Analysis of the Obama Speeches Corpus using XSLT and XQuery

Sample Analysis by Franziska Horn

1 Introduction

US president Barack Obama is well known for his rhetorical skills. The CNN journalist Stephanie Holmes (2008), for instance, pinpoints his "ability to captivate and inspire audiences with his powerful speeches". In the following his speeches collected in a corpus are analysed using XSLT and XQuery[1]. XSLT can be applied to transform XML into Html, plain text or another XML-document. The frequency of elements in an XML-document, for instance, the number of words or sentences can be calculated and presented, e.g. as html-output, with XQuery.

The aim of this article is to emphasise the possibilities provided by XSLT and XQuery for transformation and analysis of language data in XML format without focussing on the concrete realization by the respective scripts.

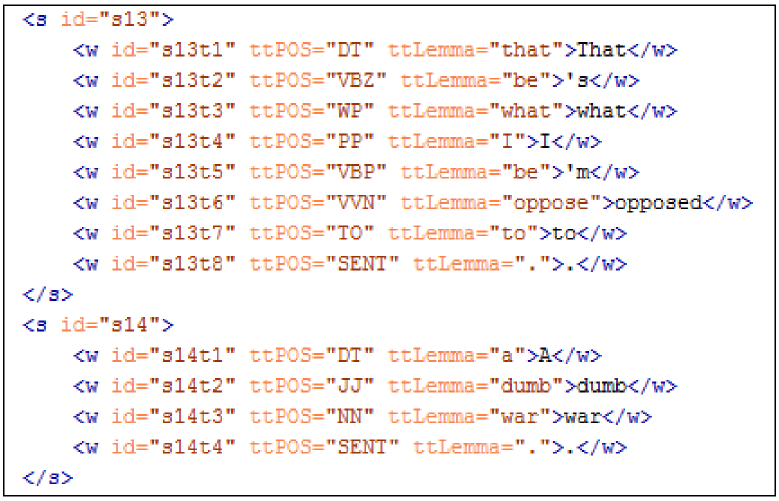
2 Obama Speeches Corpus – Collection and Preprocessing

The Obama Speeches Corpus is a collection of 94 speeches given by Barack Obama. The corpus contains, for instance, his *Inaugural Speech* (2009), the *Election Night Victory Speech* (2008) or his speech about the Patriot Act (2005).

We receive the texts from the Obama Speeches website. In a next step, the speeches were segmented into sentences and tokenized. In the following, the TreeTagger was applied for part-of-speech tagging and lemmatization. The output format was XML structured conform to the TEI (Text Encoding Initiative) guidelines for language data. These preprocessing steps were organized and performed by a tool developed by Sabine Bartsch and David O' Sullivan.

The advantage of XML as output format is its sustainability and the further options for analysis and transformation using XSLT and XQuery. Figure 1 shows a section of the xml-file to Obama's *Speech Against the Iraq War*.

[1] The w3schools tutorials can be recommended as introduction to XML, XSLT and XQuery.



```
<s id="s13">
    <w id="s13t1" ttPOS="DT" ttLemma="that">That</w>
    <w id="s13t2" ttPOS="VBZ" ttLemma="be">'s</w>
    <w id="s13t3" ttPOS="WP" ttLemma="what">what</w>
    <w id="s13t4" ttPOS="PP" ttLemma="I">I</w>
    <w id="s13t5" ttPOS="VBP" ttLemma="be">'m</w>
    <w id="s13t6" ttPOS="VVN" ttLemma="oppose">opposed</w>
    <w id="s13t7" ttPOS="TO" ttLemma="to">to</w>
    <w id="s13t8" ttPOS="SENT" ttLemma=".">.</w>
</s>
<s id="s14">
    <w id="s14t1" ttPOS="DT" ttLemma="a">A</w>
    <w id="s14t2" ttPOS="JJ" ttLemma="dumb">dumb</w>
    <w id="s14t3" ttPOS="NN" ttLemma="war">war</w>
    <w id="s14t4" ttPOS="SENT" ttLemma=".">.</w>
</s>
```

Figure 1 Example of the XML-file

The tag <*s*> means sentence, the attribute *id* represents the number of the sentence. The tag <*w*> refers to word and the attribute *id* represents the respective position of the word in the sentence. The attributes *ttPOS* and *ttLemma* are the corresponding pos-tags and lemmas for each word.

