УДК 81'322
DOI: 10.17223/19986645/63/8

М.И. Солнышкина, Г.М. Гатиятуллина

ИСТОРИЯ РАЗВИТИЯ КОРПУСНОЙ ЛИНГВИСТИКИ (НА ПРИМЕРЕ АНГЛОЯЗЫЧНЫХ КОРПУСОВ)[1]

Предложена авторская периодизация формирования и развития англоязычных корпусов, базирующаяся на принципах Г. Кеннеди (1998), в соответствии с которой выделяем четыре основных периода: доэлектронный – до 1960-х гг. (архивы), I – с 1960-х по 1990-е гг. (корпусы), II – с 1990-х по 2000 г. (мегакорпусы), III началось в 2000-е гг. (гигакорпусы). Предложено описание периодов разработки программного обеспечения корпусов: программ-конкордансеров и автоматизированной обработки текстов.

Ключевые слова: *история лингвистики, корпусы текстов, корпусная лингвистика, поколения корпусов, классификация корпусов.*

Корпусная лингвистика как наука зародилась в конце 1970-х гг., однако методы, лежащие в ее основе, были известны с XIII в. Так, в зависимости от объема и принципов отбора текстов в развитии корпусов выделяют несколько периодов [1, 2][2]. Эпоха **доэлектронных корпусов** началась в XIII в. и завершилась к началу 1960-х гг. [1–3]. Понятие «корпус» в его лингвистическом значении появилось только к концу доэлектронной эпохи, поскольку им признавалось отдельное религиозное или литературное произведение или собрание сочинений одного автора, к которому вручную составлялся конкорданс[3], формируемый преимущественно для теологических, литературоведческих и лексикографических исследований.

А. Круден называет конкордансами словарь или указатель к Библии, в котором все слова, использованные в «боговдохновенном писании», расположены в алфавитном порядке, а также указано место, в котором употребляется данное слово, чтобы можно было легко найти стих с этим словом и сравнить несколько значений, в которых оно употребляется [4]. Все конкордансы доэлектронной эпохи отличались от современных и представляли собой некий указатель места употребления слова или словосочетания. Такого рода конкордансы также именуют каталогами или алфавит-

[1] Исследование выполнено при финансовой поддержке Российского научного фонда, грант № 18-18-00436.

[2] Здесь и далее перевод с английского выполнен авторами статьи.

[3] В настоящее время конкордансом называют алфавитный список всех употреблений конкретного слова в корпусе. Обязательным является также указание контекста слова, как правило, от двух до пяти, реже семь слов справа и слева от искомого слова [2. P. 42]. Определяя задачи конкорданса, П. Бейкер, А. Харди и Т. Макинери обращаются к учению Дж.Ферса (1957) о коллокациях как «реальных словах» в привычном окружении. По их мнению, конкорданс призван определить наиболее частотные коллокации [Ibid. P. 36].

ными указателями (indexes), а сам процесс – индексацией (indexing) [1]. Конкорданс состоит из «узловых слов» (node words) и контекста их употребления [5]. Объем контекста конкорданса обычно ограничивался восемью – десятью словами, поэтому объем конкорданса к Библии составил 2 370 000 словоупотреблений и по объему превышал объем Библии [6].

Первый конкорданс был составлен в XIII в. монахом Антонием Падуанским к латинской версии Библии V в. «Vulgate». Этот конкорданс назывался *Concordantiae Morales*. Примерно в то же время в Париже кардинал Гуго де Сен-Шер с помощью монахов прихода Святого Джеймса за два года составили алфавитный указатель слов к Библии Vulgate [7. P. 3]. Также известны попытки создания конкорданса в XV в. Джоном Марбеком [6. P. 2]. В 1737 г. А. Круден опубликовал первое издание «Полного конкорданса к Святому Писанию», в котором узловое слово располагалось на отдельной строке, а далее следовало указание названия книги и главы в Библии, где употреблено данное слово [4]. В тексте цитаты узловое слово сокращено до первой буквы. Все цитаты представлены на отдельной строке. Левосторонний и правосторонний контексты не превышают двух – пяти слов. «Полный конкорданс к Святому Писанию» А. Крудена также содержит полную цитату из Библии с данным словом (рис. 1).

DRY
Ezek.17.24[1]. Made d.tree flourish
a. devour every d.tree
30. 12 I will make the rivers dry
37.2 bones d. || 4. O ye d. bones

Рис. 1. A. Cruden A Complete Concordance to Holy Scriptures (1737)

А. Круден отдельно выделял словоформы: например, глагол «dry» и его форма прошедшего времени «dried» указывались отдельно. В качестве узловых в конкордансе А. Крудена выделялись как однословные единицы (существительное, глагол), так и многословные (устойчивые сочетания) (рис. 2).

DRY ground	**DRY verb**
Gen.b.13 face of the ground was d.	Job. 12.13. waters, they d.up
E.red. 14.16. on d.gr. in sea	flame d.up his branches
Josh.3.17. Riests stood firm on d.ground in Jordan	Isa. 42.15. will d.up herbs, pools
2. Kin.2.8. Elisha went over on d.g.	d. up thy rivers || 50.2. sea
Psal. 107.33. water-springs into d.g.	**DRIED**

Рис. 2. Конкордансы к устойчивым словосочетаниям А. Крудена

[1] Элементы метаразметки, включающие название книги, главу и стих, представлены на каждой строке.

В 1890 г. Дж. Стронг публикует «Исчерпывающий конкорданс к Библии» (*Strong's Exhaustive Concordance of the Bible*), в котором приведены этимологические сведения для 8 674 слов из Ветхого Завета, корни которых происходят из иврита, и 5 624 слова с корнями греческого происхождения в Новом Завете. К каждому слову дается информация о количестве (частотности) и месте употребления [8].

После публикации первого издания конкорданса А. Крудена в 1737 г. по такому же принципу стали составляться конкордансы к произведениям великих писателей. Так, важной работой для развития корпусной лингвистики стал «Конкорданс к произведениям У. Шекспира во всех редакциях» (1 787) А. Беккета, поскольку в нем помимо информации о месте употребления того или иного слова (пьесы, акта и действия) был представлен отрывок произведения, в котором употреблялось данное слово (рис. 3). Узловое слово содержало все словоформы. Например, вместе со словом «dream» указана и форма множественного числа «dreams». Объем иллюстрирующего отрывка по усмотрению автора мог содержать от одной до пяти строк [9. P. 167–183].

DREAM

My spirits as in a dream are all bound up *Tempest,* A.1, S.2

– we are such stuff
As dreams are made on, and our little life
Is rounded with sleep. *Tempest*, A.4, S.1

– Dreams are toys:
Yet, for this once, yea, superstitiously,
I will be squar'd by this *Winter's Tale*, A.3, S.3

I have heard (but not believ'd) the spirits of the dead
May walk again; if such things be, thy mother
Appeared to me last night; for ne'er was dream
So like a walking. *Winter's Tale* A.3, S.3

Рис. 3. A. Becket “A Concordance to Shakespeare suited to all the editions” (1787)

Известны также конкордансы к произведениям У. Шекспира, составленные М. Коуден-Кларк (1847) и С. Ойскотом (1790). Статья конкорданса С. Ойскота содержит следующие зоны: узловое слово, контекст, а также место употребления данного слова (пьеса, акт, сцена, страница, колонка и строчка). Узловое слово также содержит все словоформы (рис. 4) [10].

		A.	S.	P.	C.	L.
Disorder, that hath spoil'd us, befriend us now	*Henry v.*	5	5	533	1	45
– Fear frames disorder, and disorder where it should guard	*2 Henry vi*	5	2	601	2	29
– But his own disorders deferv'd much less advancement	*Lear.*	2	4	944	2	53
Disparage. I will disparage her no farther	*M.Ado About Noth*	3	2	133	2	59
– not the faith thou dost not know	*Mids. Night's Dream*	3	2	186	2	31

Рис. 4. S. Ayscough “Dramatic works with Explanatory notes” (1790)

Конкоданс, предлагаемый М. Коуден-Кларк, также создан по типу конкорданса А. Крудена, однако как и в конкордансе С. Ойскота, узловое слово представляет все словоформы (рис. 5) [11].

FEMALE – poor females mad *Mid.N's Dream,* iii.2.
the female ivy so enrings the …..... – iv.1
a female: or for thy ….. *Love's I...Lost*, 3.1. (letter)
the boy is fair, of female favour.. *As you like it,* iv.3.
of this female, which in the common … – v.1.
abandon the society of this female …. – v.1.

Рис. 5. M. Cowden-Clarke (1845) The Complete Concordance to Shakespeare: Being a Verbal Index to All the Passages in the Dramatic Works of the Poet[1]

Традиция составления конкордансов вручную к произведениям художественной литературы сохранялась вплоть до 1995 г. и была реализована в следующих работах: Конкоданс к «Секретному агенту» Дж. Конрада The Concordance to Conrad's The Secret Agent (Bender, 1979), Конкорданс к «Дейзи Миллер» Генри Джеймса A Concordance to Henry James's Daisy Miller (Bender, 1987), Конкорданс к полному собранию пьес и поэм Т.С. Эллиота A Concordance to the Complete Poems and Plays (Dowson, 1995) [12. P. 169].

На рубеже XIX и XX вв. было организовано несколько проектов по сбору эмпирического материала для лексикографических целей. На их основе были составлены «Словарь американского варианта английского языка» под редакцией Н. Вебстера (*Noah Webster's An American English Dictionary)* (1828) и «Оксфордский словарь английского языка» (*The Oxford English Dictionary, OED*) (1884). Для создания исследовательской базы «Оксфордского словаря» две тысячи читателей-добровольцев собрали около пяти миллионов цитат общим объемом примерно 50 миллионов словоупотреблений для того, чтобы проиллюстрировать значения и употребление 414 825 слов в словаре. На основе собранных текстов английской диалектной речи Дж. Райт составил «Словарь английских диалектов» *The English Dialect Dictionary* (1898–1905) [1].

Эмпирический материал О. Есперсена, который включал фрагменты из произведений О. Хаксли, Дж. Остин, У. Черчилля, Ч. Дарвина, Г. Филдинга, Э. Хемингуэя, Р. Киплинга, Дж. Локка, Г. Менкена, П. Шилли, Дж. Пристли, Х. Уолпола, В. Вульф, имел особое значение для преподавания практической грамматики английского языка, основанной на дескриптивных, не предписывающих принципах [13].

Поворотным моментом в истории развития конкордансов стала разработка методики использования ключевых слов (key words) в системе Keyword out of context (KWOC) ключевых слов вне контекста или Keyword in title ключевые слова в названии (1856) А. Крестадоро для систематизации каталогов в государственной библиотеке г. Манчестера. В 1958 г. Х.П. Лун доработал данную методику и ввел в компьютерную технологию под названием keywords in context (KWIC) «ключевые слова в контексте», в соответствии с которой ключевое слово располагалось в центре, а линии конкорданса можно было расположить слева или справа от ключевого сло-

[1] В иллюстрациях сохранена пунктуация первоисточника.

