

Annotated Corpora vs. Raw Web Page Collections

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

Marina Santini

Marina.Santini@itri.brighton.ac.uk

University of Brighton


UK

AAACL/ICAME, 12-15 May 2005,
Ann Arbor, MI

The Web: A Ready-Made Corpus?

- A massive quantity of documents freely available.
- Can it be used as a ready-made corpus?
- A case study: *Investigation of text types in web pages.*


Web Pages: New Types of Documents?



Products

The company

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 the setting and emission.

e-mail
molinsdefel.com

iform

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 plan reviews, etc., and coordination with Community Development and Building department 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 of manager;
- Assisting

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

Search

Prospective Students | Current Students | Researchers | Employers | Alumni | Outreach | Faculty and Staff

Financial Aid | Career Services | School Info. | Systems Synthesis | Student Events | Directories

Course Descriptions

Course Schedules

Exam Schedules

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

TA Evaluation Questions	No.	Avg.	% Low	% High
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
able?	13	4.38	0.0	100.0
g of this TA?	19	3.68	10.5	63.2


[HOME](#) // [CLASSIFIEDS](#) // [NWSOURCE](#) // [FORUMS](#) // [MONEY](#) // [WEATHER](#) // [HOME DELIVERY](#)

Art Thiel

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.

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

Centre for Environmental Informatics

Environmental Reporting Clearinghouse

Social and Ethical Reporting Clearinghouse

University of Sunderland
Environmental Report

Environmental Education

Sakha Republic, Russia

The Study

- I. A corpus-based approach: the web pages of the SPIRIT collection
- II. A number of text typologies suggested by previous studies
- III. Linguistic features
- IV. Tagger/Parser
- V. Algorithm

I. Web Page Collection

The working corpus is made of 200 random English web pages coming from a raw Web crawl, the SPIRIT collection, carried out by the University of Waterloo in 2001.

Centre for Environmental Informatics

Environmental Reporting Clearinghouse

University of Sunderland Environmental Report

Environmental Education

Alumni Directory

If you would like to have your name listed here, please fill out the [Alumni Registration form](#).

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

A

[Abbott, David M.](#) 1996, 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

[Amava, Lana C.](#) 1992, BS - Social Science

[Anderson, Elizabeth Ann](#) 2000, BS - Criminal Justice

[Anderson, Phyllis Mullins](#) 1993, BBA

Search

Prospective Students

Current Students

Researchers

Employers

Alumni

Outreach

Faculty and Staff

Financial Aid

Career Services | School Info | Systems Synthesis | Student Events | Directories

Contact Application

Course Descriptions

Course Schedules

Exam Schedules

Course Evaluations

CMU Courses

Online Registration

TA, Ahumada-Lobos
Semester: Summer
Course: 90-803, E

TA Evaluation of
TA, Ahumada-Lobos
Semester: Summer
Course: 90-803, E

TA, Ahumada-Lobos
Semester: Summer
Course: 90-803, E

TA, Ahumada-Lobos
Semester: Summer
Course: 90-803, E

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.

Traditional Text Types on the Web?

Werlich (1976) text typology:

1. Narration
2. Description
3. Exposition
4. Argumentation
5. Instruction

Broad text typology:

1. Nominal
2. Verbal

Biber (1988) text typology:

1. Involved production
2. Informational production
3. Narrative concern
4. Explicit reference
5. Situation-dependent reference
6. Overt expression of persuasion
7. Abstract information
8. Online informational elaboration

Typology of web registers

Towards a typology of web registers: A multi-dimensional analysis (Biber 2004)

1. Personal, involved narration
2. Persuasive/argumentative
3. Advice??
4. Abstract/technical discourse

II. Features of Werlich's Text Types

1. Narration: past tense, time indicators, location indicators.
2. Description: present tense, location indicators, adjectives, high type/token ratio.
3. Exposition: explicatory formulae, low sentence length, many paragraphs.
4. Argumentation: 1st and 2nd pers. sing. and plur., terms like in my opinion, in our view, according to me, conjuncts, concessive adverbial subordinators, such as therefore, thus, etc.
5. Instruction: imperatives, second person singulars and plurals.

II. Features of Biber's Text Types

1. Involved production: private verbs, contractions, present tense verbs, 1st and 2nd pers. 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, final prepositions .
2. Informational production: nouns, word length, prepositions, type/token ratio, attributive adjectives.
3. Narrative concern: past tense verbs, third person pronouns, perfect aspect verbs, public verbs, synthetic negation, present participial clause.
4. Explicit reference: WH relative clauses, nominalizations, phrasal coordination .
5. Situation-dependent reference: time adverbials, place adverbials, adverbs.
6. Overt expression of persuasion: infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals .
7. Abstract information: conjuncts, passives, past participial clauses, other adverbial subordinators.
8. Online informational elaboration: THAT clauses, demonstratives.

II. Features of Broad Text Types

1. Nominal: noun is the main bearer of information, therefore all features connected to nouns are included here, for instance noun phrases, prepositional complements, pre-modifiers of a nominal, determiners, etc.
2. Verbal: 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, nonfinite auxiliary predicators, etc.

III. Tagger/Parser

Connexor by Tapanainen and Järvinen (1997).