3 Analysis

3.1 Type/Token Ratio

The corpus contains 202 392 running words and the number of types is 12 256. The type/token ratio is about 16.5 per cent.

Types	202392
Tokens	12256
Type/Token Ratio	16.5

Table 1 TTR of the corpus

The type/token ratio of the whole corpus differs clearly from the ones calculated for each single speech. Here are some examples:

Speech	**Types**	**Tokens**	**Type/Token Ratio**
Death of Rosa Parks (2005)	446	982	45,42
Literacy and Education in a 21st-Century Economy (2005)	930	3090	30,10
Election Night Victory Speech (2008)	709	2261	31,36

Table 2 TTR of single speeches

An overview of the type/token ratio of all single speeches in the corpus is given in the end of this document. The list shows slight differences in the vocabulary variation between the texts. The majority of speeches has a type/token ratio between 28 and 40 per cent – clearly higher than the type/token ratio of the whole corpus. This difference might be based on the repetition of topics, for instance the Iraq War and his presidential campaign, in Obama's speeches and in consequence, the repetition of words belonging to those recurring topics.

3.2 POS-Tags

Table 3 shows the distribution of the different pos-tags within the corpus. The tag set used by the TreeTagger is explained on this website.

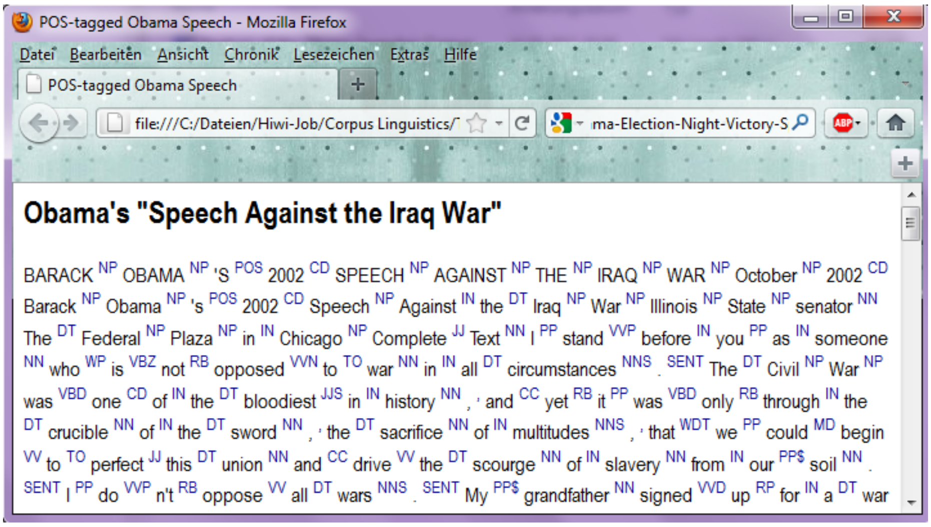
POS-Tag	Occurrences	POS-Tag	Occurrences
NN	25983	**VB**	976
IN	19543	**VBD**	956
DT	18439	**RP**	925
PP	11702	**JJR**	806
JJ	11695	**VHZ**	640
NNS	11196	**NPS**	635
RB	9627	**POS**	634
NP	8611	**RBR**	476
VV	8216	**EX**	463
CC	8019	**VBN**	462
,	7985	**VH**	422
SENT	7950	**JJS**	393
TO	5962	**VHD**	298
PP$	3726	**PDT**	207
MD	3581	**RBS**	173
VVP	3574	**SYM**	140
VVN	3279	**$**	122
VVG	3153	**VBG**	116
VBZ	2851	**UH**	60
VVD	2652	**FW**	49
IN/that	2281	**VHG**	35
:	2237	**WP$**	29
CD	1995	**VHN**	16
VBP	1734	**"**	14
WP	1616	**)**	11
VVZ	1586	**(**	8
WDT	1456	**``**	3
WRB	1420	**LS**	2
VHP	1221		

Table 3 Distribution of the pos-tags

An interesting observation is that nouns (NN for singular, NNS for plural) account for more than 20 per cent of the whole corpus. This wide distribution can be explained by the formal character of the speeches: they are usually prepared in advance in form of a monolog, thus rules of dialog ad-hoc conversations do not apply. Therefore, the frequency of nouns can be seen as a feature indicating the complexity of speeches.

