ПЕРЕВОДЫ

УДК 81’38:004.738.1
ББК 81
DOI: https://doi.org/10.18500/2311-0740-2019-1-21-22-33

Марина Сантини
Брайтон, Великобритания

Веб-страницы, типы текстов и лингвистические характеристики: некоторые вопросы

С текстологической точки зрения веб – место, в котором сосредоточено огромное количество документов. В вебе практически всё может быть рассмотрено как «документ» или, что является более подходящим термином, как «веб-страница». То количество текстов, которое представлено в вебе, превышает все мыслимые пределы. Более того, веб дик и неконтролируем. Это становится ясным, если мы сравним «приручённый» источник мира бумажных текстов, такой как Британская Национальная Библиотека, и «неукрощённый» английский веб. В данном эмпирическом исследовании были изучены текстовые типологии случайной коллекции предварительно необработанных веб-страниц, не взятых из корпуса предварительно обработанных и отобранных документов. Было установлено, что текстуальность веб-страниц может отличаться от текстуальности линейных документов (не имеет значения, бумажных или электронных). Новая текстуальность усложняет автоматическое извлечение особенностей и применение средств АОТ. Также было установлено, что текстовые типологии, которые уже предложены исследователями, можно применить не для всех типов веб-страниц. На спорные вопросы, отмеченные в данной работе, нельзя ответить однозначно. В данный момент времени моим предложением остаётся то, что их следует учитывать при анализе результатов применения любого автоматического метода к веб-страницам.

Ключевые слова: *корпусная лингвистика, веб-страницы, текстовые типы, средства АОТ, текстуальность, традиционные документы.*

Marina Santini
Brighton, UK

Web Pages, Text Types, and Linguistic Features: Some Issues

From a textual point of view, the web is a huge reservoir of documents. On the web virtually everything can be seen as a ‘document’ or better a ‘web page’. The sheer amount of texts available is just overwhelming. Furthermore, the web is mainly wild and uncontrolled. This becomes clear if we compare a ‘tamed’ resource of the paper world, like the British National Library, and the ‘untamed’ English Web. In: this empirical study, I investigated text typologies in a random sample of raw web pages, and not in a corpus of preselected and pre-processed documents. I realized that the textuality of web pages might be dissimilar from the textuality of linear documents (whether paper or electronic documents). This new textuality makes automatic feature extraction and application of NLP tools more troublesome. I also realized that the text typologies already available in the literature might not cover all web page types. The issues pointed out in this study do not have an easy solution. For the time being, my suggestion is to keep them in mind when assessing results from any automatic approach to web pages.

Keywords: *corpus linguistics, web pages, text types, NLP tools, textuality, traditional documents.*

Оригинал: Santini Marina. Web pages, text types, and linguistic features: Some issues. In: ICAME JOURNAL: Computers in English Linguistics. No. 30. April 2006, pp. 67–86.

Сведения об авторе: *Сантини Марина, доктор филологии, научный сотрудник Института компьютерных исследований.*
Место работы: Брайтонский университет.
E-mail: *marinaromestockholm@gmail.com*
https://orcid.org/0000-0002-5737-8149

About the author: *Santini Marina, Doctor of Philology, Researcher of the Institute of Computer Science. Place of employment: University of Brighton, UK.*
E-mail: *marinaromestockholm@gmail.com*
https://orcid.org/0000-0002-5737-8149

Введение

С ростом Сети огромное количество документов, в особенности веб-страниц, стало доступно для изучения лингвистами (в частности, специалистами по корпусам). Веб-страницы можно считать новым видом документа,

в значительной степени более непредсказуемым и индивидуализированным, чем те, что выпущены на бумаге. В то время как большое число документов на бумаге все еще имеет линейную структуру, что отражается в традиционных электронных корпусах, таких как Британский национальный корпус (BNC), веб-страницы обладают визуальной организацией, позволяющей им совмещать несколько функций или несколько текстов с разными коммуникативными задачами в одном документе. Например, пространство веб-страницы может быть разделено на несколько частей, организованных списками ссылок – в основном изолированными именными структурами или глагольными элементами [1: 186–187] – и отрывками текста, распределенными по основному блоку документа, такими как кнопки навигации, меню, объявления, поля поиска, которые размещены на разных участках единой страницы. В дополнение к этому «эффект осуществления гиперсвязи» [2, 3], интерактивность и мультифункциональность [4] могут повлиять на внутреннюю смысловую структуру текста, которая также опирается на использование изображений и других графических элементов. Хотя использование разных видов шрифтов, цветов и размеров, так же как и использование приемов верстки, наподобие организации текста в колонки, линий, разделяющих части документов, рисунки и т. д., само по себе не ново (ср. [5] на предмет детального описания роли и языка, и инструментов верстки в формировании типов документов), новостная статья, разделенная на колонки и имеющая в своем составе заголовки, все еще не теряет своих специфических языковых и текстуальных особенностей при размещении в корпусе, таком как BNC. Этого нельзя сказать о многих веб-страницах, поскольку визуальная структура страницы, включающей в свой состав новостную статью, в большинстве случаев не может быть сведена на нет или просто проигнорирована без потери важной информации (ср. [6, 7]). Веб-страница – своего рода хранилище, из которого читатель выбирает нужную ему / ей информацию. Искусственное отделение того, что обычно называют основным блоком, от остальной страницы – решение необоснованное и в большей части случаев неоправданное. В итоге веб-страницы обычно более сложны и запутанны, чем традиционные бумажные или электронные документы. И хотя на веб-странице все элементы не обязательно неразрывно связаны друг с другом или находятся в линейной связи, они составляют единое неделимое целое.

