Web pages, text types, and linguistic features: Some issues

Marina Santini
University of Brighton, UK

1 Introduction

With the growth of the Web a massive quantity of documents, namely web pages, are freely available for (corpus-)linguistic studies. Web pages can be considered as a new kind of document, much more unpredictable and individualized than paper documents. While the linear organization of most paper documents is still reflected in traditional electronic corpora, such as the British National Corpus (BNC), web pages have a visual organization that allows the inclusion of several functions or several texts with different communicative purposes in a single document. For example, the space on a web page can be divided into different sections, organized by lists of links – mainly isolated noun structures or verbal elements (Haas and Grams 2000: 186–187) – and snippets of text scattered around the main body of the document, such as navigational buttons, menus, ads, and search boxes, that are visually dislocated in different areas of a single page. Additionally, the effect of hyperlinking (Haas and Grams 1998; Crowston and Williams 1999), interactivity and multi-functionality (Shepherd and Watters 1999) can affect the textuality of web pages, which heavily rely also on the use of images and other graphical elements. Although the use of fonts of different types, sizes, and colours, as well as the use of formatting devices, like columns, lines separating different sections of a document, pictures, etc. is not new (cf. Waller 1987 for a detailed description of the role of both language and typography in the formation of document types), a newspaper article organized in columns and headlines does not lose its specific linguistic and textual characteristics when it is included in a corpus like the BNC. The same is not true for many web pages, because the visual structure of a web page incorporating a newspaper article in most cases cannot be flattened out or ignored without losing important information (cf. Ihlström and Lundberg 2003; Ihlström and Åkesson 2004). A web page can be considered as a sort of container from where the reader picks up the information s/he needs. Artificially separating what is considered to be the main body from the rest is an arbitrary operation and it would not make sense in many cases, for example for web pages similar to those shown in Figure 6 and Figure 8. In sum, web pages tend to be more complex and more mixed than traditional paper or electronic documents. On a web page, not all the elements necessarily belong together, but they all contribute to form a whole, even without any linear progression.

Web pages are not only noisy at textual level. They also contain lots of physical noise. On a raw web page, i.e. a web page downloaded from the web without any pre-processing, many irregularities can be found, especially if the page has an HTML format. Unpredictable punctuation, typos, grammar mistakes, exotic names, extra-linguistic elements, such as HTML tags and code snippets, can make the use of NLP tools and automatic extraction of linguistic features hard. In particular, it is difficult to regularize HTLM mark-up, first because HTML syntax is permissive, and second because HTML is written by humans using different coding styles. Even when HTML code is written with software packages, such as Microsoft Frontpage, Micromedia Dreamweaver, or Microsoft Word, these programs partly use dissimilar coding conventions. Cleaning or standardizing utilities, such as the freeware TidyHTML, have low power in this tangle of disparate HTML annotations.

In this scenario, an interesting question would be whether the text typologies suggested so far by (corpus)-linguists might still apply to web pages. With this purpose in mind, I designed a simple experiment to verify if existing text typologies were still suitable for the web. As the identification of these text typologies is based on linguistic features, in this paper I would like to focus on some issues that arose when I tried to automatically extract these features from a random sample of web pages. There are no ready-made or easy solutions for these issues. The purpose of this paper is to point them out and invite the corpus-linguistic community to further discussions and investigations. For the time being, my suggestion is to be cautious when assessing results coming out from any automatic approach to web pages.

The paper is organized as follows: Section 2 presents a short overview of what is usually meant by the label “text types”; Section 3 briefly describes a few studies dealing with web pages and text types; Section 4 illustrates the experimental setting of the study; Section 5 analyses six main issues brought about when dealing with text types and web pages; in Section 6 some conclusions are drawn.

2 Text types

Traditionally, text types refer to rhetorical categories, like *narration, description, exposition* and *argumentation*. The identification of text types is deeply rooted in our culture (Faigley and Meyer 1983), but the number and the labels of these rhetorical categories vary according to the linguist's orientation and preferences. For example, Werlich (1976) analyses five text types (*narration, description, exposition, argumentation* and *instruction*), Beaugrande and Dressler (1981) propose seven text types (*descriptive, narrative, argumentative, scientific, didactic, literary* and *poetic*), Adam (1992) analyses five text types (*récit, description, argumentation, explication* and *dialogue*).

Since the publication of Biber's work on linguistic variation across speech and writing (Biber 1988), the term "text types" has entered corpus linguistics. His work is by now a classic of statistical corpus-based approach, and has influenced also European standards for large language resources, such as the EAGLES guidelines on text typology (EAGLES 1996). Biber (1988) makes a distinction between genre – which later becomes register (Biber 1995: 9) – and text types. In his view, genre is influenced by cultural and external criteria, whereas text types can be derived from the texts themselves, irrespective of their genre. In other words, while external criteria follow distinctions and classifications already available in the culture, Biber establishes a typology of texts based on internal linguistic criteria only, which are interpreted with reference to external functions. Biber (1988: 102–103) suggests the following textual dimensions: *involved production, informational production, narrative concern, explicit reference, situation-dependent reference, overt expression of persuasion, abstract information,* and *online informational elaboration*.

However, the clear-cut distinction between genre/register and text types is not universally accepted or adopted. Some scholars use the label "text types" to indicate instrumental or practical genres, as opposed to literary genres (e.g. Görlach 2004). Others use "text types" and "genres" interchangeably, as synonyms (e.g. Stubbs 1996; Karlgren 2000). Finally, others (e.g. Kilgarriff and Grefenstette 2003) use the term "text types" without any further indication on how this label should be interpreted in the context in which they use it.