As opposed to nouns, list markers (LS) only occur twice within the whole corpus. The text type can also be seen as a reason for this observation. Speeches are given in front of an audience. They are not written to be published primarily in a newspaper or online. List markers, e.g. (1), are more useful in written texts. They are used to present and order facts clearly for the reader, for instance, in academic writing. Therefore, list markers occur extremely rare in speeches.

The distribution of pos-tags within Obama's *Speech Against the Iraq War* hold 2002 can be visualised in an html-document generated using XSLT. Figure 2 presents an extract of the visualisation.


POS-tagged Obama Speech - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
POS-tagged Obama Speech
file:///C:/Dateien/Hiwi-Job/Corpus Linguistics/
ma-Election-Night-Victory-S

Obama's "Speech Against the Iraq War"

BARACK NP OBAMA NP 'S POS 2002 CD SPEECH NP AGAINST NP THE NP IRAQ NP WAR NP October NP 2002 CD Barack NP Obama NP 's POS 2002 CD Speech NP Against IN the DT Iraq NP War NP Illinois NP State NP senator NN The DT Federal NP Plaza NP in IN Chicago NP Complete JJ Text NN I PP stand VVP before IN you PP as IN someone NN who WP is VBZ not RB opposed VVN to TO war NN in IN all DT circumstances NNS . SENT The DT Civil NP War NP was VBD one CD of IN the DT bloodiest JJS in IN history NN , , and CC yet RB it PP was VBD only RB through IN the DT crucible NN of IN the DT sword NN , , the DT sacrifice NN of IN multitudes NNS , , that WDT we PP could MD begin VV to TO perfect JJ this DT union NN and CC drive VV the DT scourge NN of IN slavery NN from IN our PP$ soil NN . SENT I PP do VVP n't RB oppose VV all DT wars NNS . SENT My PP$ grandfather NN signed VVD up RP for IN a DT war

Figure 2 Extract POS-tagged Obama speech

3.3 Lemmas

The ten most frequent lemmas are displayed in table 4. The high number of functional words can be noticed. The first lemma, which refers to a noun, is *people* at position 45. This lemma occurs “only” 579 times.

Lemma	Occurrences
the	9002
,	7985
.	7634
be	7095
and	6146
to	5906
of	4814
that	4255
a	3817
in	3374

Table 4 Ten most frequent lemmas

In addition to frequently distributed words it can be interesting for the investigation which lemmas occur less frequent. 3796 hapax legominas account for only 2 per cent of the whole corpus. Examples are the lemmas *emancipation* and *beverage*.

3.4 Combined Queries

More complex queries that combine the search for pos-tags and lemmas are also possible. One example is shown in table 5. It presents the 5 most frequent personal pronouns in the corpus.

Personal Pronoun	Occurrences
we	2978
I	2044
it	1863
you	1336
they	1141

Table 5 Five most frequent pronouns

Overall, 12 428 personal pronouns can be found in the corpus. An interesting observation is that about a quarter of them are 1st person plural ones. This high number can be explained by Obama's intention to form an entity of speaker and audience. In addition, the pronoun can be used to underline his belonging to the Democratic Party and its goals and convictions.

The 1st singular pronoun is also relatively frequent – about every 10th personal pronoun is one. The occurrences can be understood as a mean of Obama to emphasise his position and to distance himself from political opponents.

4 Conclusion

As we have seen XSLT and XQuery can be applied to generate basic corpus statistics. Examples are the number of types, tokens and the type/token ratio as well as the distribution of lemmas and pos-tags.

The analysis demonstrates how useful XQuery scripts can be for counting units in a single text or in whole corpus and for presenting the results sorted. Furthermore, calculations for linguistic measures, for instance, the type/token ratio are possible. XSLT was used to visualise one of Obama's Speeches part-of-speech-annotated by generating an html-document.