Веб-страницы создают шум не только на текстуальном уровне. Также они содержат шум физический. На необработанной веб-странице, т. е. на странице, загруженной из Сети без предобработки, может быть найдено множество неправильностей, особенно если у страницы HTML-формат. Непредсказуемая пунктуация, опечатки, грамматические ошибки, необычные имена, экстралингвистические элементы, такие как HTML-теги и куски кода, могут значительно затруднить использование инструментов автоматической обработки текстов (АОТ) и автоматического извлечения лингвистической информации. Особенно трудно упорядочить HTML-разметку, во-первых, потому что HTML-синтаксис не стандартизирован, а во-вторых, потому что HTML пишут люди, у каждого из которых разный стиль программирования. Даже программное обеспечение, такое как Microsoft Frontpage, Micromedia Dreamweaver или Microsoft Word, частично использует непохожие друг на друга стандарты оформления кода. Утилиты для чистки или стандартизации кода, такие как бесплатная TidyHTML, оказываются практически бесполезны в случаях кардинально различающихся HTML-аннотаций.

При таком развитии событий возникает интересный вопрос. Могут ли разработанные на данный момент лингвистикой (в частности, корпусной) текстовые типологии быть применены к веб-страницам? Чтобы это понять, я провела простой эксперимент, целью которого было проверить, подходят ли существующие текстовые типологии к документам в Сети. Поскольку идентификация этих текстовых типологий основана на особенностях языка, в данной статье я хотела бы сосредоточиться на некоторых вопросах, которые возникли, когда я попыталась автоматически извлечь эти характеристики из случайной подборки веб-страниц. Никакого простого или уже готового ответа на эти вопросы нет. Цель статьи – озвучить их и обратить на них внимание сообщества корпусных лингвистов; подготовить почву для дальнейших дискуссий и исследований. И пока решение не найдено, моим советом будет проявлять осторожность при интерпретации результатов, полученных от любого вида автоматической обработки веб-страниц.

Статья организована следующим образом: п. 1 представляет собой краткий обзор того, что обычно называют «тип текста»; в п. 2 кратко описывается ряд исследований, в которых исследовалась взаимосвязь веб-страниц и типов текста; п. 3 содержит условия проведения эксперимента; в п. 4 анализируются шесть ключевых вопросов, возникающих при исследовании взаимоотношений типов текста и веб-страниц; некоторые выводы представлены в конце статьи.

1. Типы текста

Традиционно типы текста относят к риторическим категориям, таким как повествование, описание, представление и аргументация. Выделение типов текстов уходит корнями глубоко в культуру [8], но количество и конкретные наименования сопоставимых с ними риторических категорий различаются в зависимости от предпочтений и научного направления, в области которого работает лингвист. Например, Werlich [9] выделяет пять типов текста (повествование, описание, представление, аргументация и руководство), Beaugrande и Dressler [10] предлагают ввести семь типов текста (описательный, повествовательный, аргументирующий, научный, дидактический, литературный и поэтический), Adam [11] анализирует пять типов текста (декламация, описание, аргументация, экспликация и диалог).

В сфере корпусной лингвистики термин «типы текста» принят с момента публикации работы Biber на тему языковой изменчивости в речи и письме [12]. Этот труд на данный момент считается классикой статистического направления с опорой на корпуса и значительно повлиял на европейские стандарты хранения больших массивов реализаций языка, такие как методические рекомендации по текстовой типологии EAGLES [13]. В работе [12] проводится четкое разграничение между жанром, который позже становится регистром [14], и типом текста. С точки зрения автора, жанр испытывает влияние культурных и иных экстралингвистических факторов, в то время как информация о типе может быть извлечена непосредственно из текста, вне зависимости от жанра. Другими словами, если внешние критерии разделения жанров следуют разграничениям и классификациям, которые уже есть в культуре, то типология текстов по Biber базируется исключительно на собственно лингвистических критериях, интерпретируемых через связь с внешними функциями. В работе Biber [12: 102–103] предложены следующие параметры текстов: участие в продуцировании, продуцирование информации, вовлеченное повествование, эксплицитная ссылка, ситуационно-зависимая ссылка, явное выражение убеждения, абстрактное информирование и мгновенное уточнение информации.

Однако абсолютно строгая дифференциация между жанром / регистром и типом текста признана или принята не повсеместно. Некоторые ученые используют понятие «тип текста», чтобы разграничить жанры речи и жанры литературы (ср. [15]). Другие оперируют терминами «жанр текста» и «тип текста» как синонимами (ср. [16, 17]. Наконец, третьи просто используют термин без дальнейшего разъяснения, как это наименование следует понимать в контексте, в котором оно используется.

В данной работе я следую традициям риторики и корпусной лингвистики. Точнее, я пытаюсь выяснить, подходят ли и могут ли быть применены к веб-страницам те типы текста, что предложил Biber [12], и те, что происходят из традиционной риторической классификации, принятой Werlich [9].

2. История вопроса

Исследования, представленные ниже, все еще не завершены или же основаны на предварительных данных по изучению типов текста веб-страниц. И хотя в каждом из них используется разный подход, все они основаны на корпусах.

TyPWEB [18, 19], франкоязычный проект, чьей открытой целью является продолжение работы Biber, предлагает общие методологические и практические принципы профилирования веб-страниц, конечная задача которых – разработать как можно более подробную типологию, для того чтобы разграничить личные и коммерческие сайты. Анализируются как лингвистические характеристики веб-страниц, так и особенности представления (расположение элементов, изображений, гиперссылок и т. д.). Хотя проект все еще в разработке, некоторые результаты уже доступны. На данный момент получены некоторые интересные выводы. Так, структура коммерческого сайта сложнее структуры личного. На главной странице коммерческого сайта больше ссылок, чем на других страницах; у личных сайтов этого не наблюдается. Использование личных местоимений первого и второго лица варьируется между коммерческими и личными сайтами, и т. д. Разрывая с традицией исключительно индуктивного подхода, постулируемой Бибером, TyPWEB предлагает двойной подход к профилированию веб-сайтов, основанный как на дедукции с определением категорий *apriori*, так и на индукции, где категории извлекаются непосредственно из данных.