ва, включая необходимый контекст [14. P. 151]. Формат KWIC дает возможность составить список коллокаций слова в алфавитном порядке, а также список частотности каждого словоупотребления. П. Бейкер, А. Харди и Е. Макинери считают термин конкорданс синонимичным термину «ключевые слова в контексте» (key words in context, KWIC).

Электронный конкорданс *Index Tomisticus* общим объемом более 10,6 миллиона словоупотреблений, созданный монахом Р. Бусой к трудам Фомы Аквинского, стал первой работой, в которой были применены элементы машинной обработки текстов [15]. Конкорданс создавался в течение пяти лет: с 1962 по 1966 г. Для удобства работы с конкордансом и его краткости Р. Буса решил представить в нем к качестве ключевого слова только лемму, или заголовочное слово, со всеми ее словоформами. Для этого он осуществил лемматизацию текстов, которая проходила в два этапа: объединение всех словоформ с флексиями под одной леммой и прикрепление кода с соответствующей частью речи для каждой леммы и ее словоформы. Лемматизация проводилась на основе Латинского машинного словаря *Lexicon Electronicum Latinum*, который Р. Буса и десять священников составляли в течение двух лет. Электронный словарь представлял собой таблицу с леммами, на основе которой компьютер осуществлял лемматизацию текстов. Данный метод работы на основе электронного словаря или списка позже во многом определил принцип электронной обработки текстов. В 1973 г. был опубликован первый том *Index Tomisticus*, в 1970-е гг. было опубликовано более 40 томов *Index Tomisticus* с алфавитными указателями, таблицами с указанием частотности слов и др. [17].

Последним корпусом доэлектронной эпохи стал смешанный корпус устной и письменной речи Р. Кверка «Обзор практического употребления английского языка» *The Survey of English Usage, SEU,* Р. Кверка, разработанный в Лондонском университете [16]. Р. Кверк называл собранный исследовательский материал «исходным материалом» или «текстами». Я. Свартвик утверждает, что в 1960 г. термин «корпус» почти не употреблялся и на конференции ученые долго спорили о множественном числе слова «корпус» (*corpuses*, *corpora* или даже *corpi*) [17. P. 15]. Данный корпус оказался наиболее хорошо структурированным и систематическим корпусом доэлектронной эпохи. Устная и письменная формы речи были представлены текстами различных жанров, при этом источниками служили как сфера формального, так и неформального общения. Корпус состоял из 200 фрагментов текстов, каждый объемом 5000 словоупотреблений. Данный корпус ознаменовал собой переход из доэлектронной эпохи в электронную.

Таким образом, в доэлектронную эпоху были созданы все предпосылки перехода к корпусам электронной эпохи. Были разработаны первые конкордансы, которые понимались как синоним словарей и указателей. Первые конкордансы имели огромное значение для дальнейшего развития корпусной лингвистики, поскольку в составе статьи конкорданса обязательными считались указание искомого слова, места его употребления,

контекст использования зафиксированных единиц языка. Кроме того, была разработана система иллюстраций контекста в конкордансе «ключевое слово в контексте». В корпусах отсутствовали единый принцип сбора текстов, единые правила составления конкордансов. Их объем и источники также сильно различались: корпусом могли быть тексты священных книг (переводы Библии, произведения богословов), а также отдельные произведения художественной литературы. С современной точки зрения, такого рода тексты являются не корпусами, а архивами или собраниями отдельных текстов. Отсутствовал также и сам термин «корпус».

Электронная эпоха (с 1960-х гг. по настоящее время). С. Йоханссон утверждает, что, несмотря на уже опубликованные в 1960-х гг. работы Р. Бусы и появление первого электронного корпуса, ученые стали активно интересоваться корпусной лингвистикой лишь в 1970-е гг. [18. P. 39]. По его мнению, настоящая корпусная лингвистика зародилась именно в 1970-е гг. с созданием первых лабораторий и центров, в которых над общими проблемами лингвистики и способами обработки текстов стали работать лингвисты и программисты. Центры компьютерной лингвистики, нацеленные на сбор, хранение и обработку текстов корпуса, были открыты в Италии, США, Англии, Германии, Канаде, Франции, Швеции, Норвегии. К середине 1970-х гг. были созданы первые базы для хранения и распространения электронных корпусов: Оксфордский архив машиночитаемых текстов OTA (Oxford Text Archive) (1976) и Международный архив электронных текстов современного английского языка ICAME (International Computer Archive of Modern English) (1977).

Корпусы первого поколения. В начале 60-х гг. XX в. впервые появились электронные корпусы. Первым электронным корпусом признан так называемый «Брауновский корпус» (The Brown corpus), названный по имени университета США The Brown University, штат Род-Айленд. Его название официально включало термин «корпус». Группа ученых под руководством Г. Кучеры и Н. Френсиса работала над созданием корпуса в период с 1961 по 1964 г. [19]. В создании данного корпуса также приняли участие Р. Кверк, П. Оконнор и Дж. Керролл, а также Филипп Б. Гоув, редактор третьего издания словаря Уэбстера [1]. Брауновский корпус был корпусом письменной американской английской речи и содержал один миллион словоупотреблений из 500 текстов, изданных только в 1961 г. В корпусе представлены следующие пятнадцать жанров письменной речи американского варианта английского языка: газетные статьи, научные труды, объявления, книги о хобби, религиозная литература, биография, эссе, художественная литература (детективы, приключения и вестерны, научно-популярная литература, любовные романы, фельетоны). Тексты в «Брауновском корпусе» наносились на перфокарту, которая содержала информацию о месте расположения текста, его названии, а также о количестве строк в тексте.

В 1968 г. Ф. Бэгли впервые ввел термин «метаразметка» (metadata) для обозначения всех данных о текстах в корпусе [20. P. 195]. С середины 1960-х гг. появились первые программы-конкордансеры на основе KWIC:

«Атлас создания конкорданса и подсчетов корпуса» (COCOA, COunt and COncordance Generation Atlas) (1967) и «Коллокации» (CLOC, CoLOCation) (1978) [5. P. 2]. При их создании машинная обработка текстов сопровождалась ручной разметкой, т.е. «прикреплением» кода (или тега) к единице текста с информацией о ней [2. P. 154]. Об автоматической разметке текста стали говорить, когда в 1971 г. Б. Грин и Дж. Рабин написали программу автоматизированной разметки текстов TAGGIT, первая апробация которой представляла собой разметку Брауновского корпуса. TAGGIT осуществляла разметку при помощи 86 тегов, выделяющих в тексте знаменательные и служебные слова, знаки препинания и отдельные морфемы. Программа «не снимала омонимию», и 23% слов в корпусе оказались размеченными одновременно несколькими тегами [3].

В 1978 г. А. Эллегард осуществил синтаксическую разметку части Брауновского корпуса вручную: было выделено три уровня синтаксической разметки – простые предложения внутри сложных предложений (clause structures in sentences), составляющие клаузальных конструкций (constituent structures of clauses), часть речи каждого слова (word class of individual word). После нескольких лет проверок и исправлений работа по частеречной разметке Брауновского корпуса в 1979 г. была завершена. Б. Грин и Дж. Рубин опубликовали все данные о морфологическом анализаторе TAGGIT с тем, чтобы другие ученые могли ее доработать и усовершенствовать [18. P. 46]. Программы-конкордансеры первого поколения COCOA и CLOC создавались для каждого отдельного компьютера и отдельной задачи, т.е. всякий раз «приходилось заново изобретать колесо» [3. С. 35]. Именно эта проблема поставила необходимость создания конкордансеров следующего, второго поколения. Ученые считают, конец 1970-х гг. временем официального признания термина «корпусная лингвистика» [17. P. 12].

В 1980-х гг. продолжается доработка и усовершенствование программы TAGGIT, в 1983 г. в университете Ланкастера группа ученых под руководством грамматиста Дж. Лича и программиста Р. Гарсайда апробировала и внедрила обновленный вариант морфологического анализатора под названием CLAWS (the Constituent Likelihood Automatic Word-tagging System, *букв.* Автоматическая система разметки составляющих на основе сходства) [3].

«Брауновский корпус» стал стандартом для составления корпусов как по объему, так и по спектру представленных в нем стилей и жанров письменной речи. С публикацией «Брауновского корпуса» в середине 1970-х гг. стали появляться подобные корпусы сначала в Великобритании, потом и в других странах. Например, в 1976 г. был опубликован совместный корпус университетов Ланкастера, Осло и Бергена (The Lancaster-Oslo-Bergen corpus (LOB) (1961–1978) [21]. В начале 1990-х гг. стали создаваться аналогичные корпусы объемом не менее одного миллиона словоупотреблений, состоящие из 500 текстов пятнадцати различных жанров письменной речи. При этом в каждом тексте должно было быть представлено не менее 2000 словоупотреблений. Такими являлись, например, корпус Австралий-

ской английской речи, The Australian Corpus of English, ACE (1986), Веллингтонский корпус новозеландской английской речи, The Wellington Written English, WWE (1986), Корпус американской английской речи университетов Фрайбурга и Брауна, The Freiburg-Brown Corpus, FROWN (1991–1992), Корпус британской английской речи университетов Фрайбурга, Лондона, Осло и Бергена, The Freiburg London-Oslo / Bergen corpus, F-LOB, (1991–1992), Колхапурский корпус индийского варианта письменной английской речи, The Kolhapur corpus Indian English (1978) [1, 2]. Эти корпусы получили общее название «Семейство корпусов Браун» [22]. Различие данных корпусов состояло лишь в том, что корпусы содержали тексты одного из вариантов письменной английской речи: американского, британского, австралийского, новозеландского, индийского (таблица).

Содержание и объем корпусов Семейства Браун (The Brown Family)

Корпусы									
Код	Brown	Frown	LOB	F-LOB	Pre-LOB	Kolhapur	ACE	WWC	LCMC
Количество текстов отдельных жанров									
A	44	44	44	44	44	44	44	44	44
B	27	27	27	27	27	27	27	27	27
C	17	17	17	17	17	17	17	17	17
D	17	17	17	17	17	17	17	17	17
E	36	36	38	38	38	38	38	38	38
F	48	48	44	44	44	44	44	44	44
G	75	75	77	77	77	77	77	77	77
H	30	30	30	30	30	37	30	30	30
J	80	80	80	80	80	80	80	80	80
K	29	29	29	29	29	59	29	29	29
L	24	24	24	24	24	24	15	24	24
M	6	6	6	6	6	2	7	6	6
N	29	29	29	29	29	15	8	29	29
P	29	29	29	29	29	18	15	29	29
R	9	9	9	9	9	9	15	9	9
S	–	–	–	–	–	–	22	–	–
W	–	–	–	–	–	–	15	–	–

Код соответствует следующим жанрам: A – репортаж, B – редакторская колонка, C – обзорная статья, D – религиозный текст, E – хобби и полезные советы, F – массовая культура, G – биография и эссе, H – отчеты и документы, J – научная проза, K – художественная литература, L – детектив, M – научная фантастика, N – вестерн и приключенческий роман, P – роман и любовная проза, R – сатира и юмор, S – исторический роман, W – женский роман [24].

