Orthographic Errors in Web Pages: Toward Cleaner Web Corpora

Christoph Ringlstetter*
Klaus U. Schulz*
CIS, University of Munich

Stoyan Mihov†
Bulgarian Academy of Science, Sofia

Since the Web by far represents the largest public repository of natural language texts, recent experiments, methods, and tools in the area of corpus linguistics often use the Web as a corpus. For applications where high accuracy is crucial, the problem has to be faced that a non-negligible number of orthographic and grammatical errors occur in Web documents. In this article we investigate the distribution of orthographic errors of various types in Web pages. As a by-product, methods are developed for efficiently detecting erroneous pages and for marking orthographic errors in acceptable Web documents, reducing thus the number of errors in corpora and linguistic knowledge bases automatically retrieved from the Web.

1. Introduction

The automated analysis of large corpora has many useful applications (Church and Mercer 1993). Suitable language repositories can be used for deriving models of a given natural language, as needed for speech recognition (Ostendorf, Digalakis, and Kimball 1996; Jelinek 1997; Chelba and Jelinek 2002), language generation (Oh and Rudickny 2000), and text correction (Kukich 1992; Amengual and Vidal 1998; Strohmaier et al. 2003b). Other corpus-based methods determine associations between words (Grefenstette 1992; Dunning 1993; Lin et al. 1998), which yields a basis for computing thesauri, or dictionaries of terminological expressions and multiword lexemes (Gaizauskas, Demetriou, and Humphreys 2000; Grefenstette 2001).

From multilingual texts, translation lexica can be generated (Gale and Church 1991; Kupiec 1993; Kumano and Hirakawa 1994; Boutsis, Piperidis, and Demiros 1999; Grefenstette 1999). The analysis of technical texts is used to automatically build dictionaries of acronyms for a given field (Taghva and Gilbreth 1999; Yeates, Bainbridge, and Witten 2000), and related methods help to compute dictionaries that cover the special vocabulary of a given thematic area (Strohmaier et al. 2003a). In computer-assisted language learning (CALL), mining techniques for corpora are used to create individualized and user-centric exercises for grammar and text understanding (Schwartz, Aikawa, and Pahud 2004; Brown and Eskenazi 2004; Fletcher 2004a).

By Zipf's law, most words, phrases, and specific grammatical constructions have a very low frequency. Furthermore, the number of text genres and special thematic

* Funded by German Research Foundation (DFG)
† Funded by VolkswagenStiftung

Submission received: 21 January 2005; revised submission received: 3 August 2005; accepted for publication: 10 December 2005.

areas that come with their own picture of language is large. This explains that most of the aforementioned applications can only work when built on top of huge heterogeneous corpora. Since the Web represents by far the largest public repository for natural language texts, and since Web search engines such as Google offer simple access to pages where language material of a given orthographic, grammatical, or thematic kind is found, many recent experiments and technologies use the Web as a corpus (Kehoe and Renouf 2002; Morley, Renouf, and Kehoe 2003; Kilgarriff and Grefenstette 2003; Resnik and Smith 2003; Way and Gough 2003; Fletcher 2004b).

One potential problem for Web-based corpus linguistics is caused by the fact that words and phrases occurring in Web pages are sometimes erroneous. Typing errors represent one widespread phenomenon. Many Web pages, say, in English, are written by non-native speakers, or by persons with very modest language competence. As a consequence, spelling errors and grammatical bugs result. The character sets that are used for writing Web pages are often not fully adequate for the alphabet of a given language, which represents another systematic source for inaccuracies. Furthermore, a small number of texts found in the Web is obtained via optical character recognition (OCR), which may again lead to garbled words. As a consequence of these and other error sources, the Web contains a considerable number of "bad" pages with language material that is inappropriate for corpus construction.

In one way or the other, all the aforementioned applications are affected by these inadequacies. While the problem is probably not too serious for approaches that merely collect statistical information about given language items, the construction of dictionaries and related linguistic knowledge bases—which are, after all, meant to be used in different scenarios of automated language processing—becomes problematic if too many erroneous entries are retrieved from Web pages. Obviously, in computer-assisted language learning it is a principal concern that words and phrases from the Web that are presented to the user are error free.

In discussions we found that problems resulting from erroneous language material in Web pages for distinct applications are broadly acknowledged (see also Section 4.4 of Kilgarriff and Grefenstette [2003]). Still, to the best of our knowledge, a serious analysis of the frequency and distribution of orthographic errors in the Web is missing, and no general methods have been developed that help to detect and exclude pages with too many erroneous words. In this article we first report on a series of experiments that try to answer the following questions:

1. What are important types of orthographic errors found in Web pages?
2. How frequent are errors of a given kind? For a given error level (percentage of erroneous tokens) τ, which percentage of Web pages exceeds error level τ?
3. How do these figures depend on the language, on the thematic area, and on the genre of the Web pages that are considered? How do these figures depend on the document format of the Web pages that are considered?

We then look at the problem indicated above.

4. Which methods help to automatically detect Web pages with many orthographic errors?
   Which methods help to mark orthographic errors found in Web pages?

To answer questions 1–3, we retrieved and analyzed a collection of large English and German corpora from the Web, using suitable queries to Web search engines. In our error statistics we wanted to distinguish between (1) "general" Web pages collected without any specific thematic focus on the one hand and Web pages from specific thematic areas on the other hand, and (2) between Web pages written in HTML and Web documents written in PDF. To cover the first difference, for both languages we retrieved two general corpora as well as a number of corpora for specific thematic areas. All these corpora only contain HTML pages. A parallel series of general corpora was collected that are composed of PDF documents. Details are provided in Section 2.

Special Vocabulary. Web pages often contain tokens that do not belong to the standard vocabulary of the respective language. Typical categories are, for example, special names, slang, archaic language, expressions from foreign languages, and special expressions from computer science/programming. Classification and detection of special vocabulary is outside the scope of the present article. Since sometimes a clear separation between special vocabulary and errors is difficult, we briefly come back to this problem in Section 5.4.

Proper Errors. Focusing on garbled standard vocabulary, tokens may be seriously damaged in an "unexplainable" way. Most of the remaining errors can be assigned to one of the four classes mentioned above:

- typing errors (i.e., errors caused by a confusion of keys when typing a document),
- spelling errors ("cognitive" errors resulting from insufficient language competence),
- errors resulting from inadequate character encoding, and
- OCR errors.

In order to estimate the number of errors of a given kind in the corpora, special **error dictionaries** were built. These dictionaries, which only list garbled words of a given language that do not accidentally represent correct words, try to cover a high number of the conventional errors of each type that are typically found in Web pages and other documents. Section 3 motivates the use of error dictionaries for error detection. Details of the construction of the error dictionaries are discussed in Section 4.

In Section 5, we estimate the number of orthographic errors in the corpora that remain undetected because they do not occur in the error dictionaries. We also estimate the percentage of correct tokens of the corpora that are erroneously treated as errors since they appear in the error dictionaries. Our results show that the number of tokens of a text that appear in the error dictionaries can be considered as a lower approximation of the number of real orthographic errors.

In Section 6, we describe the distribution of orthographic errors of the types distinguished above in the general test corpora, counting occurrences of entries of the error dictionaries. Section 7 summarizes the most important differences that arise when using PDF corpora, or corpora for special thematic areas. Section 8 presents various

results that illuminate the relationship between the error rate of a document and its genre.

In our experiments we observed in all corpora a rich spectrum of error rates, ranging from perfect documents to a small number of clearly unacceptable pages. This motivates the design of filters that efficiently recognize and reject pages with an error rate beyond a user-specified threshold. The construction of appropriate filters is described in Section 9, where we also demonstrate the effect of using these filters, comparing the figures obtained in Section 6 with the corresponding figures for filtered corpora. Filters work surprisingly well due to a Zipf-like distribution of error frequencies in Web pages.

In Section 10, we present two experiments that exemplify how the methods developed in the article may in fact help to improve corpus-based methods. The general question of *how deeply* distinct methods from computational linguistics based on Web corpora are affected by orthographic errors in Web pages and to what extent the methods developed in the article help to remedy these deficiencies are too complex to be discussed here.

The main insights and contributions are summarized in the Conclusion (Section 11) where we also comment on future work and on some practical difficulties one has to face when collecting and analyzing large corpora from the Web.

2. Corpora

The basis for the evaluations described below is a collection of corpora, each composed of Web pages retrieved with Web search engines (Google/AllTheWeb). In order to study how specific features of a language might influence the distribution of orthographic errors, all corpora were built up in two variants. The English and German variant, respectively, contain Web pages that were classified as English and German Web pages by the search engine. As described above, for both languages we collected general corpora with Web pages *without any thematic focus* and, in addition, corpora that cover five specific thematic areas to be described below. Statements on the "representativeness" of corpora derived from the Web are notoriously difficult. The composition of corpora retrieved with Web search engines depends on the kind of queries that are used, on the ranking mechanisms of the engine, and on the details of the collection strategy. We mainly concentrated on simple queries and straightforward collection strategies. Still, the large number of subcorpora and pages that were evaluated should guarantee that accidental results are avoided.

2.1 General Web Corpora

In a first attempt, we tried to obtain a general German HTML corpus using the meaningless query der die das, i.e., the three German definite articles. However, queries of this and a similar form did not lead to satisfactory results: As a consequence of Google's ranking mechanism, which prefers "authorities" (Brin and Page 1998), mainly portals of big organizations, companies, and others were retrieved. These pages are often dominated by graphical elements. Portions of text are usually small and carefully edited, which means that orthographic errors are less frequent than in other "less official" pages.

To achieve a more realistic scenario we randomly generated quintuples, each collecting five terms of the 10,000 top frequent German words. We used Google to retrieve

10 pages per query (quintuple) until we obtained 1,000 pages. A considerable number of the URLs were found to be inactive. After conversion to ASCII and a preliminary analysis of error rates with methods described below, some of the remaining pages were found to contain very large lists of general keywords, including many orthographic errors. Apparently these lists and errors were only added to improve the ranking of the page in search engines, even for ill-formed queries. We excluded these pages. The remaining documents represent the "primary" general German HTML corpus. Since we wanted to know how results depend on the peculiarities of the selected set of pages, a second series of queries of the same type was sent to Google to retrieve a "secondary" general German HTML corpus with a completely disjoint set of pages.

Similar procedures were used to obtain a primary and a secondary general English HTML corpus, a general German PDF corpus, and a general English PDF corpus. The translation from PDF to ASCII was found to be error prone, in particular for German documents (cf. Gartner 2003). Due to this process, some converted PDF documents were seriously damaged. Since we focus on errors in original Web pages (as opposed to converted versions of such pages), these files were excluded as well. We found these pages when computing error rates based on error dictionaries as described in Sections 6 and 7.

The number of Web pages and the number of normal tokens (i.e., tokens composed of standard letters only) in the resulting six corpora are shown in Table 1. Numbers (1) and (2) stand for the primary and secondary corpora, respectively.

2.2 Web Corpora for Specific Thematic Areas

We looked at the thematic areas "Middle Ages," "Holocaust," "Fish," "Mushrooms," and "Neurology." The given selection of topics tries to cover scientific areas as well as history and hobbies.

Simple Crawl. A first series of corpora was collected by sending a query with 25 "terminological" keywords mechanically found in a small corpus of the given area to the AllTheWeb search engine and collecting the answer set. For example, the queries

> *mushrooms mushroom pine edible harvesting morels harvested harvesters dried chanterelle matsutake poisonous flavor chanterelles caps fungi drying stuffing humidity varieties boletes recipes spores conifers pickers*

Table 1
Number of Web pages, number of normal tokens (tokens composed of standard letters only), and sizes in megabytes of the "general" corpora. Numbers (1) and (2) refer to primary and secondary corpora, respectively.

General corpora	Web pages	Normal tokens	Size (MB)
English HTML (1)	829	7,900,337	157
English HTML (2)	929	7,152,783	188
German HTML (1)	618	9,525,484	189
German HTML (2)	857	11,539,035	284
English PDF	570	2,193,598	393
German PDF	603	1,528,914	240

disorder disorders anxiety self hallucinations delusions anatomy cortex delusion neuroscience disturbance conscious psychotic stimulus hallucination unconsciously receptors cognitive psychoanalytic unconscious consciously stimuli ego schizophrenia impairment

were respectively used for collecting the corpora Mushrooms E and Neurology E. The ranking mechanism of AllTheWeb prefers pages containing hits for several keywords of a disjunctive query. Since this form of corpus construction is straightforward, not all pages in the resulting corpora belong to the given thematic area.

Refined Crawl. We wanted to see how results are affected when using less naive crawling methods. For the three areas "Fish," "Mushrooms," and "Neurology," the secondary corpora were retrieved using the following refined procedure: Starting from a small tagged seed corpus for the given domain, we mechanically extracted terminological open compounds for English (Sornlertlamvanich and Tanaka 1996; Smadja and McKeown 1990) and compound nouns for German. Examples are *amino group, action potential, defense mechanism* (English, neurology), *truffle species, morel areas, harvesting tips* (English, mushrooms), *Koffeinstoffwechsel,* and *Eisenkonzentration* (German, neurology). Each of these expressions was sent as a query to Google. From each answer set we collected a maximum of 30 top-ranked hits (many answer sets were smaller). For each document in the resulting corpus, the similarity with the seed corpus was controlled, using a cosine measure (in practice, almost all documents passed the similarity filter). Our method can be considered as a variant of Baroni and Bernardini's (2004) and leads to corpora with a strong thematic focus.

The statistics for all thematic corpora are summarized in Table 2. Numbers (1) and (2) stand for corpora crawled with the simple and the refined crawling strategy, respectively. The numbers indicate one interesting effect: Documents in the thematic corpora obtained with the refined crawling strategy turned out to be typically rather short. Since we only used the 30 top-ranked documents for each single query, this probably points to a special feature of Google's ranking mechanism. A manual inspection of hundreds of documents for both the simple and the refined crawl did not lead to additional insights.

3. Error Detection

For detecting orthographic errors of a particular type in texts, two naive base methods may be distinguished.

1. A representative list of errors of the respective type is created and manually checked. Each token of the text appearing in the list represents an error (**lower approximation**).
2. A spell checker or a large-scale dictionary is used to detect "suspicious" words (error candidates). For each such token *W* we manually check if *W* really represents an error and we determine its type (**upper approximation**).

For large corpora, both methods have serious deficiencies. With Method 1, only a small percentage of all errors is detected. On this basis, it is difficult to estimate the real number of errors. When using Method 2, the number of tokens that have to be manually

Table 2
Selected topics and statistics of English (E) and German (G) corpora for specific thematic areas. Numbers (1) and (2) refer to corpora crawled with the simple and the refined strategy, respectively.

Topic/Language	Web pages	Normal tokens	Size (MB)
Middle Ages E	710	5,069,796	172
Fish E (1)	510	10,090,682	266
Fish E (2)	940	547,407	22
Holocaust E	699	8,797,882	199
Mushrooms E (1)	676	7,876,067	197
Mushrooms E (2)	933	734,337	22
Neurology E (1)	624	8,765,899	197
Neurology E (2)	923	779,699	24
Middle Ages G	614	6,774,794	195
Fish G (1)	655	7,621,579	199
Fish G (2)	804	688,882	32
Holocaust G	616	5,659,924	160
Mushrooms G (1)	527	5,951,305	147
Mushrooms G (2)	614	538,575	28
Neurology G (1)	486	4,322,952	115
Neurology G (2)	323	345,070	12

checked becomes too large. In practice, a large number of error candidates represent correct tokens. This is mainly due to special names and other types of nonstandard vocabulary found in Web pages, as mentioned in the introduction.