В работе [20] представлен компонентный (многоаспектный) анализ двух тематических категорий Google («Дом» и «Наука» с несколькими субкатегориями). Многоаспектный анализ [12, 21] опирается на индуктивный статистический подход, основанный на факторном и групповом анализе, при котором категории извлекаются из данных и интерпретируются в свете внешних функций. Итог исследования – четыре ключевых аспекта (личнововлеченное повествование, убедительно-аргументирующий дискурс, совет и абстрактно-технический дискурс). В отличие от TyPWEB, в Biber [20] включены только лингвистические

характеристики (лексические, морфологические и синтаксические классы, множество из которых было извлечено тэггером или парсером) без особенностей представления.

Santini [22] также никак не касается особенностей представления, но, в отличие от Biber и по аналогии с TyPWEB, пытается соединить дедуктивный и индуктивный подходы. Дедуктивно-индуктивная модель основана на Бейесовской инференции. Она дедуктивна, поскольку опирается на ограниченное количество исходных состояний (четыре) и широко признанных текстовых типов (описание / повествование, экспликационный/-информативный тип, аргументирующий / убеждающий тип, руководство). Индуктивна же она потому, что обработка инференций основана на корпусе. Инференции, в свою очередь, основаны на вычислении величины вероятности для гипотезы (тип текста), проверенной на одном или нескольких фактах (частоты некоторых характеристик). Градации типов текста исчисляются в вероятностных величинах. Например, веб-страница может быть руководством с вероятностью 0.3, повествованием с вероятностью 0.5, относиться к информативному типу с вероятностью 0.7 и аргументирующему с вероятностью 0.9. Проще говоря, это означает, что анализируемая веб-страница может считаться скорее всего аргументирующей, с высокой степенью вероятности информативной, возможно, повествовательной, но вряд ли руководством. По данным предварительной оценки выяснилось, что градации текстовых типов, которые дает модель, в большой степени совпадают с мнениями людей, изучавших тот же материал.

Более частный, но не менее интересный для нашего исследования взгляд, представленный в Roberts [23], сосредоточен на единственном типе текста – повествовании – для единственного жанра – личной главной страницы. Помимо всего прочего, автор предлагает оригинальное истолкование гиперссылок в контексте нарратологии.

Хотя во всех этих исследованиях представлены интересные находки, при этом каждое отличается особенным подходом и набором характеристик, ни в одном из них открыто не говорится о методологии извлечения лингвистической и нелингвистической информации в процессе обработки веб-страниц. В то же время не отмечается и то, что извлечение допускало двойные истолкования или вызывало проблемы. Как уже было подчеркнуто во введении, веб-страницы могут считаться новым типом документа, более сложным при обработке, нежели традиционные документы. Следовательно, путь, по которому извлекается информация и используются средства АОТ при обработке веб-страниц, может очень сильно повлиять на результаты, особенно если исследователь полагается на статистические методы.

3. Исследование

Как уже упоминалось ранее, ряд классификаций текстовых типов для традиционных типов документов, как бумажных или электронных (п. 1), так и веб-страниц (п. 2), уже был предложен в предыдущих работах по (корпусной) лингвистике. Целью данного исследования было при помощи простейших исследовательских установок верифицировать возможность приложения двух традиционных типологий текстов – Werlich [9] и Biber [12] – к веб-страницам. К этим двум типологиям добавлено еще одно базовое противопоставление – «субстантивный vs. глагольный». Проводя эксперимент, я поняла, что результаты могут быть незначительно искажены помехами, возникающими при автоматической обработке текста. Эксперимент проводился пошагово:

- составление списка лингвистических характеристик, представляющих три текстовых типологии: Werlich, Biber и «субстантивный vs. глагольный» (прил.);
- извлечение случайного набора английских веб-страниц из электронной коллекции SPIRIT [24];
- конвертация веб-страниц из HTML в ASCII;
- обработка ASCII-версий веб-страниц средствами АОТ (теггер, парсер, разбиение на n-граммы);
- кодирование типов текстов как массивов именованных характеристик;
- кодирование веб-страниц как массивов именованных характеристик;
- сравнение массива каждой веб-страницы с массивами типов текста и выведение типа текста для веб-страницы.

Единица анализа – конкретная веб-страница как целое. Ее отдельные характеристики извлекаются, подсчитываются и сравниваются с заранее заданным списком характеристик (заранее заданным текстовым типом). ASCII-версии веб-страниц были тегированы и подвергнуты частеречной разметке (с использованием Connexor, созданного Tapanainen и Järvinen в 1997 г. [25]); n-граммы были подсчитаны в единицах словарной вариативности (при помощи бесплатных программ). Относительные и абсолютные частоты для ASCII-версий были подсчитаны скриптами на языке программирования Perl. Относительные частоты были приведены к процентному виду. Каждая веб-страница кодировалась как массив. Совпадения между заранее заданными

текстовыми типами и характеристиками веб-страницы учитывались как пересечение массивов (рис. 1). Также было возможно установить порог совпадения. Например, при выборе порога в 30% извлекались только те характеристики, чьи нормализованные значения равнялись тридцати процентам. Повышение или понижение порога влияло на количество характеристик, включенных в массив, представляющий веб-страницу. При низком пороге (скажем, 20%) в массив веб-страницы было бы включено больше характеристик, и совпадение с несколькими заранее установленными типами текстов более вероятно. При этом при установлении порога выше (допустим, 80%) количество лингвистических характеристик, извлекаемых из текста, сокращается, как и возможность совпадения с предустановленными типами текстов.

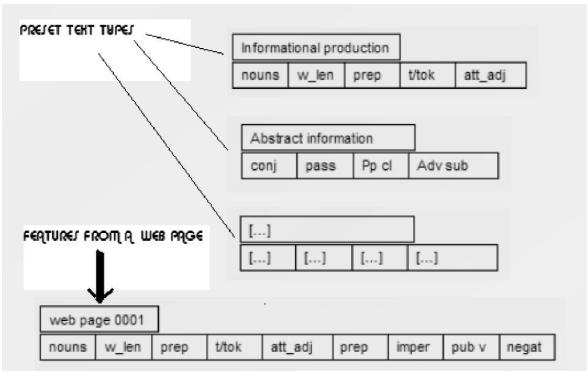


Рис. 1. Массивы (на изображении: заранее установленные типы текста, характеристики веб-страницы)
Fig. 1. Arrays (On the image: predefined text types, webpage specifications)

Эксперимент проводился с целью установить соответствие между веб-страницей как целым и одним или больше типами текста. Например, веб-страница могла классифицироваться как участие в продуцировании по Biber, аргументация по Werlich и глагольный тип текста.