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	the	the	det:>2	@DN> %>N DET
2	cat	cat	subj:>3	@SUBJ %NH N NOM SG
3	is	be	main:>0	@+FMAINV %VA V PRES SG3
4	on	on	loc:>3	@ADVL %EH PREP
5	the	the	det:>6	@DN> %>N DET
6	mat	mat	pcomp:>4	@<P %NH N NOM SG
7	.	.		
8	<s>	<s>		

IV. Methodology

- Web pages converted into Ascii files
- Parsing of the Ascii files
- Frequency counts
- Normalization to percentage
- Text types were coded as arrays of linguistic features
- Intersection of arrays
- Threshold


Informational production				
nouns	w_len	prep	t/tok	att_adj

Abstract information			
conj	pass	Pp cl	Adv sub

[...]			
[...]	[...]	[...]	[...]

web page 0001								
nouns	w_len	prep	t/tok	att_adj	prep	imper	pub v	negat

Issue 1: Elements of text coded as images



The company

Products

Customers

Products

Customers

e-mail
media@molinsderet.com

ifor
rm

Products

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.



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.

[...]

Issue 2: Headings

Audio-visual	audio-visual	attr:>2	@A>	%>N	A	ABS			
Division	division	attr:>3	@A>	%>N	N	NOM	SG		
M&D	m&d		@OBJ	%NH	Heur	N	NOM	SG	
i	i	subj:>6	@SUBJ	%NH	PRON	PERS	NOM	SG1	
Associats	associats	mod:>4	@APP	%NH	Heur	N	NOM	SG	
assumes	assume	main:>0	@+FMAIN	%VA	V	PRES	SG3		
production	production		@OBJ	%NH	N	NOM	SG		
and	and	cc:>7	@CC	%CC	CC				
broadcasting	broadcasting	cc:>7	@I-OBJ	%NH	N	NOM	SG		
realization	realization		@OBJ	%NH	N	NOM	SG		
of	of	mod:>10	@<NOM-C	%N<	PREP				
all	all	ad:>13	@AD-A>	%E>	ADV				
kind	kind	pcomp:>11	@<P	%NH	A	ABS			
of	of	mod:>13	@<NOM-C	%N<	PREP				
radio	radio	attr:>16	@A>	%>N	N	NOM	SG		
programmes	programme	pcomp:>14	@<P	%NH	N	NOM	PL		
,	,								
including	include	man:>6	@-FMAIN	%VA	ING				
any	any	det:>20	@DN>	%>N	DET				
class	class	obj:>18	@OBJ	%NH	N	NOM			
of	of	mod:>20	@<NOM-C	%N<	PREP				
programme	programme	pcomp:>21	@<P	%NH	N	NOM	SG		
,	,								
duration	duration	cc:>22	@<P	%NH	N	NOM	SG		
and	and	cc:>24	@CC	%CC	CC				
periodicity	periodicity	cc:>24	@<P	%NH	N	NOM	SG		
.	.								
<p>	<p>								

Issue 3: Lists

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.

Centre for Environmental Informatics



Environmental Reporting
Clearinghouse



Social and Ethical Reporting
Clearinghouse



University of Sunderland
Environmental Report



Environmental Education



Sakha Republic, Russia

Issue 4: Proper Nouns

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

Issue 5: Tabular Text

The screenshot shows a university website with a green navigation bar. The top bar contains buttons for 'Search', 'Prospective Students', 'Current Students', 'Researchers', 'Employers', 'Alumni', 'Outreach', and 'Faculty and Staff'. Below this is a secondary bar with links for 'Home', 'Contact Application', 'Financial Aid', 'Career Services', 'School Info.', 'Systems Synthesis', 'Student Events', and 'Directories'. On the left side, there is a vertical menu with links for 'Course Descriptions', 'Course Schedules', 'Exam Schedules', 'Course Evaluations', 'CMU Courses', and 'Online Registration'. The main content area displays two TA evaluation sections for 'TA: Ahumada-Lobo, Ivico'.

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

TA Evaluation Questions	No.	Avg.	% Low (182)	% High (485)
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 (182)	% High (485)
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

Issue 6: Mixed Text

[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



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](#)

Conclusions (I)

- Even though most of the text types returned by my algorithm are, broadly speaking, correct, we saw that feature extraction is troublesome and counts can be unreliable (issues 2, *headings*; and issue 3, *lists*).
- It seems that grammatical and lexical features alone are not enough to derive a text typology for web pages (issue 4, *proper nouns*; and issue 4, *tabular text*). Proper nouns should be identified and HTML markup tags should be interpreted in a functional way and combined with linguistic features.
- Some types of web pages (issue 3, *lists*; issue 4, *directories*; and issue 5, *tabular text*) do not fit well into traditional text types.

Conclusions (II)

- One important fact that should be taken into account is the mixed nature of a text (issue 6, *mixed text*). A text can be a mixture of different forms of expressions and different communicative purposes; it rarely corresponds to a single text type. This is especially true for Web pages, which are visual objects, mostly with a non-linear organization of the text(s), and several purposes or functions.

Conclusions (III)

- It seems that creating a corpus from the Web entails several issues, because web pages bring about a number of novelties.
- What if we wish to encode web pages in a national corpus? Are we going to keep all the navigational elements together with all other non-linear items? Shall we keep the HTML coding? Shall we keep the visuality of a web page?

I leave all these issues open to your suggestions and hints.

Thank you for your attention!