We decided to use a third strategy, which can be considered as a synthesis and compromise between the above two approaches. As a starting point, we took standard dictionaries of English, $D(English)$; German, $D(German)$; French, $D(French)$; and Spanish, $D(Spanish)$; and a dictionary of geographic entities, $D(Geos)$; a dictionary of proper names, $D(Names)$; and a dictionary of abbreviations and acronyms, $D(Abbreviations)$.[1] The number of entries in the dictionaries is described in Table 3. The German dictionary contains compound nouns, which explains the large number of entries.

From these standard dictionaries, we derived special error dictionaries that were used in the experiments described later. First, for each of the four error types mentioned above we manually collected a number of general patterns that "explain" possible mutations from correct words to erroneous entries. In a second step, these patterns were used to garble the words of the given background dictionaries. Third, garbled words that were found to correspond to correct words (entries of the above dictionaries) were excluded (**filtering step**). Collecting the remaining erroneous strings, we obtained large error dictionaries for each type of orthographic error.

Experiments described in Section 5 show that our error dictionaries cover the major part of all orthographic errors occurring in the English and German Web pages. At

1 These dictionaries are nonpublic. They have been built up at the Centre for Information and Language Processing (CIS) during the last two decades (Maier-Meyer 1995; Guenthner 1996). Each entry comes with a frequency value that describes the number of occurrences in a 1.5-terabyte subcorpus of the Web from 1999. Dictionaries for French and Spanish were included to improve the filtering step. Suitable dictionaries for other languages were not available.

Table 3
Size of background dictionaries.

Dictionary	Number of entries
D(*English*)	315,300
D(*German*)	2,235,136
D(*French*)	85,895
D(*Spanish*)	69,634
D(*Geos*)	195,700
D(*Names*)	372,628
D(*Abbreviations*)	2,375

the same time, the number of tokens that are *erroneously* treated as errors due to the unavoidable incompleteness of the filtering step remains acceptable. On this basis, an estimate of the number of conventional orthographic errors occurring in Web pages is possible, counting the number of occurrences of entries of the error dictionaries.[2] Before we comment on these points, we describe the construction of the error dictionaries in more detail. In the remainder of the article, by D_{conv} we denote the union of all the conventional dictionaries listed above.

4. Construction of Error Dictionaries

For the construction of error dictionaries, the most important error patterns for each type of error were determined. For typing errors and errors caused by character encoding problems, error patterns were obtained analytically. For spelling errors and Optical Character Recognition (OCR) errors, important mutation patterns were collected empirically. As a general rule, all error dictionaries were restricted to entries of length >4. Many tokens of length ≤ 4 occurring in texts represent acronyms, special names, and abbreviations, and it is difficult to mechanically distinguish between this special kind of vocabulary and errors.

4.1 Error Dictionaries for Typing Errors

Typing errors can be partitioned into transpositions, deletions, substitutions, and insertions. Transpositions of two letters occur if two keys are hit in the wrong order. Deletions result if a key is not properly pushed down. Substitutions occur if a neighbor key is pressed down instead of the intended one. Horizontal and vertical shifts of fingers may be distinguished. If a finger hits the middle between two keys, a neighbor key may be pressed in addition to the intended one. The wrong letter may occur before or after the correct letter.

Transpositions, deletions, substitutions, and insertions cover most of the typing errors discussed in the literature (Kukich 1992). We ignored homologous errors, that is, substitutions that are traced back to a confusion of the left and right hand. Since

2 Note that we do not capture **false friends**, that is, garbled strings that accidentally represent correct words of the dictionary. Detection of false friends is known to be notoriously difficult and outside the scope of this article.

there are many possible positions for both hands, this kind of error leads to large confusion sets.

Since we did not find other patterns in the texts, only mutation variants that are exclusively composed of standard letters (as opposed to digits and other special symbols) were taken into account. Furthermore, since typing errors in general do not affect the first letter of a word,[3] we left this letter unmodified. We analyzed the number of mutated variants of a given word. Both for the American and for the German keyboard we have approximately $16l$ variants for a word of length l. This shows that the above patterns for typing errors are very productive. It is not possible to garble all the words of our background dictionary for constructing the error dictionaries. For the generation of the dictionary of English typing errors, $D_{err}(English,typing)$, we took the 100,000 entries of the English background dictionary with the highest frequency. Applying the above mutation patterns we generated 10,785,675 strings. After removal of duplicates and deletion of words in D_{conv} (filtering step), we obtained $9,427,051$ entries for the dictionary $D_{err}(English,typing)$.

The same procedure was used for creating the dictionary of German typing errors, $D_{err}(German,typing)$. Since the average length of German words is large, we obtained $13,656,866$ entries.

4.2 Error Dictionaries for Spelling Errors

English. In order to find the most characteristic patterns for English spelling errors, a bootstrapping procedure was used to compute an initial list of errors. We started with the misspelled English words *verry, noticable, arguement*, and *inteligence*. For each term we retrieved 20 Web documents. After conversion to ASCII we computed the list of all normal tokens occurring in these documents. The resulting list was sorted by frequency, and words in D_{conv} were filtered out. After a manual selection of new errors with high Google counts, the procedure was iterated until we did not find new erroneous words with high frequency. During the bootstrapping procedure, we also found Web pages that listed some "common misspelled words" of English. The most frequent errors mentioned in these lists were also added. Table 4 presents some strings that were found with a large number of Google hits.[4]

Most of the errors that we found can be traced back to a rule set partially described in Table 5. The full rule set contains 95 rules. We applied each rule to $D(English)$, introducing one error at the first possible position, for each entry of the appropriate form. As a result we obtained a list with 1,223,128 garbled strings. After applying the standard filtering procedure, we obtained the dictionary $D_{err}(English,spell)$ of English spelling errors with 1,202,997 entries.

German. Similarly as for English, we built an initial error list. Bootstrapping was started with the misspelled German terms *nähmlich, addresse, resourcen*, and *vorraus*. Table 6 shows some of the resulting German words, the misspelled variant, and the number of Google hits of the garbled version. From the initial error list, we obtained a set of 65 rules partially described in Table 7. We applied these rules to $D(German)$, introducing one error for each entry of the appropriate form. Each rule was applied to each entry using the first possible position for mutation. For example, for the lexical entry *Adresse* of

3 A phenomenon often discussed in the literature; see, for example, Kukich (1992), page 388f.

4 It is well-known that the number of Google hits for a phrase can vary from one day to the next.

Table 4
Some frequently misspelled English words and the number of Google hits of the correct and misspelled forms.

Word	Google hits	Transformation	Misspelled variant	Google hits
accommodate	5,800,000	mm → m	accomodate	559,000
category	109,000,000	teg → tag	catagory	525,000
definitely	10,800,000	itely → ately	definately	1,270,000
independent	25,700,000	dent → dant	independant	523,000
millennium	10,500,000	nn → n	millenium	2,540,000
occurrence	4,640,000	rr → r	occurence	279,000
receive	57,000,000	ie → ei	recieve	1,260,000
recommend	31,400,000	mm → m	recomend	707,000
separate	26,300,000	ara → era	seperate	1,340,000

the German standard dictionary we obtained the following error terms: *adrese, ahdresse, adrehsse, addresse, adrresse*. As a result we obtained a list with 19,265,271 strings. The large size is mainly caused by the rules for reduplication of consonants, which are not restricted by word context. The filtering procedure led to an error dictionary with 18,970,716 entries.

Table 5
Rule set (incomplete) for the generation of English spelling errors with examples for each transformation class.

Deletion of doubled consonants					
cc	→	c	occasionally	→	ocasionally
nn	→	n	drunkenness	→	drunkeness
Deletion of consecutive consonants					
mn	→	m	column	→	colum
rh	→	r	rhythm	→	rythm
Deletion of doubled vowels					
ee	→	e	exceed	→	exced
uu	→	u	vacuum	→	vacum
Deletion in vowel pair					
aison	→	ason	liaison	→	liason
ou	→	o	mischievous	→	mischievos
ievous	→	evious	mischievous	→	mischevious
Deletion of silent vowels					
?ed	→	?d	maintained	→	maintaind
Substitution of consonants					
sede	↔	cede	supersede	→	supercede
dent	↔	dant	independent	→	independant
Substitution of vowels					
itely	→	ately	definitely	→	definately
teg	→	tag	category	→	catagory
Insertion/reduplication of consonants					
$\kappa \in \{c,d,f,l,n,m,p,r,s,t\}$	→	$\kappa\kappa$	always	→	allways
Transposition of consonants					
ght	→	gth	right	→	rigth
Transposition of vowels					
ie	↔	ei	believe	→	beleive

Table 6
Some frequently misspelled German words and the number of Google hits of the misspelled version.

Word	Google hits	Transformation	Misspelled version	Google hits
Weihnachten	5,450,000	ih → i	Weinachten	99,600
Adresse	8,040,000	d → dd	Addresse	676,000
Videothek	581,000	th → t	Videotek	18,300
Kamera	10,900,000	mm → m	Kammera	14,200
deshalb	8,330,000	s → ss	desshalb	33,900
ziemlich	2,970,000	i → ih	ziehmlich	48,900
ekelig	20,600	lig → lich	ekelich	17,200
nämlich	1,620,000	ä → äh	nähmlich	53,800
Maschine	1,840,000	i → ie	Maschiene	28,300
direkt	18,200,000	ek → eck	direckt	20,600
danach	5,100,000	n → nn	dannach	46,200
voraus	1,960,000	r → rr	vorraus	214,000

4.3 Error Dictionaries for OCR Errors

As a starting point we used a list of typical OCR errors that we found in a corpus with 200 pages of OCR output (Ringlstetter 2003). Error types are shown in Table 8.

Table 7
Rule set (incomplete) for the generation of German spelling errors. The symbol ^t means that t is not the preceding letter.

Deletion of doubled consonants					
dd	→	d	Kuddelmuddel	→	Kudelmuddel
mm	→	m	Kommando	→	Komando
Special rules for deletion of consonants					
mn	→	m	Kolumne	→	Kolume
äh	→	ä	ähnlich	→	änlich
Deletion of vowels					
ie	→	i	ziemlich	→	zimlich
aa	→	a	Aal	→	Al
Substitution of consonants					
nt	→	nd	eigentlich	→	eigendlich
rd	→	rt	Standard	→	Standart
Substitution of vowels					
ä	→	e	Empfänger	→	Empfenger
era	→	ara	Temperatur	→	Temparatur
Insertion/reduplication of consonants					
[aeiouäöü]	↔	[aeiouäöü]h	viel	→	viehl
[aeiouäöü]k	→	[aeiouäöü]ck	direkt	→	direckt
κ ∈ {d,f,l,n,m,p,r,t}	→	κκ	Großbritannien	→	Großbrittannien
^tz	→	tz	Schweiz	→	Schweitz
Insertion of vowels					
i	→	ie	Maschine	→	Maschiene
Shifting					
äu	→	aü	äußerst	→	aüßerst
llel	→	lell	parallel	→	paralell

Table 8
List of typical OCR errors.

Character substitutions	Character merges	Character splits
l → i	rn → m	m → rn
i → l	ri → n	n → ri
g → q	cl → d	ü → ii
o → p		
l → t		
v → y		
y → v		
o → c		
e → c		
l → 1		

English. The error dictionary D_{err}(*English,ocr*) was generated by applying to the entries of D(*English*) the transformation rules listed in Table 8. The transformation of D(*English*) with its 315,300 entries led to a list of 1,697,189 entries. The filtering procedure where we erase words from D_{conv} led to the error dictionary D_{err}(*English, ocr*) with 1,532,741 entries. Table 9 shows some of the most frequent English words, the transformation result, and the number of Google hits of the garbled variant.

German. When scanning German texts, vowels *ä*, *ö*, and *ü* are often replaced by their counterparts *a*, *o*, *u*. However, even more frequently, this kind of replacement occurs as the result of a character encoding problem (see below). Since we wanted to avoid having our statistics for OCR errors being heavily overloaded with errors caused by character encoding problems, we did not add these patterns to the list of typical OCR errors for German texts. This means that we applied to D(*German*) the transformation rules mentioned in Table 8. The transformation of D(*German*) with its 2,235,136 entries led to a list of 11,623,989 strings. After filtering, we obtained the error dictionary D_{err}(*German,ocr*) with 10,608,635 entries. Table 10 shows some frequent German words, the transformation result, and the number of Google hits of the garbled variant.

Table 9
Some members of the top 1,000 most frequent English words transformed by typical OCR error transformations and the number of Google hits of a garbled version.

Word	Transformation	Garbled result	Google hits
company	m→rn	cornpany	1.220
from	m → rn	frorn	5,310
government	rn → m	governrnent	705
many	m → rn	rnany	541
market	m → rn	rnarket	282
more	m → rn	rnore	707
most	m → rn	rnost	1,540
only	y → v	onlv	4,080
said	d → cl	saicl	172
system	m → rn	systern	2,060
time	m → rn	tirne	2,090
will	ll → 11	wi11	3,570

Table 10
Some members of the top 1,000 most frequent German words transformed by typical OCR error transformations and the number of Google hits of a garbled version.

Word	Transformation	Garbled result	Google hits
Dipl-Ing	l → i	Dipi-Ing	213
über	ü → ii	iiber	2,360
vorne	rn → m	vome	1,110
davon	o → p	davpn	96
lager	g → q	laqer	164
ferner	rn → m	femer	841

4.4 Error Dictionaries with Erroneous Character Encoding of German Words

In character sets used for the encoding of Web pages, often the German letters *Ä, Ö, Ü, ä, ö, ü*, and *ß* ("sharp s") are not available. In many of these cases, vocals are replaced, following the substitution scheme (**e-transformation**):

$$\ddot{A} \rightarrow Ae, \ddot{O} \rightarrow Oe, \ddot{U} \rightarrow Ue, \ddot{a} \rightarrow ae, \ddot{o} \rightarrow oe, \ddot{u} \rightarrow ue.$$

In other Web pages, the aforementioned vocals are replaced using the following scheme:

$$\ddot{A} \rightarrow A, \ddot{O} \rightarrow O, \ddot{U} \rightarrow U, \ddot{a} \rightarrow a, \ddot{o} \rightarrow o, \ddot{u} \rightarrow u.$$

This transformation, which is typically found in Web pages written by non-native speakers of German, will be called **ϵ-transformation**.

Table 11 shows some transformed terms of the top 1,000 German words and gives the number of Google hits for correct and incorrect spellings. The right-hand side of the table gives the corresponding numbers for PDF documents. The numbers show that misspellings caused by e-transformation are a widespread phenomenon. Note that the quality of PDF corpora is much better in this respect.

When applying the e- or ϵ-transformation, letter *ß* is typically replaced by *ss* (**s-transformation**). For two reasons, the distinction between *ß* and *ss* is a delicate matter. Since the Swiss spelling is *ss*, a string representing an erroneous German word may be a correct Swiss word. To make things even more complicated, the correct spelling of many German words has been changed during the so-called "Rechtschreibereform" some years ago, which affected the selection between *ß* and *ss* (e.g., *Mißverständnis* became *Missverständnis*). Still (and unofficially), the old spelling variant is broadly used. In what follows, a token written with *ss* that is officially written with *ß* is treated as an error.