4. Ключевые вопросы

Большинство веб-страниц, проанализированных с порогом в 50%, причислены к субстантивному типу. Однако анализ данных по частотности показал, что программы автоматического извлечения не столь точны, как ожидалось, и некоторые проблемы, связанные с обработкой текста, вероятно, негативно повлияли на итоговый результат. В подразделах я вкратце опишу следующие шесть ключевых вопросов, а именно (рис. 2):

1. Части текста, вставленные в код как изображения.
2. Заголовки.
3. Списки.
4. Имена собственные
5. Таблицы.
6. Смешанный тип текста.

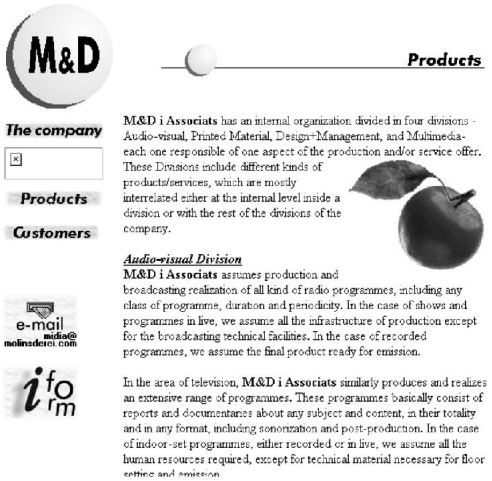


Рис. 2. Части текста, вставленные в код как изображения
Fig. 2. Text coded as images

4.1. Ключевой вопрос 1: Части текста, вставленные в код как изображения

Если сравнить ASCII-версию (рис. 3) и HTML-версию (см. рис. 2.), то можно увидеть, что ряд элементов исчез из ASCII-версии, например, все размещенное на левой стороне, как и основной заголовок. Это происходит, когда части текста кодируются как изображения, вставленные в HTML-код. Трудно найти для подобных потерь информации простое решение.

M&D i Associats has an internal organization divided in four divisions – Audio-visual, Printed Material, Design+Management, and Multimedia- each one responsible of one aspect of the production and/or service offer. These Divisions include different kinds of products/services, which are mostly interrelated either at the internal level inside a division or with the rest of the divisions of the company.
Audio-visual Division
M&D i Associats assumes production and broadcasting realization of all kind of radio programmes, including any class of programme, duration and periodicity. In the case of shows and programmes in live, we assume all the infrastructure of production except for the broadcasting technical facilities. In the case of recorded programmes, we assume the final product ready for emission.
[...]

Рис. 3. ASCII-версия веб-страницы, продемонстрированной на рис. 2
Fig. 3. ASCII version of the web page shown in Fig. 2

4.2. Ключевой вопрос 2: Заголовки

У второго абзаца на рис. 3 есть заголовок (выделен полужирным), который не идентифицируется парсером как независимая единица анализа (рис. 4). Неправильные выходные данные обусловлены тем, что парсер соединяет

udio-visual	audio-visual	attr:>2	@A>	%>N	A	ABS	
ivision	division	attr:>3	@A>	%>N	N	NOM	SG
&D	m&d		@OBJ	%NH	Heur	N	NOM	SG
	i	subj:>6	@SUBJ	%NH	PRON	PERS	NOM	SG1
ssociats	associats	mod:>4	@APP	%NH	Heur	N	NOM	SG
ssumes	assume	main:>0	@+FMAIN	%VA	V	PRES	SG3	
roduction	production		@OBJ	%NH	N	NOM	SG	
nd	and	cc:>7	@CC	%CC	CC			
roadcasting	broadcasting	cc:>7	@I-OBJ	%NH	N	NOM	SG	
ealization	realization		@OBJ	%NH	N	NOM	SG	
f	of	mod:>10	@<NOM-(	%N<	PREP			
ll	all	ad:>13	@AD-A>	%E>	ADV			
ind	kind	pcomp:>11	@<P	%NH	A	ABS		
f	of	mod:>13	@<NOM-(	%N<	PREP			
adio	radio	attr:>16	@A>	%>N	N	NOM	SG	
rogrammes	programme	pcomp:>14	@<P	%NH	N	NOM	PL	
	.							
ncluding	include	man:>6	@-FMAIN	%VA	ING			
ny	any	det:>20	@DN>	%>N	DET			
lass	class	obj:>18	@OBJ	%NH	N	NOM		
f	of	mod:>20	@<NOM-(	%N<	PREP			
rogramme	programme	pcomp:>21	@<P	%NH	N	NOM	SG	
	,							
uration	duration	cc:>22	@<P	%NH	N	NOM	SG	
nd	and	cc:>24	@CC	%CC	CC			
eriodicity	periodicity	cc:>24	@<P	%NH	N	NOM	SG	
	.							
p>	<p>							

Рис. 4. Заголовок не анализируется как отдельная синтаксическая единица
Fig. 4. The heading is not analysed as a separate syntactic unit

заголовок со следующим предложением. Заголовки редко оформляются пунктуационным знаком в конце или состоят из грамматически оформленного предложения (за исключением вопросов, например, «Как вы создаёте свой интранет?»). В результате заголовки представляют большую проблему для разделителя или парсера. Решением проблемы может стать предварительная обработка. Например, HTML-тег <h#> (парсеры могут распознать HTML-тег для заголовков, но обычно просто не интерпретируют их) может быть применён, чтобы создать искусственную границу между предложениями. Однако, как показано на рис. 2, заголовок может быть тегирован совершенно иначе, как <p><u>Audio-visual Division</u>
. Типы текста на рис. 2 – описание по Werlich, продуцирование информации по Biber, субстантивный тип текста.