Of course the results generated by XSLT and XQuery are based on the annotations produced during the preprocessing. As a consequence, it is necessary to know basically how tools, e.g. the tokenizer, lemmatizer or part-of-speech tagger work and how reliable their results have to be estimated. This knowledge facilitates the recognition of possible annotation problems and to interpret the results generated by XSLT and XQuery correctly.

Besides the numerous possibilities for transforming and analysing a single text or a whole corpus the advantage of both XSLT and XQuery is the reuse of scripts and the relatively fast generation of results even if your corpus contains a large amount of Obama Speeches.

5 References

Holmes, Stephanie (2008): Obama: Oratory and originality. (May 17th 2011, 7:00pm). http://news.bbc.co.uk/2/hi/americas/7735014.stm.

TEI Consortium (2011), eds. *Guidelines for Electronic Text Encoding and Interchange*. (June 20th 2011, 10:00pm). http://www.tei-c.org/P5/.

Obama Speeches (2011), (June 20th 2011, 10:00pm). http://obamaspeeches.com/.

W3school (2011). XML Tutorial. (May 17th 2011, 06:00pm). http://www.w3schools.com/xml/default.asp.

W3school (2011). XSLT Tutorial. (May 17th 2011, 6:00pm). http://www.w3schools.com/xsl/default.asp.

W3school (2011). XQuery Tutorial. (May 17th 2011, 6:00pm). http://www.w3schools.com/xquery/default.asp.

6 Type/Token Ratio of each document

Title of the Speech	Types	Tokens	Type/Token Ratio in %
001-2002-Speech-Against-the-Iraq-War-Obama-Speech.htm.txt	415	948	43,78
011-Amendment-to-Provide-Meals-Phone-Service-to-Wounded-Veterans-Obama-Speech.htm.txt	383	940	40,74
012-Abraham-Lincoln-Presidential-Library-and-Museum-Obama-Speech.htm.txt	368	832	44,23
013-SIUC-College-of-Agriculture-50th-Anniversary-Obama-Speech.htm.txt	649	1978	32,81
014-National-Press-Club-Obama-Speech.htm.txt	820	2704	30,33
015-NAACP-Fight-for-Freedom-Fund-Dinner-Obama-Speech.htm.txt	858	2891	29,68
016-Rockford-Register-Star-Young-American-Awards-Obama-Speech.htm.txt	459	1230	37,32
017-America-Nuclear-Non-Proliferation-Policy-Remarks-Obama-Speech.htm.txt	684	1923	35,57
018-Abraham-Lincoln-National-Cemetery-Obama-Speech.htm.txt	610	1567	38,93
021-Nomination-of-Justice-Janice-Rogers-Brown-Obama-Speech.htm.txt	913	3255	28,05
022-Pritzker-School-of-Medicine-Commencement-Obama-Speech.htm.txt	1007	3022	33,32
024-Literacy-and-Education-in-a-21st-Century-Economy-Obama-Speech.htm.txt	930	3090	30,10
025-American-Legion-Conference-Obama-Speech.htm.txt	637	1790	35,59
026-Foreign-Operations-Appropriations-Bill-and-the-Avian-Flu-Obama-Speech.htm.txt	509	1180	43,14
028-AFL-CIO-National-Convention-Obama-Speech.htm.txt	851	2642	32,21
029-Statement-on-Hurricane-Katrina-Relief-Efforts-Obama-Speech.htm.txt	462	1085	42,58
030-Resources-for-the-Future-Obama-Speech.htm.txt	963	3158	30,49
031-Confirmation-of-Judge-John-Roberts-Obama-Speech.htm.txt	565	1609	35,11
032-Avian-Flu-Obama-Speech.htm.txt	460	1086	42,36
033-Teaching-Our-Kids-in-a-21st-Century-Economy-Obama-Speech.htm.txt	1090	4122	26,44
034-Death-of-Rosa-Parks-Obama-Speech.htm.txt	446	982	45,42
035-Chicago-White-Sox-Obama-Speech.htm.txt	401	890	45,06
036-Non-Proliferation-and-Russia-The-Challenges-Ahead-Obama-Speech.htm.txt	784	2201	35,62
037-Sex-on-TV-4-Report-Obama-Speech.htm.txt	769	2403	32,00
038-National-Womens-Law-Center-Obama-Speech.htm.txt	811	2792	29,05
039-Robert-F-Kennedy-Human-Rights-Award-Ceremony-Obama-Speech.htm.txt	814	2210	36,83
040-Moving-Forward-in-Iraq-Chicago-Council-on-Foreign-Relations-Obama-Speech.htm.txt	1309	4783	27,37
041-The-PATRIOT-Act-Obama-Speech.htm.txt	408	945	43,17
042-From-the-Road-Speaking-with-American-Troops-in-Iraq-Obama-Speech-Podcast.htm.txt	725	2432	29,81
043-From-the-Road-Israel-and-the-Palestinian-territories-Obama-Podcast.htm.txt	523	1272	41,12
044-Remarks-Honest-Leadership-and-Open-Government-Obama-Podcast.htm.txt	488	1199	40,70
045-Meeting-on-Iraq-with-President-Bush-Obama-Speech.htm.txt	367	904	40,60
046-Confirmation-of-Judge-Samuel-Alito-Jr-Obama-Speech.htm.txt	301	677	44,46
047-Lobbying-Reform-Summit-National-Press-Club-Obama-Speech.htm.txt	1031	2845	36,24