We built two error dictionaries respectively representing errors introduced via e-transformation and ϵ-transformation. All vowels of the form *ä, ö, ü* (or upper-case variants) in the German dictionary were replaced by their images under the respective transformation. Letter *ß* occurring in the entries was categorically replaced by *ss*. For the e-transformation we obtained a list of 436,198 strings. The filtering procedure led to an error dictionary $D_{err}(German, enc\text{-}e)$ with 432,987 entries.

Applying the ϵ-transformation and the usual filtering step, we generated the error dictionary $D_{err}(German,enc\text{-}\epsilon)$ with 407,013 entries. A considerable number of well-formed words was generated and filtered out. The rules of German morphology yield a

partial explanation: For so-called strong verbs some paradigmatic forms only differ by a mutation of vowels (*möchte-mochte*).

An extra error dictionary D_{err}(*German,enc-s*) was built by replacing *ß* by *ss* in German dictionary entries without occurrences of vocals *Ä, Ö, Ü, ä, ö, ü*. The dictionary has 42,340 entries.

4.5 Summary and Maximal Error Dictionaries

Using the union of all error dictionaries for both languages, we constructed the maximal error dictionaries D_{err}(*English,all*) and D_{err}(*German,all*). Table 12 summarizes the sizes of all error dictionaries.

5. Error Overproduction and Underproduction

Before we analyze the number of tokens in the corpora that represent entries of the error dictionaries, we comment on the limitations of this kind of analysis. Obviously, not all orthographic errors of a given type occur in the respective error dictionary (underproduction). On the other hand, some tokens classified as errors by the error dictionary might in fact be correct words (overproduction) due to the incompleteness of

Table 11
Most frequent German words with vowels ä, ö, ü; frequencies of correct spelling and frequency after applying e-transformation. Frequencies are counted in arbitrary Web pages (left part of the table) and in PDF documents in the Web.

Word	Norm.	Transf.	Percentage	PDF norm.	PDF transf.	Percentage
für	19,000,000	5,140,000	27.05	4,050,000	30,900	0.76
über	17,800,000	2,330,000	13.08	3,610,000	16,000	0.44
können	14,500,000	290,000	2.00	1,790,000	3,960	0.22
müssen	7,420,000	177,000	2.38	1,090,000	2,060	0.18
wäre	3,500,000	173,000	4.94	590,000	631	0.11
fünf	2,470,000	291,000	11.78	541,000	570	0.10
könnte	2,900,000	165,000	5.69	570,000	618	0.11
hätten	815,000	43,100	5.28	234,000	315	0.13
dafür	3,580,800	124,000	3.46	814,000	865	0.11
würde	3,770,000	162,000	4.30	601,000	693	0.11

Table 12
Size of error dictionaries.

Error dictionary	Entries	Error dictionary	Entries
D_{err}(*English,typing*)	9,427,051	D_{err}(*German,typing*)	13,656,866
D_{err}(*English,spell*)	1,202,997	D_{err}(*German,spell*)	18,970,716
D_{err}(*English,ocr*)	1,532,741	D_{err}(*German,ocr*)	10,608,635
		D_{err}(*German,enc-e*)	432,987
		D_{err}(*German,enc-ϵ*)	407,013
		D_{err}(*German,enc-s*)	42,340
D_{err}(*English,all*)	11,884,284	D_{err}(*German,all*)	43,688,771

the final filtering step in the construction of the error dictionaries. From the construction of the error dictionaries we may expect that incompleteness/underproduction is mainly caused by

- missing patterns for spelling errors and OCR errors, and
- the fact that we do not seriously damage words when constructing the error dictionaries.

For both English and German, to estimate under/overproduction of the error dictionaries, the primary general HTML corpus was split into four subclasses. The class Best contains all documents where the number of hits (tokens representing entries of the maximal error dictionary) per 1,000 tokens is $\leq$1. For class Good (Bad, Worst, respectively), the number of hits per 1,000 tokens is 1–5 (5–10, >10, respectively). The number of documents in each class is found in Tables 13 and 14.

5.1 Estimating Underproduction

To estimate underproduction of the English error dictionaries, the English general HTML corpus was split into subfiles, each containing 300 tokens. We then randomly selected such subfiles and analyzed the proper errors found in these portions. Since we wanted to avoid an unbalanced selection where most errors are from the document class Worst, a maximum of three errors from each subfile was used for the analysis. Error candidates were found with the help of a spell checker and using our standard dictionaries as a second control. Slang and special vocabulary were not used for the statistics. We also excluded errors where two words were merged. We found that most of the latter errors were caused by the conversion process from HTML to ASCII. Each candidate was manually controlled; in difficult cases we consulted Merriam-Webster Online. We continued the search until 1,000 proper errors were isolated. From these, 624 (62.4%) turned out to be entries of the maximal English error dictionary.

Table 13 refines these statistics and shows the number of errors and the percentage of errors found in the error dictionary for the four quality classes of documents. As a tendency, recall of the error dictionary is better in "bad" documents.

The same procedure was used for German and confirmed this tendency. From 1,000 errors in the German general HTML corpus, 638 (63.80%) were found in the maximal German error dictionary. The statistics for the four quality classes of documents is presented in Table 14.

5.2 Estimating Overproduction

In our first experiment with English texts we found that a considerable number of hits corresponded to special names introduced in the documents. Many of these names are artificial (e.g., *Hitty*). In order to avoid all difficulties with special names we decided to restrict the error analysis in English texts to words starting with a lowercase letter. In each of the four classes, 1,000 hits of this form were randomly selected. We then manually checked which of these tokens represent correct words, reading contexts and consulting Merriam-Webster Online in difficult cases.

The results are presented in Table 15 and show a clear tendency. The percentage of proper errors is larger in documents with a large number of hits. In the class Worst, 95%

Table 13
Underproduction of the maximal error dictionary in the primary English general HTML corpus.

Document class	Documents	Errors found	Entries of error dict.	%
Worst	24	248	166	66.93
Bad	39	194	131	67.53
Good	226	389	242	62.21
Best	540	169	85	50.29

Table 14
Underproduction of the maximal error dictionary in the primary German general HTML corpus.

Document class	Documents	Errors found	Entries of error dict.	%
Worst	50	389	307	78.92
Bad	42	166	101	60.84
Good	297	385	201	52.21
Best	229	60	29	48.33

of all hits are proper errors; in the class Best, only 60% of the hits represent orthographic errors. Most of the remaining hits could be assigned to one of the following categories: correct standard expressions (missing entries of the standard dictionaries), names and geographic expressions, foreign language expressions, archaic and literary word forms, and abbreviations. The number of hits in each category is found in Table 15. The large number of standard words among the hits in the class Best is caused by an incompleteness of our English dictionary, which does not always contain both the British and the American spelling variants.

In the German general HTML corpus, where we could not restrict the experiment to tokens starting with a lowercase letter, a more shallow picture is obtained (Table 16). For the classes Best (61% proper errors), Good (62% proper errors), and Worst (88% proper errors), results are similar to the English case and confirm the above-mentioned general tendency. Due to the large number of names, foreign language expressions, and archaic/literary word forms that are found in class Bad, we here have only 56% proper errors. The results show that overproduction could be considerably reduced when filtering error dictionaries with better standard dictionaries for geographic entities, personal names, foreign language expressions, and archaic and literary word forms.

Table 15
Overproduction of the maximal error dictionary in the English general HTML corpus.

Document class	Best	Good	Bad	Worst
Hits	1,000	1,000	1,000	1,000
Percentage proper errors	72	86	89	95
Proper errors	722	856	894	952
Standard words	206	31	21	5
Personal names and geographic entities	23	35	24	27
Foreign language expressions	32	42	36	12
Archaic and literary word forms	9	28	1	1
Abbreviations	8	6	24	2

5.3 Summary So Far

From the above percentages we obtain a naive estimate for the ratio between the real number of errors and the number of hits of the error dictionaries, which is presented in Table 17. The results show that the number of hits can be seen as a lower approximation of the real number of errors. The ratio between both numbers is larger for English. It does not differ dramatically between the distinct quality classes. However, since both over- and underproduction are larger for "good" documents, error estimates for these classes come with a larger degree of uncertainty.

5.4 Difficulties

The above analysis turned out to be much more time-consuming and difficult than it might appear. One problem is caused by the fact that nonstandard vocabulary and errors do not represent disjoint categories. Orthographic errors are sometimes "abused" as slang expressions. A separation between archaic/foreign language expressions and orthographic errors is often only possible when taking the sentence context into account. These and other examples explain that demarcation issues are sometimes difficult to solve. The construction of special dictionaries for slang, foreign language expressions, special names, and archaic word forms represents an important step for future work. Using these dictionaries in the filtering step of the construction of the error dictionaries, overproduction may probably be reduced in a significant way. Furthermore, these dictionaries should help to detect Web pages with nonstandard vocabulary of a particular type.

Table 16
Overproduction of the maximal error dictionary in the German general HTML corpus.

Document class	Best	Good	Bad	Worst
Hits	1,000	1,000	1,000	1,000
Percentage proper errors	61	62	56	88
Proper errors	615	624	564	884
Standard words	126	123	47	3
Names and geos	201	147	193	49
Foreign language expressions	31	46	103	37
Archaic and literary word forms	18	44	82	24
Abbreviations	9	16	11	3

Table 17
Naive estimates of the ratio between the real number of errors and the number of hits of the error dictionaries for distinct quality classes.

English				German			
Best	0.72/0.5029	=	1.43	Best	0.61/0.4833	=	1.26
Good	0.86/0.6221	=	1.38	Good	0.62/0.5221	=	1.19
Bad	0.89/0.6753	=	1.32	Bad	0.56/0.6084	=	0.92
Worst	0.95/0.6693	=	1.42	Worst	0.88/0.7892	=	1.12

6. Distribution of Orthographic Errors in the General HTML Corpora

We define the **error rate** of a text with respect to an error dictionary D_{err} as the average number of entries of D_{err} that are found among 1,000 tokens of the text. In this section we describe the distribution of error rates for all types of errors in the general HTML corpora. Experiments for other corpora are summarized in the following section. The results of the previous section indicate that the error rate represents a reasonable lower approximation for the real number of errors per 1,000 tokens in the document. Incompleteness of the rule sets for generating spelling errors and OCR errors should be kept in mind. Recall that in English documents, only words starting with a lowercase letter are taken into account.

6.1 Distribution of Error Rates for Orthographic Errors

In the first test, we consider orthographic errors, that is, errors of arbitrary type. Accordingly, error rates for documents are computed with respect to the maximal error dictionaries. For a coarse survey, as in the previous section, we distinguish four quality classes Best, Good, Bad, Worst, respectively, covering pages with error rates in the intervals $[0, 1)$, $[1, 5)$, $[5, 10)$, and $[10, \infty)$.

English. The histograms in Figure 1 show the percentage of documents in each class in the primary (left-hand side) and secondary (right-hand side) English corpora. Remarkably, the differences between the two corpora are almost negligible. In both cases, most documents belong to class Best; only a small percentage of documents belongs to classes Bad and Worst.

Table 18 presents the average error rate for various document classes. As to the length of documents in the corpora, drastic differences exist. We did not find a correlation between document length and error rates, with the following eye-catching exception: small (larger) documents of an excellent quality tend to have an error rate 0 (close to 0, but >0).[5] In order to avoid a dominating influence of long documents, we simply computed the arithmetic mean of all error rates obtained for the single documents. The class Best 80% collects 80% of all documents with lowest error rate, and similarly for the class Best 90%.

Note that a significant difference exists between the average rate for all documents (2.47, 2.24, respectively) and the means for the Best 80% classes (0.67, 0.68, respectively). These numbers point to an effect that will be found again in other figures and experiments: The large majority of all documents in the corpora have a very good quality. Yet, at the "bad end" of the spectrum we find a considerable number of unacceptable documents with a very large number of errors. The phenomenon becomes even more apparent in Figure 2 (left diagram) where we depict the error rates of all documents. In what follows we often describe mean error rates for all documents *and* for the class Best 80%. When comparing distinct corpora, the two values help to see if deviations concern the class of all documents or if they are rather caused by a small number of "bad" documents.

Note also that all corresponding average error rates obtained for the primary and secondary corpora are almost identical. This gives at least some evidence to the conjec-

5 This explains the special effect seen in Figures 14 and 15 where the refined crawl produces many short documents.

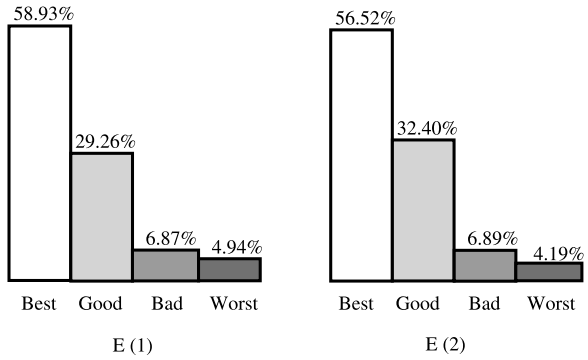


Figure 1
Percentage of documents in the four quality classes for the primary (left-hand side) and secondary (right-hand side) English corpora. The four quality classes cover distinct error rates for orthographic errors.

ture that for corpora crawled with similar queries and collection strategies, error rates will not differ too much. As we see next, the situation for the German corpora is more complex.

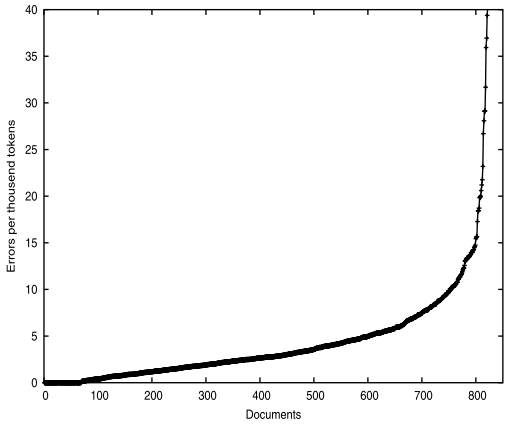
German. The histogram in Figure 3 shows the percentage of documents in each class of the primary (left-hand side) and secondary (right-hand side) German corpora. A large number of documents belongs to class Good. We now find a larger difference between the primary and secondary corpora. Several phenomena might be responsible. As mentioned above, for the German corpora we did not restrict the analysis to tokens starting with a lowercase letter. Hence, documents with many names can cause special effects and lead to differences between corpora. Second, errors caused by encoding of special characters represent an important extra source for errors in German documents where numbers may vary from one corpus to another. This is seen in Table 20 where we analyze all error types occurring in the primary and secondary German corpus. The means for e-transformation are 0.62 for the primary corpus and 1.40 for the secondary corpus.

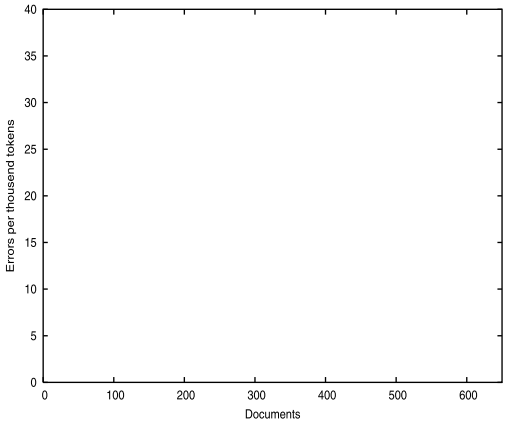
The average error rates obtained for distinct documents classes of the German corpus, which are presented in Table 19, show that

- for all classes we have more errors than in the English documents, and
- for different corpora, sometimes nontrivial deviations must be expected.