4.3. Ключевой вопрос 3: Списки

Рис. 5 демонстрирует очень распространённый способ организации текста – списки. Здесь могут возникнуть проблемы как с парсингом, так и со стилеметрией, в частности подсчётом средней длины предложения. У списков есть целый ряд особенностей (ср. [26]). Например, вводное предложение списка обычно семантически неполно; в нём либо не хватает точки в конце, либо в конце поставлено двоеточие. Помимо этого, позиции списка могут представлять отдельные слова или целые предложения, оформленные или неоформленные пунктуационно. На рис. 5 сверху дан короткий список, каждая позиция которого содержит несколько слов без пунктуации, а также более длинный список, вводное предложение которого заканчивается двоеточием. Позиции последнего – длинные предложения, заканчивающиеся точкой с запятой или точкой. Парсер подобные конструкции сбивают с толку. В дополнение к этому недостоверными оказываются и данные о средней длине предложения, которые обычно основаны на частоте точек, вопросительных и восклицательных знаков, а также ряда других символов. Решением может стать использование тега <li>, для того чтобы установить искусственные границы предложений. Но что тогда делать с семантически неполным вводным предложением или же длинным списком с двоеточием в конце?

Типы текста, полученные на выходе для рис. 5, – руководство по Werlich и субстантивный тип текста.

Другой пример нехудожественной страницы продемонстрирован на рис. 6. Это список из позиций, распределённых по странице без маркирования, нумерации или знаков пунктуации. Визуально эта неполнота не ощущается, поскольку строки текста воспринимаются как отдельные сущности. Внутренняя структура HTML – таблица.

Тип текста, получаемый на выходе для рис. 6, – субстантивный.

Рис. 5. Списки
Fig. 5. Lists

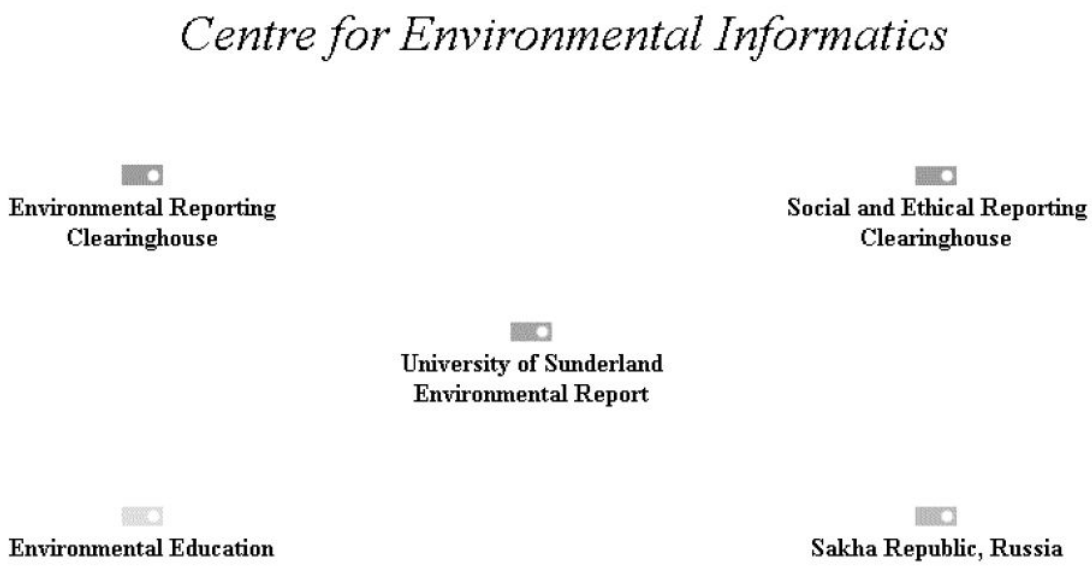


Рис. 6. Распределённый список
Fig. 6. Scattered list

4.4. Ключевой вопрос 4: Имена собственные

Тип веб-страниц, показанный на рис. 7, часто встречается в сети. На странице список имён и некоторые личные детали. Возможно, утилита для распознавания именованных сущностей и аббревиатур окажется полезнее парсера или теггера в этом случае.

Тип текста, возвращённый для рис. 7, – субстантивный.

4.5. Ключевой вопрос 5: Таблицы

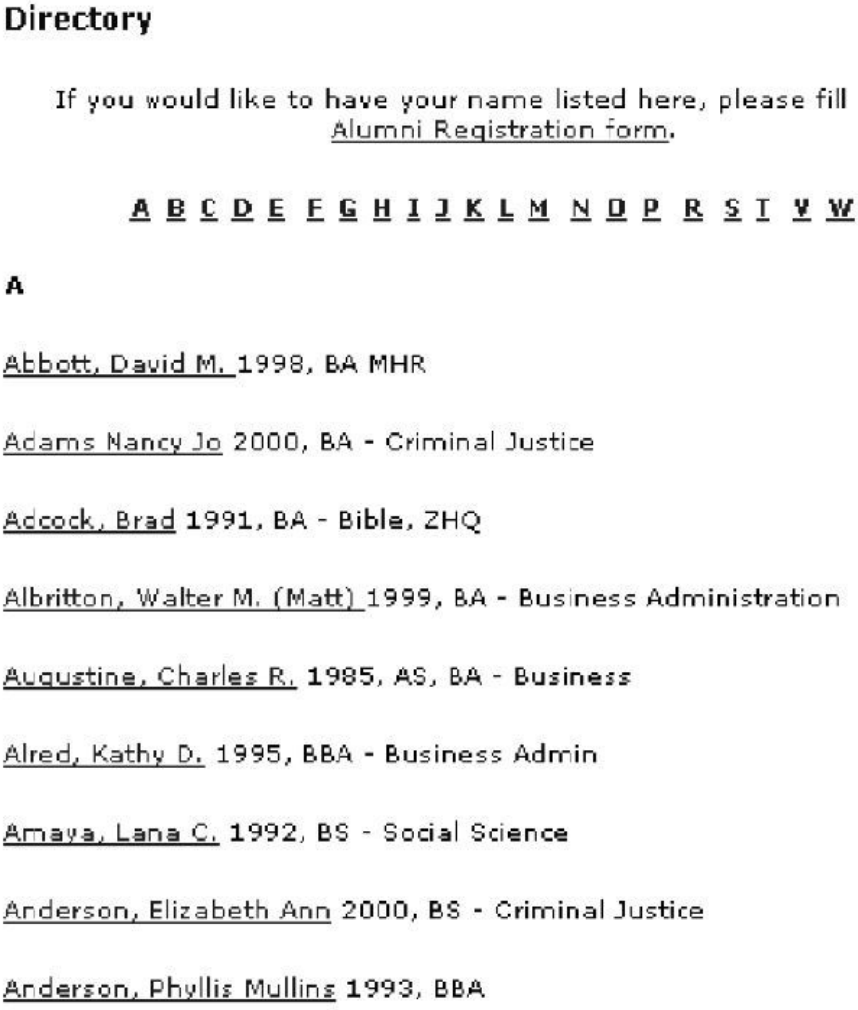
Таблицы, похожие на продемонстрированные на рис. 8, также часто встречаются в Сети. Считалось, что их трудно анализировать с лингвистической точки зрения, но, к счастью, это представление меняется (ср. [27–29]).