Корпусы устной речи. Корпусы устной речи появились значительно позже письменных, их впервые начали публиковать в 1990-е гг.

Корпус London-Lund (LLC) был разработан в период с 1975 по 1990 г. Я. Свартвиком, Р. Кверком, С. Гринбаумом и К. Хофландом на основе двух проектов: корпус SEU (1959–1989) (см. доэлектронную эпоху) и Корпус устной английской речи (SSE, 1975). Корпус LLC состоит из 100 затранскрибированных текстов устной монологической и диалогической ре-

чи по 5000 словоупотреблений каждый. Диалогическая речь зафиксирована в текстах разговорного стиля между друзьями и коллегами, в беседах и телефонных разговорах. Монологическая речь представлена спонтанной (комментарии и рассказы), а также подготовленной речью, не читаемой с листа [22. P. 408–409]. Помимо грамматической разметки тексты в корпусе размечены на просодическом уровне, т.е. содержат информацию о тоновых единицах, начале звука (onset), места ядра (слова, синтагмы), направлении ядерных тонов (восходящий, нисходящий, ровный, восходяще-нисходящий), высоте тона, паузе (короткая и длинная), ударении (обычное и выделенное). Тексты из проекта SEU имеют детальную просодическую разметку: указания на различный уровень громкости и темпа (быстрая, прерывистая, манерно-растянутая), модификации качественных характеристик голоса (высота, ритм, напряжение и т.д.), дополнительные характеристики (шепот, хрип) [23].

Источником корпуса устной английской речи (The Spoken English Corpus, SEC) общим объемом 53 000 словоупотреблений послужили тексты эфиров радиовещания, записанные в период с 1984 по 1987 г. и характеризующиеся жанровым многообразием: комментарии, новости, лекции для небольшой аудитории, лекции для большой аудитории, радиопередачи на религиозные темы, включая литургии, репортажи о светской жизни, телефонные разговоры с радиослушателями и др. [22].

Одним из первых размеченных (или аннотированных) корпусов устной английской речи является также машиночитаемый вариант корпуса SEC, MARSEC (Machine readable spoken English corpus) (1992–1994) – совместный проект Лаборатории компьютерных исследований английского языка (The Unit for Computer Research on the English Language, UCREL), университетов Ланкастера и Лидза, а также научного центра IBM в Винчестере. MARSEC в отличие от SEC был доработан фонологической разметкой: были размечены паузы, длина слова во временном отрезке, звуковое содержание, а также тоновое ударение [Ibid. P. 408–409].

С разработкой Брауновского корпуса появилось понятие «референтный корпус», которым стали характеризовать все перечисленные корпусы, поскольку исследователи проверяли свои предположения и теории (так называемые “intuitive data”) с помощью этих корпусов. Референтный корпус определяли как корпус, создаваемый для проведения частотного анализа текстов, а также для сравнения текстов большого спектра жанров или источников [2. P. 137]. Именно в этот период было доказано, что объем в миллион словоупотреблений нерепрезентативен для изучения низкочастотных слов, поскольку они могут отсутствовать в корпусе [1].

Кроме того, в этот период начинает формироваться ряд устных корпусов для распознавания и синтеза устной речи, разрабатываемых по заказу Агентства Министерства обороны США по передовым научно-исследовательским проектам (Defense Advanced Research Projects Agency, DARPA).

В 1984 г. компанией Texas Instruments была собрана база данных устной английской американской речи TI-DIGITS, которая содержала 77 зачитанных вслух цифровых последовательностей. В качестве дикторов выступили 111 мужчин, 114 женщин, 50 мальчиков и 51 девочка. Данный корпус был создан для автоматического распознавания цифровых последовательностей в устной речи [24, 25].

В 1990 г. для акустико-фонетических исследований, разработки и оценки автоматических систем распознавания речи был создан корпус устной слитной речи TIMIT Acoustic-Phonetic Continuous Speech Corpus. В разработке корпуса принимали участие Массачусетский технологический институт (MIT), Стэнфордский научно-исследовательский институт (SRI) и компания Texas Instruments. Корпус содержит тексты на восьми основных диалектах устной английской американской речи 630 дикторов (70% мужчин и 30% женщин), которые зачитывали вслух по десять предложений. Для тестирования систем распознавания речи корпус TIMIT включает три типа текстов: диалектные (1 260 предложений), фонетически насыщенные (compact), т.е. покрывающие весь фонематический состав и отдельные сочетания фонем, представляющие определенную трудность распознавания (3 150 предложений), и фонетически разнообразные тесты (diverse) с повтором каждой фонемы в различном контексте (1 890 предложений). Для третьей части корпуса TIMIT использовались тексты Брауновского корпуса, а также из диалогов театральных постановок того времени. Данный корпус включает орфографическую, подробную фонетическую транскрипцию, а также транскрипцию каждого отдельного слова с временной соотнесенностью. Каждый диктор зачитывал пять предложений из подкорпуса с фонетически насыщенными текстами, три предложения из подкорпуса с фонетически разнообразными текстами и по два предложения из подкорпуса диалектных текстов. Корпус TIMIT поделен на две части: 20–30% корпуса составляет оценочно-тестовая часть и 70–80% – тренировочная. Повтор предложений и дикторов как в тестовой, так и в тренировочной частях был минимизирован. Тестовая часть была также поделена на две части: основная оценочная подборка Core Test Set (192 текста, произнесенных 24 дикторами: 16 мужчинами и 8 женщинами) и подборка для заключительной оценки Complete Test Set (1 344 предложений или 168 дикторов (112 мужчин и 56 женщин) по 8 предложений). Тренировочная часть включает весь языковой материал, не вошедший в тестовую часть. Тренировочная часть содержит 4 620 предложений, зачитанных 462 дикторами (73% дикторов корпуса) [26].

Корпус Управление ресурсами (Resource management corpus) (1988) для тестирования систем распознавания слитной речи включает более 25 000 высказываний более 160 респондентов, говорящих на различных региональных диалектах американского варианта английского языка. Корпус включает два подкорпуса: RM1 и RM2. Подкорпус RM1 состоит из трех частей. Тренировочная часть с подбором говорящего (Speaker-dependent) включает речь 12 лиц, каждый из которых зачитывает вслух

600 «тренировочных» предложений на двух диалектах и десять предложений для «быстрой адаптации» (rapid adaptation sentences). 600 предложений подобраны таким образом, что они покрывают 97% лексического материала корпуса. Общий объем данного подкорпуса составляет 7 344 предложения. Подкорпус “Speaker independent” содержит 3 360 предложений, зачитанных вслух 80 лицами на двух диалектах, и по 40 предложений, взятых из основного корпуса RM. Тестовая часть RM содержит 1 600 предложений, зачитанных вслух двумя дикторами. Тестовая часть снабжена диагностическим и оценочным программным обеспечением. Подкорпус RM2 представляет собой дополненную версию подборки RM1 Speaker-dependent. Подкорпус содержит 10 508 предложений, зачитанных двумя мужчинами и двумя женщинами (по 2 652 предложения каждый). В данный подкорпус вошли 600 стандартных тренировочных предложений из подкорпуса RM1, 2 диалектных предложения, 10 предложений быстрой адаптации, 1800 дополнительных тренировочных предложений, 120 дополнительных предложений для промежуточных испытаний (development-test sentences), 120 оценочных предложений (evaluation test sentences) [27].

Корпус информационной службы (Air Travel Information Service Corpus, ATIS) (1990) был разработан для изучения спонтанной речи и синтеза речи. Корпус также делится на тренировочную и тестовую части. ATIS содержит тексты разговора людей с автоответчиком “I would like a ticket to...”, “I want to fly to Boston from New York next week”. На основе данного корпуса позже были созданы диалоговые системы, которые могли ответить на вопросы типа “Does Air Canada fly from Toronto to Dallas?” [28].

Данные корпусы, разработанные по военному заказу, показали возможность обучения машин автоматическому распознаванию речи и дали новые термины: токенизация (разделение слитной речи на отдельные слова), сегментация (разделение слитной речи на предложения и синтагмы), парсер (синтаксический анализатор), нормализация (приведение к фонетической норме слов, произнесенных с различными индивидуальными особенностями говорящего) на основе временной соотнесенности фразы (time alignment).

Характеризуя типы корпусов, Г. Кеннеди утверждает, что все корпусы текстов отдельных жанров различных исторических эпох, тексты речи представителей отдельных профессиональных сообществ, возрастных групп либо региональных диалектов являются примерами корпусов первого поколения, поскольку их цель заключается в изучении речи отдельной формы языка, а не языка в целом во всем его многообразии [1]. Таким образом, согласно его классификации мультимедийные корпусы, которые стали разрабатываться с середины 2000-х гг., вне зависимости от их технической составляющей считаются корпусами первого поколения, так как являются специальными корпусами и преимущественно репрезентируют отдельные жанры устной речи.

В 1960–1990-е гг. постепенно формируются требования к корпусам: обязательным стало привлечение текстов письменной речи общим объемом до миллиона словоупотреблений. Однако при этом привлекались пре-

имущественно тексты наиболее распространенных жанров письменной речи, объем каждого фрагмента текста составлял примерно 2 000 словоупотреблений. Характерным признаком этого времени является также тот факт, что корпусы содержали не полные тексты письменной речи, а фрагменты с фиксированным объемом слов.

1970-е гг. стали определяющими в развитии корпусной лингвистики: появились центры и лаборатории по разработкам электронных средств обработки текстов. Методика KWIC позволила систематизировать форму представления конкорданса, позднее появились первые программы-конкордансеры, такие как COCOA (COunt and COncordance Generation Atlas) и CLOC (CoLOCation). Электронная обработка корпусов поставила перед учеными проблему точности электронной обработки текстов, которая давала хорошие результаты только совместно с ручной разметкой.