Table 18
Mean error rate for arbitrary orthographic errors in various document classes; results for the general English HTML corpus.

Document class	Best	Good	Bad	Worst	Best 80%	Best 90%	All
E (1)	0.30	2.31	8.83	23.23	0.67	1.06	2.47
E (2)	0.27	2.19	6.77	21.61	0.68	1.03	2.24





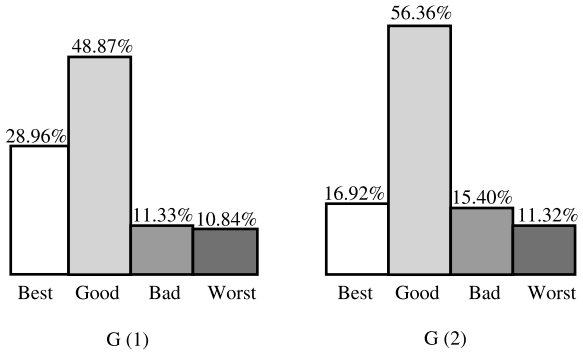


Figure 3
Percentage of documents in the four quality classes for the primary (left-hand side) and secondary (right-hand side) German corpora. The four quality classes cover distinct error rates for orthographic errors.

corpora; results are presented in Table 20. The tendency observed earlier for orthographic errors was confirmed: the difference between the two English corpora (mean 0.39 versus mean 0.38) is negligible; for the two German corpora, the difference is larger (mean 0.45 versus mean 0.58).

OCR Errors. The diagrams in Figure 6 show that most documents do not contain any OCR errors. Of course this result is not surprising. Probably not all errors that contribute to the two diagrams are really caused by wrong character recognition. Although some of the documents with the highest errors were explicitly marked to contain scanned

Table 19
Mean error rate for arbitrary orthographic errors in various document classes; results for the general German HTML corpus.

Document class	Best	Good	Bad	Worst	Best 80%	Best 90%	All
G (1)	0.41	2.61	7.30	15.15	1.89	2.58	3.86
G (2)	0.48	2.57	7.21	24.38	2.40	3.09	5.40

Table 20
Mean of error rates for all error types in primary and secondary general HTML corpora.

Error type	Mean error rate English corpus HTML (1)	Mean error rate English corpus HTML (2)	Mean error rate German corpus HTML (1)	Mean error rate German corpus HTML (2)
arbitrary	2.47	2.24	3.86	5.40
typographic	2.31	2.03	2.15	2.79
spelling	0.39	0.38	0.45	0.58
OCR	0.06	0.07	0.13	0.18
e-transformation	0.003	0.004	0.62	1.40
ϵ-transformation	0.02	0.01	0.19	0.24
s-transformation	0.00003	0.00	0.76	0.96

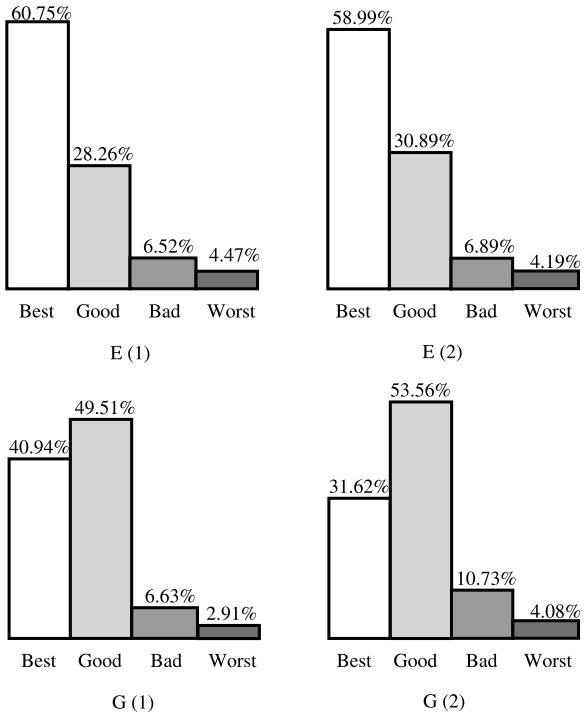


Figure 4
Typographic errors: the percentage of documents in the four quality classes in the general English (upper part) and German (lower part) HTML corpora. Quality classes refer to error rates for typographic errors.

text, it is natural to assume that the total number of such documents in the corpus is very small. Ambiguous error types might explain some of the errors found in Figure 6; see the discussion below. As a matter of fact, the number of OCR errors will grow when analyzing corpora with many OCRed pages.

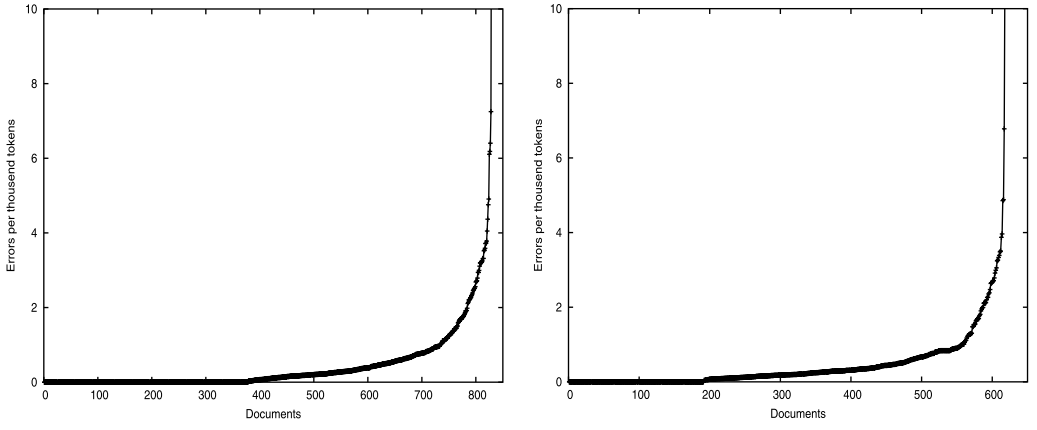


Figure 5
Distribution of error rates for *spelling* errors in the primary English (left diagram, mean error rate 0.39) and German (right diagram, mean error rate 0.45) general HTML corpora. In the left (right) diagram, one document with error rate 14.95 (11.31) is omitted.

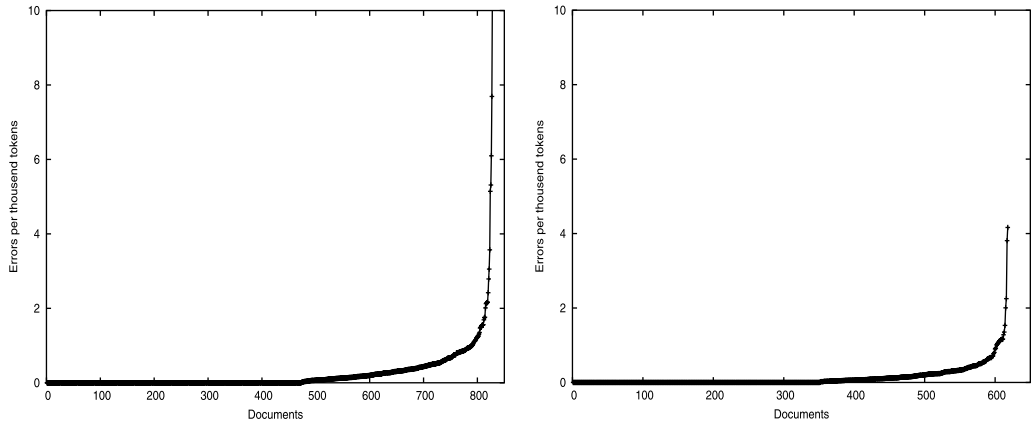


Figure 6
Distribution of error rates for *OCR errors* in the primary English (left diagram, mean error rate 0.06) and the German (right diagram, mean error rate 0.13) general HTML corpora.

e-transformation and ϵ-transformation. Figures 7 and 8 show some interesting differences between the use of both transformations in German Web pages. In the primary German corpus, e-transformation errors are concentrated in a small class of documents (documents with rank >480) where we have a nontrivial number of occurrences, leading to a mean error rate of 0.62. The mean error rate for ϵ-transformation is much smaller (0.19). Still, there are *more* documents containing an ϵ-transformation error. This indicates that e-transformation is applied more systematically. The small plateau in Figure 7 is generated by some portion of text that was found in several documents. The error rates that arise when applying e-transformation in a completely systematic way are typically larger. In the corpus we found some documents of this kind; since the rates are too high, these documents are not depicted in the figure.

We also looked for e- and ϵ-transformation errors in the documents of the English general HTML corpus. These errors, which mutate German words, only occur

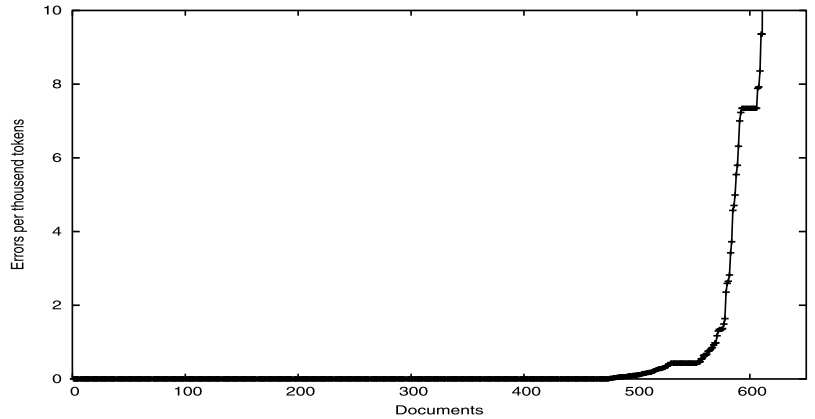


Figure 7
Distribution of error rates for *e-transformation* in the primary German general HTML corpus. Mean: 0.62. Here 7 documents with error rates ranging from 13.16 to 34.10 are omitted.

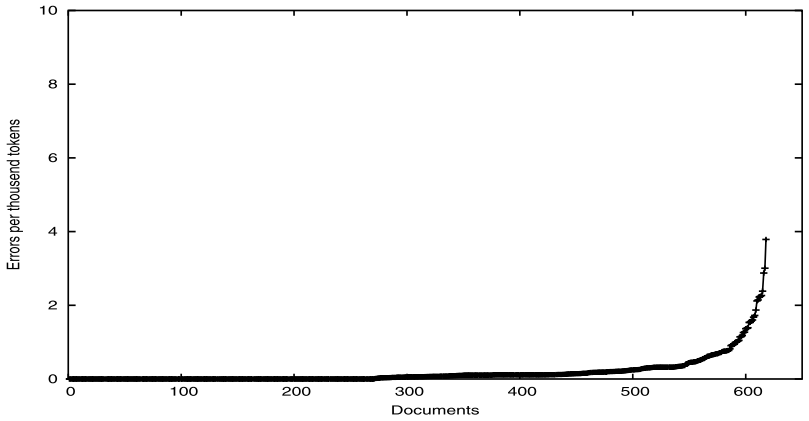


Figure 8
Distribution of error rates for ϵ-*transformation* in the primary German general HTML corpus. Mean 0.19.

in a small number of English documents. Whereas German writers strongly prefer the e-transformation in situations where the correct characters are not available, we find a clear preference for the ϵ-transformation in the English documents.

s-Transformation. Figure 9 shows the distribution of error rates for s-transformation in the primary German general HTML corpus. Since the corpus contains some Swiss documents, where "ß" is categorically written "ss" (cf. Section 4.4), the high mean (0.76) has to be relativized.

Overview. Table 20 summarizes the error rates of all types of errors in the general HTML corpora. The numbers show that not all errors can be traced back to a unique error type.

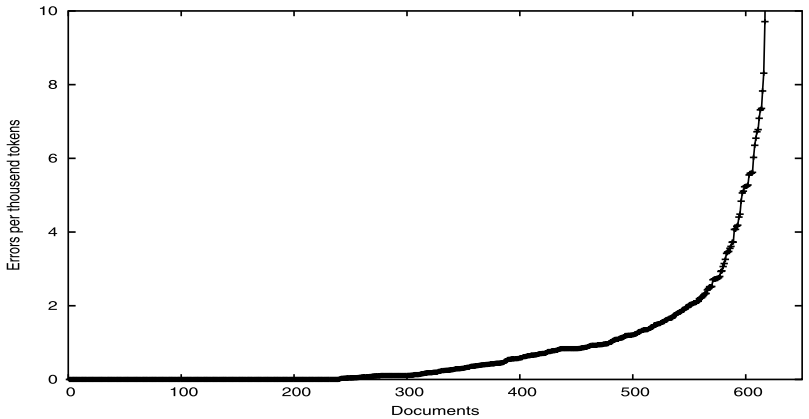


Figure 9
Distribution of error rates for *s-transformation* in the primary German general HTML corpus. Mean 0.76. One document with error rate 11.46 is omitted.

6.3 Summary So Far

For both languages, the large majority of all documents has a small number of orthographic errors. On the other hand, at the "bad end" of the spectrum, a considerable number of unacceptable documents with high error rates is found. Mean values for error rates are strongly influenced by the latter documents; the average error rate for the Best 80% class is usually much lower. The latter rate should also be considered when comparing results obtained for two corpora.

Phenomena observed in English corpora seem to be more stable than those for German: Results obtained for the primary and the secondary English general HTML corpus are almost identical. Differences between the two German corpora may partially be explained by names occurring in texts and by special character encoding problems. Table 20 illustrates this effect, showing the mean error rates for all error types in the primary and secondary HTML corpora.

The most important error class are typographic errors. In the German documents, e-transformation and s-transformation represent another typical error source. Whereas the number of spelling errors is significant, OCR errors do not play an essential role.

Interestingly, the basic form of the distribution curves in Figure 2 is found again in all corresponding curves for other error types and other corpora (see also Figures 14 and 15); although the absolute numbers for error rates and details are of course distinct. The close similarity of all distribution curves is striking and gives some evidence to the assumption that relevant features of the error rate distribution are stable, regardless of the corpora that are investigated.

7. Differences for Special Corpora

We summarize the error rates that we found in PDF corpora and in corpora for special thematic fields. In Figures 14 and 15, we present a small selection of distribution curves for error rates. Similarities of the distribution curves mentioned in the previous section should also be noted.

7.1 Distribution of Orthographic Errors in the General PDF Corpora

Figure 10 presents the mean error rates for distinct error types found in the general PDF and (primary) HTML corpus for English. The results show that PDF documents in general have a better quality than HTML documents. Whereas we have a mean error rate of 2.47 for orthographic errors in the HTML documents, the corresponding mean is only 1.38 for PDF. For the Best 80% documents the means are 0.67 (HTML) and 0.38 (PDF).