048-Supreme-Court-Nomination-of-Samuel-Alito-Obama-Podcast.htm.txt	261	620	42,10
049-Hurricane-Katrina-Child-Assistance-Amendment-Obama-Speech.htm.txt	380	976	38,93
050-Foreign-Relations-Committee-Lugar-Obama-legislation-S1949-Obama-Speech.htm.txt	264	526	50,19
052-Darfur-Current-Policy-Not-Enough-Obama-Speech.htm.txt	461	1246	37,00
053-Floor-Statement-S2271-PATRIOT-Act-Reauthorization-Obama-Speech.htm.txt	359	795	45,16
054-Energy-Security-is-National-Security-Governors-Ethanol-Coalition-Obama-Speech.htm.txt	1015	3355	30,25
055-Debate-on-Ethics-Reform-Obama-Speech.htm.txt	927	2370	39,11
056-Lobbying-Reform-Meals-Amendment-Obama-Speech.htm.txt	322	788	40,86
057-21st-Century-Schools-for-a-21st-Century-Economy-Obama-Speech.htm.txt	990	3512	28,19
059-Improving-Chemical-Plant-Security-Obama-Podcast.htm.txt	432	957	45,14
060-Energy-Independence-and-the-Safety-of-Our-Planet-Obama-Speech.htm.txt	1182	3938	30,02
061-Immigration-Reform-Obama-Speech.htm.txt	586	1676	34,96
062-Southern-Illinois-University-School-of-Medicine-Commencement-Obama-Speech.htm.txt	712	2267	31,41
063-Updates-on-Darfur-Immigration-Gas-Prices-Obama-Podcast.htm.txt	583	1659	35,14
064-Amendment-to-Stop-No-Bid-Contracts-for-Gulf-Coast-Reconstruction-Obama-Speech.htm.txt	507	1326	38,24
065-Immigration-Rallies-Obama-Podcast.htm.txt	635	1644	38,63
066-A-Real-Solution-for-High-Gas-Prices-Obama-Speech.htm.txt	490	1202	40,77
067-EMILYS-List-Annual-Luncheon-Obama-Speech.htm.txt	951	3533	26,92
068-Honoring-Our-Commitment-to-Veterans-Obama-Speech.htm.txt	485	1538	31,53
071-Employment-Verification-Amendment-for-the-Immigration-Bill-Obama-Speech.htm.txt	334	723	46,20
072-Opposition-to-the-Amendment-Requiring-a-Photo-ID-to-Vote-Obama-Speech.htm.txt	405	925	43,78
073-General-Michael-Hayden-Nomination-Obama-Speech.htm.txt	406	939	43,24
074-University-of-Massachusetts-at-Boston-Commencement-Address-Obama-Speech.htm.txt	1027	3546	28,96
075-Federal-Marriage-Amendment-Obama-Speech.htm.txt	320	784	40,82
076-Network-Neutrality-Obama-Podcast.htm.txt	293	625	46,88
077-Take-Back-America-Obama-Speech.htm.txt	985	3594	27,41
078-Katrina-Reconstruction-Obama-Podcast.htm.txt	393	1005	39,10
079-Northwestern-University-Commencement-Address-Obama-Speech.htm.txt	1033	3523	29,32
080-Iraq-Debate-Obama-Speech.htm.txt	515	1273	40,46
081-Call-to-Renewal-Keynote-Address-Obama-Speech.htm.txt	1460	5194	28,11
082-Campus-Progress-Annual-Conference-Obama-Speech.htm.txt	662	2019	32,79
083-Statement-of-Support-for-Stem-Cell-Research-Obama-Speech.htm.txt	663	1546	42,88
084-Support-of-HR-9-the-Voting-Rights-Act-Obama-Speech.htm.txt	673	1716	39,22
085-Vote-against-the-Gulf-of-Mexico-Energy-Bill-Obama-Speech.htm.txt	587	1507	38,95
086-AFSCME-National-Convention-Obama-Speech.htm.txt	736	2390	30,79
087-Xavier-University-Commencement-Address-Obama-Speech.htm.txt	885	2998	29,52
088-An-Honest-Government-A-Hopeful-Future-Obama-Speech.htm.txt	1109	3614	30,69
089-Energy-Independence-A-Call-for-Leadership-Obama-Speech.htm.txt	760	2030	37,44