К середине 1970-х гг. с развитием техники и, как следствие, доступности записи звучащей речи начали формироваться корпусы для более широкого спектра исследовательских целей. В 1980-х гг. разработан морфологический анализатор текстов CLAWS (the Constituent Likelihood Automatic Word-tagging System). К 1990-м гг. были опубликованы два корпуса устной речи, при этом спектр представленных жанров не был богат и сводился к следующим: беседы в неформальной обстановке, разговоры по телефону, радио, выступления на лекции. Объем корпусов также значительно уступал письменным. Создание корпусов устной речи поставило вопросы адекватной транскрипции и разметки. Корпусы устной речи также составлялись в военных целях для разработки систем распознавания и синтеза живой звучащей речи. В данный период закрепилось современное толкование значений таких терминов, как «корпус», «корпусная лингвистика», «разметка», «метаразметка», «конкордансер», «морфологический анализатор». При изучении устной речи появились термины «токенезация», «токены», «сегментация», «нормализация», «временная соотнесенность» (time alignment).

Корпусы второго поколения, мегакорпусы. В начале 1980-х гг. был разработан язык разметки текстов, или метаязык SGLM (Standard Generalized Markup Language, *букв.* Единый стандартный язык разметки), который представляет собой набор тегов, стандартизирующий разметку текстов [2. P. 149]. Данный формат оставался эталонным до 2007 г., когда ему на смену пришел упрощенный формат XML с более унифицированной и строгой формой разметки для предотвращения дублирования разметки, как это имело место в SGML [Ibid. P. 71; 3. С. 76–77].

В 1990-х гг. ученые Университета Ланкастера разработали ряд программ для следующих уровней разметок: разметка анафорических референтных связей (1992), просодическая разметка (1993), семантическая разметка (1993), (2004), художественно-стилистическая (1996 и 2004), прагматическая разметка (2003) и разметка ошибок говорящих (1999, 2003) [3. P. 78, 83; 29].

Изучение устной речи показало необходимость исследования описания прагматики высказывания, поскольку смысл высказывания в полной мере

может быть понят и представлен при условии фиксации речи (текста) в прагматическом контексте с указанием повышения или понижения голоса, жестикуляции, движения головы и др. [30, 31]. Прорывной явилась разработка программы ELAN (EUDICO Linguistic Annotator, 2006), позволяющая размечать тексты на уровне жестов, однако решение этой проблемы подняло вопрос этики [32, 33].

Т. Макинери и А. Харди утверждают, что 1990-е стали эпохой программ-конкордансеров второго поколения. Конкордансеры второго поколения работали на платформе IBM, поэтому могли использоваться на персональных компьютерах, поддерживающих операционную систему IBM. Конкордансеры второго поколения, такие как Micro-OCP (1988), Longman Mini-Concordancer (1989), Kaye concordancer (1990), также работали на основе методики KWIC и осуществляли следующие функции: составление алфавитного списка конкордансов с контекстным окружением слов справа и слева, составление списка слов корпуса, элементарные описательные статистические данные, такие как подсчет словоупотреблений, соотношение количества слов и словоупотреблений (type-token ratio). Совмещение функций отрицательно сказалось на мощности и производительности конкордансеров второго поколения. В качестве дополнительных причин такого положения указываются следующие: отсутствие единого формата, стандартов представления символов и разметок [3. P. 40].

В 1987 г. на конференции в Колледже Вассара в г. Пафкипси, штат Нью-Йорк, было основано сообщество Инициатива по кодированию текстов (Text Encoding Initiative, TEI), которое поставило проблему разработки единых стандартов составления, транскрипции и разметки корпусов [34]. Появление большого количества корпусов, созданных на основе различных типов текстов, привело к необходимости создания единого свода правил, в котором бы содержались все правила по сбору, транскрипции и аннотации текстов как устного, так и письменного дискурсов. Кроме того появились вопросы этики и передачи авторских прав. Так, если в 1970-е гг. использование скрытых микрофонов для записи речи, указание личных имен и адресов считалось приемлемым, то к 1990-м гг. использование подобных методов стало вызывать вопросы [1. P. 76–78; 3. P. 60–69]. Таким сводом правил стали выпущенные Инициативой TEI документы TEI (Text Encoding Initiative Principles[1]) [2. P. 157].

В 1991 г. некоммерческая компания «Уникод консорциум» разработала стандарт кодирования символов Уникод (Unicode) для ASCII (American Standard Code for Information Interchange), предназначенный для всех типов письменных языков мира, а также для кодирования непечатных символов

[1] В период с 1990 по 2018 г. Инициатива TEI опубликовала пять редакций данного документа с соответствующей нумерацией P1–P5. В редакциях P1–P3 (1990–1999) SGML был рекомендованным языком разметки. В редакции P4 (2002) составителям предоставлялся выбор между SGML и XML. В редакции P5 (2007) единственно рекомендованным языком разметки является XML. С ноября 2007 г. документ TEI стал обновляться дважды в год [35, 36].

(транскрипции, математических формул и др.). В настоящее время UTF-8 является наиболее распространенной спецификацией Unicode [2, 37, 38].

Попытки стандартизации составления корпусов были также предприняты Европейской консультационной группой по стандартам обработки языка – Expert Advisory Group on Language Engineering Standards (EAGLES) (1993), которая предложила свой стандарт сбора и разметки текстов в корпусе Corpus Encoding Standard (CES) (1998), имевший в своей основе сначала язык разметки SGML (1998), в настоящее время – язык разметки XML – XCES (2000) [2. P. 50].

Для решения вопроса о необходимости стандартизации разметок для всех языков в четвертой редакции TEI P4 (2002) составителям предоставлялся выбор между SGML и более строгим и унифицированным языком разметки XML. В пятой редакции TEI P5 (2007) единственно рекомендованным языком разметки является XML [3].

В 1993 г. Дж. Лич опубликовал максимы для составления метаразметки, т.е. метатекста, или текста о тексте, с указанием полной экстралингвистической информации. По мнению Дж. Лича, метаразметка должна соответствовать установленным требованиям и включать следующую информацию о критериях и источниках отбора текстов: 1) возможность доступа к исходному варианту материала; 2) отдельное хранение метатекста от основного текста; 3) перечисление всех использованных принципов разметки в отдельном документе; 4) доступность информации об авторах разметки и основные характеристики разметки (ручная / автоматизированная[1], программное обеспечение и т.д.); 5) понимание разметки как авторской интерпретации, ее относительности; 6) обязательное изложение в разметке максимально полной информации о тексте на основе общепринятых лингвистических принципов; 7) недопустимость признания ни одной разметки как эталонной [39].

Во вторую эпоху развития корпусной лингвистики с конца 1990-х гг. по 2000-е гг. были разработаны и внедрены конкордансеры третьего поколения (WordSmith 0.4 (1996), MonoConc (2000), AntConc (2005)). Данные программы характеризуются способностью обрабатывать большой объем текстов любой письменности, а также выполнять сложный статистический анализ. Кроме того, программы-конкордансеры начала XXI в. отличает их высокая функциональность: одна программа способна быстро составить список ключевых слов, конкордансы, выполнить частотный анализ и анализ коллокаций [3. P. 35].

Таким образом, с начала 1990-х гг. технические возможности позволили ученым компилировать и разрабатывать корпусы больших объемов. Цель данных корпусов состояла в охвате большого спектра форм языка, манифестируемых как в письменной, так и в устной речи, представляя таким образом все многообразие языка. Стало возможным автоматически

[1] До сих пор автоматически размеченные тексты проходят процедуру post-tagging – ручную выверку разметки.

размечать устные корпусы на просодическом, фонетическом, морфологическом, лексическом, синтаксическом и дискурсивном уровнях. Более того, появился целый ряд программ для автоматизированной обработки конкордансов. Г. Кеннеди [1], П. Бейкер, А. Харди, Т. Макинери [2. P. 35] называют корпусы, разработанные в период с конца 1980-х гг., корпусами второго поколения, или мегакорпусами, поскольку их объем приблизился к 100 миллионам словоупотреблений. К таким корпусам традиционно относят сеть корпусов Логман, The Longman Corpus Network (1991), Банк английского языка, The Bank of English, BoE (1993), Британский национальный корпус, The British National Corpus, BNC (1994), Американский национальный корпус, The American National Corpus, ANC (2008).

Одним из наиболее масштабных проектов, разработанных в конце 1980-х гг., стала Collins Birmingham University International Language Database (Международная база данных языка при Бирмингемском университете и компании Коллинз), или Корпус COBUILD. Корпус создавался группой ученых под руководством Дж. Синклера. В проекте использована так называемая Бирмингемская коллекция текстов (The Birmingham Collection of Texts), включающая 20 миллионов словоупотреблений текстов письменной и устной речи. Объем основного корпуса составил 7,3 миллиона словоупотреблений, а объем так называемого «резервного корпуса» – 13 миллионов словоупотреблений. Корпус на 75% состоит из текстов письменной речи, на 25% – устной речи. Корпус COBUILD содержит тексты, опубликованные в период с 1960-х гг. до 1982 г. Письменная речь преимущественно представлена прозаическими художественными текстами. В корпусе зафиксирована устная кодифицированная речь, в которой используется только общеупотребительная неспециальная лексика. 75% устной речи – речь мужчин старше 16 лет, 25% – речь женщин. 20% корпуса составляют тексты американского варианта английского языка. По мнению С. Йохансона, проект COBUILD был прорывным для своего времени по ряду причин: 1) объем корпуса превышал 20 миллионов словоупотреблений; 2) источниками служили полные тексты, а не короткие фрагменты; 3) он был наиболее репрезентативным и включал тексты устной и письменной речи различных жанров. COBUILD стал самым объемным корпусом своего времени и лег в основу Словаря английского языка издательства Коллинз, The Collins COBUILD Dictionary of English (1987) [40].

По завершении проекта COBUILD в 1991 г. Дж. Синклер стал писать о том, что объем корпусов должен быть максимально большим [41]. В 1990-х гг. ученый объявил о проекте по расширению корпуса COBUILD и созданию на его основе корпуса «Банк английского языка» (The Bank of English, BoE). Цель нового проекта состояла в создании «динамического» корпуса объемом несколько сот миллионов словоупотреблений, который непрерывно пополнялся бы новыми текстами английской устной и письменной речи. Такого рода корпус также именовался «мониторный корпус», поскольку ожидалось, что подобный корпус поможет отслеживать изменения, происходящие в языке [1. P. 47; 2. P. 65, 116]. Как и COBUILD, корпус

BoE состоит на 75% из текстов письменной речи и 25% – устной речи, при этом 70% являются текстами британского варианта английского языка, 20% – американского варианта и 10% – других национальных вариантов английского языка. К 1997 г. объем корпуса «Банк английского языка» составил 300 миллионов словоупотреблений. Корпус стал впервые по-настоящему динамичным: ежегодно в состав корпуса добавляли новые тексты. Г. Кеннеди пишет, что подобный тип корпуса поставил перед учеными новые задачи обработки текстов: ежемесячно из каждой газеты-источника поступало до 2,5 миллиона словоупотреблений [1. P. 47]. И хотя разработчики еще не были до конца уверены в целесообразности использования мониторных корпусов, тем не менее корпусы COBUILD и BoE сформировали новый стандарт в составлении корпусов – сбалансированность и репрезентативность. Сбалансированность как принцип формирования корпуса, по мнению П. Бейкера, может быть реализована только в больших референтных корпусах, в которых должна быть представлена как устная, так и письменная формы высокого, формального и низкого регистров [2. С. 18]. В настоящее время корпус называется Word Banks Online и содержит 259,4 миллиона словоупотреблений британского английского языка (41,4 миллиона словоупотреблений устной речи) и 189,4 миллиона словоупотреблений американского английского языка (33,1 миллиона словоупотреблений устной речи) [42].