In this paper, I follow the rhetorical and corpus-linguistic tradition. More specifically, I investigate whether the text types suggested by Biber (1988) and those coming from the traditional rhetorical partition adopted by Werlich (1976) are suitable and applicable to web pages.

3 Background

The studies presented below are still ongoing or are preliminary investigations on text types of web pages. Although they use different approaches, they are all corpus-based.

TyPWEB (Beaudouin *et al.* 2001a, 2001b), a project for the French language that explicitly aims at extending Biber's work, provides a methodological and practical framework for website profiling, where the final goal is to develop a fine-grained typology to discriminate between personal and commercial websites. Discriminating features include both linguistic features and presentational features (layout, images, hyperlinks, etc.). Although the project is still ongoing, some results are available. Current findings show a number of interesting traits. For instance, the structure of commercial web sites is more complex than the structure of personal websites; the home page of a commercial web site has more links than other pages in the same website, while the home page of personal website does not show such a difference; the use of personal pronouns (1st and 2nd person pronouns) differs between personal and commercial websites, and so on. Breaking with Biberian tradition of a pure inductive approach, TyPWEB proposes a double approach to website profiling, a deductive approach, where categories are defined *a priori*, and an inductive approach, where categories are derived from the data itself.

Biber (2004) presents a multi-dimensional analysis of two topical categories from Google (Home and Science, including several subcategories). The multi-dimensional analysis (Biber 1988, 1989) relies on an inductive statistical approach based on factor analysis and cluster analysis, where categories are derived from data and interpreted in the light of external functions. The study returns four dimensions (*personal-involved narration*, *persuasive-argumentative discourse*, *advice??* (sic), and *abstract-technical discourse*). Unlike TyPWEB, Biber (2004) includes only linguistic features (lexical, morphological and syntactic classes, many of them extracted using a tagger and a parser), without any presentational traits.

Santini (2005) does not include any presentational features either, but unlike Biber and similar to TyPWEB, the author tries to combine the deductive approach with the inductive approach. The deductive-inductive model is based on Bayesian inference. It is deductive because it starts from a limited number of four broad and widely acknowledged text types (*descriptive/narrative, explicatory/informational, argumentative/persuasive, instructional*). It is also inductive because the inferential process is corpus-based. Inferences are based on the calculation of the probability value for a hypothesis (a text type) given one or more pieces of evidence (the frequencies of some features). Gradations of text types are returned in terms of probability values. For example, a web page can be: 0.3 instructional, 0.5 narrative, 0.7 informational, and 0.9 argumentative. Simply put, this means that the web page under analysis is considered to be predominantly argumentative, highly informational, moderately narrative, but hardly instructional. From a preliminary evaluation, it turned out that the gradations of text types returned by the model are largely consistent with those returned by human subjects.

More peripheral in our perspective but nonetheless intriguing, Roberts (1998) concentrates on a single text type – *narration* – of an individual genre – the personal home page. Among other things, the author suggests an original interpretation of hyperlinks in terms of narratology.

Although all these studies report interesting findings, showing different approaches and using different feature sets, they do not explicitly say how web pages were processed to extract linguistic and non-linguistic features; nor do they clearly state if the extraction was troublesome or controversial. As highlighted in the Introduction, web pages can be considered a new type of document, more difficult to handle than traditional documents. Therefore, the way in which features are extracted and NLP tools are applied to web pages can deeply affect the results, especially when relying on statistical techniques.

4 The study

As mentioned earlier, a number of text types have already been suggested by previous (corpus-)linguistic studies for traditional documents, either paper or electronic documents (Section 2) and for web pages (Section 3). In the study presented here the aim was to verify with simple heuristics whether two well-established text typologies – Werlich (1976) and Biber (1988) – were still suitable for web pages. To these two typologies, I added two broad text types, Nominal vs. Verbal. While carrying out the experiment, I realised that results can be disturbed by issues encountered during automatic feature extraction. The experiment includes the following steps:

- listing linguistic features representing three text typologies: Werlich’s, Biber’s, and Nominal vs. Verbal (see Appendix for a breakdown);
- extracting a random sample[1] of English web pages from the SPIRIT collection (Joho and Sanderson 2004);
- converting web pages from HTML version into ASCII format;

- submitting the ASCII version of web pages to NLP tools (tagger, parser, and n-grams);
- coding text types as arrays of nominal features;
- coding web pages as arrays of nominal features;
- comparing each web page array against text types arrays, and outputting text types for a web page.

An individual web page as a whole was taken as unit of analysis. Features were extracted from it, counted and compared with preset lists of features (preset text types). The ASCII version of a web page was tagged and parsed (using Connexor by Tapanainen and Järvinen 1997), and word n-grams (freeware utility) were computed as a measure of vocabulary variation. Frequency counts and ratios on the ASCII versions were computed by Perl scripts. Frequency counts were normalized to percentage. Each web page was coded as an array. The matching between the preset text types and the features coming from a web page was computed as an intersection of arrays (see Figure 1). It was also possible to specify a threshold. For example, with a threshold of 30 percent, only features with a normalized values 30 percent were extracted. A lower or a higher threshold affects the number of features included in the array representing the web page. With a low threshold (say 20 percent) more features are included in the web page array, so the match with several preset text type arrays is more likely. Instead, when a higher threshold is specified (say 80 percent), the number of linguistic features extracted from the web page decreases, and the possibility of a match with preset text types is reduced.