In principle, the same tendency was observed in the documents of the parallel German corpora. However, special effects polluted the picture. As we mentioned in Section 2.1, the conversion of the German PDF documents to ASCII is very error prone. Although we excluded all converted documents that were obviously garbled by the conversion, we also found in the remaining documents examples of errors that were caused by the conversion process. In this sense, the error rates in the original PDF documents are probably smaller. Mean error rates are 2.15 (HTML) versus 2.04 (PDF) for typographic errors, 0.45 versus 0.41 for spelling errors, 0.13 versus 0.09 for OCR errors, 0.62 versus 0.07 for e-transformation errors, and 0.19 versus 0.16 for ϵ-transformation errors. Since the conversion tool categorically replaces letter "ß" by

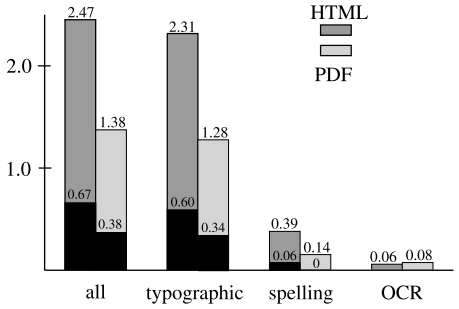


Figure 10
PDF versus HTML: mean error rates for distinct error types in the general corpora (English). Black rectangles describe mean error rates for the Best 80% subclass.

"ss", a very high number of s-transformation errors led to the effect that the overall mean error rate for the German PDF (3.95) is even larger than the rate for the German HTML (3.86).

7.2 Distribution of Orthographic Errors in Distinct Thematic Corpora

Figure 11 describes the average error rates for orthographic errors and spelling errors in the English corpora. In almost all thematic areas, mean error rates are larger than the corresponding means in the general corpora; the differences are significant and remarkable. With a mean error rate of 2.05 (0.30) for orthographic (spelling) errors, the English Neurology corpus is very clean and represents an exception. For the Fish corpus, even the mean error rate for the Best 80% subclass is 2.72. We conjecture that corpora that are collected without a special thematic focus often contain a large number of "professional" and carefully edited Web pages. Web pages for special thematic areas

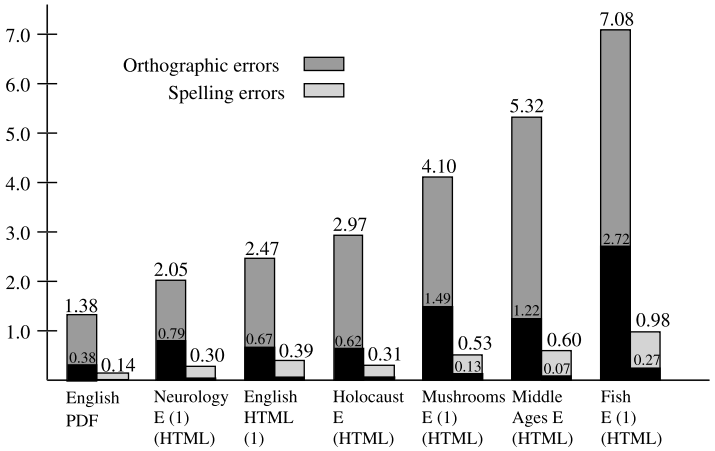


Figure 11
Thematic corpora versus general corpora: mean error rates for orthographic errors and spelling errors in distinct English corpora. All results refer to the primary thematic corpora crawled with the simple strategy (cf. Section 2.2). Black rectangles represent mean error rates for the Best 80% subclass.

are perhaps less "publicity oriented." Furthermore, as a rule of thumb, documents in thematic fields related to hobbies (e.g. Fish) contain more orthographic errors than documents in scientific fields (Holocaust, Neurology). Corpora with a focus on history seem to occupy a middle position.

In the German corpora we have the means for orthographic/spelling error rates presented in Table 21; numbers in brackets refer to the Best 80% subclass. The second column shows that, by and large, the ranking order for thematic areas induced by mean error rates observed in the English corpora is found again in the German part. The German corpus Neurology, with its high error rate, does not follow this line. The high means for the Best 80% subclasses in the German corpora are remarkable and show that the low quality is not caused by a small number of bad documents.

7.3 Differences between the Two Crawling Strategies

Table 22 summarizes the differences for the English corpora retrieved with the simple strategy on the one hand and the corpora retrieved with the refined strategy on the other hand. Numbers represent average error rates for the corpora. Numbers in brackets refer to the Best 80% subclass.

Surprisingly, all corpora crawled with the refined strategy always have a better (smaller) average error rate than those retrieved with the simple strategy, pointing to a significant difference between the two types of collection strategies. An analysis of the document genres found in the two types of corpora presented in Section 8 offers a good explanation; see Table 26.

Table 21
Mean error rates for orthographic errors and spelling errors in thematic German corpora.

German	Orthographic errors	Spelling errors
General PDF	3.95 (2.31)	0.41 (0.06)
Neurology G (1) (HTML)	6.94 (4.48)	0.51 (0.26)
General HTML (1)	3.86 (1.81)	0.45 (0.16)
Holocaust G (HTML)	4.97 (3.03)	0.50 (0.27)
Mushrooms G (1) (HTML)	7.91 (3.69)	0.78 (0.32)
Middle Ages G (HTML)	7.84 (4.30)	0.96 (0.38)
Fish G (1) (HTML)	9.34 (4.47)	1.35 (0.52)

Table 22
Dependency of mean error rates on the crawling strategy for distinct English thematic corpora.

	Orthographic errors		Spelling errors	
English	(1) Simple crawl	(2) Refined crawl	(1) Simple crawl	(2) Refined crawl
Fish E	7.08 (2.72)	3.39 (0.35)	0.98 (0.27)	0.47 (0)
Mushrooms E	4.10 (1.49)	2.58 (0.32)	0.52 (0.13)	0.50 (0)
Neurology E	2.05 (0.79)	1.77 (0.25)	0.30 (0.05)	0.26 (0)

Table 23
Dependency of mean error rates on the crawling strategy for distinct German thematic corpora.

	Orthographic errors		Spelling errors	
German	(1) Simple crawl	(2) Refined crawl	(1) Simple crawl	(2) Refined crawl
Fish G	9.34 (4.67)	7.71 (3.31)	1.35 (0.52)	1.00 (0.17)
Mushrooms G	7.91 (3.69)	8.51 (3.50)	0.78 (0.32)	0.76 (0.08)
Neurology G	6.94 (4.48)	7.08 (2.86)	0.51 (0.26)	0.47 (0.00)

Figures 14 and 15 show that the corpora crawled with the refined strategy have a large number of documents with error rate 0. This special effect is caused by the large number of short documents that are obtained. For example, the mean length of all the documents with error rate 0 in the corpus Fish E (2) is 322 (number of lowercase normal tokens), whereas the average length of the documents in the corpus Fish E (1) is 14,196 (cf. Table 2).

The relative order between the three thematic fields was not affected by the crawling strategy. For both crawls, the Neurology corpus achieves the best results, followed by Mushrooms and Fish. The excellent quality of the Best 80% classes obtained with the refined crawl are remarkable.

For the German variant of the corpora, as Table 23 shows, a more shallow picture is obtained. For two thematic areas, the simple crawl even leads to lower error rates, although the difference is small. The ranking order between the three thematic areas obtained from the two crawls is not the same.

Figure 14 presents the error rates for orthographic errors in the English HTML corpora Fish, Mushrooms, and Neurology, comparing the simple strategy (left-hand side diagrams) with the refined strategy (right-hand side diagrams). Figure 15 gives the error rates for spelling errors in the German HTML corpora Fish, Mushrooms, and Neurology, again comparing the simple and the refined strategies.

7.4 Summary So Far

PDF corpora were found to have lower error rates. Corpora covering pages from non-scientific thematic areas often have higher error rates than corpora crawled without a fixed thematic focus. Error rates in the corpora are influenced by the crawling strategy. For English texts, refined crawling strategies that collect pages with a strong thematic focus seem to produce better corpora.

8. Error Rates and Document Genre

Classifying Web documents by genre (Kessler, Nunberg, and Schütze 1997; Finn and Kushmerick 2003; Dimitrova et al. 2003) represents one possible way to improve Web search techniques. Web-based corpus linguistics may benefit from these techniques since they enable a better control of the kind of language material that is added to a collection. In this section we want to see which kind of correlation exists between the error rates observed in a document and its genre. After manual inspection of

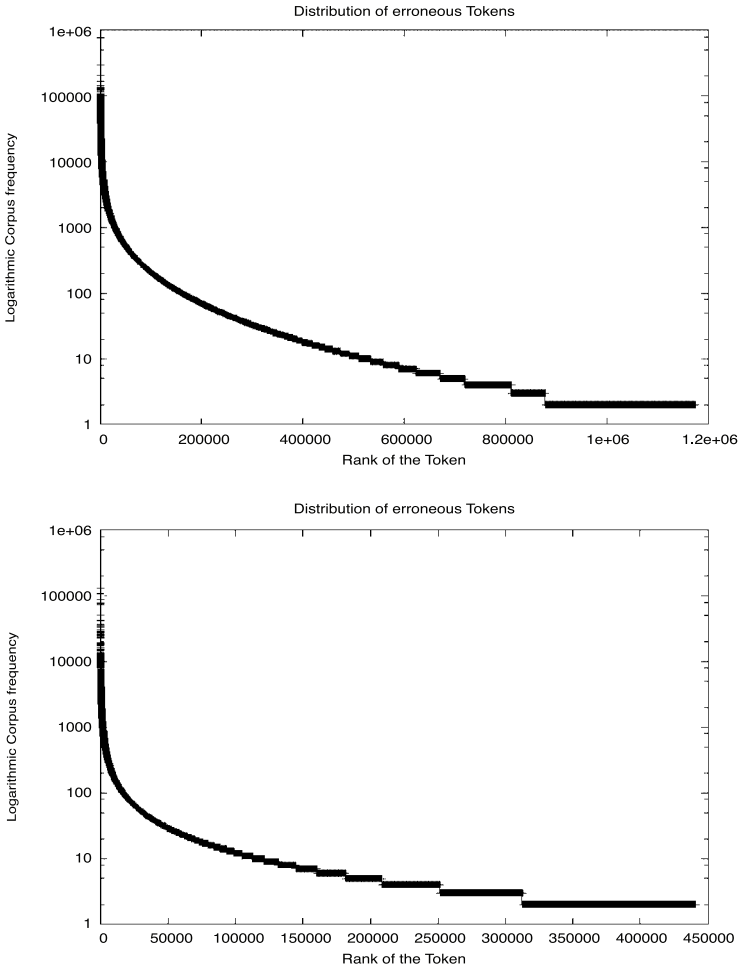


Figure 12
Zipf curves with logarithmic frequencies for English (upper diagram, 1,175,894 entries) and German (lower diagram, 454,709 entries) ranked error lists. The diagrams respectively illustrate the frequency of particular orthographic errors in English and German Web pages from a 1.4-terabyte subcorpus of the Web.

hundreds of Web pages, we decided to use the following set of document genres for our investigations:

- The class Prof contains all Web pages with professional texts from organizations, enterprises, and administrations. Also, scientific texts, professional literature, and fiction are added to this class.
- The class Priv contains private homepages and texts written from a personal point of view. A clue term is the personal pronoun *I*. Texts of this form may dominate in a Web page run by an organization. In this case, the page was classified as Priv.
- The class Chat contains chat and related collections of private statements and contributions such as guest books, mailing lists, and so forth.

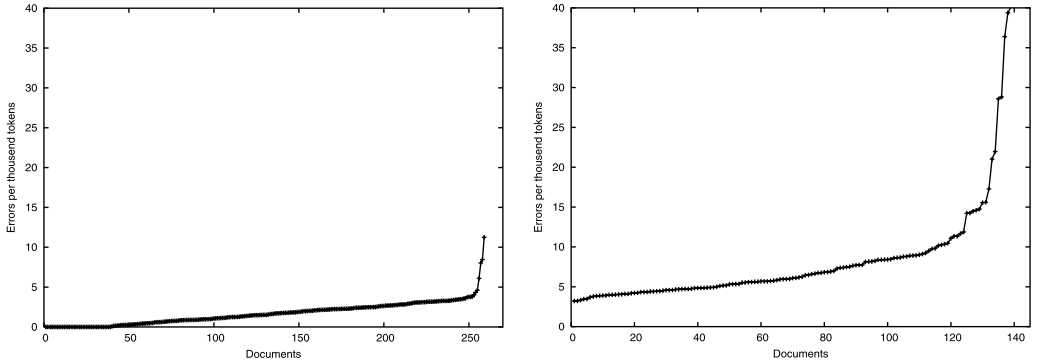


Figure 13
Distribution of error rates in documents (**passed/rejected**) by the filter $\mathcal{F}_3$ for threshold $\mu = 5$ (English test corpus). The left (right) diagram describes the distribution of documents passed (rejected) by the filter. The average error rate for accepted (rejected) documents is 1.08 (16.81).

- The class Junk contains documents where the language is "polluted," for example, by large lists of erroneous keywords, lists of foreign language expressions, dominating subparts only consisting of program code, archaic language, and so forth.
- The class Other contains all other documents. In practice we tried to assign to each document one of the above four classes, and most documents in the class Other are (almost) empty files.

Even with this small number of classes, separation issues were sometimes difficult to address. We did not introduce finer subclasses since we expected that the number of ambiguous and problematic cases would be multiplied.

Our experiments on document genre were restricted to English corpora. We looked at the primary general English HTML corpus and on the English corpora for the domains Fish, Mushrooms, and Neurology. For each of the latter three domains, both the corpus obtained with the simple crawling strategy and the corpus retrieved with the refined crawl were taken into account. Hence, a total of 7 corpora were investigated.

8.1 Genre Distribution of the Four Quality Classes

For each corpus, all documents in the classes Worst and Bad were manually classified, assigning one of the classes Prof, Priv, Chat, Junk, or Other to the document. From the classes Good and Best, 100 documents were randomly selected and classified in the same way. Table 24 presents the classification results for the primary English general HTML corpus. Not surprisingly, classes Chat and Junk dominate at the bad end of the quality spectrum, whereas class Prof dominates for good documents. The same tendency was found for all corpora, although the percentage of Prof documents in distinct quality classes was often larger. To add one further typical example, Table 25 presents the result for the corpus Fish E (1) retrieved with the simple crawling strategy. Note that even for the Bad class, 50.62% of the documents are of type Prof.