Тип текста, восстановленный для рис. 8, – эксплицитная ссылка.

4.6. Ключевой вопрос 6: Смешанные тексты

Ещё один часто встречающийся тип веб-страниц продемонстрирован на рис. 9. На этой страницы есть основная статья и другие строки текста, обрамляющие основное тело. Семантически эти строки не относятся к основному телу, лишь предоставляют дополнительную информацию читателю. Предполагая, что каждый тип текста выполняет некую коммуникативную функцию, сколько мы сможем выделить типов текста на этой странице? Как минимум 3: комментарий (основная статья), информативный список (заголовки на правой стороне) и заголовок (позиции списка на левой стороне)

Единственный тип текста, восстановленный для рис. 9, – участие в продуцировании по Biber.



Directory

If you would like to have your name listed here, please fill out our Alumni Registration form.

A B C D E F G H I J K L M N O P R S T V W

A

Abbott, David M. 1998, BA MHR

Adams Nancy Jo 2000, BA - Criminal Justice

Adcock, Brad 1991, BA - Bible, ZHQ

Albritton, Walter M. (Matt) 1999, BA - Business Administration

Augustine, Charles R. 1985, AS, BA - Business

Alred, Kathy D. 1995, BBA - Business Admin

Amaya, Lana C. 1992, BS - Social Science

Anderson, Elizabeth Ann 2000, BS - Criminal Justice

Anderson, Phyllis Mullins 1993, BBA

Рис. 7. Имена собственные
Fig. 7. Proper nouns

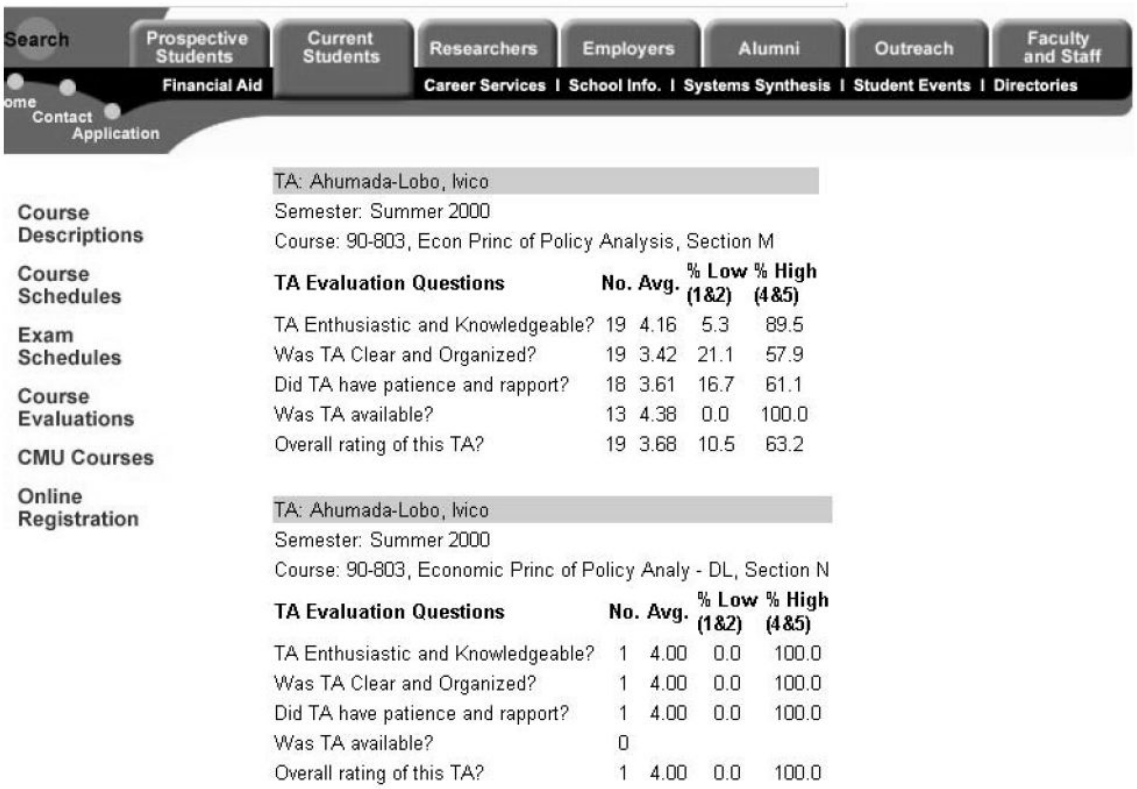


Рис. 8. Таблица
Fig. 8. Tabular structure

Рис. 9. Смешанный текст
Fig. 9. Mixed text

Выводы

Список ключевых вопросов может быть дополнен, однако я остановлюсь здесь и сделаю ряд выводов. Даже несмотря на то что большинство типов текста на выходе определены в общем верно, автоматическое извлечение особенностей веб-страниц вызывает проблемы, например:

- некоторые элементы текста веб-страниц могут быть потеряны, когда текст кодируется в изображениях (п. 4.1);
- средства АОТ могут быть ненадёжны, когда используются на ASCII-версиях веб-страниц без предварительной обработки, и полученные результаты могут быть ошибочны (п. 4.2 и 4.3). Отсутствие знаков препинания может представлять серьёзное затруднение для некоторых утилит, таких как парсеры, которые опираются на границы предложений. По-видимому, использование пунктуации в конце предложений на веб-страницах может отличаться от такового в линейных документах. По сути, поскольку веб-страницы – документы визуальные, элементы текста могут иметь различный шрифт и цвет, быть размещены беспорядочно по всей странице. Форматирование и распределение – визуальные способы, нивелирующие отсутствие границ между предложениями (например, см. рис. 6);
- по всей видимости, грамматические и лексические особенности по отдельности не могут служить полноценной опорой для типологии текста на веб-страницах. Другие особенности, такие как имена собственные и таблицы, нужно подвергнуть процедуре идентификации (п. 4.4 и 4.5). В то время как имена собственные могут быть распознаны утилитами по идентификации именованных сущностей, с таблицами всё ещё возникают затруднения.

Что же касается типов текста, я могу заключить следующее:

- для некоторых веб-страниц невозможно с точностью определить тип текста (например, см. рис. 6–8). В этом случае индуктивные методы, такие как компонентный (многоаспектный) анализ, могут помочь в выделении и интерпретации особенностей;
- точно следует обратить внимание на тексты смешанного типа (п. 4.6). Текст может представлять собой смешение различных средств выразительности и коммуникативных актов, слабо коррелировать с идеальным / идеализированным типом [10: 181 ff.]. Подобное особенно верно в случае веб-страниц, являющихся визуальными объектами, по большей части нелинейной организации, включающими в себя различные коммуникативные задачи. В этом случае метод, основанный на более точном анализе текстуальности веб-страницы (ср. [22]), окажется более эффективным.

Резюмирую. Я поставила сложную задачу – анализ текстовых типологий на случайной подборке необработанных веб-страниц, взятых не из корпуса заранее выбранных и обработанных документов. Я выяснила, что текстуальность веб-страниц может отличаться от текстуальности линейных документов (бумажных или электронных). Эта новая текстуальность значительно затрудняет автоматическое извлечение особенностей и обработку текста. Также я выяснила, что уже разработанные типологии могут не включить в себя все типы веб-страниц.

У ключевых вопросов, представленных в статье, нет простого решения. На данный момент я предлагаю помнить о них при обработке результатов от любой автоматической обработки веб-страниц. Поскольку веб-страницы как большой источник текстов не могут быть отвергнуты (корпусной) лингвистикой, необходимы дальнейшее обсуждение и изучение.

Примечания

1. Случайная коллекция из 1000 неклассифицированных веб-страниц из коллекции SPIRIT может быть загружена внизу следующей страницы: http://www.itri.brighton.ac.uk/~Marina.Santini

2. В особенности не были включены удаление *that,* использование *do* как частицы и нефразовое соположение.

3. В особенности не были включены конструкции с синтаксическим выносом.

4. В особенности не были включены разделённые связки.

Приложение

Типы и особенности текста по Werlich

Типы текста по Werlich основаны на квантитативном анализе бумажных документов [9]. Werlich выделяет пять типов текста: описание, повествование, экспозицию, аргументацию, инструкцию.

Особенности описания: настоящее время, показатели места, прилагательные, высокий коэффициент лексического разнообразия.

Особенности повествования: прошедшее время, показатели времени и места.

Особенности экспозиции: эксплицирующие формулы, малая длина предложений, большое число абзацев.

Особенности аргументации: конструкции, такие как *по моему мнению, с нашей точки зрения, согласно нашему мнению*, союзы, концессивные адвербиальные подчинительные связки (личные местоимения первого и второго лица).

Особенности инструкции: императивы и личные местоимения второго лица.

Типы и особенности текста по Biber

Типы текста по Biber основаны на квантитативно-статистическом многоаспектном анализе [12, 21]. Biber предлагает следующие типы текста: участие в продуцировании, продуцирование информации, вовлечённое повествование, эксплицитную ссылку, ситуационно-зависимую ссылку, явное выражение убеждения, абстрактное информирования, и мгновенное уточнение информации [12: 102–115].

Особенности участия в продуцировании: личные глаголы, сокращения, глаголы в настоящем времени, местоимения первого и второго лица, аналитическое отрицание, указательные местоимения, общая эмфатика, местоимение ОНО, БЫТЬ как ключевой глагол, каузативное подчинение, дискурсивы, неопределённые местоимения, общие ограничители, усилители, определительные предложения, модальные глаголы возможности, WH-клаузы, предлоги в конце.

Особенности продуцирования информации: существительные, длина слов, предлоги, коэффициент лексического разнообразия и относительные прилагательные.

Особенности вовлечённого повествования: глаголы в прошедшем времени, местоимения третьего лица, глаголы совершенного вида, глаголы, предваряющие прямую речь, синтетическое отрицание, и клаузы, включающие причастие настоящего времени.

Особенности эксплицитной ссылки: определительные WH-клаузы, номинализации и соположение сочетаний.

Особенности ситуационно-зависимой ссылки: адвербиальные конструкции времени и места, наречия.

Особенности явного выражения убеждения: инфинитивы, модальные глаголы предсказания, глаголы убеждения, подчинение через условные предложения, модальные глаголы облигаторности.

Особенности абстрактного информирования: союзы, пассивный залог, клаузы, содержащие причастия прошедшего времени, а также другие адвербиальные подчинители.

Особенности мгновенного уточнения информации: клаузы, подчиняемые через ЧТО, а также указательные местоимения.

Явное выражение убеждения, абстрактное информирование и мгновенное уточнение информации.