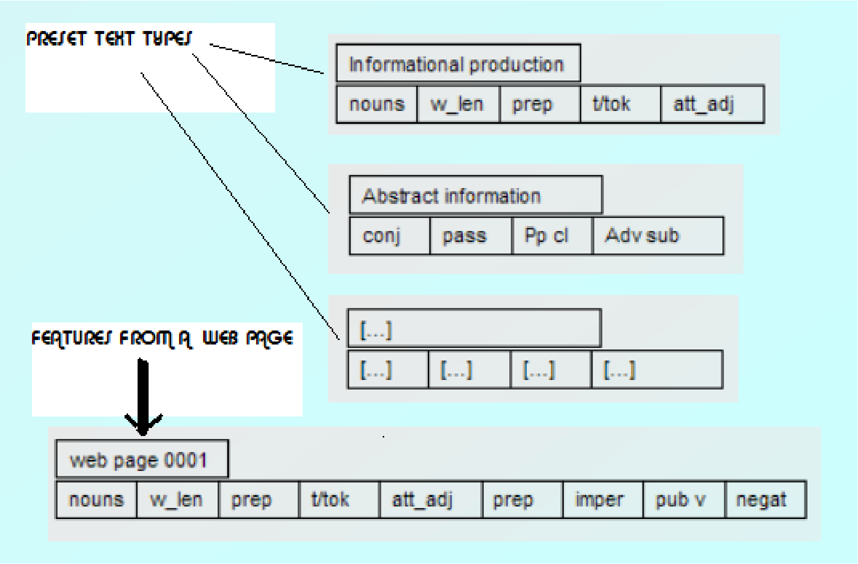


Figure 1: Arrays

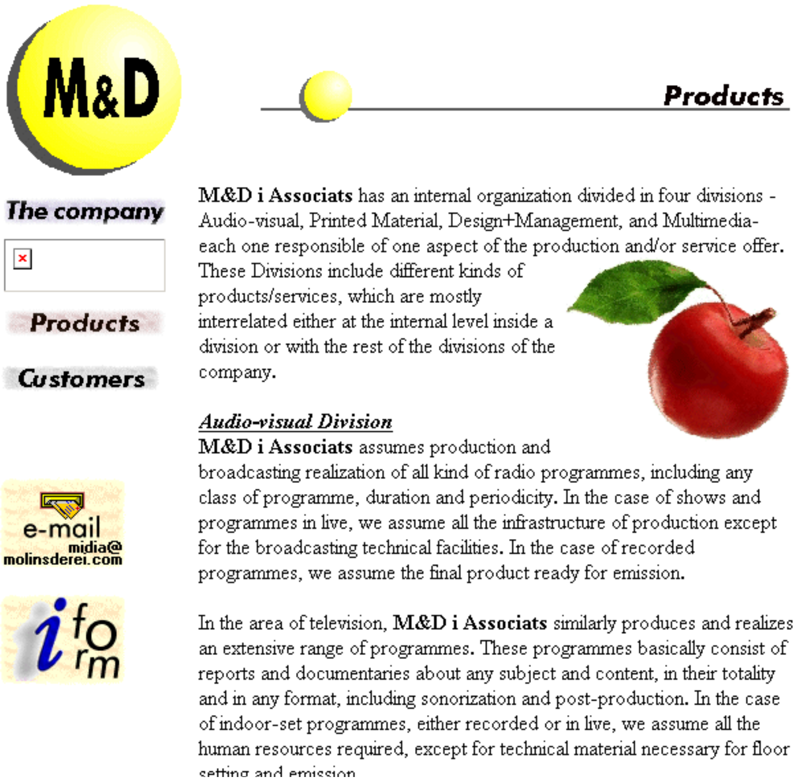
The experiment was designed to assign one or more text types to a web page as a whole. For example, a web page could be classified as *Biber's involved production*, *Werlich's argumentation*, and *Verbal text type*.

5 Issues

Most of the web pages analysed with a threshold of 50 percent fell into the *Nominal text type*. However, the analysis of the frequency logs showed that automatic feature extraction was not as smooth as one would expect, and a number of issues related to text processing were likely to undermine the final results. In the following subsections, I briefly describe six issues, namely:

1. Elements of text coded as images
2. Headings
3. Lists
4. Proper nouns
5. Tabular text
6. Mixed text

5.1 Issue 1: Elements of text coded as images



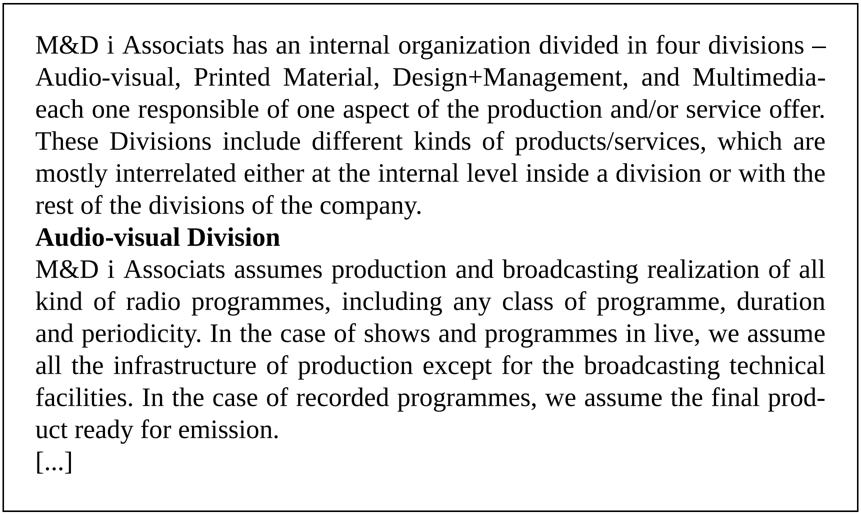
M&D i Associats has an internal organization divided in four divisions - Audio-visual, Printed Material, Design+Management, and Multimedia- each one responsible of one aspect of the production and/or service offer. These Divisions include different kinds of products/services, which are mostly interrelated 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.

In the area of television, **M&D i Associats** similarly produces and realizes an extensive range of programmes. These programmes basically consist of reports and documentaries about any subject and content, in their totality and in any format, including sonorization and post-production. In the case of indoor-set programmes, either recorded or in live, we assume all the human resources required, except for technical material necessary for floor setting and emission

Figure 2: Text coded as images



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.

[...]

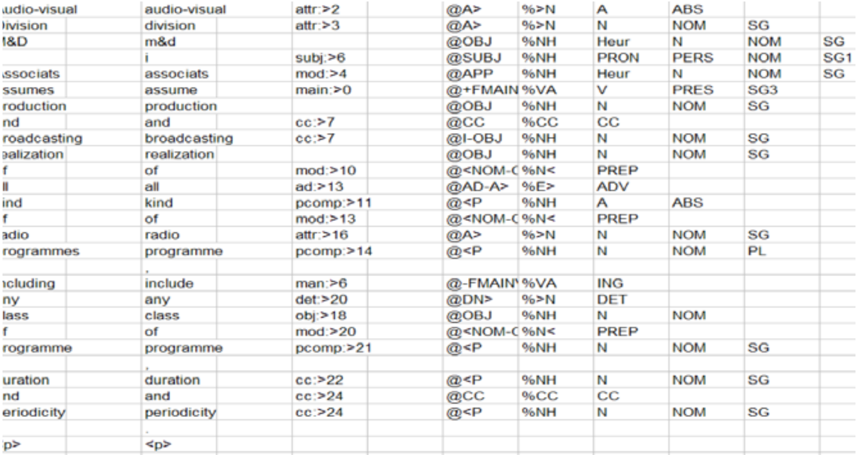
Figure 3: ASCII version of the web page shown in Figure 2

If the ASCII version (Figure 3) and the HTML version (Figure 2) are compared, you can see that some elements have disappeared from the ASCII version, i.e. all the items listed on the left-hand side together with the main heading. This happens when elements of text are coded as images embedded in the HTML code. It is difficult to find an easy solution for this first loss.

5.2 Issue 2: Headings

The second paragraph in Figure 3 has a heading (highlighted in grey) that is not detected as an independent unit of analysis by the parser (see Figure 4). A wrong syntactic analysis is returned because the parser sticks the heading together with the following sentence. Headings rarely end with a punctuation mark and rarely consist of a standard grammatical sentence (exceptions are questions, like “How Do You Create Your Intranet?”). Therefore, headings can defeat the sentence splitter of a parser. One solution is substantial pre-processing. For example, the HTML tag for headings `<h#>` (parsers can ignore HTML tags, but usually they cannot interpret them) could be employed to create an artificial sentence boundary. However, on the web page shown in Figure 2, the heading is tagged as: `<p><u><b><i>Audio-visual Division</i></b></u><br>`, which shows how the use of HTML tags can be unpredictable.

The text types returned for Figure 2 are *Werlich´s description*, *Biber´s informational production*, and *Nominal text type*.



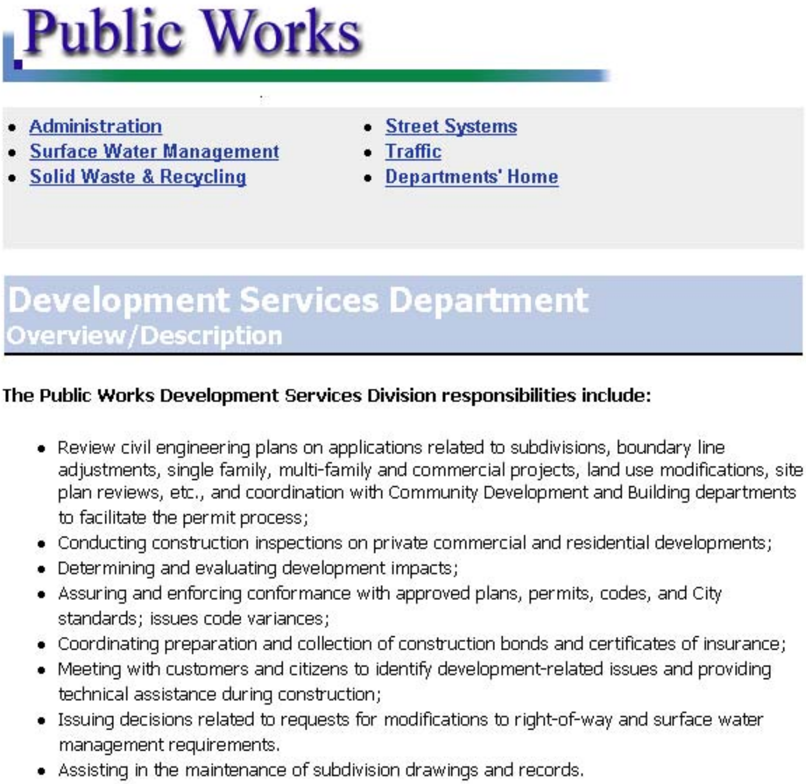
ıudio-visual	audio-visual	attr:>2	@A>	%>N	A	ABS		
)ivision	division	attr:>3	@A>	%>N	N	NOM	SG	
1&D	m&d		@OBJ	%NH	Heur	N	NOM	SG
	i	subj:>6	@SUBJ	%NH	PRON	PERS	NOM	SG1
ıssociats	associats	mod:>4	@APP	%NH	Heur	N	NOM	SG
ssumes	assume	main:>0	@+FMAIN	%VA	V	PRES	SG3	
roduction	production		@OBJ	%NH	N	NOM	SG	
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	
	,							
ıcluding	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>							