Обосновывая необходимость репрезентативности корпуса, Д. Байбер пишет, что поскольку понятие «общий язык» есть абстрактная категория, а язык – это система различных жанров или стилей, референтный корпус должен включать все стили и жанры речи, а также территориальные говоры и диалекты. Говоря о социальной представленности языка, Д. Байбер утверждает, что в корпусах необходимо фиксировать территориальный и региональный диалекты, социолекты и профессиональные языки. Кроме того, Д. Байбер заявляет, что язык должен быть представлен в историческом ракурсе, т.е. включать тексты всех исторических эпох [43. P. 12, 246–250; 44]. Таким образом, репрезентативность рассматривается Д. Байбером как представленность в корпусе текстов широкого спектра жанров и функциональных стилей.

Исследователи признают, что полную репрезентативность достичь невозможно [43–46]. П. Бейкер пишет, что понятие репрезентативности тесно связано с понятием валидности или соответствием полученных данных реальному состоянию языка в данной сфере употребления [2. P. 140]. Д. Байбер считает, что репрезентативность корпуса связана со сбалансированностью, пропорциональной представленностью жанров и стилей языка всех слоев общества, которая соответствует существующей в реальности [43. P. 246–250].

Д. Байбер также выдвигает два вида репрезентативности текстов в корпусе: лингвистический (представленность всех грамматических и лексических форм в тексте, жанре и корпусе) и ситуационный фактор (представленность ситуаций) [39]. В соответствии с точкой зрения Дж. Синклера он

утверждает, что главным критерием отбора текстов для корпуса должны стать внешние или экстралингвистические факторы (коммуникативные ситуации), а не фактор представленности той или иной грамматической конструкции или лексемы в тексте (их он называет внутренними, или лингвистическими факторами) [40]. После публикации революционных работ Дж. Синклера и Д. Байбера репрезентативность стала обязательным условием для создания корпуса.

Еще одним мегакорпусом, формирование которого было начато в конце 1980-х гг. группой под руководством Д. Саммерс, является Сеть корпусов издательства Лонгман, The Longman Corpus Network. Данная сеть корпусов в настоящее время является коммерческой базой данных, состоящей из пяти основных корпусов: 1) Лонгманский корпус речи изучающих английский язык, The Longman Corpus of Learners' English (10 миллионов словоупотреблений); 2) Лонгманский корпус письменной американской английской речи, The Longman Written American Corpus (100 миллионов словоупотреблений); 3) Лонгманский корпус устной американской английской речи, The Longman Spoken American Corpus (5 миллионов словоупотреблений); 4) Совместный корпус письменного английского языка, издательства Логман и Ланскастерского университета The Longman / Lancaster English Language Corpus (30 миллионов словоупотреблений) и 5) Лонгманский корпус устной британской английской речи, The Spoken British Corpus (10 миллионов словоупотреблений) [49]. Г. Кеннеди пишет, что хотя каждая из частей Сети корпусов Лонгман (The Longman Corpus Network) была собрана для специальной цели, объединенный корпус стал мощным инструментом, в котором зафиксировано большое разнообразие текстов различных жанров речи, созданных носителями и неносителями английского языка. Данный тип корпусов использовался для создания словарей и учебников по коммуникативной грамматике английского языка. Позднее корпус устной английской речи вошел также в состав устной части Британского национального корпуса [48].

Британский национальный корпус (British National Corpus, BNC) составлялся с 1991 по 1995 г. в Оксфордском и Ланкастерском университетах. Целью проекта явилось создание сбалансированного и репрезентативного корпуса устной и письменной английской речи для академических, лексикографических и коммерческих целей. Корпус объемом 100 миллионов словоупотреблений включает 10% транскриптов устной речи и 90% текстов письменной речи второй половиной XX в. 75% текстов письменной речи – тексты информативного жанра: научные статьи и монографии, политические, деловые, культурные (музыка, театр) и светские новости, религиозные и философские тексты, статьи из журналов о спорте и домоводстве. 25% корпуса – произведения художественной литературы. Сбалансированная устная часть корпуса разделена на так называемые «контекстуальные» и «демографические» тексты. «Контекстуальная» часть (“the context-governed texts”) подкорпуса устной английской речи содержит тексты различных жанров и стилей устной речи: научно-информативный

стиль (лекции, новости, обсуждения в классе, научные консультации); деловой (торговые выставки, встреча с профсоюзами, медицинские, юридические и профессиональные консультации, интервью); публичный (проповеди, политическая речь, заседания советов, парламентские чтения, судебные слушания); досуг (спортивные комментарии, разговоры после ужина, собрания в клубах, звонки радиослушателей). В «Демографическом подкорпусе» (“Demographic texts”) представлены тексты устных записей региональных диалектов (южный, центральный и северный диалекты) английского языка. Для записей диалектов по возрастному, половому, социальному и территориальному признакам были отобраны 124 добровольца из южных, центральных и северных графств Британии. Тексты в корпусе были размечены с помощью программы автоматизированной разметки CLAWS 5 Tagset[1], разработанной в университете Ланкастера. Разметка текстов осуществлена при помощи языка разметки SGML по стандартам TEI [24, 48, 49].

Еще одним амбициозным проектом своего времени стал Корпус национальных вариантов английского языка – The International Corpus of English (ICE), разработанный в Университетском колледже Лондона под руководством С. Гринбаума в 1996 г. Цель проекта состояла в сборе текстов региональных вариантов английского языка. Подкорпусы включают тексты устной и письменной речи региональных вариантов английского языка Британии (ICE-GB), Восточной Африки, Индии, Новой Зеландии, Сингапура, Канады, Гонконга, Ямайки, Филиппин, США, Камеруна, Фиджи, Ирландии, Кении, Мальты, Малайзии, Пакистана, Сьерра Леоне, Шри Ланки, Тринидада и Тобаго. В качестве респондентов избирались лица старше 18 лет, получившие среднее школьное образование в англоязычной школе. Все подкорпусы содержат 60% текстов письменной речи и 40% транскриптов устной речи. Подкорпус диалогической речи включает следующие жанры устной речи: частные беседы (личные встречи и телефонные разговоры) и публичные (уроки, беседы на радио и телевидении, теле- и радиоинтервью, парламентские дебаты, деловые переговоры, очные ставки. Подкорпус монологической речи разделен на две части. Первая включает высказывания спонтанной речи (комментарии, речь на демонстрациях и в суде). Вторая часть содержит подготовленную читаемую с листа речь (теле- и радионовости, теле- и радиобеседы (ток-шоу). Морфологическая разметка выполнена на основе программы CLAWS7 (C7) Tagset[2], семантическая – с помощью программы UCREL Semantic Analysis System (USAS)[3].

---

[1] CLAWS 5 tagset – пятая версия программы автоматической частеречной разметки CLAWS, имеющая 57 тегов для лексического списка, списка суффиксов и списка фразеологизмов [50].

[2] CLAWS 7 tagset – седьмая версия программы автоматической частеречной разметки CLAWS, автоматически размечающая 137 тегов для лексического списка, списка суффиксов и списка фразеологизмов [51].

[3] UCREL Semantic Analysis System (USAS) – программа автоматической семантической разметки [52].

С 2006 г. в состав корпуса начинают включать аудиозаписи речи. В подкорпусе ICE-Gb (1996) выполнена частеречная и лексическая разметки. Подкорпусы (сингапурского, индийского, филиппинского, новозеландского) вариантов английского языка не размечены [24, 53].

Таким образом, корпусы второго поколения – это корпусы объемом не менее ста миллионов словоупотреблений, цель которых предполагает репрезентацию всего многообразия письменной и устной речи. Составители стремились представить как можно больше жанров и стилей устной и письменной речи различных слоев населения. Как правило, это корпусы, доступные онлайн, собранные и размеченные по требованиям TEI. Национальные корпусы стали мониторными и составлялись на основе принципов репрезентативности отбора текстов и по правилам, характерным для корпусов второго поколения.

В 1990-е гг. в качестве нового образца корпусов использовался Британский национальный корпус, а стандартом составления корпусов стал TEI, который рекомендовал язык разметки SGML. В период с 1987 по 2004 г. были разработаны правила сбора корпусов, составления метаразметки, а также программы автоматизированной разметки текстов.

Корпусы третьего поколения, или гигакорпусы. Начало 2010-х гг. ознаменовано появлением больших технических возможностей: разработаны конкордансеры четвертого поколения BNCweb (2009), CQPweb (2012), SketchEngine (2013), Wmatrix (2013), функционально схожие с конкордансерами третьего поколения. Конкордансеры четвертого поколения были разработаны с целью решения следующих проблем: ограниченная мощность персональных компьютеров, несовместимость операционных систем персональных компьютеров и правовые ограничения распространения корпусов. Для решения правовых вопросов и упрощения процедуры получения доступа корпусы перешли на онлайн-версии, что увеличило скорость обработки запросов и расширило количество пользователей. Непосредственный доступ стал доступен через веб-браузер, снабженный онлайн-поиском [3. P. 35; 54]. Четвертое поколение конкордансеров работает онлайн и позволяет осуществить контрастивный анализ небольшого частного корпуса с корпусами BNC или текстами из Интернета. М. Девис называет конкордансеры четвертого поколения гибридными корпусами, поскольку их интерфейс представляет собой некое общее поле для создания корпуса и проведения частотного анализа на морфемном, лексическом, синтаксическом и фразовом уровнях [55].

Тенденция к увеличению объема корпусов продолжилась и после 2000-х гг. А. Мауранен [56] С. Кублер и Х. Цинсмайстер [57. P. 10] характеризуют данное поколение девизом «чем больше корпус, тем лучше», а Л. Флауэрдью первой начинает именовать данную эпоху эпохой поколения гигакорпусов[1] [58]. В это время появился ряд новых корпусов (COCA, Google

[1] Гигакорпусы (от греч. гига – миллиард) – корпусы объемом несколько миллиардов словоупотреблений.