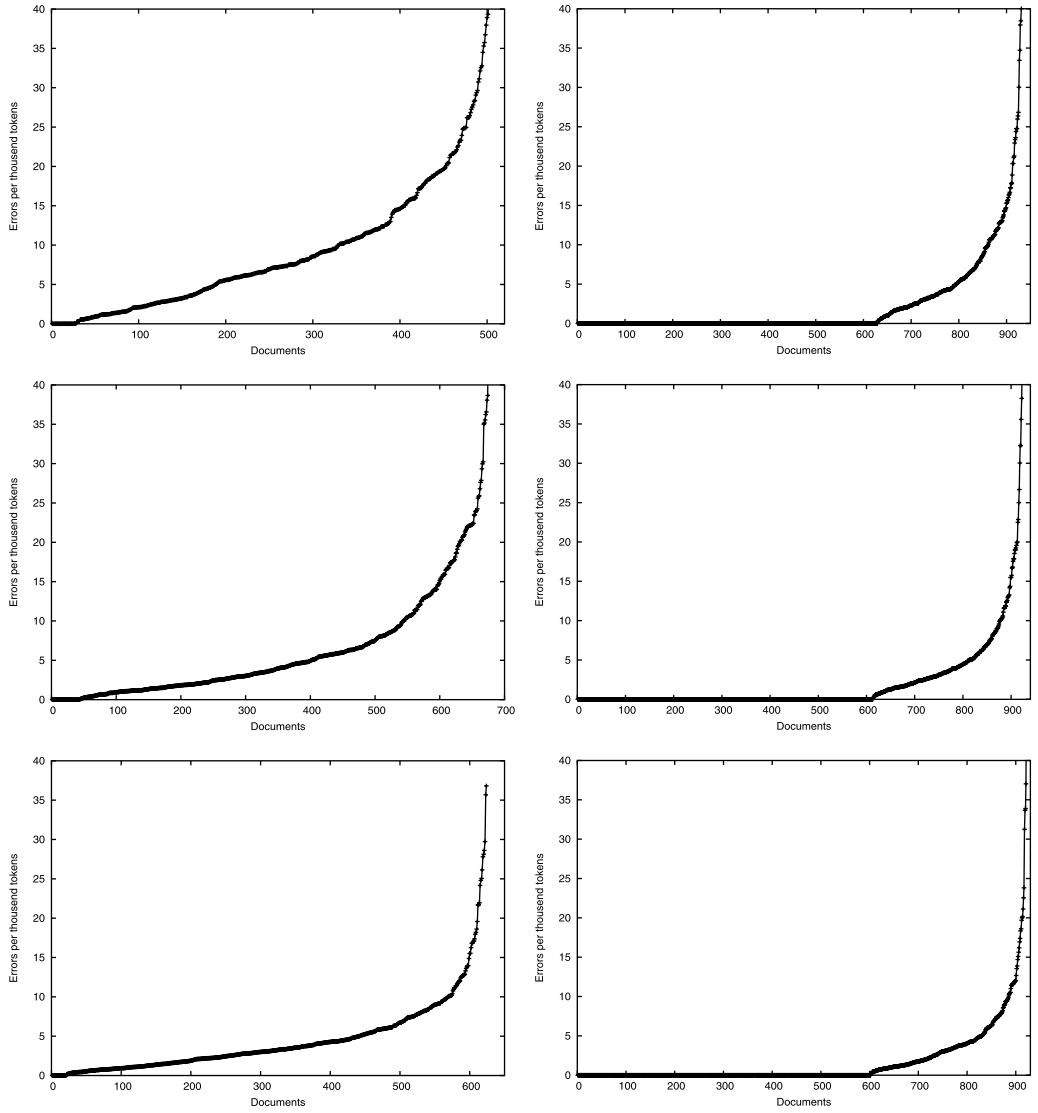


Figure 14
Distribution of error rates for *arbitrary orthographic errors* in the 6 English HTML corpora: Fish E (1) and Fish E (2) (upper diagrams), Mushrooms E (1) and Mushrooms E (2) (middle), and Neurology E (1) and Neurology E (2) (bottom diagrams). Letters (1) (diagrams on the left-hand side) refer to corpora retrieved with the simple crawling strategy. Letters (2) (diagrams on the right-hand side) stand for the refined crawling strategy. From the refined crawl (right-hand sides) a large number of documents without any error hit is obtained. Corpora crawled with the refined strategy typically contain a large number of short documents (cf. Sections 2.2 and 7.3), and short documents of good quality often have an error rate 0. A comparison along the vertical dimension illuminates differences between the three thematic areas: corpora Fish E contain more errors than corpora Mushrooms E, which contain more errors than the corpora Neurology E. Mean error rates are 7.08/3.39 [Fish E (1)/Fish E (2)]; 4.10/2.58 [Mushrooms E (1)/Mushrooms E (2)]; and 2.05/1.77 [Neurology E (1)/Neurology E (2)]. In the diagrams, some documents with high error rates are omitted to simplify scaling.

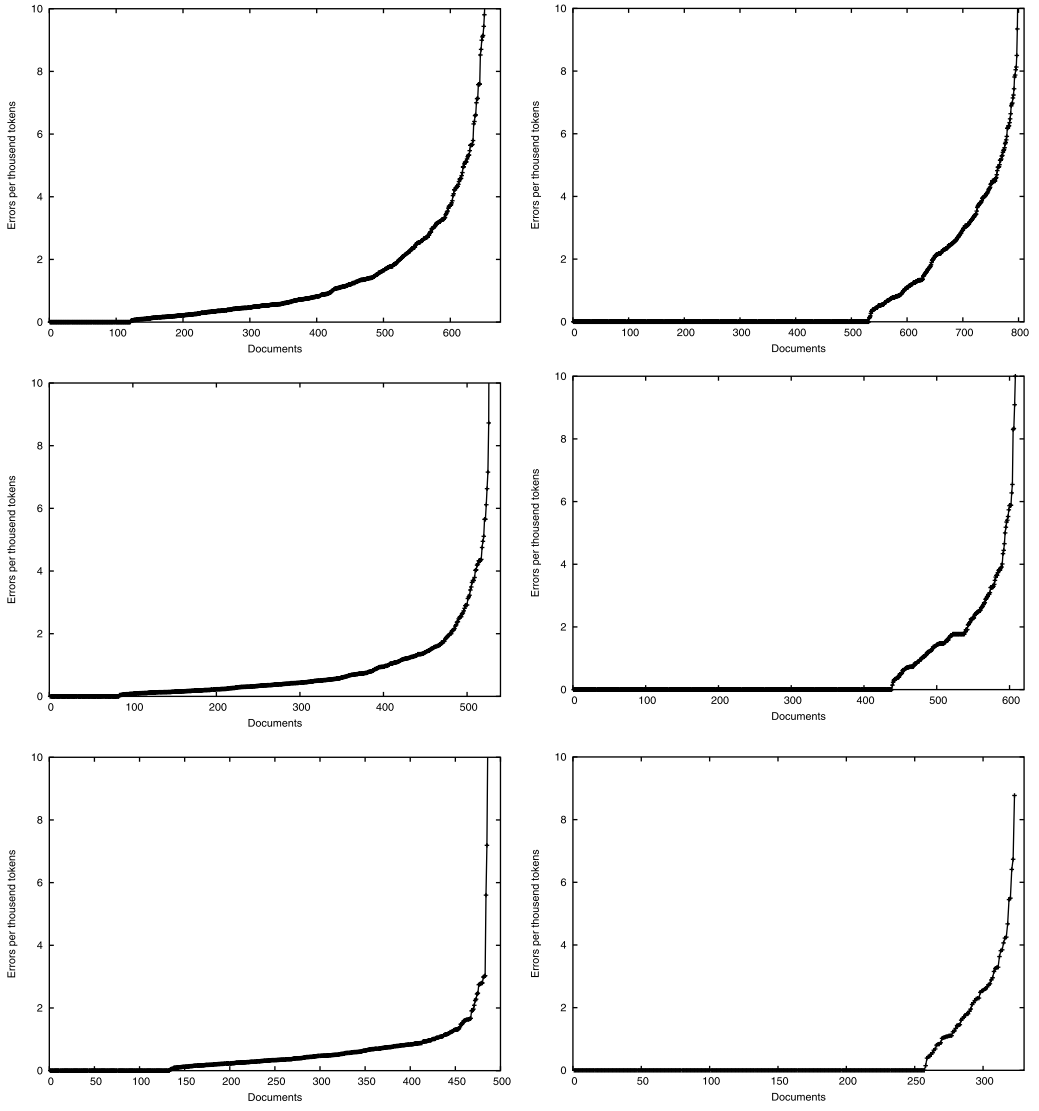


Figure 15
Distribution of error rates for *spelling errors* in the 6 German HTML corpora: Fish G (1) and Fish G (2) (upper diagrams), Mushrooms G (1) and Mushrooms G (2) (middle), and Neurology G (1) and Neurology G (2) (bottom diagrams). Letters (1) (diagrams on the left-hand side) refer to corpora retrieved with the simple crawling strategy. Letters (2) (diagrams on the right-hand side) stand for the refined crawling strategy. The latter strategy leads to a large number of short documents without any hits in the error dictionaries. See the discussion in Section 7.3. Similarly as for English HTML, corpora Fish G contain more errors than corpora Mushrooms G, which contain more errors than the corpora Neurology G. Mean error rates are 1.35/1.00 [Fish G (1)/Fish G (2)]; 0.78/0.76 [Mushrooms G (1)/Mushrooms G (2)]; and 0.51/0.47 [Neurology G (1)/Neurology G (2)]. In the diagrams, some documents with high error rates are omitted to simplify scaling.

Table 24
Genre distribution of the four quality classes for the primary general English HTML corpus.

English HTML (1)	Worst (%)	Bad (%)	Good (%)	Best (%)
Chat	42.31	56.41	24.00	1.00
Junk	38.46	5.13	1.00	0.00
Priv	3.85	10.26	14.00	9.00
Prof	15.38	28.20	61.00	90.00
Other	0.00	0.00	0.00	0.00

Table 25
Genre distribution of the four quality classes for the corpus Fish E (1).

Fish E (1)	Worst (%)	Bad (%)	Good (%)	Best (%)
Chat	37.39	20.99	4.00	3.00
Junk	26.09	6.17	0.00	6.00
Priv	8.70	22.22	9.00	3.00
Prof	27.82	50.62	84.00	84.00
Other	0.00	0.00	3.00	4.00

8.2 Genre Distribution: Simple Crawl versus Refined Crawl

The analysis of genres presented in Table 26 illuminates an important difference between the thematic corpora retrieved with the simple and the refined crawling strategy: In the latter corpora, the percentage of documents of type Chat and Junk is lower; differences are significant. At the same time, corpora retrieved with the refined strategy contain more documents of type Prof. We conjecture that the open compounds that were used in the queries for the refined crawl (cf. Section 2.2) represent a kind of "high-level language expressions" that are typically used in a professional or scientific context. With the above background, it is not surprising that the refined crawling strategy leads to better error rates.

8.3 Error Rates for Genres

Table 27 presents estimates for the mean error rates obtained for the distinct document genres in the seven corpora. These numbers represent estimates since not all documents

Table 26
Composition of corpora retrieved with the simple (1) and the refined (2) crawling strategies. The refined strategy (2) helps to avoid documents of type Chat and Junk, attracting documents of type Prof at the same time.

Crawls	Fish E (1) (%)	Fish E (2) (%)	Mushr. E (1) (%)	Mushr. E (2) (%)	Neur. E (1) (%)	Neur. E (2) (%)
Chat	13.86	2.69	8.63	3.52	3.87	2.87
Junk	9.10	0.88	5.40	3.15	2.97	0.11
Priv	8.79	16.13	12.70	11.96	7.49	2.44
Prof	66.03	80.30	73.27	80.01	82.83	94.58
Other	2.22	0.00	0.00	1.36	2.84	0.00

Table 27
Mean error rates (estimates) for distinct document genres in seven corpora.

Crawls	English HTML (1)	Fish E (1)	Fish E (2)	Mushr. E (1)	Mushr. E (2)	Neur. E (1)	Neur. E (2)
Chat	6.90	13.05	14.29	10.71	6.27	4.94	11.22
Junk	27.31	23.61	59.05	12.37	16.00	4.59	3.15
Priv	2.82	7.85	3.16	3.34	3.37	3.79	5.89
Prof	1.26	3.68	2.04	2.94	1.20	1.67	1.31

of the classes Good and Best were classified. In all corpora, the mean error rate for class Prof is better than the rate for class Priv, which is better than the rate for class Chat. The results indicate that the error rate of a document might be an interesting feature for genre classification: High error rates typically point to documents of the genres Junk and Chat; excellent error rates typically point to documents of type Prof. Results for the Neurology corpora indicate that "scientific Chat/Junk" may come with low error rates.

8.4 Summary So Far

An obvious correlation exists between the genre of a document and its error rate. Error rates might be used as one feature for genre classification. The analysis of genres helps one to understand the differences between corpora retrieved with distinct crawling strategies and the error rates observed in the corpora.

9. Filtering Methods

The figures seen in the previous sections show that corpora collected from the Web typically contain a non-negligible number of documents with an unacceptable number of orthographic errors. We now turn to the question of how to use error dictionaries for recognizing and filtering Web pages with a high percentage of errors, thus excluding them from the corpus construction process. The question of what should be considered as a "high percentage" has to be answered for each application. Generally speaking we would like to exclude at least those documents that are found at the right end of the diagrams presented in the previous sections.

Definition

By a **filter**, we mean a pair $\mathcal{F} = \langle D, \rho \rangle$ consisting of an error dictionary, D, and a **filter threshold**, ρ. The filter **rejects** a text document (Web page) T iff the average number of entries of D that are found among 1,000 tokens of T exceeds ρ.

As a matter of fact, we may use the maximal error dictionaries for filtering. For some applications, small error dictionaries, which occupy less space and are easier to handle, may be advantageous. The results presented below show that when one uses a more rigid filter threshold ρ, the filtering effect achieved with "small" error dictionaries is very similar to the effect when using the maximal error dictionaries. With an obvious interpolation, this observation supports the assumption that the

incompleteness of our maximal error dictionaries does not seriously reduce their filtering capacities.

9.1 Distribution of Error Frequencies

Since error dictionaries are necessarily incomplete in the sense that not all possible errors can be covered, it is natural to ask if filters of the above-mentioned form can work. We shall see below that even filters with small error dictionaries are useful. The reason is that the frequency of orthographic errors in the Web follows a Zipf-like[7] distribution. Since a relatively small number of erroneous tokens already covers a substantial number of all error occurrences, it should not be surprising that even small error dictionaries help to identify pages with many errors. In Figure 12, we show the logarithmic frequencies of errors in a 1.4-terabyte subcorpus retrieved from the Web in 1999 ("Web-in-a-box"). The upper diagram shows the distribution of all errors from the maximal English error dictionary, $D_{err}(English,all)$, in English Web pages. Only errors with at least two occurrences are covered. Similarly the lower diagram shows the distribution of errors from $D_{err}(German,all)$ in German Web pages.

9.2 Basic Filter Scenario

Suppose we are given a collection of Web pages, $\mathcal{C}$. We may fix a **user-defined threshold** μ in terms of the average number of errors per 1,000 tokens that we are willing to accept in a document to be added to our corpus. A document where the average number of errors per 1,000 tokens does not exceed μ is called **acceptable**, other documents are called **unacceptable**. In practice, since we cannot count real errors, a token is considered erroneous if and only if it occurs in one of our error dictionaries. In Section 5, we have seen that the number of entries of the error dictionary found in a text yields a lower approximation for the real number of errors.

In terms of information retrieval, acceptable documents can be considered as relevant documents that we would like to retrieve for "query" μ. To extend this analogy, we define the **answer set** of a filter $\mathcal{F}$ w.r.t. $\mathcal{C}$ as the set of all documents in $\mathcal{C}$ that are passed by $\mathcal{F}$. With these notions we may now define the parameters' precision and recall.

Definition

Let μ, $\mathcal{C}$, and $\mathcal{F}$ as above. The **precision** of $\mathcal{F}$ with respect to μ and $\mathcal{C}$ is the percentage of acceptable documents in the answer set of $\mathcal{F}$. The **recall** of $\mathcal{F}$ with respect to μ and $\mathcal{C}$ is the number of acceptable documents in the answer set of $\mathcal{F}$ divided by the number of all acceptable documents in $\mathcal{C}$.