Figure 4: The heading is not analysed as a separate syntactic unit

5.3 *Issue 3: Lists*

Figure 5 shows a very common textual organization, lists. Lists can raise problems for stylometric measurements, such as average sentence length, and for syntactic parsing. Lists present a number of special traits (cf. Bouayad-Agha *et al.* 1999). For instance, the introductory sentence of a list is usually semantically incomplete, and either lacks a final punctuation mark or ends with a colon. Also, the items in a list might be single words or full sentences, with or without ending punctuation. In Figure 5, there is a short list at the top, each item containing a few words without punctuation, and a longer list with an introductory sentence ending with a colon in the body. The items of this latter list have long sentences ending either with a semicolon or with a full stop. The parser is confused by this structure. Additionally, average sentence length ratios – usually based on full stops, question and exclamation marks, and a few other symbols – are misleading. One solution would be the exploitation of the mark-up tag `<li>` to set artificial sentence boundaries. But what about the semantically incomplete introductory sentence of the longer list ending with a colon?

The text types returned for Figure 5 are *Werlich´s instruction* and *Nominal text type*.



Public Works

- Administration
- Surface Water Management
- Solid Waste & Recycling
- Street Systems
- Traffic
- Departments' Home

Development Services Department
Overview/Description

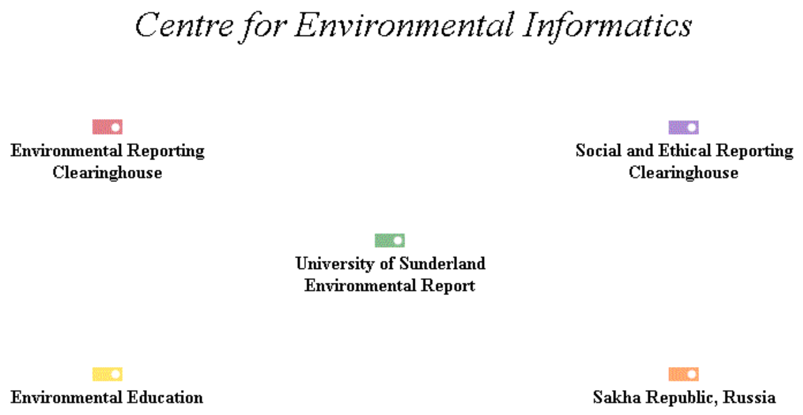
The Public Works Development Services Division responsibilities include:

- Review civil engineering plans on applications related to subdivisions, boundary line adjustments, single family, multi-family and commercial projects, land use modifications, site plan reviews, etc., and coordination with Community Development and Building departments to facilitate the permit process;
- Conducting construction inspections on private commercial and residential developments;
- Determining and evaluating development impacts;
- Assuring and enforcing conformance with approved plans, permits, codes, and City standards; issues code variances;
- Coordinating preparation and collection of construction bonds and certificates of insurance;
- Meeting with customers and citizens to identify development-related issues and providing technical assistance during construction;
- Issuing decisions related to requests for modifications to right-of-way and surface water management requirements.
- Assisting in the maintenance of subdivision drawings and records.

Figure 5: Lists

Another example of a page that is not prose is shown in Figure 6. It is a list of items scattered over the page without bullets, numbering, or ending punctuation. Visually, this lack is not felt, because the strings of text are perceived as separate entities. The underlying HTML structure is a table.

The text type returned for Figure 6 is *Nominal text type*.