Books Ngram) (см. ниже), объем которых составил несколько миллиардов словоупотреблений. Большой объем корпусов позволил проводить частотные исследования более масштабно и изучать коллокации, состоящие из трех, четырех и более слов. Такого рода коллокации Д. Байбер [59] и К. Хайленд [60] называют «лексическими пучками» (lexical bundles), где одно слово может быть переменным. Например, в коллокациях из пяти слов *in the beginning of the, in the end of the, in the form of the* переменным является третье слово. Впоследствии эти коллокации получили название n-граммы, где биграммы – это коллокации, состоящие из двух слов, триграммы – коллокации, состоящие из трех слов, а n-граммы – это коллокации, состоящие из n слов [61]. В настоящее время фиксация подобных коллокаций стала возможной благодаря созданию больших гигакорпусов, часто рассматриваемых как сама сеть Интернет (Google Ngram, Google Books, COCA и др.). Кроме того, подобные корпусы предлагают возможность построения графиков частотности n-грамм для различных периодов времени с 1800 до 2010 г.

В 2008 г. был опубликован Корпус современной американской английской речи (The Corpus of Contemporary American English (COCA), общий объем которого на данный момент составляет примерно 400 миллионов словоупотреблений. Корпус содержит тексты устной и письменной речи. Письменная речь представлена такими жанрами, как художественная литература: короткие рассказы и пьесы из литературных журналов, детская литература, первые главы книг, опубликованные с 1990 г., а также сценарии к фильмам (113 миллионов словоупотреблений); тексты из популярных журналов взяты из Time, Cosmopolitan, Men's Health, Good Housekeeping, Fortune, Christian Century, Sports Illustrated (118 миллионов словоупотреблений); тексты жанра газетной статьи взяты из 10 газет со всей Америки: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle (114 миллионов слоупотреблений); тексты жанра научная статья взяты из 100 рецензируемых журналов по различным областям науки (112 миллионов словоупотреблений)[1] [62]. В корпусе COCA объем текстов устной речи составляет 118 миллионов словоупотреблений. Данный подкорпус содержит транскрипты, видео- и аудиозаписи широкого спектра радио- и телепередач: *All Things Considered* (радиостанция NPR), *Newshour* (телеканал PBS), *Good Morning America* (телеканал ABC), *Today Show* (телеканал NBC), *60 Minutes* (телеканал CBS), *Hannity and Colmes* (телеканал Fox). Корпус COCA является динамичным и ежегодно пополняется на 20 миллионов словоупотреблений. Частеречная разметка текстов осуществляется при помощи программы CLAWS. К корпусу прилагается программа-конкордансер WordAndPhrase [Ibidem].

В 2009 г. опубликован корпус оцифрованных текстов книг Google Books Ngram Viewer, в котором представлены тексты более одного миллиарда электронных книг, опубликованных в период с 1500 по 2008 г.

[1] На данный момент объем корпуса увеличен до 520 миллионов словоупотреблений.

В 2011 г. объем корпуса Google N-gram Corpus превысил 200 миллиардов словоупотреблений [63]. В 2014 г. выпущена вторая версия корпуса Google Books, в которой письменный американский дискурс на английском языке представлен 155 миллиардами словоупотреблений, а британская английская речь – 34 миллиардами словоупотреблений [64]. В корпусе Google Books кроме текстов на английском языке в значительно меньшем объеме представлены тексты на 6 языках: испанском, французском, русском, немецком, итальянском и иврите [Ibidem].

Корпус Global Web-based of English (GloWbE) (2013), как и корпус второго поколения ICE, ставит целью представить как можно больше региональных вариантов английского по всему миру. Этот корпус содержит тексты веб-страниц и веб-сайты 20 региональных вариантов английского языка. Объем корпуса GloWbe превышает объем корпуса ICE в 100 раз: его объем составляет 1,9 миллиарда словоупотреблений [65].

Объем корпуса News on the Web (NOW) (2016) на данный момент превышает 5,7 миллиарда словоупотреблений. Авторы пишут, что корпус содержит англоязычные тексты с «2012 г. по вчерашний день» [64]. Ежедневно объем корпуса пополняется текстами на 4–5 миллионов словопотреблений. Каждую ночь с 22:00 до 1:00 тексты загружаются в корпус: программа HTTrack считывает интернет-адреса (URL) из ресурса Google News и загружает в корпус 9–10 тысяч текстов, затем при помощи программы JusText повторяющиеся и шаблонные тексты удаляются. Разметка и лемматизация текстов осуществляется с помощью программы CLAWS 7, тексты добавляются к основному составу корпуса. На сайте можно, например, отследить самое популярное слово дня или года [66].

Появление мега- и гигакорпусов показало, что большие референтные корпусы непригодны для изучения речи отдельных профессий или жанров речи, поскольку большие корпусы, несмотря на их огромный размер, содержат преимущественно тексты наиболее распространенных жанров устной и письменной речи [47, 53, 56, 58, 67]. В конце 1990-х – начале 2000-х гг. было доказано, что принципы репрезентативности специальных корпусов соблюдаются при значительно меньших объемах, поскольку частотность как терминов, так и нейтральных слов остается стабильной и равномерной [46, 47]. В этой связи Л. Флауэрдью пишет, что репрезентативность необходимо рассматривать как более важный аспект, чем объем корпуса, и для корпусов письменной профессиональной речи объем может варьироваться от 20 000 до 250 000 словоупотреблений [58]. Характеризуя различия устных и письменных корпусов, А. Кестер утверждает, что устный корпус объемом миллион словоупотреблений считается большим корпусом, а корпус письменной речи объемом 5 миллионов словоупотреблений считается маленьким [67]. Л. Флауэрдью уточняет, что письменные корпусы объемом меньше 250 000 словоупотреблений принято считать небольшими [58]. Если же рассматривать специальный корпус, то количество текстов, как правило, будет варьировать от семи до одиннадцати. А. Кестер считает, что количество текстов одного жанра или типа дискурса должно состав-

лять минимум пять текстов. Если количество текстов меньше пяти, то корпус не является репрезентативным. А. Кестер также отмечает, что тексты, записанные в одной организации, не будут репрезентативными для того или иного жанра вообще, но будут представлять данный жанр в данной организации [67].

Таким образом, этот период характеризуется слиянием методов корпусной лингвистики со Всемирной сетью: созданы программы автоматической загрузки текстов из Интернета, как в случае с корпусами NOW и GloWbE, отношение ко Всемирной сети как к корпусу (частный случай, корпус Google Books), выход самих инструментов во Всемирную сеть (SketchEngine, BNCweb). Рассуждения об n-граммах на данном этапе получили более предметный характер. Кроме того, стало возможным отслеживать развитие употребления того или иного слова на больших массивах данных, например изменение формы и значения слова в течение времени в письменной (Google Books) либо в устной речи (COCA, NOW, GloWbE). Появление корпусов с большими массивами текстов не уменьшило актуальность вопроса необходимости и репрезентативности малых корпусов профессиональной речи.

Заключение. Авторская классификация корпусов, дополняющая классификацию электронных корпусов Г. Кеннеди, имеет в своей основе два параметра: объем корпуса и принципы отбора материала. Корпусы доэлектронной эпохи (до 1960 г.) в современном представлении являются собранием текстов или архивом, в них отсутствует единая система сбора текстов, их объем и источники сильно варьируются. Эти же черты свойственны и для конкордансов того времени. В доэлектронную эпоху, были заложены основы принципов составления корпусов и формирования конкордансов. К концу доэлектронной эпохи уже существовали термины «конкорданс», «ключевые слова в контексте», «лемматизация». Развитие информационных технологий электронной эпохи (с 1960 г.) во многом определило развитие корпусной лингвистики.

Характерной чертой электронных корпусов первого поколения является их нацеленность на изучение текстов отдельных жанров и/или речи социальных групп. Они содержат фрагменты текстов длиной не более 2 000 словоупотреблений. Объем корпусов первого поколения не превышал миллиона словоупотреблений. Брауновский корпус и корпус LOB являются первыми референтными корпусами, на основе которых были проведены первые корпусные исследования лексики и грамматики устной речи. Среди наиболее актуальных вопросов того времени следует указать проблему разработки программ автоматической разметки, программ-конкордансеров. Именно в 1980-е гг. закрепились такие термины, как «корпус», «корпусная лингвистика», «разметка», «метаразметка», «конкордансер», «морфологический анализатор». При изучении устной речи также появились термины «токенезация», «токены», «сегментация», «нормализация», «синтаксический анализатор» (парсер), «временной интервал» (time alignment).

Проблема единого стандарта разметки, а также стандартизации сбора и составления корпусов была решена созданием Инициативы по кодированию текстов. Корпусы второго поколения, создаваемые по правилам TEI, в конце 1990-х гг. имели морфологическую, синтаксическую, семантическую и другие виды разметки. Середина 2000-х гг. ознаменовалась тремя достижениями: разработка программ разметки видеозаписей на уровне жестов, внедрение удобных в использовании конкордансеров второго и третьего поколений высокой производительности. Так же как и корпусы первого поколения, мегакорпусы являются референтными корпусами, однако их составление впервые базировалось на принципах репрезентативности и сбалансированности с целью представления всего многообразия языка. Они включали широкий спектр жанров письменной и устной речи различных форм языка. Главным критерием отбора признается экстралингвистический аспект, т.е. коммуникативная ситуация. BNC и ANC имели объем около ста миллионов словоупотреблений.

Объем корпусов третьего поколения, или гигакорпусов, составляет несколько миллиардов словоупотреблений (COCA, Google Books). Это динамические корпусы, объем которых постоянно пополняется новыми текстами. Они могут содержать устные или письменные тексты на нескольких языках и охватывать несколько исторических периодов. Программное обеспечение представляет возможность проследить развитие того или иного слова в различные исторические периоды, а также изучать коллокации в контексте. Появление гигакорпусов послужило основанием для создания корпусов специализированной речи, объем которых, как и корпусов первого поколения, не превышает одного миллиона словоупотреблений. Конкордансеры четвертого поколения предлагают больший спектр функций с возможностью составлять свой корпус и сравнивать полученные результаты с результатами референтных корпусов.

Литература

1. *Kennedy G.* An Introduction to Corpus linguistics. Addison Wesley Longman limited, 1998. 315 p.

2. *Baker P., Hardie A., McEnery T.* Glossary of Corpus Linguistics. Edinburgh University Press, 2006. 192 p.

3. *McEnery T., Hardie A.* Corpus Linguistics: Method, theory and practice. Cambridge university press, 2012. 312 p.

4. *Cruden A.* A Complete Concordance to Holy Scriptures of Old and New Testament. 1737. 756 p.

5. *Stubbs J.* Notes on the History of Corpus Linguistics and Empirical Semantics // Collocations and Idioms / eds by M. Nenonen, S. Niemi. Joensuu: Joensuun Yliopisto, 2007. P. 317–329.

6. *Meyer Ch.F.* Pre-electronic corpora // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. P. 1–14.

