся для обозначения одной и той же формы глагола. Несмотря на значительное количество учебников и теоретических работ в области английского глагольного словоизменения до последнего времени не было предпринято детального исчисления инвентаря словоизменительных глагольных форм на основе определенного набора дифференциальных признаков. Конечно, такое исчисление не может быть совершенным, но оно необходимо для сопоставления родного и иностранного языка в процессе обуче-

[1] *Бобкова Т., Перебийніс В., Сорокін В.* Частотні словники паралельних текстів // Людина. Комп'ютер. Комунікація: Збірник наукових праць. Львів: Вид. Національного університету «Львівська політехніка», 2008. С. 158–160.

ния, и может быть положено в основу квантитативной типологии языков.

Такое исчисление английских глагольных форм было предпринято авторами пособия *Морфология английского глагола*[1]. Варианты глагольных форм были выявлены в результате исследования значительного массива английских текстов (более 12 млн. словоупотреблений) художественной прозы, драматургии, а также текстов научного и публицистического стилей. Словоизменительная парадигма английского глагола, исчисленная на основе заданных дифференциальных признаков, включает 526 инвариантных форм[2].

В данной работе для кодирования глаголов в украинских текстах используются буквенные коды, соответствующие окончаниям форм, а в английских – буквенно-цифровые коды. Так, например, список форм английского инфинитива насчитывает 44 инварианта (Г1–Г44):

– *to go, not to go, go, not go, to be going, not to be going, be going, not be going, to have gone, not to have gone, have gone, not have gone, to have been going, not to have been going, have been going, not have been going, to be taken, not to be taken, be taken, not be taken, to have been taken, not to have been taken, have been taken, not have been taken, to get taken, not to get taken, get taken, not get taken, to be getting taken, not to be getting taken, be getting taken, not be getting taken, to have got taken, not to have got taken, have got taken, not have got taken, to be gone, not to ge gone, be gone, not be gone, to have been gone, not to have been gone, have been gone, not have been gone.*

Список «инговых» форм глагола насчитывает 16 инвариантов (Г45–Г60):

– *going, not going, having gone, not having gone, being taken, not being taken, having been taken, not having been taken, getting taken, not getting taken, having got taken, not having got taken, being gone, not being gone, having been gone, not having been gone.*

---

[1] *Мілих Н.Г., Перебийніс В.С., Рукіна Е.П.* Морфологія англійського дієслова. Київ: Либідь, 1995. 120 с.

[2] *Перебийніс В.І.* Варіативність словозмінних форм англійського дієслова // Вісник Київського лінгвістичного університету. Серія Філологія. Т. 3. № 1. 2000. С. 13–19.

Естественно, что не все инварианты глагольных форм будут представлены в анализированных текстах официальных документов НАТО. Результаты анализа функциональных характеристик глаголов показывают, что в массиве текстов объемом в 100 000 словоупотреблений встретилось 7734 глагола в 49 различных формах. При этом наиболее частотными в текстах являются инварианты «инговой» формы, инфинитива, Participle II, а среди временных форм инварианты Non-Past Indefinite Active, Non-Past Indefinite Passive, Past Indefinite Passive, и это позволяет судить как о характере информации, так и об особенностях официально-делового стиля (см. табл. 2).

Таблица 2. 10 наиболее частотных форм английских глаголов

№	**Грамматический код**	**Абсолютная частота**
1	Г45 (Ing-form)	1384
2	Г1 (Infinitive)	1163
3	Г61 (Participle II)	1055
4	Г79 (Non-Past Indefinite Active)	991
5	Г87 (Past Indefinite Active)	908
6	Г63 (Non-Past Indefinite Active)	594
7	Г218 (Non-Past Indefinite Passive)	277
8	Г3 (Infinitive)	230
9	Г224 (Non-Past Indefinite Passive)	216
10	Г251 (Past Indefinite Passive)	187

Кодирование большинства из 526 форм английского глагола в текстах документов осуществлялось в полуавтоматическом режиме с помощью контекстного анализа. Однако, существенное влияние грамматической омонимии глагольных форм (Past Indefinite Active – Participle II и др.) и лексико-грамматической омонимии (Ing-forms – Noun – Adjective и др.) требует ручного постредактирования.