Centre for Environmental Informatics

Environmental Reporting Clearinghouse

Social and Ethical Reporting Clearinghouse

University of Sunderland Environmental Report

Environmental Education

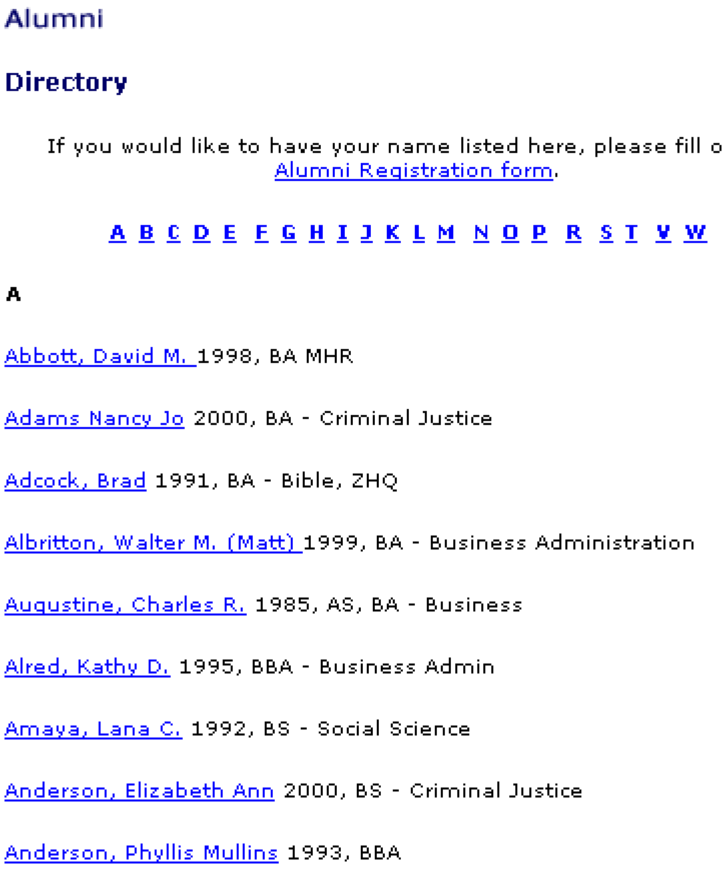
Sakha Republic, Russia

Figure 6: Scattered list

5.4 Issue 4: Proper nouns

The web page type shown in Figure 7 is very common on the web. The page contains a list of names and some personal details. Probably a named-entity and abbreviation recognition tool would be more useful than a parser or a tagger in this case.

The text type returned for Figure 7 is *Nominal text type*.



Alumni

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

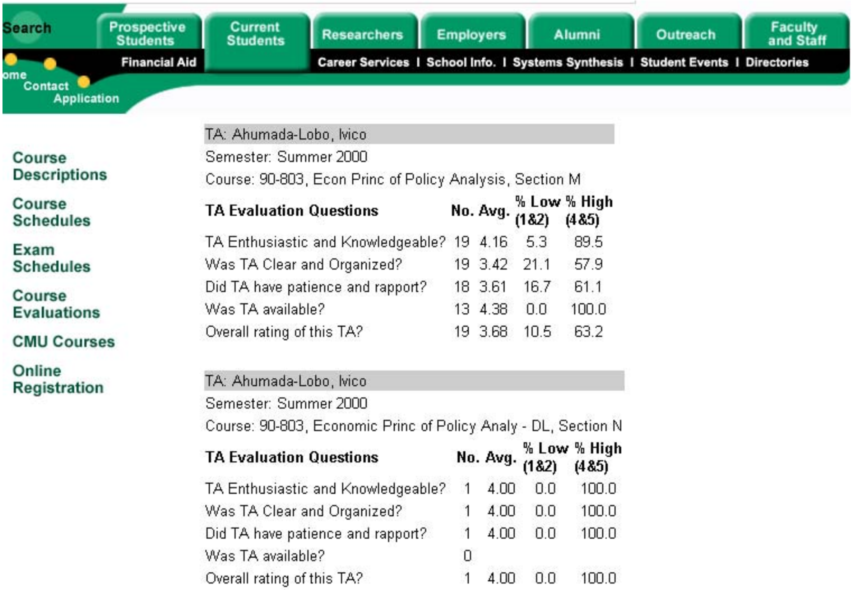
Anderson, Phyllis Mullins 1993, BBA

Figure 7: Proper nouns

5.5 Issue 5: Tabular structure

Tabular structures, similar to that shown in Figure 8, are also very common on the web. They are difficult to analyse from a linguistic standpoint, but luckily they have started receiving attention from the linguistic community (cf. Douglas and Hurst 1996; Say and Akman 1997; Hurst 2001).

The text type returned for Figure 8 is *Biber's explicit reference text type*.



TA: Ahumada-Lobo, Ivico
Semester: Summer 2000
Course: 90-803, Econ Princ of Policy Analysis, Section M

TA Evaluation Questions	No.	Avg.	% Low (1&2)	% High (4&5)
TA Enthusiastic and Knowledgeable?	19	4.16	5.3	89.5
Was TA Clear and Organized?	19	3.42	21.1	57.9
Did TA have patience and rapport?	18	3.61	16.7	61.1
Was TA available?	13	4.38	0.0	100.0
Overall rating of this TA?	19	3.68	10.5	63.2

TA: Ahumada-Lobo, Ivico
Semester: Summer 2000
Course: 90-803, Economic Princ of Policy Analy - DL, Section N