In the remainder of the section, we define and evaluate filters for the English and German general HTML corpora, which are denoted $\mathcal{C}_E$ and $\mathcal{C}_G$, respectively. We consider three user-defined thresholds: $\mu = 10$, $\mu = 5$, and $\mu = 1$. The first bound is meant to exclude a small number of documents with an extraordinary number of orthographic errors. The second bound is more ambitious. The third bound might be used in

7 Zipf's law describes the frequency of words in large corpora. It states that the i-th most frequent word appears as many times as the most frequent one divided by i^θ, for some constant $\theta \geq 1$.

situations where high accuracy is needed and we want to retrieve only documents with a negligible number of orthographic errors.

9.3 Automated Filter Construction

We define a hierarchy of filters

$$\mathcal{F}_1 = \langle D_1, \rho_1 \rangle, \mathcal{F}_2 = \langle D_2, \rho_2 \rangle, \mathcal{F}_3 = \langle D_3, \rho_3 \rangle, \ldots$$

Filters $\mathcal{F}_k$ with higher index k generally lead to better results. On the negative side, they are more complex in terms of the number of entries of D_k. In the following description we generally assume that a user-defined threshold μ has been fixed. For simplicity, we refer to the construction of filters for the English corpus, $\mathcal{C}_{\mathrm{E}}$. The same construction was used, mutatis mutandis, for $\mathcal{C}_{\mathrm{G}}$. All filters are computed automatically on the basis of training data. For training, two inputs were used.

1. *Ranked error list.* We computed a list of all entries of the maximal English error dictionary, D_{err}(*English*,*all*), that occur at least twice in the corpus Web-in-a-box (cf. Section 9.1). The list was ordered by descending frequency of occurrence, as in Figure 12. The resulting ranked error list contains 1,175,894 entries.
2. *2. Training corpus.* The corpus $\mathcal{C}_{\mathrm{E}}$ was randomly split into a training subcorpus (427 documents) and a test subcorpus (407 documents).[8] From the training corpus, all documents were excluded that did not contain at least five distinct errors from the ranked error list, leaving 384 documents.

Definition of Filters. The error dictionary D_k for filter $\mathcal{F}_k$ was defined as the minimal initial segment S of the ranked error list such that each **unacceptable** document in the training corpus contains at least k **distinct** entries of the segment S. The threshold ρ_k was defined as the minimal average number of occurrences of entries of D_k per 1,000 tokens in an unacceptable document of the training corpus. These entries need not be distinct.

Clearly, with the given threshold we achieve a precision of 100% on the training corpus. The philosophy behind this selection of a threshold is simple: We do not want to add any unacceptable document to the corpus to be built. Precision is much more important than recall, as long as a substantial number of documents is retrieved. As a matter of fact, we cannot expect a 100% precision on the test corpus. However, our results show that the loss of precision is not significant.

9.4 Filtering Results for English General HTML Corpus

In what follows we consider the three user-defined thresholds $\mu = 10$, $\mu = 5$, and $\mu = 1$. For each of the filters $\mathcal{F}_1 = (D_1, \rho_1), \ldots, \mathcal{F}_5 = (D_5, \rho_5)$, as defined earlier, Table 28 shows

8 The distinct sizes of both corpora seem to indicate that the random selection was not perfectly balanced. We ignored this problem, which does not influence the construction.

Table 28
Evaluation of filters $\mathcal{F}_k$, $1 \leq k \leq 5$, for English general HTML corpus, user-defined threshold $\mu = 10$ (top), $\mu = 5$ (middle), and $\mu = 1$ (bottom).

	$\lvert D_k \rvert$	ρ_k	P^{Train} (%)	R^{Train} (%)	P^{Test} (%)	R^{Test} (%)
$\mu = 10$						
$k = 1$	12,217	0.91	100.00	85.42	99.67	80.00
$k = 2$	21,037	1.83	100.00	89.79	99.69	84.73
$k = 3$	46,111	2.19	100.00	91.83	99.40	87.63
$k = 4$	110,201	4.63	100.00	93.87	99.71	91.31
$k = 5$	291,309	5.62	100.00	93.00	99.70	89.21
$\mu = 5$						
$k = 1$	34,322	1.23	100.00	87.42	99.34	86.00
$k = 2$	47,747	2.19	100.00	95.70	98.50	94.00
$k = 3$	90,160	3.53	100.00	98.77	97.47	97.42
$k = 4$	110,201	3.71	100.00	98.77	97.70	97.42
$k = 5$	291,309	4.83	100.00	100.00	96.15	100.00
$\mu = 1$						
$k = 1$	37,994	0.13	100.00	51.15	93.43	55.89
$k = 2$	169,507	0.49	100.00	78.35	96.75	78.16
$k = 3$	279,543	0.63	100.00	86.14	97.02	85.58
$k = 4$	299,397	0.67	100.00	90.90	97.10	87.77
$k = 5$	580,330	0.89	100.00	97.40	96.06	96.91

the size of the filter dictionary D_k (second column), the filter threshold ρ_k (third column), and the precision and recall values achieved with the filter on the training and test corpora (columns 4, 5, 6, 7).

Baselines. When treating the complete test corpus as a "naive" answer set (recall 100%), we obtain

1. for $\mu = 10$, a precision of 94.76%, corresponding to 380 acceptable and 21 unacceptable documents,
2. for $\mu = 5$, a precision of 87.28%, corresponding to 350 acceptable and 51 unacceptable documents.
3. for $\mu = 1$, a precision of 57.10%, corresponding to 229 acceptable and 172 unacceptable documents.

For $\mu = 10$, with a precision (recall) of 99.40% (87.63%) on the test corpus, the filter $\mathcal{F}_3$ represents a good compromise between size and quality. Precision is almost optimal. The answer set for the filter contains only one unacceptable document with an error rate of 13.24, which is very close to the threshold.

For $\mu = 5$, using the filter $\mathcal{F}_3$ we obtain a precision (recall) of 97.47% (97.42%). An inspection of the nine unacceptable documents in the answer set of the filter shows that they come very close to the bound $\mu = 5$. Note that error dictionaries D_1, D_2, and D_3 are larger than the corresponding dictionaries for the threshold $k = 10$ due to the larger number of unacceptable documents in the training corpus.

For $\mu = 1$, using the filter $\mathcal{F}_3$ we obtain a precision (recall) of 97.02% (85.58%). There are six unacceptable documents in the answer set, all with an error rate below 2. The numbers in Table 28 show how a more rigid (smaller) filter threshold compensates for

the reduced size of error dictionaries essentially without sacrificing precision and with a modest loss of recall. To illustrate the effect of filtering, yet from another perspective, Figure 13 presents the distribution of error rates (number of entries from the maximal English error dictionary D_{err}(*English,all*) per 1,000 tokens) in the answer set and in the set of documents rejected by the filter $\mathcal{F}_3$ constructed for the user-defined threshold $\mu = 5$. The filter was evaluated on the test subcorpus. The figure shows that almost all documents passed (rejected) by the filter have an error frequency below (beyond) 5 errors per 1,000 tokens.

9.5 Filtering Results for the German General HTML Corpus

For computing the ranked error list, a list with the frequencies of 18,624,436 tokens in German Web pages was used. Via intersection with the list of all entries of the maximal German error dictionary, D_{err}(*German,all*), we obtained a ranked error list with 454,709 entries. The training and test corpora contain 314 and 308 documents, respectively, from the German general HTML corpus. Since the results are similar to the English case, we only point to some differences. Frequencies decrease more rapidly in the German ranked error list, as may be seen in Figure 12. In the German list, the top-ranked part is dominated by e/ϵ-transformation errors and errors where the letter ß is replaced by *ss*. The 10 top-ranked entries and their frequencies are shown in Table 29. This special class of frequent errors leads to small filter dictionaries. For example, the filter dictionary for $\mu = 10, k = 5$ has 16,277 entries, and the dictionary for $\mu = 5, k = 5$ has 127,023 entries. On the other hand, the recall values achieved with the dictionaries in general are lower than in the English case.

10. Example Applications

Obviously, the methods described above are very useful for all corpus tools that visually present linguistic data from Web pages (words, n-grams, concordances, phrases, sentences, aligned bilingual material, etc.) to the user. Filters help to exclude inappropriate pages. In the remaining data, tokens that represent entries of the error dictionaries can be marked. Depending on the application, the system may then decide to suppress this material or to add a warning when presenting it. In the remainder of this section, two case studies are presented that demonstrate the usefulness of filtering techniques and error dictionaries in distinct applications.

10.1 Text Correction with Crawled Dictionaries

It has often been observed that fixed handcrafted dictionaries only have a modest coverage when applied to new texts and corpora.[9] Still, for various text processing tasks, dictionaries with high coverage are needed. The generation of **crawled dictionaries** that collect the vocabulary of appropriate Web pages is one way to obtain a better coverage. As a matter of fact, the quality of these dictionaries suffers from orthographic errors in the analyzed pages. Using the above filters helps to reduce the number of errors that are

9 Kukich (1992) describes an experiment by Walker and Amsler (1986): "Nearly two thirds (61%) of the words in the *Merriam-Webster Seventh Collegiate Dictionary* did not appear in an eight million word corpus of *New York Times* news wire text, and, conversely, almost two-thirds (64%) of the words in the text were not in the dictionary."

Table 29
Top-ranked errors in German ranked error list and their frequencies.

Entry of error list	Correct word	Error frequency
Universitaet	Universität	131,494
grossen	großen	107,904
koennen	können	107,730
knnen	können (kennen?)	87,167
heisst	heißt	76,667
andern	ändern (anderen?)	73,972
Gruss	Gruß	51,721
ausser	außer	42,410
waere	wäre	37,071
muessen	müssen	35,864

imported. In order to further improve a crawled dictionary, we may either eliminate all words that represent entries of the error dictionaries, or we may mark these words for a manual inspection. In what follows we report on an experiment in the area of lexical text correction where these techniques improved:

1. the quality of crawled dictionaries by avoiding erroneous entries,
2. the accuracy of lexical text correction achieved with these dictionaries, using a high-level text correction system (Strohmaier et al. 2003a, 2003b).

Correction Strategy. Ignoring details, we used the following correction strategy[10]: For each token[11] of the input text, the most similar words are retrieved from the dictionary as a set of correction candidates. In many cases the token will be found in the dictionary and represents a correction candidate with optimal similarity. Based on (1) the similarity between text token and correction candidate and (2) the frequency of the correction candidate in a corpus, each candidate receives a score. If the score of the best candidate exceeds a given threshold τ, the token is replaced by this candidate. In the other case, the token is left unmodified. A good balance between similarity and frequency information in the score is obtained via training. The threshold, which is also optimized via training, guarantees that the input token is only replaced if additional confidence is available that the best correction candidate in fact represents the corrected version of the token. In the experiment described below, the system was trained on a corpus for the domain Mushrooms. The evaluation corpus is from the domain Fish. Hence, the two corpora are disjoint and cover distinct thematic areas. More details on the correction system can be found in Strohmaier et al. (2003b).

Garbled Input Text for Correction. We collected 10 texts from the domain Fish, all containing a nontrivial number of errors. Texts were retrieved from the Web, using queries to Google with spelling errors, such as *fish anglers infomation realy*. We checked that the texts do not contain paragraphs that are also found in the documents of the corpora Fish E introduced in Section 2.2. The concatenation of the 10 texts was used as

10 To simplify evaluation, a fully automated variant of text correction was considered.
11 In what follows, by a token, we always mean a token composed of standard letters only.

input to the text correction system. For the evaluation, a corrected version of the full text was manually created. The full text contains 17,697 tokens of which 418 (2.36%) were found to be erroneous.

Background Dictionaries for Correction. As a baseline, a first crawled dictionary D_{crawl} with 505,652 entries was built, collecting all words from the documents in the corpus Fish E (1). A second dictionary D_{crawl}^{+F} used only those pages that were not rejected by the filter for threshold $\mu = 2$, based on the maximal English error dictionary $D_{err}(English,all)$.[12] In this case, 324 documents passed the filter, whereas 186 were rejected. In this case we obtained 291,065 entries. Deleting in D_{crawl}^{+F} all words that represent entries of $D_{err}(English,all)$, a third dictionary D_{crawl}^{+F+ED} with 269,079 entries was computed.

Note that we did *not* extend D_{crawl}^{+F} and D_{crawl}^{+F+ED} by analyzing an additional set of filtered Web pages. Hence, D_{crawl}^{+F} is in fact a subdictionary of D_{crawl}, and similarly for D_{crawl}^{+F+ED} and D_{crawl}^{+F}. This explains why the coverage of D_{crawl}^{+F} (D_{crawl}^{+F+ED}) is smaller than the coverage of D_{crawl} (D_{crawl}^{+F}); see below. With an extended filtered crawl, even better coverage and accuracy results would probably be possible.

Evaluation Results. We then compared the lexical coverage (percentage of tokens of the correct version of the input text found in the dictionary) and correction accuracy (percentage of correct tokens after automated correction) for each of the three dictionaries. The results are presented in Table 30. The accuracy of the input text is 97.64%. The fifth column gives the improvement in accuracy, taking the input text as a baseline. The last column mentions the number of erroneous tokens in the text that are found in the respective error dictionary.

Note that the use of the filtered corpus leads to a measurable improvement in correction accuracy. The second step in which we eliminate all entries of the error dictionaries in the correction dictionary leads to an additional gain.

Overproduction and Underproduction of the Error Dictionary. As mentioned above, 418 tokens of the input text represented proper errors. From these, 254 (60.77%) turned out to be entries of the maximal English error dictionary $D_{err}(English,all)$. Note that this value for underproduction is very compatible with our estimates in Section 5. Remarkably, only seven of the correct tokens of the input text occurred in the error dictionary.

Analyzing the Effect of Using Filters and Error Dictionaries. The most important error source in the correction process are erroneous tokens of the text that—by accident—represent entries of the crawled dictionaries. Using the above strategy, these **false friends** are only replaced by another word w of the correction dictionary if overwhelming frequency information is available that leads to a preference of w after computing the balanced score for similarity and frequency. The dictionary D_{crawl} contains 262 of the 418 erroneous tokens of the text. The dictionary D_{crawl}^{+F}, which collects the vocabulary of filtered pages, contains only 92 erroneous tokens. After eliminating all entries of the maximal error dictionary, the new dictionary D_{crawl}^{+F+ED} contains only 49 false friends. Note that the latter tokens represent errors not contained in the error dictionary. A very interesting additional number is the following: when eliminating in D_{crawl} all

12 Other filter thresholds for $\mu = 1, 0.5$, and 0 were also tested and led to very similar accuracy values.

Table 30
Measuring the quality of distinct dictionaries for text correction. D_{crawl} is produced by an unfiltered crawl, D_{crawl}^{+F} by a filtered crawl. For D_{crawl}^{+F+ED}, a filtered crawl is used and remaining entries of error dictionaries are eliminated.

Dictionary	Entries	Coverage (%)	Accuracy (%)	± (%)	False friends
D_{crawl}	505,652	99.08	98.45	0.81	262
D_{crawl}^{+F}	291,065	98.77	98.61	0.97	92
D_{crawl}^{+F+ED}	269,079	98.75	98.74	1.10	49

the entries that are found in D_{err}(*English,all*), the resulting dictionary contains 105 erroneous tokens of the text. This shows that the filtering step eliminates 56 (= 105 − 49) erroneous tokens of the text that are *not* found in the error dictionary and proves that a two-step procedure—first using filters for crawling pages, then eliminating entries of error dictionaries afterwards—leads to optimal results.