7. *McCarthy M., O'Keeffe A.* Historical perspective: What are corpora and how have they evolved? // The Routledge handbook of corpus linguistics / ed. by A. O'Keeffe and M. McCarthy. 2010. P. 3–13.

8. *Strong J.* Strong's Exhaustive Concordance of the Bible. 1890. 1807 p.

9. *Becket A.* A concordance to Shakespear: suited to all the editions. 1787. 470 p.

10. *Dramatic* Works with Explanatory Notes. A New Ed., to which is Now Added a Copious Index to the Remarkable Passages and Words by Samuel Ayscough. 1790. Vol. 2. 558 p.

11. *Cowden Clarke M.V.* The Complete Concordance to Shakespeare: being a verbal index to all the passages in the dramatic works of the poet. 1847. 890 p.

12. *Tribble C.* What are concordances and how are they used // The Routledge handbook of corpus linguistics / ed. by A. O'Keeffe, M. McCarthy. 2010. P. 167–183.

13. *Jespersen O.* A modern English grammar: on historical principles. 1949. 542 p.

14. *Korycinski C., Newell A.F.* Text indexing: the problem of significance // Computers and writing. State of the Art / ed. by P.O. Holt [et al.]. 1992. P. 149–171.

15. *Busa R.* The Annals of Humanities Computing: The Index Tomisticus // Computers and the Humanities. 1980. Vol. 14. P. 83–90.

16. *Quirk R.* A grammar of contemporary English. 1972. 1120 p.

17. *Svartvik J.* Corpus linguistics 25+ years // Corpus Linguistics 25 Years On / ed. by R. Faccinetti. 2007. P. 11–27.

18. *Johansson S.* Some aspects of the development of corpus linguistics in the 1970-s and 1980-s // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. P. 33–53.

19. *The Brown* Corpus. URL: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus ling/content/corpora/list/private/brown/brown.html (дата обращения: 20.06.2018).

20. *Nguen T.H., Nunavath V., Prinz A.* Big Data Metadata Management in small Grids // Big Data and Internet of Things: A Roadmap for Smart Environments. 2014. P. 189–215.

21. *The LOB* Corpus. URL: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html (дата обращения: 20.06.2018).

22. *Xiao R.* Well-known and influential corpora // Corpus Linguistics: An International Handbook / ed. by A. Ludeling, M. Kyto. 2008. P. 383–457.

23. *The LLC*. URL: http://www.helsinki.fi/varieng/CoRD/corpora/LLC/index.html (дата обращения: 20.06.2018).

24. *Lamel L., Cole R.* Spoken Language Corpora // Survey of the State of the Art in Human Language Technology. 1997. P. 338–391.

25. *TIDIGITS*. URL: https://catalog.ldc.upenn.edu/LDC93S10 (дата обращения: 20.06.2018).

26. *DARPA TIMIT* Acoustic-Phonetic Continuous Speech Corpus. CD-ROM / J.S.Garofolo [et al.]. 1993. 94 p.

27. *Resource* Management Corpus. URL: https://catalog.ldc.upenn.edu/LDC93S3C (дата обращения: 20.06.2018).

28. *Tur G.* Spoken Language Understanding: Systems for Extracting Semantic Information from Speech / ed. by G. Tur, R. De Mori. 2011. 470 p.

29. *Corpus* annotation. URL: http://ucrel.lancs.ac.uk/annotation.html (дата обращения: 20.06.2018).

30. *McNeill D.* Hand and Mind: What Gestures Reveal About Thought. Chicago : University of Chicago Press, 1992.

31. *Rowley-Jolivet E.* Visual discourse in scientific conference papers A genre-based study // English for Specific Purposes. 2002. Vol. 21, iss. 1. P. 19–40.

32. *ELAN*. URL: https://tla.mpi.nl/tools/tla-tools/elan/release-notes (дата обращения: 20.06.2018).

33. *Crawford Camiciottol B., Fortanet-Gómez I.* Multimodal Analysis in Academic Settings: From Research to Teaching. Routledge, 2015. 251 p.

34. *Lou Burnard.* The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure // Journal of the Text Encoding Initiative. June 2013. Is. 5. Online since 21 June 2013, connection on 01 April 2018. URL: http://journals.openedition.org/jtei/811; DOI: 10.4000/jtei.811

35. *TEI* Guidelines. URL: http://www.tei-c.org/Guidelines (дата обращения: 20.06.2018).

36. *Introducing* the guidelines. URL: https://tei-c.org/support/learn/introducing-the-guidelines/. (дата обращения: 20.06.2018).

37. *Meyer Charles F.* English Corpus Linguistics: An Introduction. Cambridge University Press, 2004. 168 p.

38. *Kubler H., Zinsmeister S.* Corpus linguistics and linguistically annotated corpora. 2015. 320 p.

39. *Leech G.* Corpus annotation schemes // Literary and Linguistic Computing. 1993. № 8 (4). P. 275–281.

40. *The history* of COBUILD. URL: https://www.collinsdictionary.com/cobuild/ (дата обращения: 20.06.2018).

41. *Sinclair J.* Corpus, Concordance, Collocation. Oxford University Press, 1991.

42. *Word* Bank Online (Bank of English) режим доступа. URL: https://corpus.byu.edu/coca/old/help/compare boe.asp (дата обращения: 20.06.2018).

43. *Biber D., Conrad S., Reppen R.* Corpus linguistics: Investigating language structure and use. Cambridge University Press, 1998.

44. *Biber D.* Representativeness in corpus design // Literary and Linguistic computing. 1993. Vol. 8 (4). P. 243–257.

45. *Sinclair J.* Corpus and Text – Basic Principles // Developing Linguistic Corpora: a Guide to Good Practice / ed. by M. Wynne. 2005. P. 1–16.

46. *Tognini-Bonelli E.* Corpus linguistics at work. Amsterdam : John Benjamins, 2001.

47. *The Longman* Corpus Network. URL: http://www.longmandictionariesusa.com/longman/corpus (дата обращения: 20.06.2018).

48. *The British* National Corpus. URL: http://www.natcorp.ox.ac.uk (дата обращения: 20.06.2018).

49. *Leech G.* A brief users' guide to the grammatical tagging of the British National Corpus. URL: http://www.natcorp.ox.ac.uk/docs/gramtag.html (дата обращения: 20.06.2018).

50. *UCREL* CLAWS5 tagset. URL: http://ucrel.lancs.ac.uk/claws5tags.html (дата обращения: 20.06.2018).

51. *Introduction* by word-class to the claws7 tagging scheme. URL: http://www.natcorp.ox.ac.uk/docs/claws7.html# Toc334867959 (дата обращения: 20.06.2018).

52. *UCREL* Semantic Analysis System (USAS). URL: http://ucrel.lancs.ac.uk/usas/ (дата обращения: 20.06.2018).

53. *The International* Corpus of English. URL: http://www.ucl.ac.uk/english-usage/projects/ice.htm (дата обращения: 20.06.2018).

54. *Laurence A.* A critical look at software tools in corpus linguistics // Linguistic Research. 2013. № 30 (2). P. 141–161.

55. *Davies M.* Corpora: an introduction // The Cambridge handbook of Corpus Linguistics / ed. by D. Biber, R. Reppen. Cambridge University Press, 2015. P. 11–31.

56. *Mauranen A.* Speaking professionally in L2 // Variation and change in spoken and written discourse: Perspectives from Corpus Linguistics / ed. by J. Bamford, S. Cavalereri, G. Diani. 2013. P. 5–31.

57. *Kuebler S., Zinsmeister H.* Corpus Linguistics and Linguistically Annotated Corpora. London : Bloomsbury Publishing, 2015. 320 p.

58. *Flowerdew L.* The argument for using English specialized corpora to understand academic and professional language // Discourse in professions: perspectives from Corpus Linguistics / ed. by U. Connor, T. Upton. 2004. P. 11–33.

59. *Biber D.* University Language: A Corpus-based Study of Spoken and Written Registers. Amsterdam : John Benjamins, 2006. 261 p.

60. *Hyland K.* As it can be seen: Lexical bundles and disciplinary variation // English for Specific Purposes. 2008. Vol. 27. P. 4–21.

61. *Rayson P.* Computational tools and methods for corpus compilation and analysis // The Cambridge handbook of English corpus linguistics / ed. by D. Biber, R. Reppen. Cambridge university press, 2015. P. 32–49.

62. *The Corpus* of Contemporary American English.URL: https://corpus.byu.edu/coca/ (дата обращения: 20.06.2018).

63. *The Google* Books Corpora. URL: http://www.helsinki.fi/varieng/CoRD/corpora/GoogleBooks/ (дата обращения: 20.06.2018).

64. *Google* Books. URL: https://googlebooks.byu.edu/ (дата обращения: 20.06.2018).

65. *Google* Books Ngram Viewer. URL: https://books.google.com/ngrams/info (дата обращения: 20.06.2018).

66. *GloWbE*. URL: https://corpus.byu.edu/glowbe/ (дата обращения: 20.06.2018).

67. *Koester A.* Building small specialized corpora // The Routledge handbook of corpus linguistics. 2010. P. 66–80.

The History of Corpus Linguistics (On the Example of the English Language Corpora)

Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology. 2020. 63. 132–160. DOI: 10.17223/19986645/63/8

Marina I. Solnyshkina, Galiya M. Gatiyatullina, Kazan (Volga Region) Federal University (Kazan, Russian Federation). E-mail: mesoln@yandex.ru / ggaliya-m@mail.ru

Keywords: history of linguistics, text corpora, corpus linguistics, corpus generations, corpus classification.

The aim of the research is to review the milestones in the development of corpus linguistics and present an original classification of the main periods in formation and development of English-language corpora which includes the following four periods: (a) the “pre-electronic” period or the period of text archives which lasted for over several centuries and finished in the 1960s; (b) “the first generation” covers the period from the 1960s to the mid-1990s; (c) “the second generation” period of megacorpora corresponds to the last decade of the 20th century; (d) the third generation period of gigacorpora started in the mid-2000s. The pre-electronic corpora and concordances lacked a unified system of text collection, views on representative size, and sources of corpora. In this period, there were developed the basic principles of concordance collection, the KWIC system, lemmatization. The first generation corpora were mostly compiled for the study of certain genres and/or speech of certain groups of people. These corpora typically contained texts with a limited number of tokens, usually no more than 2,000. Among the most significant achievements of that period are The Brown Corpus and the London-Oslo-Bergen corpus, the first reference corpora, which were used for lexical and grammatical studies of “language in use”, the first concordance software (CLOC, COCOA), and the first automatic tagging software (TAGGIT). By the early 1990s, the following terms were introduced, specified and defined: “corpus linguistics”, “metatext”, “tagging”, “concordancer”, “POS-tagging”, “tokenization”, “segmentation”, “parsing”. The problem of a standardized corpus, its compilation, and tagging were addressed in the project of Text Encoding Initiative (1987). The annotation patterns of that period began requiring POS, syntactic, semantic, and other tagging. Concordances of the mid-2000s became faster and more user friendly. Representativeness in corpora was achieved by the presence of texts of spoken and written speech in various communicative events. Therefore, the referential corpora of the second generation (BNC, ANC) represent the national language with a wide range of both written and spoken genres in many territorial dialects. The size of the third generation corpora or gigacorpora (COCA, Google Books) was increased to several billion tokens, and they became dynamic. The installed software enables tracking the form, meaning, and use of words and n-grams in written and spoken texts in a number of languages covering several historical periods. Modern concordances are also tools for compilation of small subcorpora and contrasting the obtained results with those of the larger corpora (BNC, COCA).