091-Floor-Statement-on-the-Habeas-Corpus-Amendment-Obama-Speech.htm.txt	754	2571	29,33
092-Military-Commission-Legislation-Obama-Speech.htm.txt	553	1522	36,33
093-Martin-Luther-King-Memorial-Groundbreaking-Ceremony-Obama-Speech.htm.txt	404	948	42,62
094-A-Way-Forward-in-Iraq-Obama-Speech.htm.txt	1307	4779	27,35
095-Race-Against-Time-World-AIDS-Day-Speech-Obama-Speech.htm.txt	1269	4165	30,47
096-Floor-Statement-on-Presidents-Decision-to-Increase-Troops-in-Iraq-Obama-Speech.htm.txt	625	1865	33,51
097-The-Time-Has-Come-for-Universal-Health-Care-Obama-Speech.htm.txt	909	3010	30,20
099-Announcement-For-President-Springfield-Illinois-Obama-Speech.htm.txt	898	3014	29,79
120-Obama-Turn-The-Page-Speech-California-Deomcratic-Convention.htm.txt	874	3179	27,49
E-Barack-Obama-Speech-Manassas-Virgina-Last-Rally-2008-Election.htm.txt	614	2224	27,61
E01-Barack-Obama-Iowa-Caucus-Night-Des-Moines-Iowa-January-03-2008.htm.txt	501	1510	33,18
E02-Barack-Obama-Super-Tuesday-Chicago-IL-February-5-2008.htm.txt	601	1916	31,37
E03-Barack-Obama-Potomac-Primary-Night-Madison-WI-February-12-2008.htm.txt	650	2164	30,04
E04-Barack-Obama-March-4-Primary-Night-Texas-and-Ohio-San-Antonio-TX-March-4-2008.htm.txt	666	2041	32,63
E05-Barack-Obama-A-More-Perfect-Union-the-Race-Speech-Philadelphia-PA-March-18-2008.htm.txt	1442	5489	26,27
E06-Barack-Obama-AP-Annual-Luncheon-Washington-DC-April-14-2008-religion-guns-pennsylvania.htm.txt	753	2525	29,82
E07-Barack-Obama-Pennsylvania-Primary-Night-Evansville-Indiana-April-22-2008.htm.txt	656	2282	28,75
E08-Barack-Obama-North-Carolina-Primary-Night-Raleigh-NC-May-6-2008.htm.txt	760	2658	28,59
E09-Barack-Obama-Final-Primary-Night-Presumptive-Democratic-Nominee-Speech-St-Paul-Minnesota-June-3-2008.htm.txt	840	2757	30,47
E10-Barack-Obama-The-American-Promise-Acceptance-Speech-at-the-Democratic-Convention-Mile-High-Stadium--Denver-Colorado-August-28-2008.htm.txt	1267	5316	23,83
E11-Barack-Obama-Election-Night-Victory-Speech-Grant-Park-Illinois-November-4-2008.htm.txt	709	2261	31,36
P-Obama-Inaugural-Speech-Inauguration.htm.txt	949	2730	34,76