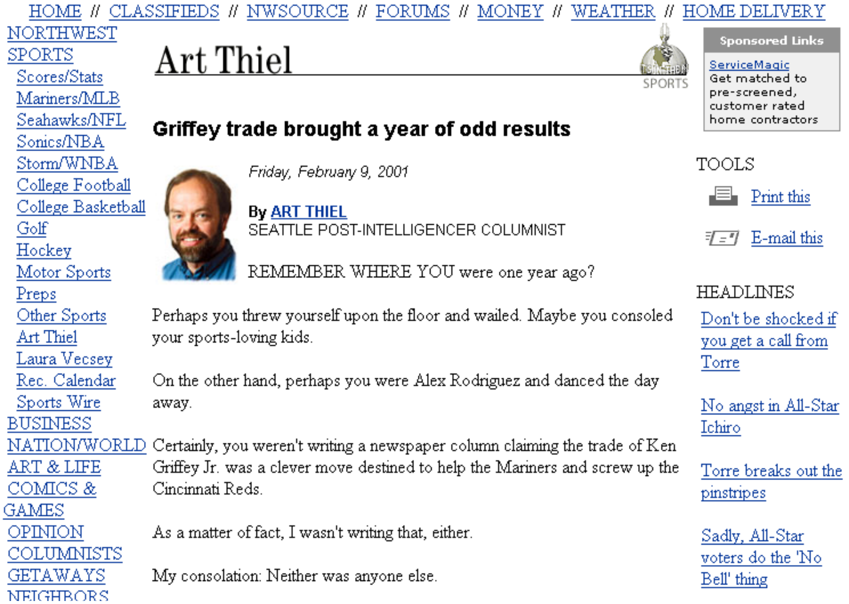
TA Evaluation Questions	No.	Avg.	% Low (1&2)	% High (4&5)
TA Enthusiastic and Knowledgeable?	1	4.00	0.0	100.0
Was TA Clear and Organized?	1	4.00	0.0	100.0
Did TA have patience and rapport?	1	4.00	0.0	100.0
Was TA available?	0			
Overall rating of this TA?	1	4.00	0.0	100.0

Figure 8: Tabular structure

5.6 *Issue 6: Mixed texts*

Another common web page type is shown in Figure 9. In this page, there is a main article and other strings of text located around the main body. Semantically, these strings do not belong to the body, but provide additional information to the reader. Assuming that every text type represents a communicative function, how many text types are included in this page? At least three: a comment (the main article), an informational list (the headlines on the right-hand side), and an index (the items on the left-hand side).

The text type returned for Figure 9 is only *Biber's involved production*.



HOME // CLASSIFIEDS // NWSOURCE // FORUMS // MONEY // WEATHER // HOME DELIVERY

NORTHWEST
SPORTS
Scores/Stats
Mariners/MLB
Seahawks/NFL
Sonics/NBA
Storm/WNBA
College Football
College Basketball
Golf
Hockey
Motor Sports
Preps
Other Sports
Art Thiel
Laura Vecsey
Rec. Calendar
Sports Wire
BUSINESS
NATION/WORLD
ART & LIFE
COMICS &
GAMES
OPINION
COLUMNISTS
GETAWAYS
NEIGHBORS

Art Thiel

SPORTS

Griffey trade brought a year of odd results

Friday, February 9, 2001

By ART THIEL
SEATTLE POST-INTELLIGENCER COLUMNIST

REMEMBER WHERE YOU were one year ago?

Perhaps you threw yourself upon the floor and wailed. Maybe you consoled your sports-loving kids.

On the other hand, perhaps you were Alex Rodriguez and danced the day away.

Certainly, you weren't writing a newspaper column claiming the trade of Ken Griffey Jr. was a clever move destined to help the Mariners and screw up the Cincinnati Reds.

As a matter of fact, I wasn't writing that, either.

My consolation: Neither was anyone else.

Sponsored Links
ServiceMagic
Get matched to pre-screened, customer rated home contractors

TOOLS
Print this
E-mail this

HEADLINES
Don't be shocked if you get a call from Torre

No angst in All-Star Ichiro

Torre breaks out the pinstripes

Sadly, All-Star voters do the 'No Bell' thing

Figure 9: Mixed text

6 Conclusions

The list of issues could be extended, but I stop here and draw some conclusions. Even though most of the text types returned by the heuristics are, broadly speaking, correct, automatic feature extraction from web pages is troublesome. For instance:

- Some text elements of web pages are lost when text is coded as images (Section 5.1).
- NLP tools might be unreliable when run on the ASCII version of a web page without any pre-processing, and counts can be misleading (Sections 5.2 and 5.3). The lack of final punctuation can be a big challenge for some NLP tools like parsers, which usually rely on sentence boundaries. It seems that the use of final punctuation on web pages might differ from that of linear documents. In fact, since web pages are visual documents, text elements can have different font and colour, and can be dislocated anywhere within the page. Formatting and position are visual devices that make the use of sentence delimitation redundant (for example, see Figure 6).

- It seems that grammatical and lexical features alone are not enough to derive a text typology for web pages. Other features, such as proper nouns and tabular structure, need to be identified (Sections 5.4 and 5.5). While proper nouns can be detected with a named-entity recognition utility, it remains difficult to analyse a tabular structure.

As for text types, I notice that:

- Some web pages do not fit well into existing text types (for example, see Figure 6, Figure 7 and Figure 8). In this case, an inductive approach, such as the multi-dimensional analysis, would help in highlighting and interpreting novelties.
- One important fact that should be taken into account is the mixed nature of a text (Section 5.6). A text can be a mixture of different forms of expressions and different communicative acts; it rarely corresponds to an ideal or idealized type (Beaugrande and Dressler 1981: 181 ff.). This is especially true for web pages, which are visual objects, mostly with a non-linear organization, including several communicative purposes. In this case, an approach that could return a more fine-grained analysis of the textuality of a web page (e.g. Santini 2005) would be more suitable.

In summary, I tried the difficult task of investigating text typologies in a random sample of raw web pages, and not in a corpus of pre-selected 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 presented in this paper 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. As web pages represent a huge textual reservoir that cannot be neglected by the (corpus-)linguistic community, further discussions and investigations are needed.