References

1. Kennedy, G. (1998) *An Introduction to Corpus linguistics*. Addison Wesley Longman limited.

2. Baker, P., Hardie, A. & McEnery, T. (2006) *Glossary of Corpus Linguistics*. Edinburgh University Press.

3. McEnery, T. & Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

4. Cruden, A. (1737) *A Complete Concordance to the Holy Scriptures of Old and New Testament*. London.

5. Stubbs, J. (2007) Notes on the History of Corpus Linguistics and Empirical Semantics. In: Nenonen, M. & Niemi, S. (eds) *Collocations and Idioms*. Joensuu: Joensuun Yliopisto. pp. 317–329.

6. Meyer, Ch.F. (2008) Pre-electronic corpora. In: Ludeling, A. & Kyto, M. (eds) *Corpus Linguistics: An International Handbook*. Walter de Gruyter. pp. 1–14.

7. McCarthy, M. & O'Keeffe, A. (2010) Historical perspective: What are corpora and how have they evolved? In: O'Keeffe, A. & McCarthy, M. (eds) *The Routledge Handbook of Corpus Linguistics*. Routledge. pp. 3–13.

8. Strong, J. (1890) *Strong's Exhaustive Concordance of the Bible*. The Methodist Book Concern.

9. Becket, A. (1787) A Concordance to Shakespear: Suited to all the Editions. Printed for G.G.J. and J. Robinson.

10. Shakespear, W. (1790) *Dramatic Works with Explanatory Notes.* A New Ed., to which is Now Added a Copious Index to the Remarkable Passages and Words by Samuel Ayscough. London : Printed for John Stockdale.

11. Cowden Clarke, M.V. (1847) *The Complete Concordance to Shakespeare*: being a verbal index to all the passages in the dramatic works of the poet. Bickers and Son.

12. Tribble, C. (2010) What are concordances and how are they used. In: O'Keeffe, A. & McCarthy, M. (eds) *The Routledge Handbook of Corpus Linguistics*. Routledge. pp. 167–183.

13. Jespersen, O. (1949) *A Modern English Grammar: On Historical Principles*. Copenhagen: George Allen & Unwin Ltd.

14. Korycinski, C. & Newell, A.F. (1992) Text indexing: the problem of significance. In: Holt, P.O. et al. (eds) *Computers and Writing. State of the Art*. Springer. pp. 149–171.

15. Busa, R. (1980) The Annals of Humanities Computing: The Index Tomisticus. *Computers and the Humanities.* 14. pp. 83–90.

16. Quirk, R. (1972) *A Grammar of Contemporary English.* Addison-Wesley Longman Ltd.

17. Svartvik, J. (2007) Corpus linguistics 25+ years. In: Faccinetti, R. (ed.) *Corpus Linguistics 25 Years On*. Rodopi. pp. 11–27.

18. Johansson, S. (2008) Some aspects of the development of corpus linguistics in the 1970-s and 1980-s. In: Ludeling, A. & Kyto, M. (eds) *Corpus Linguistics: An International Handbook*. Walter de Gruyter. pp. 33–53.

19. *The Brown Corpus.* [Online] Available from: https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html. (Accessed: 20.06.2018).

20. Nguen, T.H., Nunavath, V. & Prinz, A. (2014) Big Data Metadata Management in Small Grids. In: Bessis, N. & Dobre, C. (eds) *Big Data and Internet of Things: A Roadmap for Smart Environments*. Springer. pp. 189–215.

21. *The LOB Corpus.* [Online] Available from: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html. (Accessed: 20.06.2018).

22. Xiao, R. (2008) Well-known and influential corpora. In: Ludeling, A. & Kyto, M. (eds) *Corpus Linguistics: An International Handbook*. Walter de Gruyter. pp. 383–457.

23. Varieng. (n.d.) *The LLC*. [Online] Available from: http://www.helsinki.fi/varieng/CoRD/corpora/LLC/index.html. (Accessed: 20.06.2018).

24. Lamel, L. & Cole, R. (1997) Spoken Language Corpora. In: Varile, G.B. et al. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press. pp. 338–391.

25. Leonard, G.R. & Doddington, G.R. (1993) *TIDIGITS*. [Online] Available from: https://catalog.ldc.upenn.edu/LDC93S10. (Accessed: 20.06.2018).

26. Garofolo, J.S. et al. (1993) *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. CD-ROM. Gaithersburg, MD.

27. *Resource Management Corpus*. [Online] Available from: https://catalog.ldc.upenn.edu/LDC93S3C. (Accessed: 20.06.2018).

28. Tur, G. (2011) *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley and Sons.

29. UCREL. (n.d.) *Corpus Annotation*. [Online] Available from: http://ucrel.lancs.ac.uk/annotation.html. (Accessed: 20.06.2018).

30. McNeill, D. (1992) *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press.

31. Rowley-Jolivet, E. (2002) Visual discourse in scientific conference papers A genre-based study. *English for Specific Purposes*. 21 (1). pp. 19–40.

32. *ELAN*. [Online] Available from: https://tla.mpi.nl/tools/tla-tools/elan/release-notes. (Accessed: 20.06.2018).

33. Crawford Camiciottol, B. & Fortanet-Gómez, I. (2015) *Multimodal Analysis in Academic Settings: From Research to Teaching*. Routledge.

34. Burnard, L**.** (2013) The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure. Journal of the Text Encoding Initiative. 5. [Online] Available from: http://journals.openedition.org/jtei/811. DOI: 10.4000/jtei.811

35. TEI. (n.d.) *TEI Guidelines*. [Online] Available from: http://www.tei-c.org/Guidelines. (Accessed: 20.06.2018).

36. TEI. (n.d.) *Introducing the Guidelines*. [Online] Available from: https://tei-c.org/support/learn/introducing-the-guidelines/. (Accessed: 20.06.2018).

37. Meyer, Ch.F. (2004) *English Corpus Linguistics: An Introduction*. Cambridge University Press.

38. Kubler, H. & Zinsmeister, S. (2015) *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic.

39. Leech, G. (1993) Corpus annotation schemes. *Literary and Linguistic Computing*. 8 (4). pp. 275–281.

40. Collins. (n.d.) *The History of COBUILD*. [Online] Available from: https://www.collinsdictionary.com/cobuild/. (Accessed: 20.06.2018).

41. Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford University Press.

42. *Word Bank Online (Bank of English)*. [Online] Available from: https://corpus.byu.edu/coca/old/help/compare_boe.asp. (Accessed: 20.06.2018).

43. Biber, D., Conrad, S. & Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

44. Biber, D. (1993) Representativeness in corpus design. *Literary and Linguistic Computing*. 8 (4). pp. 243–257.

45. Sinclair, J. (2005) Corpus and Text – Basic Principles. In: Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books. pp. 1–16.

46. Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

47. *The Longman Corpus Network*. [Online] Available from: http://www.longmandictionari-esusa.com/longman/corpus. (Accessed: 20.06.2018).

48. *The British National Corpus*. [Online] Available from: http://www.natcorp.ox.ac.uk. (Accessed: 20.06.2018).

49. Leech, G. (n.d.) *A Brief Users' Guide to the Grammatical Tagging of the British National Corpus*. [Online] Available from: http://www.natcorp.ox.ac.uk/docs/gramtag.html. (Accessed: 20.06.2018).

50. UCREL. (n.d.) *UCREL CLAWS5 tagset*. [Online] Available from: http://ucrel.lancs.ac.uk/claws5tags.html. (Accessed: 20.06.2018).

51. UCREL. (1996) *Introduction by word-class to the claws7 tagging scheme*. [Online] Available from: http://www.natcorp.ox.ac.uk/docs/claws7.html#_Toc334867959. (Accessed: 20.06.2018)

52. *UCREL Semantic Analysis System (USAS)*. [Online] Available from: http://ucrel.lancs.ac.uk/usas/. (Accessed: 20.06.20180.

53. *The International Corpus of English*. [Online] Available from: http://www.ucl.ac.uk/english-usage/projects/ice.htm. (Accessed: 20.06.2018).

54. Laurence, A. (2013) A critical look at software tools in corpus linguistics. Linguistic Research. 30 (2). pp. 141–161.

55. Davies, M. (2015) Corpora: an introduction. In: Biber, D. & Reppen, R. (eds) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press. pp. 11–31.

56. Mauranen, A. (2013) Speaking professionally in L2. In: Bamford, J., Cavalereri, S. & Diani, G. (eds) *Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistics*. Amsterdam: Benjamins. pp. 5–31.

57. Kuebler, S. & Zinsmeister, H. (2015) *Corpus Linguistics and Linguistically Annotated Corpora*. London: Bloomsbury Publishing.

58. Flowerdew, L. (2004) The argument for using English specialized corpora to understand academic and professional language. In: Connor, U. & Upton, T. (eds) *Discourse in the Professions: Perspectives From Corpus Linguistics*. Amsterdam: Benjamins. pp. 11–33.

59. Biber, D. (2006) *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

60. Hyland, K. (2008) As it can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*. 27. pp. 4–21.

61. Rayson, P. (2015) Computational tools and methods for corpus compilation and analysis. In: Biber, D. & Reppen, R. (eds) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press. pp. 32–49.

62. *The Corpus of Contemporary American English*. [Online] Available from: https://corpus.byu.edu/coca/. (Accessed: 20.06.2018).

63. *The Google Books Corpora*. [Online] Available from: http://www.helsinki.fi/varieng/CoRD/corpo-ra/GoogleBooks/. (Accessed: 20.06.2018).

64. *Google Books*. [Online] Available from: https://googlebooks.byu.edu/. (Accessed: 20.06.2018).

65. *Google Books Ngram Viewer*. [Online] Available from: https://books.google.com/ngrams/info. (Accessed: 20.06.2018).

66. *GloWbE*. [Online] Available from: https://corpus.byu.edu/glowbe/. (Accessed: 20.06.2018).

67. Koester, A. (2010) Building small specialized corpora. In: O'Keeffe, A. & McCarthy, M. (eds) *The Routledge Handbook of Corpus Linguistics*. Routledge. pp. 66–80.