10.2 Generating Translation Data from Parallel Corpora

Parallel texts represent an important resource for automatic acquisition of bilingual dictionaries. Since only a small number of large parallel corpora are available, which are moreover specialized both with respect to form and contents, the Web represents an important archive for mining parallel texts (Resnik and Smith 2003). When building up bilingual dictionaries for machine translation, or when presenting parallel phrases to users, correctness is an important issue. Hence, it is interesting to see how error dictionaries help to reduce errors in parallel corpora. Our methods can be applied to any kind of parallel corpus. For our experiments we used the freely available Europarl corpus.[13] The corpus covers the proceedings of the European Parliament 1996–2001 in 11 official languages of the European Union. We only analyzed the English and German versions of the parallel texts. The 488 documents in the corpus are of an excellent quality. Our goal was to find English and German texts with a nontrivial number of errors (if any) and to detect these errors. Since the overproduction of error dictionaries in very accurate texts is high, the problem is challenging. The maximal error dictionaries for the two languages were used to determine the error rate of each document. Table 31 shows the twenty documents with the highest error rates for both the English and the German subcollection of the corpora. Columns 4 and 5 describe the number of tokens that represent entries of the respective error dictionary and the number of real errors among these hits. The results show that when analyzing very accurate texts, the error rate is not always a safe indicator for a corresponding number of real errors. Still, the experiment isolates 246 real errors, only looking at 40 documents. When collecting translation correspondences, we may simply discard all phrases/sentences with a hit in an error dictionary, together with their aligned counterparts. Many translation pairs with errors will be avoided. Given the length of the documents, the number of hits of the error dictionaries is small, hence the loss of recall is not essential. In this way our

13 The corpus, which was also used by Koehn, Och, and Marcu (2003), is available at http://www.isi.edu/koehn/europarl/.

Table 31
English (E) and German (G) documents of the Europarl corpora, sizes, error rates w.r.t. maximal English and German error dictionaries, numbers of hits of the error dictionaries, and numbers of real errors among hits.

Documents	Tokens	Error rate	Hits	Real errors	Percentage
ep-96-09-20.txt (E)	9,945	1.31	13	2	15.38
ep-97-04-24.txt (E)	8,074	0.99	8	8	100.00
ep-97-09-19.txt (E)	3,230	0.93	3	0	0.00
ep-97-02-21.txt (E)	5,830	0.86	5	5	100.00
ep-99-01-28.txt (E)	5,347	0.75	4	0	0.00
ep-97-06-25.txt (E)	20,012	0.70	14	11	78.57
ep-96-07-19.txt (E)	4,383	0.68	3	3	100.00
ep-97-04-23.txt (E)	21,930	0.64	14	14	100.00
ep-97-12-04.txt (E)	9,463	0.63	6	6	100.00
ep-99-02-12.txt (E)	5,426	0.55	3	3	100.00
ep-00-03-29.txt (E)	22,252	0.54	12	12	100.00
ep-96-07-17.txt (E)	34,381	0.52	18	14	77.77
ep-99-03-10.txt (E)	31,509	0.51	16	0	0.00
ep-00-11-15.txt (E)	35,167	0.48	17	1	5.88
ep-97-04-10.txt (E)	16,653	0.48	8	6	75.00
ep-97-05-15.txt (E)	20,942	0.48	10	2	20.00
ep-97-10-20.txt (E)	8,601	0.46	4	4	100.00
ep-97-04-11.txt (E)	6,857	0.44	3	1	33.33
ep-99-01-15.txt (E)	9,193	0.43	4	0	0.00
ep-96-06-18.txt (E)	32,768	0.43	14	6	42.86
ep-03-01-13.txt (G)	15,926	2.57	41	2	4.89
ep-97-05-16.txt (G)	12,344	1.94	24	15	62.50
ep-02-09-02.txt (G)	14,845	1.62	24	1	4.16
ep-98-11-05.txt (G)	15,035	1.46	22	3	13.64
ep-99-01-28.txt (G)	6,798	1.32	9	0	0.00
ep-02-04-25.txt (G)	10,842	1.29	14	4	28.57
ep-97-10-02.txt (G)	13,650	1.25	17	9	52.94
ep-99-07-20.txt (G)	2,431	1.23	3	0	0.00
ep-00-03-15.txt (G)	34,904	1.20	42	31	73.81
ep-96-06-21.txt (G)	8,474	1.18	10	9	90.00
ep-96-06-17.txt (G)	9,408	1.17	11	2	18.18
ep-99-04-16.txt (G)	8,667	1.15	10	9	90.00
ep-96-04-19.txt (G)	8,694	1.15	10	2	20.00
ep-00-12-15.txt (G)	6,964	1.15	8	3	37.50
ep-00-09-08.txt (G)	4,374	1.14	5	0	0.00
ep-96-07-04.txt (G)	10,975	1.09	12	11	91.66
ep-01-04-05.txt (G)	26,941	1.08	29	20	68.96
ep-97-06-09.txt (G)	11,152	1.08	12	12	100.00
ep-97-07-14.txt (G)	11,180	1.07	12	5	41.66
ep-97-07-18.txt (G)	10,392	1.06	11	10	90.90

methods may help to improve the generation of translation data even from collections of very accurate parallel texts.

11. Conclusion

In this article we investigated the distribution of orthographic errors of distinct types in the English and German Web. Experiments based on a variety of very large error dictionaries showed that Web corpora typically contain a non-negligible number of pages

with an unacceptable number of orthographic errors. Typing errors represent the most important subclass. In German Web pages, errors resulting from character encoding problems represent another important category. In our experiments, PDF documents were found to contain less orthographic errors than HTML documents, and corpora covering specific thematic areas were found to contain more errors than collections of "general" pages without such a focus. Some differences were remarkable; in particular, our corpora for special thematic areas related to hobbies contain many pages with a high number of orthographic errors. We also found that mean error rates are influenced by the collection strategy. Specific crawling strategies help to avoid chat and junk while attracting professional documents. Since document genre and error rates are correlated, refined crawling strategies may help to reduce mean error rates.

Error dictionaries, even subdictionaries of modest size, can be used as filters that help to detect and eliminate pages with many orthographic errors. Filters with user-defined thresholds work well for both languages. Obviously, the possibility of deleting pages with many orthographic errors and of marking all entries of error dictionaries in the remaining documents opens a wide range of interesting applications in distinct areas of corpus linguistics. To exemplify possible applications we showed how to improve the quality of Web-crawled dictionaries for text correction. With these filtered dictionaries, higher values for correction accuracy were obtained than with those directly obtained from Web crawls. In a second experiment, we showed how error dictionaries may be used to improve the automated collection of translation correspondences, avoiding translation pairs with orthographic errors.

Going beyond corpus linguistics, it might be interesting to design (special modes of) Web search engines where the error rate of a given document is used as one parameter in the ranking of answers. In many search scenarios, answer documents with a large number of orthographic errors appear to be less reliable, and the user might wish to concentrate on "professional" or carefully edited Web pages.

In our practical work we found that the collection and analysis of very large Web corpora is difficult for many reasons. For example, it is not clear how to treat pages with artificial vocabulary that is only introduced to obtain a better ranking. We learned that often these junk lists are intensionally enriched with many orthographic errors to obtain a better ranking, in particular for erroneous queries. In our experiments, some of these pages were found immediately, looking at error rates, and excluded. Later, when inspecting documents for genre classification, other less eye-catching examples were found. Some portions of text occurred in several documents. The conversion of Web pages into ASCII represents a potential source for new errors. In particular the conversion of German PDF documents to ASCII turned out to be very error prone. Nonstandard vocabulary (special names, foreign language expressions, archaic language, programming code, slang, etc.) is another source that makes various pages inappropriate for corpus construction.

One step for future work is the development of special dictionaries for frequent foreign language expressions, archaic language, programming code, and slang. Special dictionaries for these expressions would not only help to detect and exclude pages with a high amount of nonstandard vocabulary, but they could also be used as additional filters in the construction of error dictionaries. The results in Section 5.2 indicate that the overproduction of our error dictionaries could be reduced in a significant way by eliminating entries that represent expressions of the earlier-mentioned type. As a matter of fact, new types of spelling errors were found during the experiments described earlier. It might be interesting to enlarge the error dictionaries for spelling errors, taking the new patterns into account.

We also found that enlarged error dictionaries that store with each garbled entry the correct word from which it was derived are very useful for error correction. In contrast to our first intuitions, the number of ambiguities arising from this correction strategy is small, and the predictive power of enlarged error dictionaries is high. More details on text correction with error dictionaries will be given in a forthcoming paper.

Acknowledgments

The authors thank the anonymous referees of *Computational Linguistics*. Their remarks and suggestions helped to improve the contents and presentation of the article. Special thanks to Annette Gotscharek and Uli Reffle for all their help.

References

Amengual, Juan Carlos and Enrique Vidal. 1998. Efficient error-correcting viterbi parsing. *IEEE Transactions on PAMI*, 20(10):1–109.

Baroni, Marco and Silvia Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon.

Boutsis, Sotiris, Stelious Piperidis, and Iason Demiros. 1999. Generating translation lexica from multilingual texts. *Applied Artificial Intelligence*, 13(6):583–606.

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117.

Brown, Jonathan and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. In *Proceedings of the InSTIL/ICALL2004 Symposium on Computer Aided Language Learning*, pages 25–28, Venice.

Chelba, Ciprian and Frederick Jelinek. 2002. Recognition performance of a structured language model. In *Proceedings of Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99)*, pages 1567–1570, Budapest.

Church, Kenneth W. and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24.

Dimitrova, Maya, Nicholas Kushmerick, Petia Radeva, and Joan Jose Villanueva. 2003. User assessment of a visual Web genre classifier. In *Third International Conference on Visualization, Imaging, and Image Processing*, Malaga.

Dunning, Ted. 1993. Accurate models for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Finn, Aidan and Nicholas Kushmerick. 2003. Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Text Style and Synthesis*, Acapulco. *Journal of the American Society for Information Science and Technology* (in press).

Fletcher, William H. 2004a. Facilitating the compilation and dissemination of ad-hoc web corpora. In Guy Aston, Silvia Bernardini, and Dominic Stewart, editors, *Corpora and Language Learners*, number 17 in Studies in Corpus Linguistics. John Benjamins Publishing Company, Amsterdam.

Fletcher, William H. 2004b. Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton, editors, *Corpus Linguistics in North America 2002.* Rodopi, Amsterdam.

Gaizauskas, Robert, George Demetriou, and Kevin Humphreys. 2000. Term recognition in biological science journal articles. In *Proceedings of the Workshop on Computational Terminology for Medical and Biological Applications, 2nd International Conference on Natural Language Processing (NLP-2000)*, pages 37–44, Patras.

Gale, William A. and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of Fourth DARPA Workshop on Speech and Natural Language*, pages 152–157, Pacific Grove, CA.

Gartner, Hans-Jürgen. 2003. Extraktion von semantischer Information aus Layout-orientierten Daten. Master's thesis, Technical University of Graz.

Grefenstette, Gregory. 1992. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 89–97, Copenhagen.

Grefenstette, Gregory. 1999. The WWW as a resource for example-based MT tasks. Paper presented at ASLIB "Translating and the Computer" conference, London.

Grefenstette, Gregory. 2001. Very large lexicons. In *Computational Linguistics in the Netherlands 2000: Selected Papers from the Eleventh CLIN Meeting*, Language and Computers, Amsterdam.

Guenthner, Franz. 1996. Electronic lexica and corpora research at CIS. *International Journal of Corpus Linguistics*, 1(2):287–301.

Jelinek, Frederick. 1997. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA.

Kehoe, Andrew and Antoinette Renouf. 2002. WebCorp: Applying the web to linguistics and linguistics to the Web. In *Poster Proceedings of the 11th International World Wide Web Conference, WWW02*, Honolulu.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automated detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction. *Computational Linguistics—Special Issue on the Web as Corpus*, 29(3):333–348.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton.

Kukich, Karen. 1992. Techniques for automatically correcting words in texts. *ACM Computing Surveys*, 24(4):377–439.

Kumano, Akira and Hideki Hirakawa. 1994. Building a MT dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, pages 76–81, Kyoto.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 17–22, Columbus, OH.

Lin, Shian-Hua, Chi-Sheng Shih, Meng Chang Chen, Jan-Ming Ho, Ming-Tat Ko, and Yueh-Ming Huang. 1998. Extracting classification knowledge of internet documents with mining term associations: A semantic approach. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 241–249, Melbourne, Australia.

Maier-Meyer, Petra. 1995. *Lexikon und automatische Lemmatisierung*. Ph.D. thesis, CIS, University of Munich.

Morley, Barry, Antoinette Renouf, and Andrew Kehoe. 2003. Linguistic research with the XML/RDF aware WebCorp tool. In *Poster Proceedings of the 12th International World Wide Web Conference, WWW03*, Budapest.

Oh, Alice H. and Alexander I. Rudickny. 2000. Stochastic language generation for spoken dialogue systems. In *ANLP/NAACL 2000 Workshop on Conversational Systems*, pages 27–32, Seattle.

Ostendorf, Mari, Vassilios V. Digalakis, and Owen A. Kimball. 1996. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions Speech and Audio Processing*, 4(5):360–378.

Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics - Special Issue on the Web as Corpus*, 29(3):349–380.

Ringlstetter, Christoph. 2003. OCR-Korrektur und Bestimmung von Levenshtein-Gewichten. Master's thesis, LMU, University of Munich.

Schwartz, Lee, Takako Aikawa, and Michel Pahud. 2004. Dynamic language learning tools. In *Proceedings of the InSTIL/ICALL2004 Symposium on Computer Aided Language Learning*, pages 107–110, Venice.

Smadja, Frank A. and Kathleen R. McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252–259, Pittsburgh, PA.

Sornlertlamvanich, Virach and Hozumi Tanaka. 1996. The automatic extraction of open compounds from text corpora. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 1143–1146, Copenhagen.

Strohmaier, Christian, Christoph Ringlstetter, Klaus U. Schulz, and Stoyan Mihov. 2003a. Lexical postcorrection of OCR-results: The web as a dynamic secondary dictionary? In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 03)*, pages 1133–1137, Edinburgh.

Strohmaier, Christian, Christoph Ringlstetter, Klaus U. Schulz, and Stoyan Mihov.

2003b. A visual and interactive tool for optimizing lexical postcorrection of OCR results. In *Proceedings of the IEEE Workshop on Document Image Analysis and Recognition, DIAR'03*, Madison, WI.

Taghva, Kazem and Jeff Gilbreth. 1999. Recognizing acronyms and their definitions. *International Journal on Document Analysis and Recognition*, 1(4):191–198.

Walker, Donald E. and Robert A. Amsler. 1986. The use of machine-readable dictionaries in sublanguage analysis. In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum, Hillsdale, NJ, pages 69–83.

Way, Andy and Nano Gough. 2003. wEBMT: Developing and validating an example-based machine translation system using the world wide web. *Computational Linguistics—Special Issue on the Web as Corpus*, 29(3):421–458.

Yeates, Stuart, David Bainbridge, and Ian H. Witten. 2000. Using compression to identify acronyms in text. In *Proceedings of the Conference on Data Compression*, page 582, Snowbird, UT.