Notes

1. A random sample of 1,000 unclassified web pages from the SPIRIT collection is downloadable from http://www.itri.brighton.ac.uk/~Marina.Santini/, bottom of the page.

2. The following features were not included: *that*-deletion, *do* as a pro-verb and non-phrasal co- ordination.
3. The following feature was not included: pied-piping construction.
4. The following feature was not included: split auxiliary.

References

Adam, Jean-Michel. 1992. *Les textes: types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: Nathan.

Beaudouin, Valérie, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe and Marie Pasquier. 2001a. TyPWeb: décrire la toile pour mieux comprendre les parcours. *Colloque International sur les Usages et les Services des Télécommunications, e-Usages*, no pagination, Paris.

Beaudouin, Valérie, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe and Marie Pasquier. 2001b. *Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands*.

Available at http://www.atala.org/je/010428/TyPWeb.ppt

Beaugrande, Robert-Alain and Wolfgang Dressler. 1981. *Introduction to text linguistics*. London-NY: Longman.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27: 3–43.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

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

Bouayad-Agha Nadjet, Donia Scott and Richard Power. 2000. Integrating content and style in documents: A case study of patient information leaflets. *Information Design Journal* 9: 2–3, 161–176.

Crowston, Kevin and Marie Williams. 1999. The effects of linking on genres of web documents. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, no pagination, Hawaii, USA.

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

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

Faigley, Lester and Paul Meyer. 1983. Rhetorical theory and readers' classification of text types. *Text* 3: 305–325.

Görlach, Manfred. 2004. *Text types and the history of English.* Berlin-NY: Mouton de Gruyter.

Haas, Stephanie and Erika Grams. 1998. Page and link classifications: Connecting diverse resources. In *Proceedings of Digital Libraries '98*, 99–107, Pittsburgh, USA.

Haas, Stephanie and Erika Grams. 2000. 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* 51 (2): 181–192.

Hurst, Matthew. 2001. 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.

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

Ihlström, Carina and Jonas Lundberg. 2003. The online news genre through the user perspective. In *Proceedings of the 36th Hawaii International Conference on System Science*, no pagination, Hawaii, USA.

Joho, Hideo and Mark Sanderson. 2004. The SPIRIT collection: An overview of a large web collection. *SIGIR Forum* 38: 2, no pagination.

Karlgren, Jussi. 2000. Stylistic experiments for information retrieval. Thesis submitted for the degree of Doctor of Philosophy, Stockholm University.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on the web as a corpus. *Computational Linguistics* 29 (3): 333–347.

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

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

Say, Bilge and Varol Akman. 1997. Current approaches to punctuation in computational linguistics. *Computers and the Humanities* 30 (6): 457–469.

Shepherd, Michael and Carolyn Watters. 1999. The functionality attribute of cybergenres. In *Proceedings of the 32nd Hawaii International Conference on System Science,* no pagination. Hawaii, USA.

Stubbs, Michael. 1996. *Text and corpus analysis.* Oxford: Blackwell Publishers.

Tapanainen, Pasi and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing,* 64–71, Washington, USA.

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

Werlich, Egon. 1976. *A text grammar of English.* Heidelberg: Quelle and Meyer.

Appendix

Werlich’s text types and features

Werlich's text types are based on a qualitative analysis of paper documents (Werlich, 1976). Werlich outlines five text types for: *description, narration, exposition, argumentation*, and *instruction*.

Selected features for *Description*: present tense, location indicators, adjectives, and high type/token ratio.

Selected features for *Narration*: past tense, time indicators, and location indicators.

Selected features for *Exposition*: explicatory formulae, low sentence length, and high number of paragraphs.

Selected features for *Argumentation*: terms like *in my opinion, in our view, according to me,* conjuncts, concessive adverbial subordinators, and 1^{st} and 2^{nd} person pronouns.

Selected features for *Instruction*: imperatives, and second person pronouns.

Biber’s text types and features

Biber’s text types are derived with a quantitative-statistical analysis, the multidimensional analysis (Biber, 1988, 1989). Biber suggests the following text types: *involved production, informational production, narrative concern, explicit reference, situation-dependent reference, overt expression of persuasion, abstract information,* and *online informational elaboration* (Biber, 1988: 102–115).

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 prepositions[2].

Features of *informational production*: nouns, word length, prepositions, type/token ratio, and attributive adjectives.

Features of *narrative concern*: past tense verbs, third person pronouns, perfect aspect verbs, public verbs, synthetic negation, and present participial clause.

Features of *explicit reference*: WH relative clauses, nominalizations, and phrasal coordination[3].

Features of *situation-dependent reference*: time adverbials, place adverbials, and adverbs.

Features of *overt expression of persuasion*: infinitives, prediction modals, suasive verbs, conditional subordination, and necessity modals[4].

Features of *abstract information*: conjuncts, passives, past participial clauses, and other adverbial subordinators.

Features of *online informational elaboration*: THAT clauses, and demonstratives.

Nominal and verbal text types and features

Features of *Nominal text type*: nouns are the main bearers of information, therefore all features connected to nouns are included here, for instance noun phrases, prepositional complements, pre-modifiers of a nominal, determiners, etc.

Features of *Verbal text type*: verbs and their attributes represent the core features of this text type, together with many other verbal features, for instance verb particles, finite auxiliary predicators, non-finite auxiliary predicators, etc.