Features of *involved production*: private verbs, contractions, present tense

verbs, 1st and 2nd person pronouns, analytic negation, demonstrative pronouns, general emphatics, pronoun IT, BE as main verb, causative subordination, discourse

particles, indefinite pronouns, general hedges, amplifiers, sentence relatives,

WH questions, possibility modals, WH clauses, and final prepositions2.

Субстантивные и глагольные типы и особенности текста

Характеристики субстантивного типа текста: информативны в основном существительные, включаются все характеристики, связанные с существительными, например, именные словосочетания, пропозициональные дополнения, пропозициональные зависимые от имён, детерминанты и др.

Характеристики глагольного типа текста: ключевыми характеристиками являются глаголы и их атрибуты, вместе с другими глагольными характеристиками, например, глагольными частицами, финитными предикатами-связками, нефинитными предикатами-связками и др.

REFERENCES

1. Haas S., Grams E. Readers, authors, and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science*, 2000, vol. 51, no. 2, pp. 181–192.

2. Haas S., Grams E. Page and link classifications: Connecting diverse resources. *Proceedings of Digital Libraries '98*, Pittsburgh USA, 1998, pp. 99–107.

3. Crowston K. Williams M. The effects of linking on genres of web documents. *Proceedings of the 32nd Hawaii International Conference on System Sciences*, Hawaii, USA, 1999, no pagination.

4. Shepherd M., Watters C. The functionality attribute of cybergenres. *Proceedings of the 32nd Hawaii International Conference on System Science* Hawaii, USA, 1999, no pagination.

5. Waller R. The typographic contribution to language. Thesis submitted for the degree of Doctor of Philosophy, University of Reading, UK, 1987.

6. Ihlström C., Lundberg J. The online news genre through the user perspective. *Proceedings of the 36th Hawaii International Conference on System Science*, no pagination, Hawaii, USA, 2003.

7. Ihlström C. Åkesson M. Genre characteristics – a front page analysis of 85 Swedish online newspapers. *Proceedings of the 37th Hawaii International Conference on System Science*, Hawaii, USA, 2004, no pagination.

8. Faigley L., Meyer P. Rhetorical theory and readers' classification of text types. *Text*, 1983, vol. 3, pp. 305–325.

9. Werlich E. *A text grammar of English.* Heidelberg, Quelle and Meyer, 1976.

10. Beaugrande R.-A., Dressler W. *Introduction to text linguistics.* London, New York, Longman, 1981.

11. Adam J.-M. *Les textes : types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris, Nathan, 1992.

12. Biber D. *Variation across speech and writing.* Cambridge, Cambridge University Press, 1988.

13. Eagles 1996. *EAGLES preliminary recommendations on text typology.* EAGLES Document EAG-TCWG-TTYP/P, Version of June, 1996. Available at: http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html.

14. Biber D. *Dimensions of register variation.* Cambridge, Cambridge University Press, 1995.

15. Görlach M. *Text types and the history of English.* Berlin, New York, Mouton de Gruyter, 2004.

16. Stubbs M. *Text and corpus analysis.* Oxford, Blackwell Publishers, 1996.

17. Karlgren J. *Stylistic experiments for information retrieval.* Thesis Diss. Doct. Sci. (Philos.). Stockholm University, 2000.

18. Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., Pasquier M. TyPWeb: décrire la toile pour mieux comprendre les parcours. *Colloque International sur les*

Usages et les Services des Télécommunications, e-Usages, no pagination, Paris, 2001.

19. Beaudouin V., Fleury S., Habert B., Illouz G., Licoppe C., Pasquier M. *Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands*. 2001. Available at: http://www.atala.org/je/010428/TyPWeb.ppt.

20. Biber D. *Towards a typology of web registers: A multi-dimensional analysis*. Invited lecture, Conference on Corpus Linguistics, Perspectives for the future. University of Heidelberg, Germany, 2004.

21. Biber D. A typology of English texts. *Linguistics*, 1989, vol. 27, pp. 3–43.

22. Santini M. Automatic Text Analysis: Gradations of text types in web pages. *Proceedings of the Tenth ESSLLI Student Session*, Edinburgh UK, 2005, pp. 276–285

23. Roberts G. The home page as genre: A narrative approach. *Proceedings of the 31st Hawaii International Conference on System Science*. Hawaii, USA, 1998, no pagination.

24. Joho H., Sanderson M. The SPIRIT collection: An overview of a large web collection. *SIGIR Forum*, 2004, vol. 38, no. 2, no pagination.

25. Tapanainen P., Järvinen T. A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington USA, 1997, pp. 64–71.

26. Bouayad-Agha N., Scott D., Power P. Integrating content and style in documents: A case study of patient information leaflets. *Information Design Journal*, 2000, vol. 9, no. 2–3, pp. 161–176.

27. Douglas S., Hurst M. Layout and language: Lists and tables in technical documents. In: *Proceedings of SIGPARSE Workshop on Punctuation in Computational Linguistics*. Santa Cruz, 1996, pp. 19–24.

28. Say B., Akman V. Current approaches to punctuation in computational linguistics. *Computers and the Humanities*, 1997, vol. 30, no. 6, pp. 457–469.

29. Hurst M. Layout and language: Challenges for table understanding on the web. In: *Proceedings of the 1st International Workshop on Web Document Analysis*, no pagination, Seattle, USA, 2001.

Перевод И. А. Афанасьева

Статья поступила в редакцию 18.09.2017

БИБЛИОГРАФИЧЕСКОЕ ОПИСАНИЕ СТАТЬИ

Сантини М. Веб-страницы, типы текстов и лингвистические характеристики : некоторые вопросы // Жанры речи. 2019. № 1 (21). С. 22–33. DOI: https://doi.org/10.18500/2311-0740-2019-1-21-22-33

For citation

Santini M. Web Pages, Text Types, and Linguistic Features : Some Issues. *Speech Genres*, 2019, no. 1 (21), pp. 22–33 (in Russian). DOI: https://doi.org/10.18500/2311-0740-2019-1-21-22-33